A Practical Robotic Grasping Method by Using 6-D Pose Estimation With Protective Correction

Hui Zhang, *Member, IEEE*, Zhicong Liang, Chen Li, Hang Zhong, *Member, IEEE*, Li Liu, *Member, IEEE*, Chenyang Zhao, Yaonan Wang, *Senior Member, IEEE*, and Q. M. Jonathan Wu, *Senior Member, IEEE*

***Abstract*— The pose estimation is the critical technology in industrial robot. Nowadays, many machine vision-based approaches have applied the technology and achieved excellent results. However, the rapid detection of the pose estimation in complex multiscene environments is still a challenge, due to the interference of multiangle light and multibackground. To address these issues, this article proposes a practical robotic grasping method by using the 6-D pose estimation with protective correction. In this method, the synthetic dataset by self-production is used to train the improved deep object pose estimation network and then use the standard perspective-n-point algorithm to estimate the 6-DoF pose for each object instance. Meanwhile, in order to prevent grasp collisions cause by misrecognition, we propose the corrected grasping pose algorithm for protective correction by measured translation and predicted translation. Finally, the proposed grasping method has an average grasping success rate of 83.3% for the three objects under normal light, and the network for single-image detection speed has been to 1.490 frames/s. The code is available at https://github.com/aimiplus/Practical_Robotic_Grasping_Method.**

***Index Terms*—Deep learning, point cloud segmentation, pose estimation, residual block, robotic grasping, vision detection.**

Manuscript received November 30, 2020; revised February 19, 2021 and March 8, 2021; accepted April 1, 2021. Date of publication May 3, 2021; date of current version December 20, 2021. This work was supported in part by the National Key RD Program of China under Grant 2018YFB1308200, in part by the National Natural Science Foundation of China under Grant 61971071 Grant 6202780012, in part by the Changsha Science and Technology Project under Grant kq1907087, in part by the Special funds for the construction of innovative provinces in Hunan Province under Grant 2020SK3007, in part by the Postdoctoral Innovative Talent Support Program under Grant BX20200122, in part by the Hunan Key Laboratory of Intelligent Robot Technology in Electronic Manufacturing under Grant IRT2018009, in part by the Hunan Key Project of Research and Development Plan under Grant 2018GK2022, and in part by the Changsha science and technology project under Grant kq1907087. *(Corresponding authors: Hui Zhang and Hang Zhong.)*

Hui Zhang is with the School of Robotics, and the National Engineering Laboratory for Robot Visual Perception and Control Technology, Hunan University, Changsha 410012, China (e-mail: zhanghui1983@hnu.edu.cn).

Zhicong Liang, Chen Li, and Chenyang Zhao are with the College of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410012, China (e-mail: liuchen1019@stu.csust.edu.cn; zongjie@stu.csust.edu.cn; zhanghuihby@csust.edu.cn).

Hang Zhong, Li Liu, and Yaonan Wang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: zhonghang@hnu.edu.cn; liuli@hnu.edu.cn; yaonan@hnu.edu.cn).

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2021.3075836.

Digital Object Identifier 10.1109/TIE.2021.3075836

I. Introduction

THE six-degree-of-freedom (6-DoF) pose estimation is the key technology in the application of artificial intelligence, including augmented reality, autonomous driving, and robotic manipulation [1]–[3]. It can help the robot know the position and orientation of the object to perform the grasping [4]. Specifically, robots picking up target cargo from warehouse shelves are inseparable from fast and reliable pose estimation in the Amazon Picking Challenge [5] and humanoid robots can effectively help humans complete complex cooperative tasks [6]. Thanks to the maturity of camera imaging technology and the wide application of RGB images, some pose estimation methods [7]–[9] based on RGB images have achieved good results. However, the only RGB image data lack diversity and contains limited information, the design requirements for the network architecture are strict. Therefore, estimate the 6-DoF pose of an object from only RGB images effectively is still a valuable research direction and full of challenges. In this article, a single RGB image is used for the pose estimation and the depth information is only used for the robotic safe grasping correction feedback, which is not involved in the pose estimation detection.

Previously, there had been many traditional approaches to solving the 6-D pose estimation [10]. The authors in [11] and [12] proposed to detect textureless objects with template-based matching methods, but the similarity scores are usually lower when facing occlusion, truncation, etc. among objects, which may cause incorrect pose estimation. Based on this, the authors in [13] and [14] proposed to rely on the hand-made local features, and the object pose estimation was generated by the correspondence between the 2-D image and the 3-D model by feature-based methods, which could handle problems such as occlusion and truncation among objects, but hand-made features require the object with rich textures and are not robust to illumination and scene clutter.

Thanks to the outstanding performance of deep learning in tasks such as object classification [15], object detection [16], and semantic segmentation [17]. Recent methods have presented

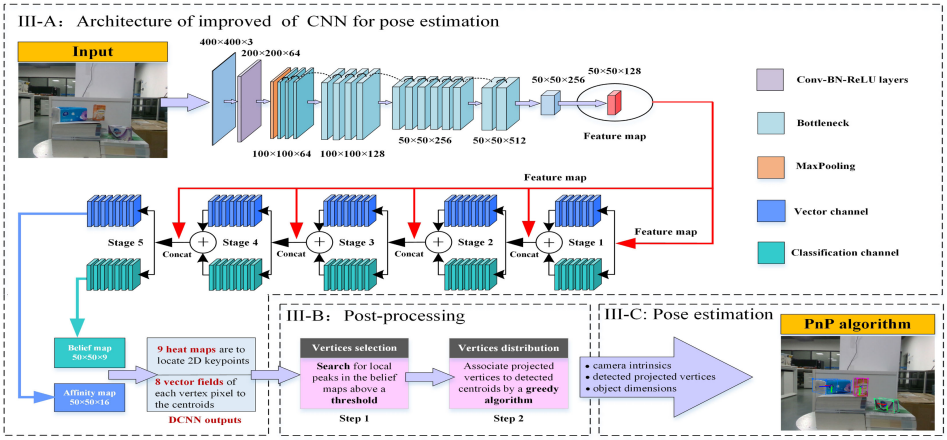


Fig. 1. Improved pose estimation algorithm called FastNet-V1. The pose estimation algorithm consists of the architecture of improved CNN for pose estimation, postprocessing and pose estimation.

deep learning techniques for the 6-D pose estimation [18]–[20], which can significantly improve the accuracy of object detection and pose estimation. Peng *et al.* [21] proposed a method for defining the key points of the feature, and designed a two-stage network to train the key points of the predicted feature, which can deal with problems such as no texture, occlusion, and truncation. Pavlakos *et al.* [22] proposed a method of using the convolutional network to predict the semantic key points of an object and combining it with a deformable shape model. Then, to estimate the continuous 6-DoF pose (3-D translation and rotation) of the object from a single RGB image. Hu *et al.* [23] proposed a subdivision-driven 6-D pose estimation framework, which uses the 2-D key point position of each visible part of the object to provide local pose prediction. The predicted confidence measure is used to merge these pose candidates into a set of reliable 3-D to 2-D correspondences, from which a reliable pose estimate is obtained. Li *et al.* [7] pointed out that there are obvious differences between rotation and translation in 6-DoF pose estimation, and proposed a coordinate-based unwrapping pose network to predict the rotation and translation of the object, respectively. A dynamic zoom in algorithm is proposed to improve the robustness of pose estimation to detection errors.

