2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
June 28-30, 2023. Seattle, Washington, USA

Image Foreground Segmentation Based on Small Data Set for Visual Servo Applications

Yan Luo, Gaoming Chen, Chao Liu, and Zhenhua Xiong

Abstract—**Extraction of features is a key process in image-based visual servo. However, existing image processing methods are difficult to segment the target foreground and cannot overcome distracting factors, such as background and illumination, resulting in reduced accuracy of feature extraction. Therefore, target foreground segmentation is a critical problem in image-based visual servo tasks. In this paper, a method for image foreground segmentation and visual servo control based on small data training is proposed. Semantic segmentation is achieved by training a small number of images. Focusing on the target artefact region and blurring the background are also achieved to avoid its influence on feature recognition, especially for industry parts. It is shown that recognition and segmentation under different lighting conditions can be obtained, reducing the interference of lighting on visual servo. Experimental results show that the proposed method is effective in visual servo control applications.**

Index Terms—**Feature extraction, foreground segmentation, small data set, visual servo.**

I. Introduction

Visual servo using machine vision and robots is a common task in industry applications. According to the usage of different error signals, visual servo can be classified into position-based visual servo (PBVS), image-based visual servo (IBVS), and hybrid visual servo (HVS). The PBVS method obtains the target geospatial position and motion parameters by image 3D reconstruction of the target, which can reflect the target motion in coordinate space more intuitively. However, its positioning accuracy is highly dependent on the calibration and attitude estimation accuracy [1]. The HVS method represents the current error with respect to the desired target by decoupling the single-response matrix into the corresponding position and rotation components. Although this method can improve the stability, it requires online estimation of the depth of each feature point, which is hard to implement in real time and also more sensitive to image noise interference [2]. The IBVS method uses the error of the features in the image as the control signal, so there is no need to obtain the spatial location of the target and the computational complexity is low. However, this method imposes higher requirements on the avoidance of interferences from background, illumination and the extraction of target features [3]. Among the three visual servo methods mentioned above, the IBVS method is more widely used. Therefore, attention needs to be paid to extracting target features more robustly.

Traditional methods of image segmentation and feature extraction include threshold segmentation, watershed algorithm [4], edge detection, corner point detection, etc. But they are more sensitive to noise and require high contrast between target foreground and background. It is worth noting that the above methods are all computed with the whole picture and cannot eliminate the interference of similar features in the background environment.

With the rise of deep learning, techniques such as target detection and semantic segmentation are rapidly developing. Target detection can distinguish targets by outputting a bounding box, while semantic segmentation provides pixel-level masks for target objects, adapts to different shapes and even occluded target objects to improve accuracy. The GrabCut algorithm [5], improved from the GraphCut algorithm, could perform foreground segmentation on colour images. This method requires the user to provide a priori target box positions, and has difficulty achieving good results on foreground segmentation of non-visible objects, which is time-consuming and less robust. From the R-CNN object detection algorithm proposed by Girshick [6] to the feature pyramid network FPN [7] proposed by Tsung-Yi Lin based on Faster RCNN, to the YOLO algorithm [8] to predict the class probability and bounding box of each grid, the speed and accuracy of object detection methods have been continuously improved. From the Olaf Ronneberger team, which improved on the FCN model and proposed the U-Net method [9], to the DeepLab family of models [10] proposed by Liang-Chieh Chen to achieve pyramidal pooling of voids in spatial dimensions through void convolution, to the Mask R-CNN algorithm [11] proposed by Kaiming He to predict the segmentation mask in a pixel-to-pixel manner, major breakthroughs in the accuracy and fine-grained granularity of semantic segmentation are obtained.

