Robot visual measurement and grasping strategy for roughcastings

Guoyang Wan1, Guofeng Wang1, Kaisheng Xing2, Yunsheng Fan1 and Tinghao Yi3

Abstract
To overcome the challenging problem of visual measurement and grasping of roughcasts, a visual grasping strategy for an industrial robot is designed and implemented on the basis of deep learning and a deformable template matching algorithm. The strategy helps realize the positioning recognition and grasping guidance for a metal blank cast in complex backgrounds under the interference of external light. The proposed strategy has two phases: target detection and target localization. In the target detection stage, a deep learning algorithm is used to recognize the combined features of the surface of an object for a stable recognition of the object in nonstructured environments. In the target localization stage, high-precision positioning of metal casts with an unclear contour is realized by combining the deformable template matching and LINE-MOD algorithms. The experimental results show that the system can accurately provide visual grasping guidance for robots.

Keywords
Industrial robot, deep learning, object detection, pose measurement, template matching

Date received: 1 February 2021; accepted: 15 February 2021

Introduction
A metal cast is a common processing object in contemporary industrial production. With the popularization of industrial robots, there have been an increasing number of robotic applications combined with vision technology in industrial production.1 The use of industrial robots for processing casts has advantages such as low labor costs, work quality improvement, and optimization of production lines.2–5 An industrial robot workstation based on machine vision can further automate production processes. Machine vision is another research hotspot in robot-integrated applications.

Compared with welding, handling, and other processes, industrial robots are rarely used in the processing of rough metal casts, where a lot of the work depends on manual operation.6 Roughcasts are characterized by multiple burrs, small deformations, and indistinct contours; moreover, the production environment is poor. These factors reduce the stability and accuracy of positioning recognition of vision systems and thus increase their application cost. Under the premise of controlling the hardware cost, it remains challenging to realize a

1 Department of Marine Electrical Engineering, Dalina Maritime University, Dalian, People’s Republic of China
2 Xinwu Economic Development Zone, Anhui Institute of Information Technology, Wuhu, People’s Republic of China
3 University of Science and Technology of China, Hefei, People’s Republic of China

Corresponding author:
Guoyang Wan, Department of Marine Electrical Engineering, Dalina Maritime University, Linhai Road No.1, Dalian 11600, People’s Republic of China.
Email: 704610266@qq.com

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
stable identification and accurate positioning of rough-casts using a vision system.

In terms of equipment, the commonly used visual guidance methods for industrial robots can be divided into two types: two-dimensional (2D) vision based on monocular cameras and three-dimensional (3D) vision based on 3D sensor. Generally, a 3D vision system can provide more detailed information of the measured object; however, compared with a 2D vision system, it has drawbacks such as high cost, low image acquisition speed, and small measurement area. Therefore, a 2D vision system is still preferred for robot vision in terms of practicality.

The main problems of 2D vision in the recognition and localization of rough metal casts are as follows:

1. **There is interference in the image acquisition process:** the processing environment of rough casts is typically characterized by dust and ambient light pollution, which decreases the quality of images obtained using vision systems. In terms of the cost, this problem cannot be solved simply by adding hardware equipment.

2. **Objects are in a cluttered background:** during storage and transportation, numerous scratches and/or stains occur on the surface of rough objects and the surrounding transportation equipment, leading to cluttering of the imaging background. The information contained in the cluttered background can be easily misjudged by the vision system.

3. **Objects have burrs and deformations:** in the process of mass production, some cast materials undergo deformation after forming. In the case of casts in multimode production lines, the die aging and pouring process can cause burr residues and size differences on the object surface. Currently, the commonly used vision-based fine localization algorithms (such as shape-matching and PatMax) in industrial environment depend on the surface profile features of the measured object. Such features are not evident because of the unprocessed nature of roughcasts. Some casts exhibit local reflection due to an uneven surface, which makes visual detection and localization even more difficult.

To overcome the above problems, this article proposes a localization and grasping strategy for industrial robots based on monocular vision. Combined with machine vision target detection and a localization algorithm, we demonstrate the loading and localization of a rough object by an industrial robot in the presence of external light interference and in the case where the object contour is indistinct. The proposed strategy is divided into two stages: target detection and target localization. In the target detection stage, a deep learning algorithm is used to detect the multiple local features of the object and complete the identification and coarse localization of the object in a cluttered background. In the target localization stage, we realize the precise localization of the object with an unclear contour using an improved template matching algorithm. In addition, the system can complete an automated production process with less manual intervention, ensuring the safety of the production process under the premise of controlling pollution.

The contributions of this article are summarized as follows:

1. A new industrial robot grasping strategy for non-structured environments is proposed. This strategy has better localization accuracy and stability than conventional methods.