However, training a deep convolutional neural network (CNN) typically requires a large amount of labeled data, including objects that are annotated with a precise 6-D pose. Compared with 2-D detection, 3-D detection based on CNNs prohibits manual labeling of data, and there is no guarantee of accuracy. Thus, synthetic data are a promising alternative for training deep convolution neural networks for the accuracy and convenience of making annotation data. Inspired by the inconvenience of collecting annotation data and the increasing popularity of 3-D models, the authors in [24]–[26] attempted to use a rendering software to inexpensively acquire a large number of synthetic images to train pose estimation networks, but synthetic data have its weakness, which the most of all is the reality gap.

Therefore, in order to bridge the reality gap for synthetic data, we carry out research on 6-D pose estimation algorithm based on deep object pose estimation (DOPE) [27], using deep learning technology and deploying to the self-designed platform to conduct in-depth practical research of robotic grasping (see Fig. 1), including the following three aspects of contributions.

1) In order to improve the robustness of identification, we adjust the strategy of making dataset according to the actual situation, and consider the model and dataset of making training under the conditions of real grasping environment including illumination imbalanced and different detection distances. Meanwhile, the influence of the combination ratio of the number of datasets on the vision detection of robotic grasping is considered.
2) At the same time, due to the low speed of detection and the large time cost of training, we improve the DOPE [27] networks by modifying the feature extraction network through residual block [28] and the optimizing branch network through reducing the convolutional calculation, which the detection speed is up to 1.490 frames/s. Therefore, there is valuable improvement compared with the original.
3) Furthermore, considering the safety risk of using the data predicted by the CNN for robotic grasping control, a corrected grasping pose (CGP) algorithm has been proposed to correct the translation of the estimated pose and to ensure the safety of the whole grasping process.

II. Related Works

A. Traditional Detection Methods

The traditional methods mentioned in this section are based on traditional machine vision. In the past, the traditional methods can be categorized into template-based methods and feature-based methods.

Template-based methods: First to construct rigid template of the object, the template of which is the rendering of the object from the 3-D object model or direction gradient histogram, then to scan different positions in the input image with template, and calculate a similarity score in each position for the best match [29]. Meanwhile, Hinterstoisser *et al.* [30] proposed a

multimodalities method, which used the gradient information of color images combined with the normal features of the object surface as the basis of template matching.

Feature-based methods: The authors in [31]–[33] proposed a scale-invariant method, which took depth information to extract or learn features that are robust to different illumination conditions and even partial occlusion.

B. Deep Learning-Based Detection Methods

The detection methods in this section are based on deep learning. Nowadays, the deep learning-based methods can be categorized into real data-based training methods and synthetic data-based training methods.

Real data-based training methods: The authors in [34] and [35] both used the CNN trained with real data to return the 3-D object coordinate position of each pixel to establish a 2D–3D correspondence directly, but 3-D coordinate regression will encounter ambiguity when dealing with symmetric objects. Tekin *et al.* [36] proposed to use the you only look once (YOLO) networks to train the projection of the 3-D model bounding box corner points on the real 2-D image with real Linemod dataset, and then the perspective-n-point algorithm is used to recover the 6-D pose of the object. Xiang *et al.* [37] proposed a new CNN for the 6-D object pose estimation, which estimates the 3-D translation of the target by locating the center of the object in the image and predicting its distance from the camera, and then regressing to a quaternion represent the 3-D rotation to estimate the object. Wang *et al.* [38] proposed an end-to-end deep learning method for estimating 6-DoF poses of known objects from RGB-D input. The embedded features were extracted from the RGB map and the depth map, and the two features were intensively merged pixel by pixel to directly predict the rotation and offset of the object. He *et al.* [39] proposed a new data-driven method for robust 6-DoF object pose estimation from a single RGB-D image. The deep Hough voting network is used to detect the 3-D key points of the object, and then the 6-D pose parameters of the object are estimated by the method of least square fitting and iterative closest point (ICP) iteration.

Synthetic data-based training methods: Tobin *et al.* [40] proposed domain randomization, a simple technique for training models on simulated images that transfer to real images by randomizing rendering in the simulator. Thus, it transferred the deep neural networks trained with synthetic images to robot control in the real world successfully. In order to bridge the reality gap, Wu *et al.* [25] used heat map of key point to bridge the real data and the synthesized data for 6-D pose estimation, which avoided the difference between the real image and the composite image statistics caused by the imperfect rendering.

C. Grasping Methods Using CNNs

Recently, deep learning has also been introduced in other subsystems of the robotic grasping system to improve grasping efficiency and accuracy [41]. Redmon and Angelova [42] use a huge neural network to replace the traditional sliding window method to predict multiple grasp coordinates. It does not require prior knowledge such as the 3-D model of the object. It has accuracy and the real-time performance. Guo *et al.* [43] proposed a shared CNN, which can complete target detection and grasp detection tasks in parallel. Park *et al.* [44] proposed an accurate robotic grasp detection algorithm using fully CNNs with high-resolution images to recover the five poses $(x; y; \alpha; w; h)$ for manipulation. Pas *et al.* [45] proposed a method for generating grasp hypotheses on any visible surface without requiring a precise segmentation of the target object. They also proposed a new grasp descriptor that incorporates surface normals and multiple views. The common point of these methods is that they use an end-to-end approach to directly return the grasping position from the input image through a CNN, without the need to locate the target object in advance.

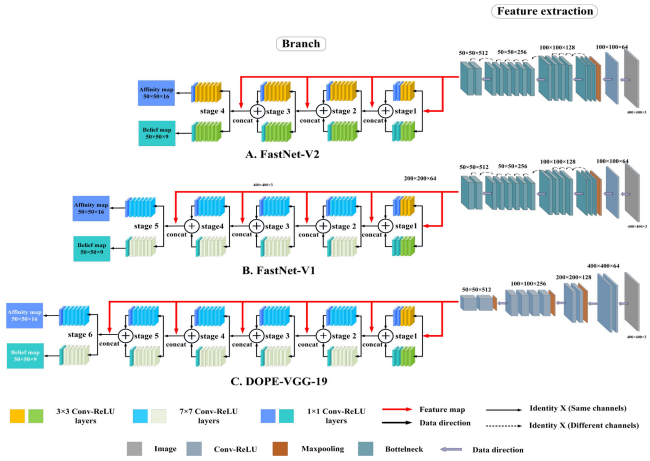


Fig. 2. Three models of branch networks.

