BoT-YOLOv8: A highly accurate and stable initial weld position segmentation method for medium-thickness plate

Zongmin Liu [1,2*], Jie Li [2], Shunlong Zhang [2], Lei Qin [2], Changcheng Shi [2], Ning Liu [3*]

1 National Research Base of Intelligent Manufacturing, Chongqing Technology and Business University, Chongqing 40006, PR China.

2 School of Mechanical Engineering, Chongqing Technology and Business University, Chongqing 400067, PR China.

3 Advanced Remanufacturing & Technology Centre (ARTC), Agency for Science, Technology and Research (A*STAR), 3 CleanTech Loop, #01-01 CleanTech Two, Singapore 637143, Republic of Singapore.

* Corresponding author.

E-mail addresses: liu_zm@ctbu.edu.cn (Zongmin Liu), Liu_Ning@artc.a-star.edu.sg (Ning Liu).

Abstract: Due to the complexity of the initial welding position types in medium-thickness plate workpieces, the welding robot struggles to achieve automatic vision guidance before welding these workpieces. To address this problem, an initial welding position segmentation method with high accuracy and stability for medium-thickness plate is proposed. This method is developed through the integration of the Bottleneck Transformer (BoT) with YOLOv8, enabling a novel framework termed BoT-YOLOv8. To optimize the filtration of redundant information in the image and enhance the model's feature representation ability, the Bottleneck Transformer is incorporated following the final bottleneck layer within the residual module of the YOLOv8 neck structure. Subsequently, to capture multi-scale information of the target features, the atrous convolution is integrated into the spatial pyramid pooling structure. Furthermore, to enhance the learning of welding position characteristics for the robotic welding system, the Hue-Saturation-Value (HSV) color space region segmentation method is employed to post-process the weld seam features. Finally, ablation experiments are conducted on the self-created weld dataset. The results demonstrate that the proposed method strikes a trade-off between detection accuracy (93.1% $mAP^{0.5}$) and detection speed (26.5 FPS) on a 12GB NVIDIA GeForce RTX 3060 GPU. Compared with the existing methods, the proposed method achieves higher detection accuracy for medium-thickness plate weld and also improves the robotic welding efficiency.

Keywords: Robot welding; Medium-thickness plate; Initial weld position segmentation; YOLOv8; Bottleneck Transformer

1. Introduction

Due to the high repeatability and capability to operate in harsh environment, welding robots have gained increasing popularity in welding tasks. However, when dealing with medium-thickness plate commonly found in large construction machinery, energy equipment, and steel structures, welding robots still encounter challenges. These challenges arise from the complexity introduced by the characteristics of medium-thickness plate, such as the large volume, irregular weld seams and three-dimensional distribution. To visually guide the initial weld position on the medium-thickness plate, the welding robot mainly relies on cumbersome teaching-playback and offline programming methods. As shown in Fig. 1, manual welding still dominates welding tasks for medium-thickness plate in actual production and manufacturing. This leads to challenges such as high costs, low efficiency, and inconsistent weld quality. To effectively overcome the automation challenges faced by welding robots, it is crucial to address autonomous robotic vision guidance in medium-thickness plate welding scenarios. However, the success of vision guidance for welding robots is highly dependent on the robot's perception precision regarding the initial welding position. This requires in-depth investigation of feature extraction technology to assist welding robots for welding medium-thickness plate.

Fig. 1 Medium-thickness plate welding site

In recent decades, researchers have proposed various solutions to tackle the technical challenges of automatic welding, mainly focusing on two key approaches: active vision and passive vision.

In the field of active vision research, line-structured light sensors have been widely adopted for their robustness and high accuracy. They have been employed for various tasks, such as initial welding position recognition, parameter extraction, welding tracking and welding pool monitoring [1, 2]. Liu et al. [3] reported an automatic extraction algorithm based on dynamic programming to identify laser inflection points, which can facilitate welding robots to determine the initial position of fillet welds. Wang et al. [4] proposed a welding recognition method that utilized structured light and worked even under challenging noisy conditions. Employing NURBS-snake model to extract the laser centerline of the weld, this method generated a dynamic region of interest. Li et al. [5] utilized the welding edge point detection operator and the Internal Propulsion Center Extraction (IPCE) algorithm to identify multiple types of welds, such as zigzag groove butt welds, V-groove butt welds, single groove butt welds, and lap welds.

While welding recognition methods employing line structured light sensors offer strong anti-interference capabilities, they are still suffering from low recognition efficiency for medium-thickness plate where a huge surface are to be recognized. Firstly, the line structured light sensor can only obtain partial 3D information such as a single line or a column of the target object at a time. For 3D information of the entire workpiece area, motion scanning or scanning from different viewpoints are required, thus resulting in a relatively slow information acquisition speed. In addition, a consistent velocity between the target object and the measuring instrument must be maintained during the scanning process. Otherwise, shape distortion may occur and negatively impact the imaging quality.