2. A new detector based on You Only Look Once (YOLOv4), which can be used in industrial areas, is proposed. This detector can realize a fast and accurate detection of similar-sized objects in non-structured environments.

3. Combined with a spring model structure, a deformable LINE-MOD algorithm is proposed for the precise localization of objects with unclear local edges.

The rest of this article is organized as follows: The second section presents the related work. The third section details the key points of the proposed grasping strategy for a particular type of object. The fourth section introduces the main experimental equipment, experimental environment, and experimental analysis. Finally, the article is concluded in the fifth section.

**Related work**

The visual localization of industrial robots has been a hot-spot in industrial research. Qin et al. proposed a robot assembly strategy that combines a 2D camera and a distance sensor to locate and grasp large-volume objects; they employed the strategy for a high-precision assembly of rectangular objects, the industrial robot automatic assembly of this kind of object is realized. Song et al. proposed a Computer-Aided Design (CAD)-based six-degree-of-freedom pose estimation design for random bin picking, which realize high precision robotic grasping to randomly place industrial objects. A binocular stereovision system was used for the 3D positioning of an object in ref. It can realize high precision positioning and grasping of large volume metal workpiece in unstructured environment. The system has been used in industrial production and achieved good results. Le and Lin proposed a 2D localization algorithm for a robot based on deep learning, which enabled an out-of-order grasping of a specific object base on the deep neural network. Literature proposed a cooperative robotics pick and place system for different objects. It solving the perceptual issues posed by cluttered environments though a custom algorithm that selects a suitable picking point.
Target detection methods can be divided into conventional feature-based methods and deep learning-based methods. In the template matching method, a surface feature area of an object is selected, and a model is established, whereby the object to be identified is found in the searched image. LINE-MOD is a gradient feature-based method proposed by Hinterstoisser in 2012. The algorithm cleverly designs matching similarity criteria and uses the Steaming SIMD Extensions (SSE) parallel computing of C2 Steaming SIMD Extensions (SSE) parallel computing of C2 Steaming SIMD Extensions (SSE) parallel computing of C2 Steaming SIMD Extensions (SSE) parallel computing of C2 Steaming SIMD Extensions (SSE). Since its proposal, LINE-MOD has received wide attention. Edward and Hebert proposed a LINE-MOD algorithm for occlusion model optimization to solve a common occlusion problem in target detection. Some scholars have integrated the LINE-MOD model patch into a random forest to jointly estimate the position and orientation of objects. Zhang et al. proposed a CT-LINE-MOD algorithm combined with the cognitive template-clustering method. The above algorithms have improved the performance of the LINE-MOD algorithm from different aspects.

Deep learning has emerged as a popular machine learning method in recent years. Since AlexNet, which was proposed by Alex et al., won the 2012 ImageNet Challenge with an absolute advantage, deep learning has been widely studied and applied in the field of image processing. Unlike traditional template matching, target detection algorithms based on deep learning do not rely on a single feature to achieve target detection, so those methods are more capable of recognizing targets in complex environments. The object detector base on deep neural network can be divided into two categories: two-stage detection and one-stage detection. In the former case, the detection frame is defined as a "from-coarse-to-fine" process, whereas it is defined as a "one-step carry out" in the latter case. Generally, a two-stage network has better detection and positioning accuracy; however, the one-stage network is more simple and the speed is higher. Representative two-stage detection networks include spatial pyramid pooling net (SPPNet), fast region based convolutional neural networks (RCNN), faster RCNN, and feature pyramid networks (FPNs). One-stage methods include YOLO, single-shot multibox detector, and RetinaNet.

In different environments, different detection objects have different characteristics. Currently, scholars began to use deep learning detector to address specific problems in different applications. In the aspect of target detection and classification, Kang and Chen tried to apply the target detection algorithm to autonomous harvesting and realized the visual classification and automatic picking of mature fruits. Literature proposed an automated visual inspection detector for trains exploring structural knowledge for train component detection and the method achieves pleasurable detection performance for the specified object. In addition, some scholars use different methods of image enhancement on the basis of target classification. A novel framework that applies deep learning to improve gas tungsten arc welding process monitoring and penetration detection using multi-source sensing images is presented in ref. The framework is capable of analyzing multiple types of optical sensing images synchronously and generative adversarial networks (pix2pix) are adopted for image denoising. Literature combining generative adversarial networks and conventional approaches for image enhancement and then the welding seam detection is realized by YOLOv3 detector. Through the analysis of the above application, according to the characteristics of the target object, the improved target detector can get better detection effect than using the existing detection network directly.