III. Methodology and Design

A. Feature Extraction Network Design

Visual geometry group VGG-19 is the feature extraction network of the DOPE method. Both VGG-16 and VGG-19 are the best performance networks in the combination mode of stacked network layer, which use the data of the upper layer to process the next layer without shared data information in the hidden layers. Therefore, the comparison of the performance between VGG-16 and VGG-19 is of great significance, which is explored the maximum performance of the pose estimation network. Meanwhile, considering that VGG series networks without any information exchange in the hidden layers, the residual network structure [28] is introduced to improve the feature extraction network and increase the opportunity for information sharing. The improved network is named FastNet-V1, as shown in Fig. 2.

B. Branch Network Design

According to the analysis, the branch network has convolution layers with a large amount of computation. As far as DOPE is concerned, the convolution layer to be calculated has 80 layers $(10 + 70)$, among which the first stage has ten convolution layers

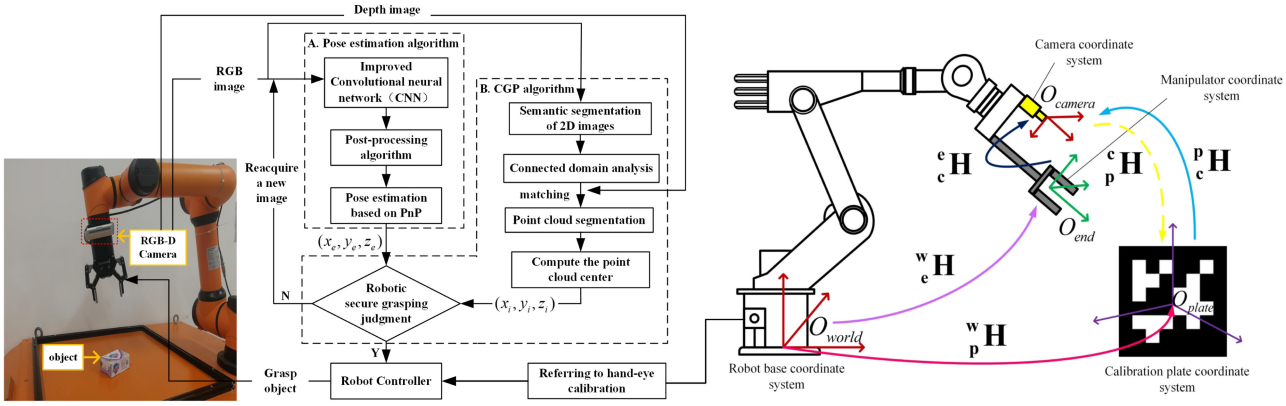


Fig. 3. Proposed grasping strategy. Improved pose estimation algorithm is combined with CGP algorithm for robotic grasping and demonstrating the hand-eye calibration principle.

and the remaining five stages have a total of 70 layers. Therefore, it undoubtedly increases the cost of network training and test time caused by enormous convolution computation. Based on the problems mentioned above, the improved FastNet-V1 network reduces the number of convolution stages for the branch network with the original sixth stage specifically. At this time, the convolution layer to be calculated has 66 layers $(10 + 56)$, and only the first five stages of the branch network need to be calculated. The transition stage of the two networks goes through one $3 \times 3 \times 128$ convolution layer and one rectified linear units (ReLU) layer.

In order to further explore the residual network performance, the second improved network named FastNet-V2. First of all, the network structure consistent with the FastNet-V1, and except that the branch network only four convolutional stages, there are 52 $(10 + 42)$ convolution layers need to calculate, and all stages of the convolution size are 3×3 result in greatly reducing convolutional computation, as shown in Fig. 2.

C. Proposed Robotic Grasping Strategy

In order to grasp objects effectively for the robotic manipulator, we propose a feasible grasping technology strategy, which the improved pose estimation algorithm with the CGP algorithm complement each other, as shown in Fig. 3 on the left. In particular, the point cloud segmentation method is used to obtain the point clustering of target object, and the point cloud center is calculated as the measured translation (the distance of the object with respect to the camera). At this time, measured translation and the predicted translation of the pose estimation are combined with the CGP algorithm, thus, the effective translation of output and the predicted rotation are output as the final estimated pose. Furthermore, the result of hand-eye calibration is used to calculate relationship from estimated pose value to the robotic control value to perform grasping.

1) Corrected Grasping Pose: The aim of CGP algorithm is to correct translation for estimated pose. First, the 2-D image segmentation of the target object using the semantic segmentation network of SegNet [46] is considered to be carried out on the mobile robot grasping platform. Therefore, the training datasets are all from the object image of the grasping platform, and the principle of semantic segmentation is not described in detail. Then, due to the scattered interference pixels in the segmentation, the coarse segmentation image of the object goes through the connected domain analysis [47]. Each separate block separated by segmentation is represented by an independent label, and the largest area of block is the final segmented region of interest. Therefore, the depth image of RealSense D435 camera is used for image matching with the segmented RGB image to further obtain the 3-D point cloud segmentation region of the object (E_i=$[x, y, z]^T$, $(i = 1, 2, 3\ldots)$). Then, the point clustering center of the object is calculated, and the point cloud of the region is combined to calculate the average value:

$$\mathrm{Cen}_i = \frac{1}{N}\sum_{j=1}^{N} E_{ij} \tag{1}$$

where cen_i represents the center of the point cloud cluster of each object, N is the total number of point clouds in the i_{th} cluster, and E_{ij} is the j_{th} point cloud coordinate of the i_{th} object. Further calculate the side length of the maximum bounding box, and take the difference of the maximum value in the three directions (x, y, z) in i_{th} point clustering as the length (l_i), width (w_i), and height (h_i) of the bounding box, respectively

$$\begin{aligned} l_i &= \max(E_{ix}) - \min(E_{ix}) \\ w_i &= \max(E_{iy}) - \min(E_{iy}) \\ h_i &= \max(E_{iz}) - \min(E_{iz}). \end{aligned} \tag{2}$$

As shown in Fig. 4, the result of the point cloud segmentation. The calculation of the object center as the measurement of translation ($\mathrm{Cen}_i = [x_i, y_i, z_i]^T$), which is added to the robotic secure grasping judgment and improved grasping feasibility.

Next, the judgment principle of robotic secure grasping is stated in detail, as shown in Algorithm 1.

Step 1: Collect two data of predicted translation (x_e, y_e, z_e) of pose estimation network and measured translation (x_i, y_i, z_i) of point cloud.