Considering the uniqueness of artifacts in industry, this paper proposes a target foreground extraction algorithm for small dataset training by fusing YOLOv7 target detection algorithm and Mask R-CNN semantic segmentation algorithm. The accuracy and robustness of target foreground extraction can be improved, and the interference of background and illumination on target feature extraction can be avoided while reducing the labeling cost of datasets. Based on the above target foreground extraction results, the workpiece features are then extracted by methods such as Hough transform linear detection and used as input in a BP neural network. The output of the corresponding control rate can be realized in visual servo task,

This work was supported in part by the Major Science and Technology Projects for Self-Innovation of FAW (20210301032GX).

Yan Luo, Gaoming Chen, Chao Liu, and Zhenhua Xiong are with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China (e-mail: luoyan99@sjtu.edu.cn; cgm1015@sjtu.edu.cn; aalon@sjtu.edu.cn; mexiong@sjtu.edu.cn)

Corresponding author: Zhenhua Xiong, Chao Liu.

978-1-6654-7633-1/23/$31.00 ©2023 IEEE

which can accelerate the visual servo process and improve the servo accuracy by pre-training.

The main contributions of this paper are: 1. For the loss function in YOLOv7, Focal-EIOU calculation is used instead of CIOU to improve the target box recognition accuracy; 2. Feature comparator algorithm is designed to realize target foreground segmentation and enhance the robustness of feature extraction 3. Combining BP neural network for servo control to speed up the visual servo process and improve the servo accuracy.

The rest of the paper is organized as follows. Section II presents the related problems and the overall framework. Section III elaborates the fused target foreground segmentation algorithm and visual servo control method. Experimental results and comparisons are given in Section IV. Section V concludes and outlooks the paper.

II. Problem Statement and Overall Framework

A. Problem Statement

Sheet metal stampings are common workpieces in the automotive industry, and its servo task using 2D industrial cameras requires the extraction of features of the target object. When extracting features using methods such as Hough detection, the robustness is poor, susceptible to interference from lighting factors and difficult to distinguish target features from complex backgrounds. However, current deep learning methods require manual annotation of large datasets for special objects, which is costly and difficult to extend. Considering that the YOLO algorithm can effectively avoid background errors and generate false positives, while the Mask R-CNN algorithm can provide pixel-level masks. Therefore, it is interesting to see how to achieve better target foreground segmentation by training on small data images and fusing the results of both methods.

B. Overall Framework

The basic framework structure of the proposed small data training based image foreground segmentation and servo control method is shown in Figure 1. Firstly, the images are passed through Mask R-CNN network and YOLOv7 network respectively to obtain Semantic Pixel and Identification Box. The standard mask of the Feature Mask is then input to the Feature Comparator and rotated to obtain the corresponding minimum target box. The minimum target box is then compared with the Identification Box size and fed back to the Feature Comparator, and when its overlap region is the largest, the output information from the Feature Comparator is used as the Segmentation Region as the pose feature. Then, combined with the position information of Semantic Pixel, the orientation features are output through Feature Comparator and the uniquely determined Feature Mask is applied directly to the image to generate Segmentation Region. finally, the obtained target foreground segmentation map is used for Hough straight line detection, extracting slope, Centre point and other features are used as input in the BP neural network to obtain the control rate and achieve visual servo.

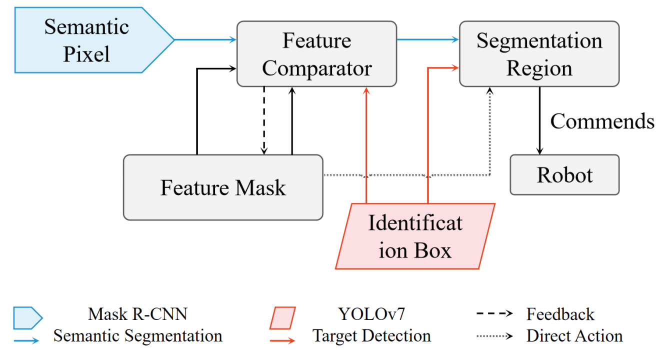


Fig. 1. Overall framework of the foreground segmentation and visual servo method. The blue pentagon is semantic segmentation and the red quadrilateral is target detection, the long dashed line indicates that the results of both are compared with the manual annotation mask, and the short dashed line indicates that the processing results are applied to the image as a region of semantic segmentation.