Compared to the traditional target detection algorithm, deep learning localization algorithm does not have high accuracy. In addition, for the same object, many target detection networks pay more attention to the target scaling than rotation. In the industrial field, especially for the visual guidance application of industrial robots, the situation is just the opposite. Those are the common problems in the visual positioning of the blank metal casts.

To address the above problems, this article proposes a visual guidance strategy for industrial robots realized by combining the advantages of both template matching and deep learning. The strategy involves first applying the deep learning method for object recognition, followed by the template matching method for fine localization. More specifically, the deep learning method is used to separate the local features of the measured object from the cluttered background, and subsequently, the template matching algorithm is used to precisely locate the local features of the object. The proposed strategy has the advantages of good generalization performance of deep learning target detection algorithm and high positioning accuracy of template matching method. Combined with the industrial robot grasping workstation, the system realizes the high-precision and stable grasping of the blank metal casts and it solves the problems of poor robustness and low positioning accuracy in the process of grasping directly using vision guidance in a traditional robot workstation.

Grasping strategy

A rough metal cast, as shown in Figure 1, is used as the object to introduce the proposed grasping strategy. The size of the object is 209 mm × 155 mm, and the weight is 7 kg. The object is made of cast iron, and there are some burrs and stains on its surface. Because of multimode production and material characteristics, there are slight deformations on the object. The object is placed on a wooden pallet for repeated use according to certain rules, and the rotation angle around the vertical axis is no more than ±5°. The production environment is a conventional casting production workshop, which has dust and ambient light pollution.
due to the grinding of casts in the workshop. The robot needs to accurately position the object in sequence and complete the grasping task in a cluttered background.

The template matching algorithm has advantages such as fast execution, stable matching effect, and high positioning accuracy. However, in cluttered and occluded environments, such algorithms exhibit low accuracy and recognition rate. A deep learning detector relies on a large amount of data for training. With sufficient training, the recognition rate of deep learning algorithms in cluttered environments is better than those of conventional methods; however, the accuracy is still insufficient. In this study, we combine the advantages of the two algorithms and propose a two-step positioning algorithm from coarse to fine (see Figure 2).

**Rough workpiece feature positioning based on YOLOv4**

The YOLO is an end-to-end target detection network proposed by Joseph in 2016. Since its introduction, the network has been favored for its high speed and accuracy. The YOLO network has been developed from v1 to v4. It has been developed on the basis of YOLOv4. Its network structure is in the following.

Compared with the previous YOLOv3 network, the YOLOv4 network has the following main changes:

**Backbone network.** The backbone network is changed from DarkNet53 to Cross Stage Partial (CSP) DarkNet53 (see Figure 3). CSPDarkNet53 uses the Res(x) Block_body structure, which can transfer more upper-layer information in the backbone network to the lower layer. Its activation function is changed from LeakyReLU to a continuously differentiable Mish function.

\[
\text{LeakyReLU} : f(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{otherwise} \end{cases} \quad (1)
\]

\[
\text{Mish} : f(x) = x \cdot \tanh(\text{softplus}(x)) \quad (2)
\]

The Mish activation function is smoother than the LeakyReLU activation function and has a comprehensive improvement in terms of training stability, average accuracy, and peak accuracy.

**Feature enhancement part.** The Spatial Pyramid Pooling Net, (SPPNet) is used, and the PAN net replaces the FPN net in Figure 1.

![Figure 1. Object in a cluttered background.](image)

![Figure 2. Grasping strategy workflow.](image)
YOLOv3. The SPP uses multilevel spatial bins, which have been shown to be robust to object deformations, and it can pool features extracted at variable scales.\textsuperscript{39} For different levels of detectors, a Path Aggregation Network (PANet) net from different backbone layers was selected as a parameter aggregation method. More accurate segmentation results can be outputted by fusing the two branches.

**Prediction output part.** YOLOv4 still retains the output part of the three different scales in YOLOv3.

The YOLOv4 network has had additional improvements, such as the proposal of a new data enhancement method Mosaic and self-adversarial training and selection of the optimal hyperparameters when applying genetic algorithms.

To make the YOLOv4 algorithm more applicable to the classification of blank metal casts, we have made the following modifications:

1. **A data set for industrial objects is established, including several common industrial objects:** The classification problem in industrial belongs to small sample learning problem. Due to the limitation of production cost, it is difficult to collect a large number of training samples to train the deep neural network model. However, the detection performance of neural network depends on the number of training samples.\textsuperscript{34} To ensure the target recognition effect of the detector, we enhance the target image to simulate the imaging effect under external interference in an industrial environment. The image enhance method include image rotation, shading adjustment, blur, and Gaussian noise (see Figure 4). The industrial data set can not only be used to detect the performance of the network but also be used as the pretraining.