Step 2: Two error thresholds (D_1 and D_2 are related to the maximum side length of the object, in this experiment, we set

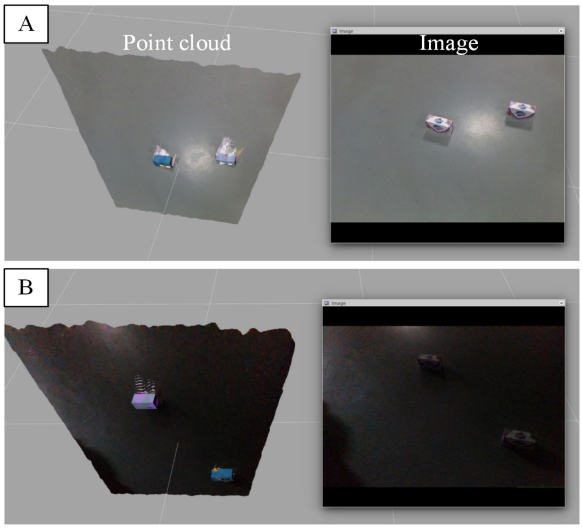


Fig. 4. Result of the point cloud segmentation under different environmental conditions, which the blue bounding box of the object is the effective region of segmentation. A: Bright environment and B: Dark environment.

Algorithm 1: Robotic Secure Grasping.

Input: Predicted translation: (x_e, y_e, z_e);
Measured translation: (x_i, y_i, z_i)
Output: The effective translation: (x_e', y_e', z_e')
Initialize D_1 = 0.01m and D_2 = 0.03m;
$D_0 = (|x_e - x_i| + |y_e - y_i| + |z_e - z_i|)$;
if $0 \le D_0 \le D_1$ **then**
  The predicted translation (x_e, y_e, z_e) is used as the output translation;
  Go to final;
else if $D_1 \le D_0 \le D_2$ **then**
  The measured translation (x_i, y_i, z_i) is used as the output translation;
  Go to final;
  else
    Request to get new pose estimation translation (x_e, y_e, z_e) and measured translation (x_i, y_i, z_i) of point cloud
final ;
return The effective translation: (x_e', y_e', z_e')

$D_1 = 0.01$˜m and $D_2 = 0.03$˜m), and the maximum threshold error should be smaller than the diameter of the object plane.

Step 3: Calculate the measurement and prediction translation error value D_0: $D_0 = (|x_e - x_i| + |y_e - y_i| + |z_e - z_i|)$.

Step 4: The comparison between the calculated error value (D_0) and the set error thresholds (D_1 and D_2) can be divided into the following three situations.

A) In the $0 < D_0 < D_1$, the pose estimation is considered to be accurate, so the direct output predicted translation((x_e, y_e, z_e)) is used as the output translation.

B) In $D_1 < D_0 < D_2$, it is considered that the pose estimation has a little deviation, but it does not cause the grasping failure. Therefore, the measured translation $((x_i, y_i, z_i))$ is used as the output translation.

C) In $D_2 < D_0$, it is considered that there is a serious deviation in the pose estimation and the grasping operation cannot be performed, so it immediately requests to acquire a new RGB image and detect the object again.

Step 5: If the situations of A and B in step 4 is got, then the effective translation of the output is combined with the predicted rotation to calculate the control value of the robotic grasping, and perform the final grasping. If the situation C is got, the grasping operation is stopped and the redetection is requested, until either A or B.

So far, this is the detailed introduction of the whole robotic secure grasping mechanism. By introducing a new visual detection channel (point cloud segmentation) to supervise the effectiveness of the pose estimation algorithm, the robotic grasping security has been improved.

2) Robotic Grasping Control: The matrix ${}^{e}_{c}\mathbf{H}$ is known, the estimated pose that needs to calculate is converted to the value of the robotic grasping control, can perform grasping. First, ${}^{c}_{o}\mathbf{H}$ is the estimated pose of the object with respect to the camera in the neural network, where is the rotation matrix ${}^{c}_{o}\mathbf{R}$ and the translation vector ${}^{c}_{o}\mathbf{t}$ of the pose estimation, and is expressed as

$$ {}^{c}_{o}\mathbf{H} = \begin{bmatrix} {}^{c}_{o}\mathbf{R} & {}^{c}_{o}\mathbf{t} \\ 0 & 1 \end{bmatrix}. \tag{3} $$

Further need to calculate the value of robotic grasping control. Three transformation relations of ${}^{e}_{c}\mathbf{H}$, ${}^{c}_{o}\mathbf{H}$ and ${}^{w}_{e}\mathbf{H}$ multiply, to obtain the transformation relation (${}^{w}_{o}\mathbf{H}$) of the object coordinate system with respect to the robot base coordinate system

$$ \begin{bmatrix} {}^{w}_{o}\mathbf{R} & {}^{w}_{o}\mathbf{t} \\ 0 & 1 \end{bmatrix} = {}^{w}_{o}\mathbf{H} = {}^{c}_{o}\mathbf{H}\,{}^{e}_{c}\mathbf{H}\,{}^{w}_{e}\mathbf{H} \tag{4} $$

where the rotation matrix ${}^{w}_{o}\mathbf{R}$ and the translation vector ${}^{w}_{o}\mathbf{t}$ are the value of robotic grasping control. Specifically, the ${}^{w}_{o}\mathbf{R}$ is the sum of rotation effects of the three axes (X_W, Y_W, and Z_W) for the robot base coordinate system $O_W - X_W Y_W Z_W$. The ${}^{w}_{o}\mathbf{R}$ is represented the rotation matrix by quaternions (x_r, y_r, z_r, w_r) and ${}^{w}_{o}\mathbf{t}$ as follows:

$$ {}^{w}_{o}\mathbf{R} = \begin{bmatrix} 1 - 2{y_r}^2 - 2{z_r}^2 & 2x_r y_r + 2w_r z_r & 2x_r z_r - 2w_r y_r \\ 2x_r y_r - 2w_r z_r & 1 - 2{x_r}^2 - 2{z_r}^2 & 2y_r z_r + 2w_r x_r \\ 2x_r z_r + 2w_r y_r & 2y_r z_r - 2w_r x_r & 1 - 2{x_r}^2 - 2{y_r}^2 \end{bmatrix} \tag{5} $$

$$ {}^{w}_{o}\mathbf{t} = \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix}. \tag{6} $$

From (5) and (6), the translation (x_t, y_t, z_t) and rotation (x_r, y_r, z_r, w_r) are imported into robotic control for grasping the object.