III. The Proposed Method

A. Loss function Focal-EIOU

The first stage of Mask R-CNN proposes the candidate target bounding box through Region Proposal Network (RPN), then goes through RoIAlign layer, reduces the pixel error using bilinear interpolation algorithm, and generates a fixed size Feature Map, and finally regresses using fully connected layer with the loss function $L = L_{cls} + L_{box} + L_{mask}$, where L_{cls} denotes the classification loss, L_{box} denotes the detection box loss, and L_{mask} denotes the average binary cross-entropy loss. However, when the training data set is small or the background interference is large, the RoI region recognition error is large and there is incomplete semantic segmentation.

Meanwhile, considering the advantages of YOLO algorithm with fast speed and less background misdetection, the loss function CIOU of YOLOv7 is improved with Focal-EIOU to achieve more accurate center point recognition. The recognized target Box is then replaced with the target Box in Mask R-CNN and used as fused features to be input to the next stage of feature comparator.

The loss function CIOU currently used by YOLOv7 takes into account three important factors: overlap area, center point distance and aspect ratio. By giving a prediction box B and a target box B^{gt}, the CIOU loss is defined as follows.

$$L_{CIOU} = 1 - IOU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{1}$$

where b and b^{gt} denote the center points of B and B^{gt}, respectively. $\rho(\cdot) = \|b - b^{gt}\|_2$ denotes the Euclidean distance between the center points. c denotes the diagonal length of the smallest closed box covering both boxes.$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$ and $\alpha = \frac{v}{(1-IOU)+v}$ denote the difference in the measured aspect ratio.

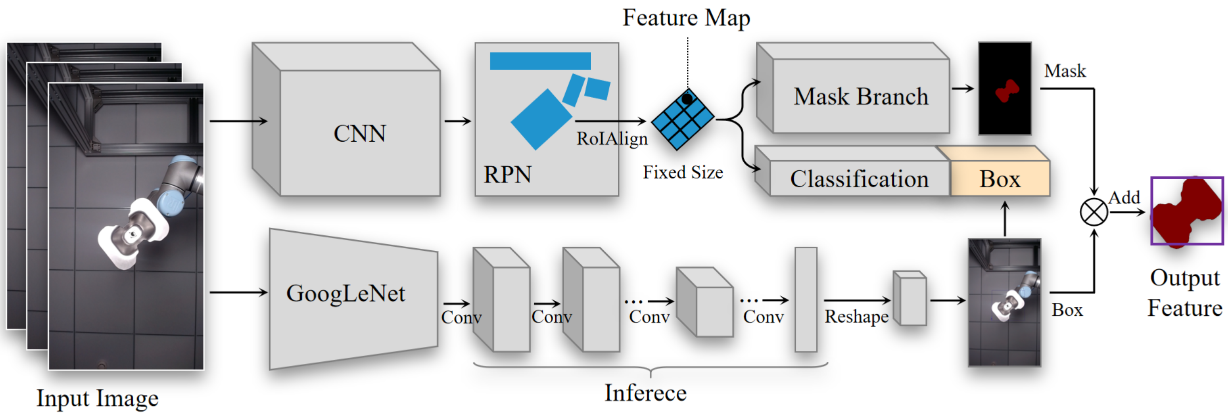


Fig. 2. Map of the fusion algorithm between Mask R-CNN and YOLOv7. CNN is the semantic segmentation network architecture and GoogLeNet is the target detection network architecture. The target box of Mask R-CNN is replaced with the target box of YOLOv7, and the features are combined and output as fused features to be fed into the feature comparator.