2. **Net structure change:** As mentioned in ref.\textsuperscript{40} the accuracy of a deep learning detector is determined by three factors: network depth, network width, and input image resolution. From the perspective of enhancing the learning ability of a CNN, Hu et al.\textsuperscript{41} proposed a CSPDenseNet structure based on DenseNet and proved that the CSPNet structure significantly improves the DenseNet201 detector. A CSPDenseNet structure is used in the detector, and the activation function is changed to the Mish function; an Squeeze-and-Excitation (SE) module is also added (see Figure 5(c)).

The SE module was added in YOLOv4 to enhance the performance of the detector. The objective of the SE module is to automatically obtain the importance of each feature channel through learning and then enhance the important features and suppress invalid features based on this importance. The SE module mainly includes a squeeze part and an excitation part. The first is the conversion operation based on the original net; we have $Ftr : X \rightarrow U, X \in \mathbb{R}^{W \times H \times C}, U \in \mathbb{R}^{W \times H \times C}$, where $Ftr$ is the feature...
transformation from input $X$ to output $U$. The calculation formula of $Ftr$ is as follows

$$u_c = v_c \ast X = \sum_{s=1}^{C} v^s_c \ast x^s$$  \hspace{1cm} (3)

where $\ast$ represents the convolution operation, $v_c$ represents the $c$th convolution kernel, and $x^s$ indicates the $s$th input. Thereafter, the squeeze is calculated as follows

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_{c}(i,j), z \in R^C$$  \hspace{1cm} (4)

Equation (4) converts the input of $H \times W \times C$ to the output of $C$. For excitation, we have

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \text{ReLU}(W_1 z)),$$

$$W_1 \in R^{C \times C}, W_2 \in R^{C \times C}$$  \hspace{1cm} (5)

The $s$ value is used to characterize the weight of the $C$ feature maps in $U$. Finally, the weights are applied to the original features on $U$

$$x_c = F_{scale}(u_c, s_c) = s_c \cdot u_c$$  \hspace{1cm} (6)

The two areas of object which name feature $M$ and feature $H$ are selected for object detection. Considering that the size of the feature area relative to the input image is relatively fixed and the output of the detector for small target detection is removed (see Figures 6 and 8).

The improved detector is named SE-YOLOv4 (see Figure 7). SE-YOLOv4 has a better recognition effect for the local features of the roughcasting object. The false detection rate during target detection can be further reduced using combined features.

**Fine localization based on deformable Line-Mod (D-LINE-MOD)**

The SE-YOLOv4 detector can well identify the object despite the influence of external interference and ambient light; however, the localization accuracy fluctuates significantly. To improve the localization and grasping accuracy of the rough object, a deformable LINE-MOD algorithm is
proposed for precise localization of the roughcast on the basis of SE-YOLOv4.

In the process of 2D image localization, gradient information is used to generate a 2D model in the LINE-MOD algorithm. The similarity between the input image and the model is calculated using the following equation

\[
\varepsilon(I, T, c) = \sum_{r \in R} \left( \max_{t \in R(c+r)} |\cos(\text{ori}(O, r) - \text{ori}(I, t))| \right)
\]  

(7)

Figure 6. (a) 13 × 13 Meshing of a workpiece and local features detected by YOLOv4; the yellow area is the feature H and the blue area is the feature M and (b) 10 × 10 meshing diagram. YOLO: You Only Look Once.

Figure 7. Improved YOLOv4 network. YOLO: You Only Look Once.
where \( R(c + r) = [c + r - \frac{T}{2}, c + r + \frac{T}{2}] \times [c + r - \frac{T}{2}, c + r + \frac{T}{2}] \) is the neighborhood of size \( T \) centered on location \( c + r \) in the input image. \( \text{ori}(O, r) \) is the gradient of position \( r \) in the model image \( O \). \( I \) is the image to be matched and \( P \) is the set of positions \( r \) to be calculated in \( O \).

Because the object surface exhibits reflection and unevenness, we may not obtain good results if the LINE-MOD algorithm is directly applied to locate the feature \( H \) of the object. Most of the improved LINE-MOD algorithms used to detect and locate objects are based on the local features of the object surface; any rigid deformation on the object is out of focus of these algorithms. Because of ambient light influence and burrs on the object surface, the change in the contour characteristics of the object is similar to nonrigid deformation during imaging. Inspired by the idea of deformable part models, \(^{42}\) a deformable LINE-MOD algorithm that can realize the precise localization of objects without evident contours is proposed. The main functions of the proposed algorithm are as follows:

1. Based on the local feature shape of the object, a gradient contour model is designed using the geometric method to avoid information loss due to external interference in the model generated directly from the image.
2. A local feature model is generated from the feature gradient contour model.
3. By using a spring model to identify any slight deformation of the contour, we can locate a rough metal casting with local contour deformation in the imaging due to ambient light.