IV. System Design and Dataset

A. System Configuration

Considering that grasping objects cover multiple categories, objects are susceptible to interference during identification, such as different light direction and intensity, and there are uncertainties in the moving path of complex dynamic scenes. Therefore, we design a mobile robot grasping platform, which includes

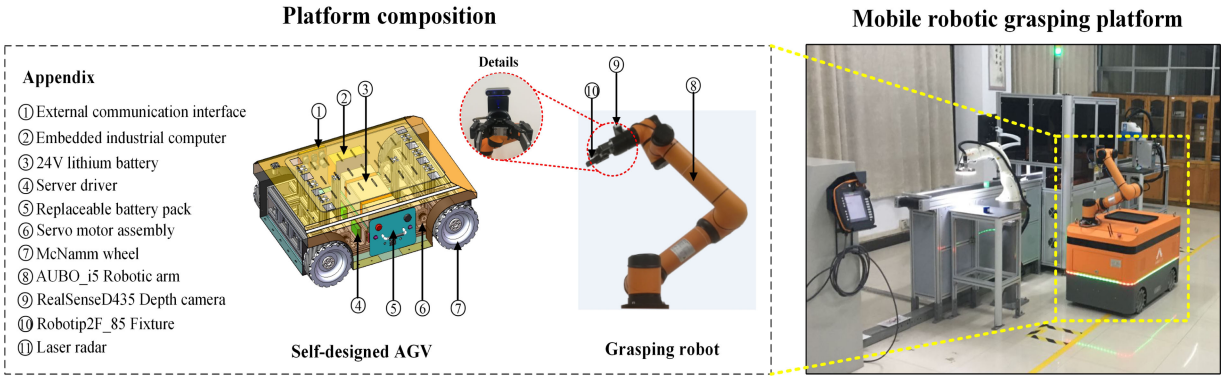


Fig. 5. Mobile robotic grasping platform. The whole platform includes the self-designed automatic guided vehicle (AGV) and grasping robot which displaying each component of the system and focusing on the visual grasping part for detail.

visual detection, grasping, and obstacle avoidance functions, and apply it to logistics sorting to meet the requirements of intelligent grasping and accurate transportation.

The grasping robot is a 6-DoF robotic arm of the AUBO-i5. The automatic operation of a robot arm with an end load of 5 kg or less can be completed, and the accuracy can reach 0.02 mm, which can meet the working task within the working range of 886.5 mm. Besides, the two-fingered manipulator of Robotiq2F-85 is installed at the end of the robotic arm, which the payload of friction gripper and gripper is 5 kg. While, the RealSense D435 depth camera is used to capture RGB images with measurement functions, which is also mounted at the end of the robotic arm. The depth module measures the image within 5 m to an accuracy of 0.01 m.

This article involves the visual grasping part of the robot, but does not involve the robot moving path planning. Therefore, the grasping experiment is implemented on the mobile grasping robot platform, as shown in Fig. 5.

B. Data Generation

Accurate pose estimation based on objects is the prerequisite of successful grasping. Based on this, synthetic data have a series of advantages such as generating a large amount of perfectly labeled data, reducing the dependence on generated and captured data, minimizing the demand for third-party data sources, being faster than manually labeled data, and generating data that is difficult to capture in the real world. Therefore, synthetic data are used in this article to train the neural network model. So far, one of the significant challenges of data generation is to bridge the reality gap. To this end, the training dataset is divided into domain randomized (DR) data and photorealistic data, and it goes through two data-generation processes, which are making model and synthetic data.

1) Making model: In order to keep the model highly similar to the real object, we make three models including soap, calcium tablets, and tissue, which satisfy the following conditions.

1) *Physical size*: To measure the real size of the object that the sizes of x, y, and z are 3.2, 9.4, and 5.8 cm for soap, respectively; the sizes of calcium tablets are 4.5, 4.5, and 8.7 cm, respectively; the sizes of tissue are 10, 13.5, and 7 cm, respectively.
2) *Surface texture*: All the models are depicted and restored one by one according to the characteristics of actual object texture and shape. On the basis of satisfying the abovementioned two conditions, models of three objects have been finished, as shown in Fig. 6.

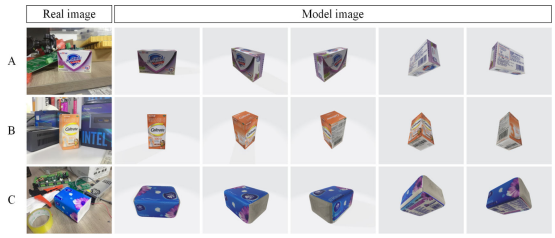


Fig. 6. Real images and model images. Model image from different visual angles in which A: Soap, B: Calcium tablet, C: Tissue.

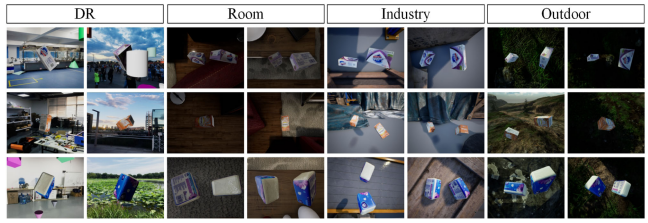


Fig. 7. DR and Photorealitic images. The images are generated in the virtual environment including DR data and Photorealistic: Photorealistic data, which contain three scenarios room, industry, and outdoor, respectively.

2) Synthetic Data: In order to consider the four challenges faced in grasping including complex background, occlusion, illumination projection, and effective grasping distance. Therefore, we mix DR data and photorealistic data to obtain the dataset needed for training, in which the DR data are used to solve the detection under multiscene detection and occlusion, while photorealistic data are used to solve the detection under the physical scene. The dataset is shown in Fig. 7.

TABLE I
NUMBER OF GENERATED TWO KINDS OF DATA

Object	DR	Photorealistic			Total
		Room	Industry	Outdoor	
Soap	4000	1333	1333	1334	8000
Calcium tablet	5000	1667	1667	1666	10000
Tissue	7000	1000	1000	1000	10000

Therefore, different number ratios are set for the training datasets of the three objects, the number of training sets is shown in Table I. Both datasets follow the number ratio of DR dataset and photorealistic dataset are 1:1, and the photorealistic dataset is concentrated in three scenes and the ratio of the number of data collected is 1:1:1. There are 10 000 data for the tissue, but the ratio of the number of data between the DR dataset and the photorealistic dataset is 7:3, and the ratio of the number of data collected by photorealistic dataset in the three scene maps is also 1:1:1. This is used to explore the differences in datasets with different number of structural proportions for robotic grasping effect.

V. EXPERIMENTS AND RESULTS

A. Experiments Design

In order to verify that the dataset made can be used for the proposed grasping strategy, finally, be able to grasp objects effectively. Two experimental verification methods are designed, including pose estimation and object detection, and robotic grasping test in the real environment, illustrating the experimental process and analyzing the results.