The gradient of v, which is with respect to w and h, is calculated as follows.

$$\begin{aligned} \frac{\partial v}{\partial w} &= \frac{8}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right) * \frac{h}{w^2+h^2} \\ \frac{\partial v}{\partial h} &= -\frac{8}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right) * \frac{w}{w^2+h^2} \end{aligned} \tag{2}$$

Although the CIOU method incases the loss of detection box scale and the loss of length and width, which increases the accuracy of the prediction box to some extent, it uses relative values to describe the aspect ratio, which is fuzzy, and does not consider the balance problem of difficult and easy samples. And EIOU is based on the advantages of CIOU, which calculates the difference value of width and height separately and replaces the aspect ratio, and also solves the balance problem of difficult and easy samples by introducing Focal Loss. the loss function calculation of EIOU is defined as follows.

$$\begin{aligned} L_{\text{EIOU}} &= L_{\text{IOU}} + L_{\text{dis}} + L_{\text{asp}} \\ &= 1 - IOU + \frac{\rho^2\left(b, b^{\text{gt}}\right)}{\left(w^c\right)^2 + \left(h^c\right)^2} + \frac{\rho^2\left(w, w^{gt}\right)}{\left(w^c\right)^2} + \frac{\rho^2\left(h, h^{gt}\right)}{\left(h^c\right)^2} \end{aligned} \tag{3}$$

where w^c and h^c are the width and height of the minimum enclosing box covering these two boxes. EIOU retains the beneficial feature part of the CIOU loss and divides the loss function into IOU loss L_{IOU}, distinct loss L_{dis} and aspect loss L_{asp}.

Considering the regression quality gradient, Focal Loss is introduced to distinguish high quality anchor boxes from low quality anchor boxes with the following equation.

$$L_{\text{Focal-EIOU}} = IOU^{\gamma} L_{\text{EIOU}} \tag{4}$$

where $IOU = |A \cap B| / |A \cup B|$, γ is the parameter controlling the degree of outlier suppression, and the value is taken as 1/2 in this paper. The accuracy of target detection is improved by modifying the calculation method of YOLOv7 loss function, and the extracted target box is replaced with the RoI detection box in Mask R-CNN to achieve feature fusion.

B. Feature comparator design

The structure of the fusion algorithm proposed in this paper is shown in Fig. 2. By rotating the manual labeling mask by different angles to obtain different minimum enclosing rectangles μ. Define the coordinates of the top left corner vertex M of the enclosing rectangle as (x_1, y_1) and the coordinates of the bottom right corner vertex N as (x_2, y_2), then, the corresponding aspect ratio δ can be calculated from the coordinates of M and N. Similarly, the coordinates of the top left corner vertex M' of the YOLOv7 target box ρ as (x'_1, y'_1) and the coordinates of the lower right vertex N' as (x'_2, y'_2), the corresponding aspect ratio δ' can be calculated from the coordinates of M' and N'.

$$\begin{cases} \delta = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\alpha}{\beta} \\ \delta' = \frac{y'_2 - y'_1}{x'_2 - x'_1} = \frac{\alpha'}{\beta'} \end{cases} \tag{5}$$

where α and α' denote the height of the enclosing box; β and β' denote the length of the enclosing box.

To calculate the repetition of the minimum enclosing box μ and the YOLOv7 target box ρ, the discrepancy function is defined as follows.

$$\sigma = |\delta - \delta'| + \sqrt{\sum_{i=1}^{2}\left(\left(x_i - x'_i\right)^2 + \left(y_i - y'_i\right)^2\right)} \tag{6}$$

Based on the calculated discrepancy function σ, the rotation angle θ can be determined so that its discrepancy function σ is minimized. It should be noted that the same minimum enclosing box covers two corresponding rotation angles θ, as shown in the following equation:

$$\begin{cases} \theta_1 = \theta \\ \theta_2 = \pi + \theta \end{cases} \tag{7}$$

The unique angle θ is then determined based on the overlap IOU of the pixel mask obtained from Mask R-CNN in the cases of θ_1 and θ_2, respectively. The Feature Comparator Design Process is shown below.