**Generation of feature points of gradient contour.** The object contour feature points of the LINE-MOD method are selected from the gradient intensity and gradient direction; however, this method is ineffective for rough objects such as the one proposed in this article. The LINE-MOD method does not obtain enough feature points on the edge of objects, and some feature points are distributed on the object surface, which cannot provide effective information for template matching calculation (see Figure 8).

A feature points generation method based on geometric features is proposed to solve the above problems. The proposed method modifies the original LINE-MOD feature extraction method based on the gradient radial strength and gradient direction to a feature point extraction method based on contour features. The steps are as follows:

1. The Sobel operator is used to calculate the image gradient intensity and gradient direction of the workpiece feature area

\[
\begin{align*}
G_x &= \text{sobel}_x(I) \\
G_y &= \text{sobel}_y(I) \\
G &= G_x^2 + G_y^2 \\
m\text{_angle} &= \tan^{-1}\left(\frac{G_y}{G_x}\right)
\end{align*}
\]

where \( I \) is the input image, \( G \) is the gradient intensity, and \( m\_\text{angle} \) is the gradient direction of the image.
2. The feature points set near the contour of the feature H region are obtained using threshold filtering and region area filtering

\[
GI = \begin{cases} 
G, & \text{if } G \geq t \\
0, & \text{otherwise}
\end{cases}
\]

where \( t \) is the threshold of the gradient intensity. The connected domain in \( GI \) is calculated after obtaining \( GI \). The largest area in \( GI \) is considered the main feature area of

\[
\text{feature point sets in different directions. An annular model that reflects the shape characteristics of the H feature of the object is designed and used for the fine positioning of the objects (see Figure 9).}
\]

The straight lines and shape features similar to the straight lines in the background are considered gradients in the same direction. Therefore, a single-direction feature point set cannot be used to construct a submodel. To improve the recognition rate of the object, three point sets with similar positions in the feature points set are combined into a submodel. There are intersections between adjacent submodels and 16 submodels are obtained. The relative position between each submodel is measured using a spring model and their relative positions are limited to a certain area.

\[
\varphi_i = (x_{i-1}, y_{i-1}, \theta_{i-1}) + (dx_{i-1}, dy_{i-1}, d\theta_{i-1})
\]

where \( \varphi_i \) is the variation range of the position of the submodel \( i \) relative to the adjacent model; \( (x_{i-1}, y_{i-1}, \theta_{i-1}) \) is the positional relationship between the center point of the submodel \( i \) and the adjacent model \( i-1 \); \( (dx_{i-1}, dy_{i-1}, d\theta_{i-1}) \) is the range of acceptable position changes between them. Each submodel allows a certain range of elastic changes to adapt to the local deformation and small-scale changes in the object due to changes in the external ambient light and/or changes in the size of the object itself and burrs, while imaging the object.

\[
S_{\text{main}} = \sum S_{ci} \times w \times k_i \quad i = 1, ..., 16
\]

When the positional relationship between the submodel and the adjacent model satisfies equation (11), \( k_i \) is 1; otherwise it is 0. \( w_i \) is the weight for each submodel, which is
determined by the number of submodels that satisfy equation (11).

**Calculation of matching position.** Because a robot uses the relative position of the object, the relative position of the matching feature area of the object can also be calculated. In the actual positioning process, the calculation method of the position of the feature area $H$ is as follows

$$P_{t} = P_{m} - P_{d}$$

$$P_{d} = \sum_{k} k_{i}(P_{i} - P_{i}')$$

$$P_{m} = \sum_{n} P_{i}$$

where $P_{t}$ is the offset of the robot grasping model position. For each submodel, $P_{i}$ is its central position; then, $P_{m}$ is the average of the central positions of each submodel, where $n$ is the number of subtemplates. $P_{i}'$ is the center position after submodel matching, and $P_{d}$ is the average error of all the submodels satisfying equation (11) during the matching process.

**Calculation of grasping pose of industrial robot**

The vision system is installed in an “eye-in-hand” manner. Zhang’s calibration method is used to complete the calibration of the relative position of the camera and the robot by imaging a fixed calibration board. For the conversion relationship between the posture measured by the vision system and the posture grasped by the robot, when the robot grasps the object

$$U^{T}_{T} + U^{T}_{1} + U^{T}_{T}$$

where $U^{T}_{T}$ is the posture of the 6th axis flange in the robot coordinate system when the robot grasps the object, that is, the robot grab posture. $T_{T}$ is the position of the object in the robot grasp coordinate system.