1) Experiment 1: Pose estimation experiment. The training network is tested in the real environment, which includes the comprehensive evaluation of the detection effect from two aspects, the detection of illumination imbalance conditions and the detection of distance difference. Among them, the feature extraction network and branch network, respectively, conduct in-depth performance research including accuracy and detection speed of pose estimation.

2) Experiment 2: Robotic grasping experiment. The proposed robotic grasping strategy is implemented to grasp three trained objects, the success rate of grasping is calculated under two lighting conditions, and the influencing factors are analyzed to summarize the superiority of the proposed method.

B. Evaluation Metrics

In the pose estimation, we use the average distance (ADD) metric as proposed in [37] for evaluation. We calculated the rotation R and translation T given to the ground truth and the corresponding estimated rotation $\tilde{R}$ and translation $\tilde{T}$ projected to the 3-D space points, and then calculated the Euclidean distance in the x, y, and z directions

$$\text{ADD} = \frac{1}{m}\sum_{x\in M}\left\|(Rx+T)-(\tilde{R}x+\tilde{T})\right\| \tag{7}$$

where M is the set of points in the 3-D model and m is the number of points, because the position of grasping refers to the centroids, and the m is 1. The 6-D pose is considered correct if the ADD is less than the predefined threshold. Refer to [27], and set the maximum threshold as 0.1 m. We use the ADD metric to show the area under accuracy-threshold curve (AUC) as shown in Fig. 8, where changing the error threshold for the ADD, and then calculating the pass rate. The AUC represents the accuracy of the corresponding pass rate.

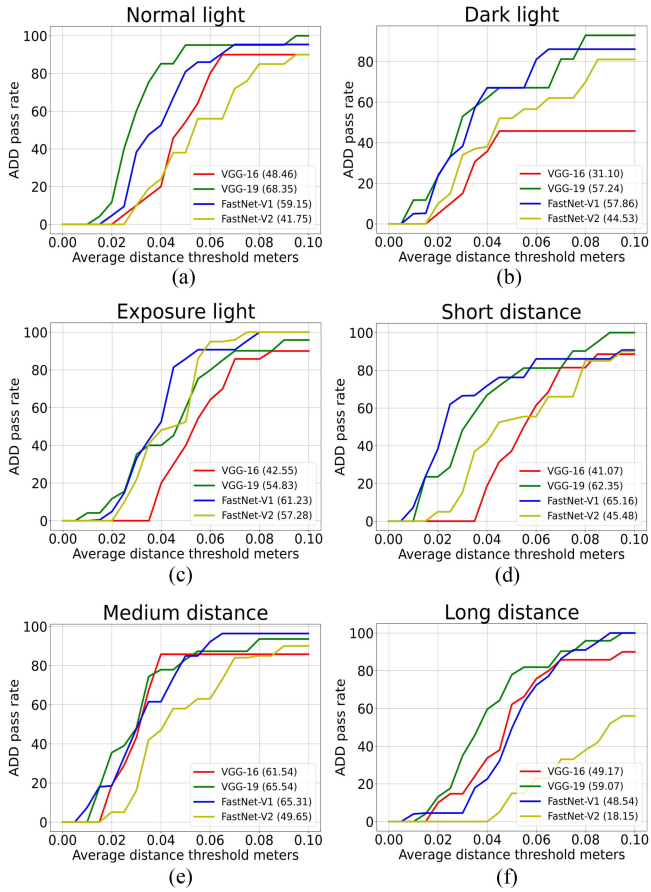


Fig. 8. Accuracy-threshold curves for four networks in difference environment conditions which are (a) normal light, (b) dark light, (c) exposure light, (d) short distance (35–45 cm), (e) medium distance (45–60 cm), and (f) long distance (60–75 cm). On self-made soap dataset. The numbers in the legend display the area under the curve (AUC). The proposed method namely FastNet-V1.

C. Analysis of Experimental Process and Results

1) Pose Estimation Experiment: In this experiment, the soap dataset is used to train four network models uniformly (namely DOPE-VGG-19, DOPE-VGG-16, FastNet-V1, and FastNet-V2), until the loss value converges, the training process is over. The test conditions include unbalanced illumination and different detection distances.

First, the performance of the improved feature extraction networks is compared, including DOPE-VGG-19, DOPE-VGG-16, and FastNet-V1. Unbalanced illumination detection: DOPE-VGG-19 performs best in normal light (AUC = 68.35). However, the FastNet-V1 detects better in dark (AUC = 57.86) and

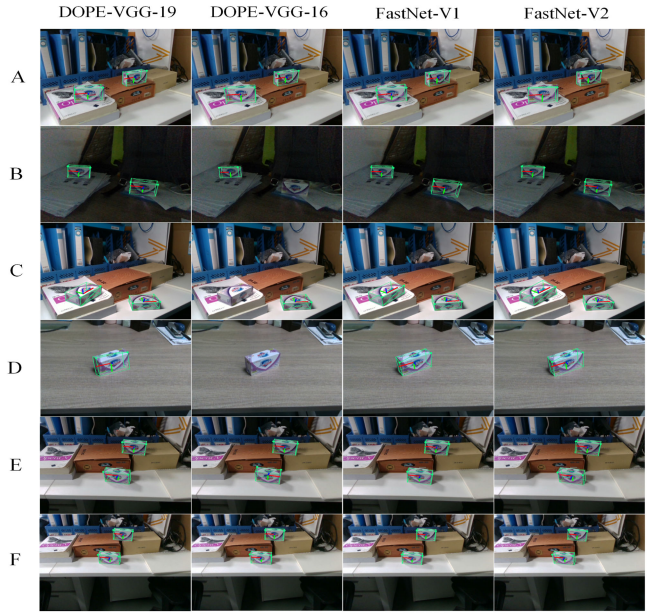


Fig. 9. Four trained models are tested under different real environment conditions. A: Normal light, B: Dark light, C: Exposure light, D: Short distance, E: Medium distance, F: Long distance.

exposed environments (AUC = 61.23). DOPE-VGG-16 performs mediocrity, with both the exposure and the dark environment missing, as shown in B and C in Fig. 9. Different detection distances: DOPE-VGG-19 comprehensively gets better results compared with other networks at medium (AUC = 65.54) and long distance (AUC = 59.07). However, FastNet-V1 is better than other networks at short distance (AUC = 65.16).

Second, the performance of the improved branch networks is compared, including DOPE-VGG-19, FastNet-V1, and FastNet-V2. Unbalanced illumination detection: FastNet-V2 in normal light, dark, and exposure of the overall detection performance is worse than FastNet-V1 and DOPE-VGG-19, only the gap of performance is small (AUC = 57.28) in the exposure condition. Different detection distances: FastNet-V2 in short distance, middle distance, and long distance detection are all worse than the rest of the both, and there is a big gap.