Algorithm 1 Feature Comparator Design Process

Input: Rotation $\boldsymbol{\theta}(\theta_1, \ldots, \theta_N)$; Mask R-CNN pixels; Desired box center; Minimum Box $(b_1, \ldots, b_N)$; IoU $(l_1, \ldots, l_N)$.
Output: Mask $(\theta, center)$.

Initializing: $Mask(pixels, \theta, center) \leftarrow \varnothing$
$\quad Mask(pixels) \leftarrow$ *Mask label*
for $i = 1$ *to* 360 **do**
$\quad (l_1, \ldots, l_N) \leftarrow$ *Minimum enclosing rectangle*
$\quad$ **if** σ *min* **then**
$\quad\quad$ Mask$(\tilde{\theta}) \leftarrow$ *Parameter* $\tilde{\theta}$ *is derived from* (7)
$\quad$ **end if**
end for
Mask $(\theta) \leftarrow$ *Mask R-CNN pixels IOU*
Mask $(center) \leftarrow$ *Desired box center*

C. Visual servo control

The image based visual servo control error expression is

$$\boldsymbol{e}(t) = \boldsymbol{s}(\boldsymbol{m}(t), \boldsymbol{a}) - \boldsymbol{s}^* \tag{8}$$

where $m(t)$ is the image data obtained by the camera, and a is additional information about the system such as camera internal parameters. $m(t)$ and a compute the resulting s called visual features, which is a series of feature values that can be extracted quantitatively, and s^* is the target value of these visual feature quantities.

The difference between different visual servo methods is mainly reflected in the different target feature s extraction. Based on the selected image features s, a velocity controller can be designed to control the movement of the robotic arm, and let the instantaneous velocity of the camera in the camera coordinate system be $V_c = (v_c, \omega_c)$, where v_c is the instantaneous linear velocity in the camera coordinate system and ω_c is the instantaneous angular velocity in the camera coordinate system. Then the rate of change of the image features with time $\dot{s}$ is related to the camera velocity V_c as

$$\dot{\boldsymbol{s}} = \boldsymbol{L_S} V_c \tag{9}$$

where $L_S \in R^{6\times k}$ is the image Jacobi matrix. Substituting Eq. (9) into Eq. (8) yields.

$$\dot{\boldsymbol{e}} = L_e V_c \tag{10}$$

where $L_e = L_S$,considering V_c as the input of the robot arm velocity control, the control equation of the visual servo can be obtained by inverse of Jacobi in Eq. (10).

$$V_C = -\lambda L_e^{+} e \tag{11}$$

The above equation is the visual servo control equation for six degrees of freedom. However, considering that the three degrees of freedom RX, RY and Z are already fixed when the servo is corrected, the image-based visual servo control can be trained using BP neural network to speed up the servo in real time.

The extracted feature data and the corresponding robotic arm parameters are normalized as test data and used as the input and output of the BP neural network for training. Then the feature data to be predicted is used as input, and the trained network is used for prediction and inverse normalization, at which time the obtained robotic arm parameters are the control parameters for servo deflection.

IV. Experiment

A. Data labeling and parameters

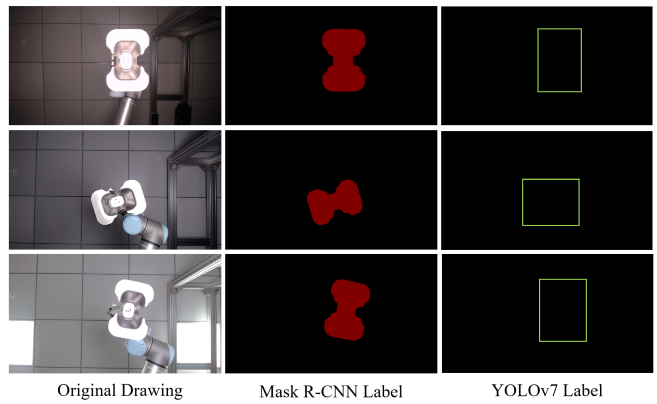


Fig. 3. Data set label. The original drawing classifies the datasets into dark, medium and strong according to the background illumination intensity.