When the robot obtains images to the same object, we have

$$U^{T}_{P} = U^{T}_{2}^{2} \times C^{T} C \times C^{T} P$$

where $U^{T}_{2}^{2}$ is the posture of the robot when the vision system takes an image, $C^{T} C$ is the posture of the vision system relative to the end of the robot flange, which can be obtained by Zhang’s hand-eye calibration method and $C^{T} P$ is the posture of the object relative to the vision system, which is directly acquired by the vision system. Through equations (14) and (15), we can obtain

$$T_{P} = \left[ U^{T}_{T} \right]^{-1} \times U^{T}_{2}^{2} \times C^{T} C \times C^{T} P$$

When the robot takes images of the workpieces in different positions, we have

$$U^{T}_{T} = U^{T}_{2}^{2} \times C^{T} C \times C^{T} P \times [T_{P}]^{-1}$$

where $C^{T} C$ and $U^{T}_{2}^{2}$ are the coordinates of the object detected by the vision system, and $U^{T}_{T}$ is the posture of the robot grasping the object (see Figure 10).

**Experiment**

**Experimental conditions and environment**

An experiment was carried out in the production workshop of a foundry. The proposed grasping strategy was used for a loading robot in the robot grinding station. The loading robot is equipped with an ER50-C10 six-axis industrial robot with a 50 kg load. A Basler brand ace-3800 series industrial camera with the resolution of 3840
$2780$ is used as part of the vision system. It can provide three-channel color image. The vision system is installed at the end effector of the six-axis industrial robot (see Figure 11). The figure shows that the robot grasping system is placed in an unstructured environment. The system visually detects and positions the object placed on the pallet. The acceptable grip error is $\pm 1.5$ mm.

**Experimental analysis of target detection based on deep learning**

We use three objects in the industrial object data set to test the SE-YOLOv4 and YOLOv4, and the detector parameters and test results are presented in Tables 1 and 2 and Figure 12. To compare detectors better, all detectors do not use pretraining model.

Recall and precision are used to evaluate the detector performance; the recall and precision is expressed as follows

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (18)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (19)
\]

TP is to predict positive class as a positive class number, FP is to predict negative classes to positive classes, and FN is to predict positive classes to negative classes.

It can be seen from Tables 1 and 2 and Figure 12 that the SE-YOLOv4 has the same recall with YOLOv4. When an object is tested individually but the precision rate is more better than YOLOv4, this proves that SE-YOLOv4 is more suitable for the detection of industrial objects. In addition, compared with YOLOv4, SE-YOLOv4 has smaller weights size and similar detection speed.

Through the experimental comparison, SE-YOLOv4 achieves better detection effect than YOLOv4 in the detection of industrial objects by improving the network structure, designing the output size, and other methods.
Experimental analysis of object in this study

A roughcasts data set comprising 400 training images and 100 test images was created for the object in this study. The images in the data set were collected from different workpieces under different lighting conditions. The data were used to train standard YOLOv4 and the improved SE-YOLOv4 detector proposed in this article; Table 3 lists the results. The accuracy means the feature area in the image is correctly detected, and error means the feature area is not detected correctly.

YOLO: You Only Look Once. The loss function can be used to assess the degree of inconsistency between the predicted value and the true value of the model. The lower the loss function, the better the robustness of the representative model. Figure 13 shows the changes in the loss functions of the standard YOLOv4 and SE-YOLOv4 when using the same data set for training. As shown, during the training process, the loss function of SE-YOLOv4 decreases faster than that of the standard YOLOv4. Because the network output in SE-YOLOv4 is changed from three channels to a single channel, fewer training functions were required. Therefore, the loss function of this network converges faster during the training process. Figure 14 shows the detection effect of SE-YOLOv4.

The above experiments show that SE-YOLOv4 can complete the detection of the two features required for the detection of the surface of the rough metal cast mentioned in this article.

Table 3. The roughcasts object in this article test result.

| Method      | (200/epoch) Accuracy | (200/epoch) Error | (400/epoch) Accuracy | (400/epoch) Error |
|-------------|----------------------|-------------------|----------------------|-------------------|
| YOLOv4      | 94                   | 6                 | 98                   | 2                 |
| SE-YOLOv4   | 96                   | 4                 | 99                   | 1                 |

Figure 12. The detect result of SE-YOLOv4 and YOLOv4. The first line is SE-YOLOv4 and the second line is YOLOv4. (a) Test result of object 1. (b) Test result of object 2. (c) Test result of object 3. YOLO: You Only Look Once.