In summary, in terms of improving the branch network, compared with the other two models, because all 7×7 convolutions are replaced with 3×3 convolutions, the calculation and complexity of the network are greatly reduced. FastNet-V2 detection speed is significantly improved, reaching 2.247 frames/s, and the detection speed of DOPE-VGG-19 reached 1.172 frames/s. Obviously there is a great improvement in speed, but the network is in limited depth due to the 3×3 convolution kernel. Insufficient ability to extract features, resulting in the overall detection effect of FastNet-V2 is lower than the average, and the detection stability is insufficient. The detection performance gap between V2 and V1 is far from meeting the requirements of practical applications, so the improvement of V2 is not enough. At the same time, it can be seen from the graph of the experimental results that the performance of FastNet-V1 and DOPE-VGG-19 are similar, but the detection speed of FastNet-V1 is better than that of DOPE-VGG-19, reaching 1.490 frames/s, increasing the

Fig. 10. FastNet-V1 tests objects in multiple scenes detection, illumination imbalanced detection, instance detection and different distance detection results. A: Random selection of detection scenarios, B: illumination imbalance detection, C: Instance detection, D: Multiple distance detection.

TABLE II
AVERAGE SUCCESS RATE OF ROBOT GRASPING UNDER DIFFERENT LIGHTING (%)

Category	Soap	Calcium tablet	Tissue	*Average*
Normal light	84.0	86.0	80.0	83.3
Dark	70.0	76.0	68.0	71.3
Average	77.0	81.0	74.0	77.3

hidden layer information exchange and reasonable reduction of branch network parameters are valuable. Therefore, from the perspective of practical applications, FastNetV1 has advantages over other comparative networks, which is conducive to improving the detection effect of robot grasping, and the detection results of FastNet-V1 in the real world are used as a reference. As shown in Fig. 10.

2) Robotic Grasping Experiment: The robotic grasping success rate is affected by a variety of complex factors. This experiment mainly explores the influence of normal light and dark environmental on the robotic grasping. The robotic arm can grasp three randomly placed objects 50 times, and the same object does not exist in the same scene. Then, calculated the success rate and the result as shown in Table II. The overall grasping process as follows: first, the camera detects the object, the robotic arm approaches the object, then it lifts and moves to the top of the placement box, and the final object is placed into the box.

We find that the overall grasping success rate of normal light is 83.3% higher than that of dark environment, because the image in dark environment is fuzzy, which is not conducive to feature detection. According to the differences in the grasping success rate of the three objects, the highest average grasping success rate is 81.0% for calcium tablets under two kinds of light, followed by 77.0% for the soap and 74.0% for the tissue. Based on the abovementioned results, we found two phenomena. On the one hand, the number of training datasets affected the grasp effect under the same environmental conditions. Specifically, the calcium tablet dataset has 2000 more data than the soap, and the composition ratio of DR data and photorealistic data is

the same. We found that adding synthetic data will not cause the model to overfit the virtual data, on the contrary, the model can learn more useful information from it, thereby shortening the gap between the virtual and the reality and increasing the success rate of grasping. On the other hand, different ratios of DR and photorealistic data will also have an impact on the grasping success rate. Specifically, the total number of calcium tablet dataset and tissue dataset is the same, but the composition ratio of DR data and photorealistic data is different. Under the same number of data, calcium tablets with a composition ratio of 1:1 are better than 7:3 tissues. It is further found that 1:1 soap is also better than 7:3 paper tissues, even if the total number of soaps is 2000 less than calcium tablets.

VI. Conclusion

In this work, a practical robotic grasping method by single image 3-D object pose estimation with protective correction was proposed. In the framework, the multilevel feature from a single image was extracted by FastNet-V1 to replace the original VGG structure in the DOPE method with a residual block. Thus, improving the ability of feature extraction. Thanks to that makes it possible to optimize the branch network and speed up to 1.490 frames/s. The CGP algorithm was added in the framework to determine whether to perform grasp by the error of estimated and measured translation compared with the set threshold. Finally, the success rate of grasping was improved and reducing robotic collision risk. The experiments demonstrate that the robotic grasping success rate of the proposed approach was 83.3% for the self-made dataset under normal light.

However, our method was limited in bridging the gap between synthetic data and real-world data. The comprehensively trained deep neural networks performed poorly in real-world environments where objects were occluded. Our further research aims to improve the branch network structure of the vector field and increase the success rate of grasping in the occlusion environment. Furthermore, trying to use instance segmentation is what we will do next, which can make our network work under more extreme conditions.

References

[1] T. Huynh-The, H. Hua-Cam, and D.-S. Kim, "Encoding pose features to images with data augmentation for 3 d action recognition," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3100–3111, May 2020.

[2] C. Liu, B. Fang, F. Sun, X. Li, and W. Huang, "Learning to grasp familiar objects based on experience and objects' shape affordance," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 49, no. 12, pp. 2710–2723, Dec. 2019.

[3] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: A hands-on survey," *IEEE Trans. Visualization Comput. Graph.*, vol. 22, no. 12, pp. 2633–2651, Dec. 2016.

[4] F. Sun, C. Liu, W. Huang, and J. Zhang, "Object classification and grasp planning using visual and tactile sensing," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 46, no. 7, pp. 969–979, Jul. 2016.

[5] N. Correll *et al.*, "Analysis and observations from the first amazon picking challenge," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018.

[6] W. Sheng, A. Thobbi, and Y. Gu, "An integrated framework for human–robot collaborative manipulation," *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2030–2041, Oct. 2015.

[7] Z. Li, G. Wang, and X. Ji, "CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7678–7687.

[8] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6 d pose estimation of objects and scenes from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3364–3372.

[9] C. Song, J. Song, and Q. Huang, "Hybridpose: 6 d object pose estimation under hybrid representations," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognition*, pp. 431–440, 2020

[10] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich, "Viewpoint-aware object detection and pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1275–1282.

[11] C. Gu and X. Ren, "Discriminative mixture-of-templates for viewpoint classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 408–421.

[12] S. Hinterstoisser *et al.*, "Gradient response maps for real-time detection of textureless objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 876–888, May 2012.

[13] D. G. Lowe *et al.*, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, vol. 99, pp. 1150–1157.

[14] V. Lepetit *et al.*, "Monocular model-based 3 d tracking of rigid objects: A survey," *Found. Trends Comput. Graph. Vis.*, vol. 1, no. 1, pp. 1–89, 2005.

[15] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7448–7458, Dec. 2020.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[17] C. C. Wong, Y. Gan, and C. M. Vong, "Efficient outdoor video semantic segmentation using feedback-based fully convolution neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5128–5136, Aug. 2020.

[18] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6 d object pose estimation using 3 d object coordinates," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.