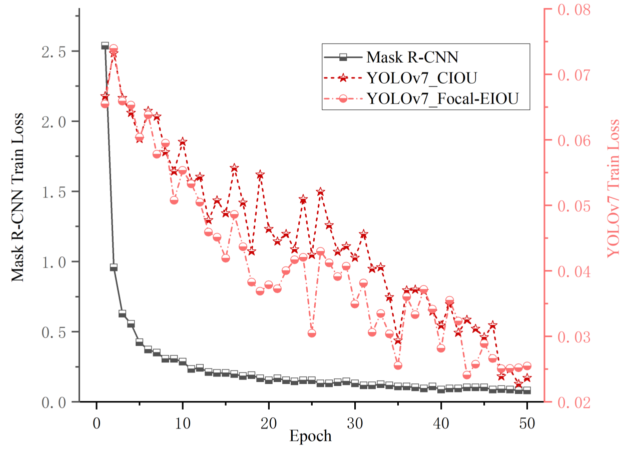
In this section, we adopt a total of 126 images containing strong, medium and dark background illumination intensity, which are divided into 75 training sets, 25 validation sets and 26 test sets. The number of training rounds is 50, and the number of batch-size is 8. The data sets are labeled with Mask R-CNN semantic segmentation and YOLOv7 target detection regions according to contour points and minimum enclosing rectangle, as shown in Fig. 3.

B. Achieving results

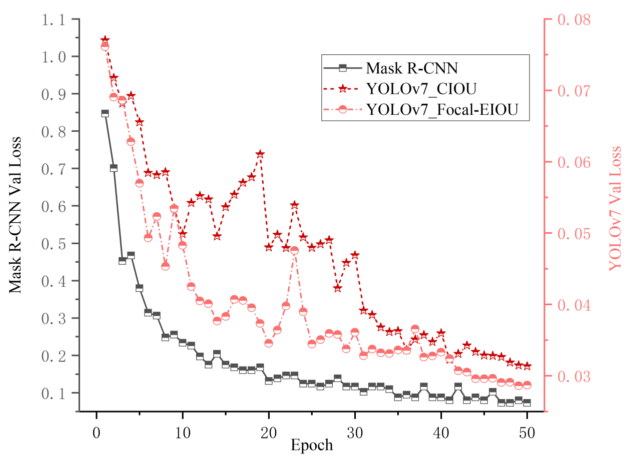
The training loss for semantic segmentation and target detection is shown in Fig. 4(a), and the verification loss for semantic segmentation and target detection is shown in Fig. 4(b). After 50 epochs, the training loss value of Mask R-CNN dropped rapidly from 2.54 to stabilize around 0.08; the training loss value obtained using the CIOU calculation method in YOLOv7 dropped from 0.067 to 0.024, and the training loss value obtained using the Focal-EIOU calculation method dropped from 0.066 to 0.025. Both YOLOv7 target recognition methods are basically the same in terms of loss values on the training dataset.

Similarly, after 50 epochs, the validation loss value of Mask R-CNN gradually decreases from 0.85 to around 0.07; the validation loss value obtained using the CIOU calculation method in YOLOv7 decreases from 0.077 to 0.031, and the validation loss value obtained using the Focal-EIOU calculation method decreases from 0.076 to 0.028. Fig. 4(b) shows that the validation loss values for the Focal-EIOU method are consistently smaller than those for the CIOU method and have a maximum difference of 0.0237.

Due to the small validation dataset, the CIOU method suffers from an increasing loss value during the validation process, which decreases further as the number of validation rounds increases. In contrast to the Focal-EIOU method, the degree of overfitting is reduced considering that it focuses on a more comprehensive set of features. Both of these methods eventually converge gradually and the latter has a lower value of validation loss.