Figure 13. Comparison between standard YOLOv4 and SE-YOLOv4 loss functions. YOLO: You Only Look Once.
De-LINE-MOD positioning experiment

We selected 80 different workpiece images obtained by the vision system under different environments and used the standard LINE-MOD algorithm and the improved De-LINE-MOD algorithm for positioning calculation for the feature area $H$. Because the proposed algorithm does not involve a 3D function, we only evaluated the 2D function of the algorithm.

Figures 15 and 16, respectively, show the matching results of the De-LINE-MOD algorithm and the standard LINE-MOD algorithm in the feature areas of different...
objects in different situations. The areas where different objects are used for matching and comparison are screened by SE-YOLOv4 positioning. To ensure the comparison effect, we set the minimum detection score to 20. As shown in Figure 15, the standard LINE-MOD algorithm has a low detection score and large fluctuations between different objects. Figure 16 shows that the LINE-MOD algorithm has only a few feature points that are correctly matched but relies on the coarse positioning calculation of SE-YOLOv4. The LINE-MOD algorithm can still successfully locate the object, which proves the effectiveness of the proposed grasping strategy. When using the De-LINE-MOD algorithm for the localization of local features, the matching score is significantly improved. Figure 14 shows that a large number of feature points are accurately matched.

To test the accuracy of the proposed strategy, a single object is placed on the ground. The object is imaged using the vision system on the robot, and the grasping posture of the object is calculated. The visual positioning results are optimized using the iterative closest point (ICP) algorithm. Table 4 lists the results.

| Method                        | X (mm) | Y (mm) | C (°) |
|-------------------------------|--------|--------|-------|
| SE-YOLOv4 + LINE-MOD + ICP    | ±4.5   | ±3.7   | ±4    |
| SE-YOLOv4 + De-LINE-MOD + ICP | ±1.2   | ±1.4   | ±0.5  |

YOLO: You Only Look Once. The proposed De-LINE-MOD algorithm has better positioning accuracy than the standard LINE-MOD algorithm. The De-LINE-MOD algorithm optimizes the extraction of the feature points and the use of variable templates in the calculation process, making the ICP iteration process more accurate. The final positioning accuracy is also better.

The above analysis shows that even under ambient light interference, the proposed algorithm can stably identify the object features in the local area.

**Grasping strategy test**

The proposed grasping strategy is applied to carry out continuous production testing on an industrial site. To meet the production cycle of the post-production process of the production line, the robot adjusts the running speed such that an object is grasped every 14 s. The loading process includes the vision system image acquisition, robot grasping pose calculation, and robot operating track running time. The overall workstation can ensure that the output per hour is no less than 200 pieces.

**Conclusions**

In this study, we focused on an industrial robot vision positioning system for rough metals in an industrial environment and developed a novel robot vision grasping guidance strategy. The proposed strategy could achieve stable
grasping of roughcastings with burrs on the surface and excessive and inconspicuous edges in unstructured environments. The proposed positioning and grasping strategy can be divided into two stages: target detection and target localization. In the target detection stage, the feature-enhanced YOLOv4 deep learning network is used to complete the identification and coarse localization of objects in a cluttered background; in the target positioning stage, combined with the elastic template theory in the deformable model, an improved LINE-MOD matching algorithm is proposed for precise positioning of the object. We also analyzed the transformation relationship between the visual coordinate system and the robot coordinate system during the grasping process of a 2D robot system. Finally, an experimental analysis was conducted, and the reliability of the proposed grasping strategy was proven by comparing it with conventional algorithms. This article provides a good solution in the form of a grasping strategy for the positioning and grasping of other similar metal casts in industrial environments.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by National Natural Science Foundation of China under Grant 51609033, Natural Science Foundation of Liaoning Province under Grant 20180520005, the Key Development Guidance Program of Liaoning Province of China under Grant 2019JH8/10100100, the Soft Science Research Program of Dalian City of China under Grant 2019J11CY014 and Fundamental Research Funds for the Central Universities under Grant 3132019005, 3132019311.