[19] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3 d poses of challenging objects without using depth," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3828–3836.

[20] Y. Cong, D. Tian, Y. Feng, B. Fan, and H. Yu, "Speedup 3-d texture-less object recognition against self-occlusion for intelligent manufacturing," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3887–3897, Nov. 2019.

[21] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4561–4570.

[22] G. Pavlakos, X. Zhou, A. Chan, K. Derpanis, and K. Daniilidis, "6-DoF object pose from semantic keypoints," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 2011–2018.

[23] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann, "Segmentation-driven 6 d object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3380–3389.

[24] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3 d model views," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2686–2694.

[25] J. Wu *et al.*, "Single image 3 d interpreter network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 365–382.

[26] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum, "Marrnet: 3 d shape reconstruction via 2.5 d sketches," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 540–550.

[27] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Proc. 2nd Conf. Robot Learn.*, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., (Proc. Mach. Learn. Res.), PMLR, vol. 87, pp. 306–316, Oct. 29–31, 2018.

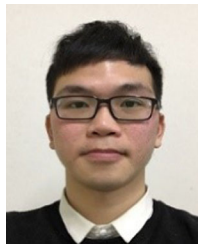
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[29] S. Hinterstoisser *et al.*, "Gradient response maps for real-time detection of textureless objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 876–888, May 2012.

[30] S. Hinterstoisser *et al.*, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 858–865.
[31] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6 d object pose and predicting next-best-view in the crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3583–3592.
[32] W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab, "Deep learning of local RGB-D patches for 3 d object detection and 6 d pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 205–220.
[33] H. Liu, Y. Cong, G. Sun, and Y. Tang, "Robust 3-d object recognition via view-specific constraint," *IEEE Trans. Syst., Man, Cybern. Syst.*, pp. 1–11, 2020, doi: 10.1109/TSMC.2020.2965729.
[34] A. Krull, E. Brachmann, F. Michel, M. Ying Yang, S. Gumhold, and C. Rother, "Learning analysis-by-synthesis for 6 d pose estimation in RGB-D images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 954–962.
[35] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6 d pose estimation of objects and scenes from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3364–3372.
[36] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6 d object pose prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 292–301.
[37] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6 d object pose estimation in cluttered scenes," 2018, *arXiv:1711.00199*.
[38] C. Wang *et al.*, "Densefusion: 6 d object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3338–3347.
[39] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "PVN3D: A deep point-wise 3 d keypoints voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11629–11638.
[40] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.
[41] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review," *Artif. Intell. Rev.*, vol. 54, pp. 1677–1734, 2021.
[42] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 1316–1322.
[43] D. Guo, T. Kong, F. Sun, and H. Liu, "Object discovery and grasp detection with a shared convolutional neural network," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 2038–2043.
[44] D. Park, Y. Seo, and S. Chun, "Real-time, highly accurate robotic grasp detection using fully convolutional neural networks with high-resolution images," 2018, *arXiv:1809.05828*.
[45] A. T. Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *Int. J. Robot. Res.*, vol. 36, pp. 1455–1473, 2017.
[46] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[47] F. Guo, J. C. Deng, and D. B. Zhou, "A connected domain analysis based color localization method and its implementation in embedded robot system," *Int. J. Image Graph. Signal Process.*, Modern Education and Computer Science Press, vol. 3, no. 5, pp. 37–43, 2011.

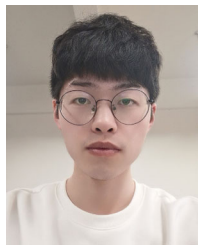
Hui Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent system from Hunan University, Changsha, China, in 2004, 2007, and 2012, respectively.

He is currently an Professor with the School of Robotics and the National Engineering Laboratory for Robot Visual Perception and Control Technology, Hunan University. He was a Visiting Scholar with Common Vulnerability Scoring System Laboratory, Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada, in 2017. His research interests include machine vision, sparse representation, visual tracking.



Zhicong Liang was born in Zhongshan, China, in 1994. He received the B.S. degree in automation from the Changsha University of Science and Technology, Changsha, China, in 2017, where he is currently working toward the master's degree in control engineering with the College of Electrical and Information Engineering.

His current research interests include image processing, deep learning, and industry inspection and so on.



Chen Li was born in Shangqiu, China, in 1996. He received the B.S. degree in automation from Haibin College, Beijing Jiaotong University, Hebei, China, in 2019. He is currently working toward the master's degree in control science and engineering with the College of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha.

His current research interests include image processing, deep learning and so on.

Hang Zhong (Member, IEEE) received the M.S. and B.S. degrees in automation in 2013 and 2016, respectively, from the College of Electrical and Information Engineering, Hunan University, Changsha, China, where he is currently working toward the Ph.D. degree in control theory and application.

His current research interests include robotics modeling and control, visual servo control, and path planning of the aerial robots.

Li Liu (Member, IEEE) was born in 1984. He received the B.S. degree in measurement and control technology and instrument from Southeast University, Nanjing, China, in 2006. He is currently working toward the Ph.D. degree in control science and engineering with the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2019.

His current research interests include robot vision measurement, robot path planning, and intelligent control.

Chenyang Zhao was born in Jiaozuo, China, in 1997. She received the B.S. degree in measurement and control technology from the Harbin University of Science and Technology, Harbin, China, in 2019. She is currently working toward the master's degree in control engineering with the College of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha, China.

Her current research interests include image processing, deep learning, and industry inspection and so on.

Yaonan Wang (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Hunan University, Changsha, China, in 1994.

He was a Postdoctoral Research Fellow with the Normal University of Defence Technology, Changsha, China, between 1994 to 1995. From 1998 to 2000, he was a Senior Humboldt Fellow in Germany. From 2001 to 2004, he was a Visiting Professor with the University of Bremen, Bremen, Germany. Between 2001 to 2020, he was the Dean of the College of Electrical and Information Engineering, Hunan University. Since 1995, he has been a Professor with Hunan University. His current research interests include intelligent control, robotics, and image processing.

Dr. Wang holds the Principle Leader with the National Engineering Laboratory of Robot Visual Perception and Control Technology, Hunan, China. He is the President of China Society of Image and Graphics, Beijing, China. He is a Fellow of Chinese Academy of Engineering.

Q. M. Jonathan Wu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990.

In 1995, he was affiliated with the National Research Council of Canada for ten years beginning, where he became a senior research officer and a group leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has authored or coauthored more than 300 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3-D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks.

Dr. Wu held the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems between 2005 to 2019. He is an Associate Editor for the IEEE TRANSACTION ON CYBERNETICS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal of Cognitive Computation*, and the *Neurocomputing*. He has served on technical program committees and international advisory committees for many prestigious conferences. He is a Fellow of Canadian Academy of Engineering.