(a) Object detection and semantic segmentation train loss graph



(b) Object detection and semantic segmentation val loss graph

Fig. 4. Target detection and semantic segmentation loss map. The loss values for Mask R-CNN are referenced to the left scale values in the figure.The loss values for YOLOv_CIOU and YOLOv_Focal-EIOU are referenced to the right scale values in the figure.

The images with different lighting backgrounds are divided into weak, medium and strong categories, and the foreground extraction is performed using the algorithm in this paper, respectively, as shown in the fourth column in Fig. 5. Comparing the original image and the processed image, it can be seen that the background is blurred and the sharpness of the foreground target is retained, and the separation and extraction of the foreground target under different lighting backgrounds is achieved.

The ablation experimental results are shown in Table I. When using the original image for straight line detection, hundreds of line segments are recognized coexist due to the interference in the background, and it is difficult to extract straight lines from the target edges; if only the Mask R-CNN semantic segmentation region is used as the foreground extraction area, the IOU ratio is about 10%, where $IOU = |A \cap B| / |A \cup B|$, and A denotes the algorithm recognition region and B denotes the actual target region. After combining Mask R-CNN with YOLOv7, the IOU recognition accuracy can be significantly improved, and the edge line segment of the target can be accurately recognized at the same time.

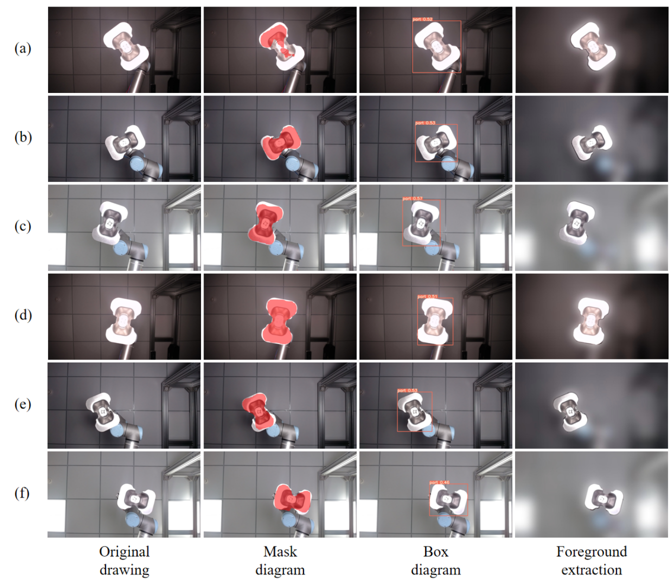


Fig. 5. Comparison of the effect of foreground target extraction. Red area in Mask diagram is the semantic segmentation result, red rectangular box in Box diagram is the target detection result, foreground extraction is the result of the target foreground segmentation algorithm in this paper.

TABLE I

COMPARISON TABLE OF ABLATION EXPERIMENTS.

Mask R-CNN	YOLOv7-CIOU	YOLOv7-Focal-EIOU	IOU	Lines
—	—	—	Nan	150∼700
√	—	—	5%∼11.7%	0
—	√	—	47%∼86%	2∼4
—	—	√	49%∼88%	2∼4
√	√	—	92%∼97%	1∼2
√	—	√	**95%∼99%**	**1∼2**

The data of X, Y and RZ degrees of freedom are trained by BP neural network, and then 30 test images are inputed, and the results are shown in Fig. 6. The robot arm movement range of X axis output is within 28 and 75 pixels. The robot arm movement range of Y axis output is within 807 and 882 pixels. The movement range of the robot arm output by RZ axis is within 2.58 and 3.68 rad. The accumulated movement error in the direction of the X and Y axes is within ± 3.5 pixels, which corresponds to the actual error of ± 0.2mm for X-axis; ± 0.5mm for Y-axis; the deviation of rotation angle is between ± 0.008 rad, which corresponds to the actual angle deviation of $\pm 0.46°$.