In the field of passive vision research, the combination of deep learning and welding robotic perception technology has emerged as a primary direction. Numerous researches have increasingly focused on the application of reinforcement learning and self-supervised learning techniques to tackle challenges in weld feature recognition and path tracking [6, 7, 8]. These approaches have demonstrated significant potential in enhancing the performance of welding robots, particularly in weld feature recognition and path tracking tasks. To further improve the efficiency of automated welding, some researchers have employed laser sensors to capture weld images. The welding robots were subsequently trained using reinforcement learning and self-supervised learning techniques on the collected image data, enabling accurate weld feature recognition and extraction [9, 10, 11]. Another group of researchers developed a weld feature extraction model using 3D point cloud methods to facilitate path planning and orientation planning for welding robots [12, 13, 14]. Furthermore, some researchers employed convolutional neural network (CNN) to identify of weld types and detect initial welding points [15, 16]. Although the weld seam recognition technology combining deep learning and passive vision has improved the efficiency in automatic welding, the weld types that can be identified in the above literature are relatively limited. Consequently, the applicability of these methods is highly constrained when handling medium-thickness plate welds complex structures.

Meanwhile, deep learning-based image segmentation methods have emerged as a prominent research focus in the field of automated welding. Several researchers have applied this technology to extract features from medium-thickness plate welds. For example, Oskar et al. [17] created an improved UNet method for solder joint segmentation of iron bars. Chen et al. [18] designed a lightweight DeepLabv3+ semantic segmentation network to segment weld images. Zou et al. [19] utilized the SOLOv2 model and replaced the backbone network of the original model with the ShuflleNetv2 network to enhance the speed of weld feature extraction. Wu et al. [20] combined semantic segmentation techniques and edge detection techniques to design a real-time segmentation network for automatic welding of medium-thickness plate. The aforementioned methods only focus on a single aspect of the precision or efficiency of the weld feature extraction. If the precision and efficiency of the robot's weld feature extraction cannot be guaranteed simultaneously during actual processing, it will lead to reduced range of applicability and increased hardware costs for the welding system. Therefore, achieving a trade-off between the precision and efficiency of weld feature extraction is crucial.

In summary, the challenges faced by current theoretical approaches and technical solutions for robotic welding of medium-thickness plate are as follows:

(1) Robotic welding for medium-thickness plate workpieces heavily relies on complex teaching-playback and offline programming methods, which cannot meet the automation standards required for medium-thickness plate welding.

(2) Most research [9-16] focuses on weld feature extraction for workpieces with simple structures, leaving a significant gap in feature extraction for more complex weld types, such as medium-thickness plate workpieces.

(3) Existing methods have not successfully attained an optimal equilibrium between efficiency and precision in the domain of initial weld position feature extraction for medium-thickness plate workpieces.

To address the aforementioned challenges, this paper proposes an initial welding position segmentation method with high accuracy and stability to facilitate the extraction of the initial weld position features in medium-thickness plate workpieces. The main contributions of this paper are stated as follows:

(1) By designing a new gradient shunt module C2f_Bottleneck_BoT and utilizing atrous convolution as the model of spatial pyramid pooling structure, a novel method based on an improved version of YOLOv8 is presented to segment and extract the features of initial weld position on medium-thickness plate.

(2) In order to obtain the initial weld position information, the HSV region segmentation method is utilized to post-process the predicted images by the presented model. As a result, the initial weld feature image is transformed into a binary image.

(3) A weld seam image dataset of medium-thickness plate is produced for training BoT-YOLOv8, and the evaluation indicators are established to measure the performance of the model. Furthermore, the effectiveness and superiority of the proposed method are verified by ablation experiment on the weld seam image dataset.

The methodology introduced in this paper is primarily designed for robotic perception challenges encountered in the medium-thickness plate's weldable area. Thereby, this paper designs a method for the initial weld position feature extraction in medium-thickness plate welding applications. Furthermore, this method can serve as a technical reference for the extraction of initial weld position feature in complex medium-thickness plate configurations. It has promising potential to reduce the time expenditure associated with the robotic welding of medium-thickness plate workpieces, consequently enhancing the efficiency of industrial manufacturing processes.

The rest of this paper is organized as follows. In Section 2, research work related to this paper is summarized. Section 3 provides a detailed introduction to the proposed method. Section 4 introduces experimental process and analysis of comparative results, following which Section 5 gives the main summaries of this article.

2. Theoretical Background

YOLOv8 is an advanced one-stage object detection algorithm, first introduced by the Ultralytics team [21]. It has emerged as the state-of-the-art (SOTA) model for object detection with the advantages of lightweight and high accuracy. The network structure of YOLOv8 is shown in Fig. 2.