ORCID iDs
Guoyang Wan https://orcid.org/0000-0002-6700-9056
Yunsheng Fan https://orcid.org/0000-0001-7828-4902

References
1. Gołda G and Kampa A. Modelling of cutting force and robot load during machining. Adv Mater Res 2014; 1036: 715–720.
2. Shi-ming J and Xi-huan H. Review of development and application of industrial robot technology. J Mech Electr Eng 2015; 32: 1–13.
3. Ma WP, Li WX, and Cao PX. Binocular vision object positioning method for robots based on coarse-fine stereo matching. Int J Autom Comput 2020; 17: 562–571.
4. Lukáš Pilný GBLD. Acoustic emission-based in-process monitoring of surface generation in robot-assisted polishing. Int J Comput Integ 2015; 23: 1–9.
5. Harada T. Design and control of a parallel robot for mold polishing. In: The 3rd International Conference on Control, Mechatronics and Automation (ICCMCA 2015), vol. 42, March 2016, pp. 03–10.
6. Ma K, Han L, Sun X, et al. A path planning method of robotic belt grinding for workpieces with complex surfaces. IEEE-ASME Trans Mechatron 2020; 25: 728–738.
7. Krishnan MG, Vijayan AT, Sankar A, et al. Performance enhancement of two-camera robotic system using adaptive gain approach. Ind Robot Int J 2019; 47(1): 45–56.
8. Ulrich M, Follmann P, Neudeck J, et al. A comparison of shape-based matching with deep-learning-based object detection. TM-Tech Mess 2019, 86(11): 685–698.
9. Oh J, Lee S, Lee C, et al. Stereo vision based automation for a bin-picking solution. Int J Control Autom Syst 2012; 10(2): 362–373.
10. Qin Z, Wang P, Sun J, et al. Precise robotic assembly for large-scale objects based on automatic guidance and alignment. IEEE Trans Instrum Meas 2016; 65: 1398–1411.
11. Song K-T, Wu C-H, Jiang S-Y, et al. CAD-based pose estimation design for random bin picking using a RGB-D camera. J Intell Robot Syst: Theory Appl 2017; 87: 455–470.
12. Wan G, Li F, Zhu W, et al. High-precision six-degree-of-freedom pose measurement and grasping system for large-size object based on binocular vision. Sensor Rev 2020; 40(1): 71–80.
13. Le T and Lin C. Bin-picking for planar objects based on a deep learning network: a case study of USB packs. Sensors 2019; 19(16): 3602.
14. D’Avella S, Tripicchio P, and Avizzano CA. A study on picking objects in cluttered environments: exploiting depth features for a custom low-cost universal jamming gripper. Robot Comput Integ Manuf 2020; 63: 101888.
15. Ulrich M, Steger C, and Baumgartner A. Real-time object recognition using a modified generalized Hough transform. Pattern Recognit 2003; 36(11): 2557–2570.
16. Hinterstoisser S, Cagniart C, Ilic S, et al. Gradient response maps for real-time detection of textureless objects. IEEE Trans Pattern Anal Mach Intell 2012; 34(5): 876–888.
17. Hinterstoisser S, Lepeitl Y, Ilic S, et al. Dominant orientation templates for real-time detection of texture-less objects. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010. IEEE.
18. Edward H and Hebert M. Occlusion reasoning for object detection under arbitrary viewpoint. IEEE Trans Pattern Anal Mach Intell 2014; 36(9): 1803–1815.
19. Tejani A, Tang D, Kouskouridas R, et al. Latent-class Hough forests for 3D object detection and pose estimation. In: European conference on computer vision, Zurich, Switzerland, 6–12 September 2014, pp. 462–477, Cham: Springer, 2014.
20. Zhang T, Yang Y, Zeng Y, et al. Cognitive template-clustering improved LINEMOD for efficient multi-object pose estimation. Cogn Comput 2020; 12: 834–843.
21. Wu X, Sahoo D, and Hoi SCH. Recent advances in deep learning for object detection. Neurocomputing 2020; 396: 39–64.
22. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer vision and pattern recognition, Columbus, OH, USA, 23–28 June 2014, pp. 580–587.
23. Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell 2016; 38(1): 142–158.
24. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 2015; 37(9): 1904–1916.
25. Girshick R. Fast R-CNN. In: 2015 IEEE international conference on computer vision (ICCV). IEEE, Santiago, Chile, 7–13 December 2015.
26. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2015; 39(6): 1137–1149.
27. Lin T-Y, Dollar P, Girshick RB, et al. Feature pyramid networks for object detection. In: CVPR, vol. 1(2), Honolulu, HI, USA, 21–26 July 2017, p. 4.
28. Redmon J, Divvala S, Girshick R, et al. You Only Look Once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016, pp. 779–788.
29. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multi-box detector. In: European conference on computer vision. Amsterdam, The Netherlands, 11–14 October 2016, pp. 21–37. Cham: Springer.
30. Lin T-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020; 42: 318–327.
31. Kang H and Chen C. Fruit detection and segmentation for apple harvesting using visual sensor in orchards. Sensors (Basel, Switzerland) 2019; 19(20): 4599.