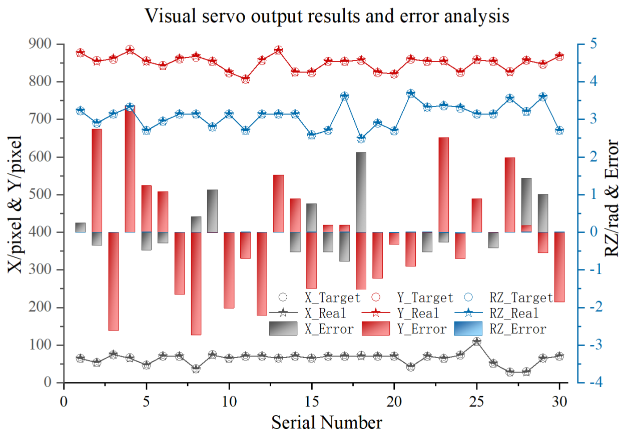


Fig. 6. Visual servo output results and error analysis. The target and actual values for the X and Y axes refer to the scale values on the left side of the graph. the target and actual values for the RZ axis refer to the scale values on the right side of the graph. the movement errors for the X, Y and RZ axes refer to the scale values on the right side of the graph.

C. Discussion

From the experimental results, it can be seen that this algorithm can effectively segment and extract foreground targets, thus reducing the interference of background illumination and other factors on foreground target recognition. In addition, fusing semantic segmentation with target detection results can circumvent the problem of imprecise segmentation mask regions due to small sample data training, especially at the edges of segmented targets, which are prone to misidentification or overfitting. To the best of our knowledge, there is no other existing method that can perform fast foreground target segmentation using small dataset training following. Although our method currently focuses on servo correction for three degrees of freedom, it can be extended in an effective way.

First, most of the parts in industry have a more stable structure, giving favorable conditions for attaching the standard mask to the original drawing by manual annotation. When the workpiece is rotated or moved around the other degrees of freedom in a small way, the standard masks can be extracted from each pose mask union of set to minimize unwanted background interference. However, when the workpiece is rotated or moved significantly, it will lead to the failure of the mask coverage area. In this case, the standard mask under different postures can be marked, and the corresponding mask can be called for different postures such as front view and side view. Second, the semantic segmentation under small data sets may have the problem of unrecognition. At this point, the target box region recognized by YOLO can be considered as the mask area directly, or the union set of two angle masks in the feature comparator design in Section 3 can be used as the final mask area. Finally, we only extracted the straight line after foreground segmentation, and the extraction of other features in different backgrounds and the effect need to be further studied.

V. Summary and Outlook

In this paper, we propose an image foreground segmentation and visual servo correction method based on small data training. The foreground target is effectively extracted through the fusion of semantic segmentation and target recognition, thus ignoring the interference of factors such as background and illumination. The training of small data sets enables the extension in industrial standard parts, providing favorable feature extraction effects for visual servo and correction of deflection. The shape and combination priority of the manual annotation mask area can be adjusted for different feature extraction tasks.

Regarding future work, the most important thing is to investigate how to improve the balance of precision, universality and speed of foreground target segmentation. Then, it can be extrapolated from general industrial parts to more flexible and diverse configurations and target features. In addition, the effectiveness of our servo-deflection control should be further tested in six degrees of freedom.

References

[1] L. Deng, "Comparison of image-based and position-based robot visual servoing methods and improvements," Ph.D. dissertation, Ph. D. Dissertation, 2003.

[2] W. Li and R. Xiong, "A hybrid visual servo control method for simultaneously controlling a nonholonomic mobile and a manipulator," *Frontiers of Information Technology & Electronic Engineering*, vol. 22, no. 2, pp. 141–154, 2021.

[3] A. Hajiloo, M. Keshmiri, W.-F. Xie, and T.-T. Wang, "Robust online model predictive control for a constrained image-based visual servoing," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 4, pp. 2242–2250, 2015.

[4] H. Ng, S. Ong, K. Foong, P.-S. Goh, and W. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *2006 IEEE southwest symposium on image analysis and interpretation*. IEEE, 2006, pp. 61–65.

[5] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.