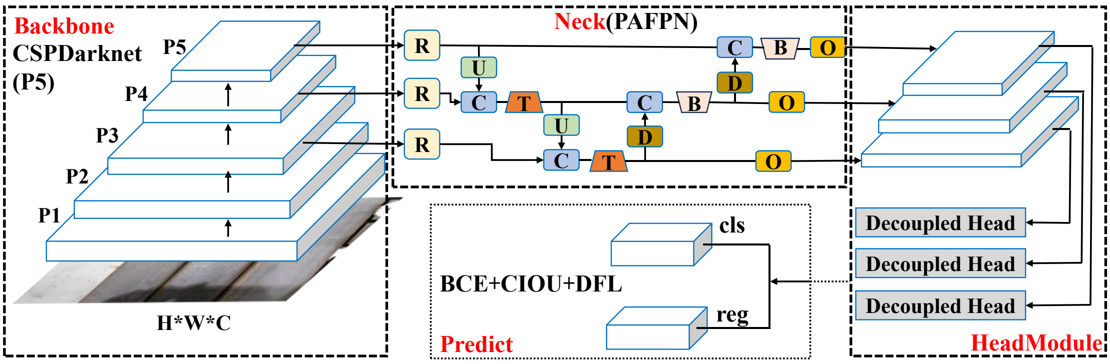


Fig. 2. YOLOv8 network structure

Fig. 2 illustrates that the YOLOv8 network consisting of four parts: Backbone, Neck, Head and Prediction layer. The Backbone of YOLOv8 still follows the design idea of CSPDarknet gradient shunt. The advantage of CSPDarknet is to improve the learning ability of the network and reduce the amount of calculation of the network without loss of accuracy, thus improving the inference speed. The Backbone contains five feature clusters, the first feature cluster P1 is a 3×3 convolution, and the feature clusters P2 to P5 have the same structure, consisting of a 3×3 convolution layer and a C2f module. The structure of C2f is shown in Fig. 3.

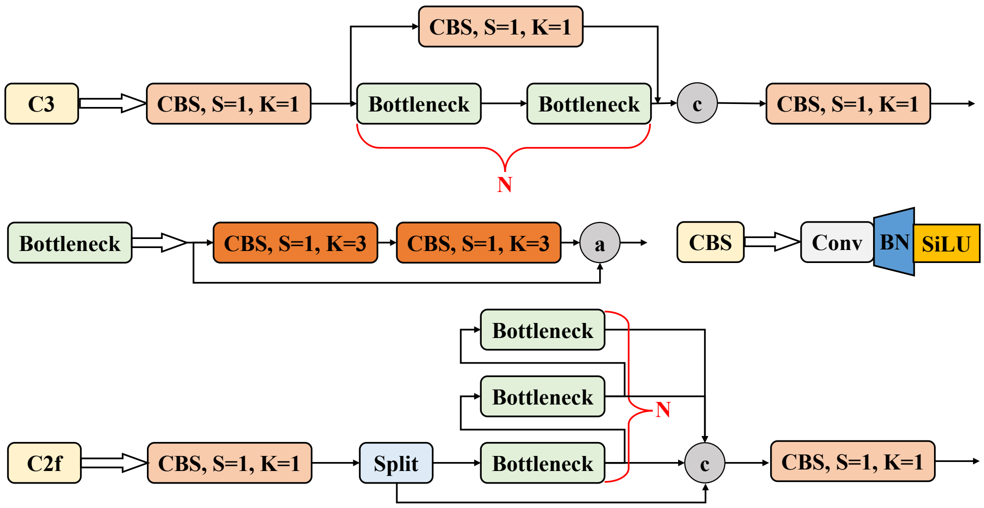


Fig. 3. Structure of C2f

It can be seen from Fig. 3 that the overall structure of C2f and C3 is similar, and both adopt the concept of CSPNet gradient shunt as well as residual structure. The main difference between C2f and C3 lies in the connection pattern. In C3, a serial connection is employed for the bottleneck layers, whereas in C2f, a parallel connection is utilized. This choice is made to achieve a balance between ensuring the lightweight of YOLOv8 and obtaining more comprehensive gradient flow information simultaneously.

The role of YOLOv8 Neck is to take the different resolution features generated in the previous step as input and output the features after efficient fusion. The Neck adopts the Path Aggregation Feature Pyramid Network (PAFPN) structure. The Path Aggregation (PA) strategy significantly reduces the number of network layers that the features of different levels have to traverse during transmission, thereby enhancing the efficiency of the network. YOLOv8 algorithm adopts Decoupled Head structure to predict classification and regression tasks separately and accelerates model convergence speed and detection accuracy by head decoupling.

The Loss function of YOLOv8 includes category classification and border regression. The Binary Cross-Entropy Loss function (BCE) is used for category classification loss, and the border regression loss is composed of Complete Intersection Over Union Loss (CIOU) and Distribution Focal Loss (DFL). The specific calculation process is as follows.

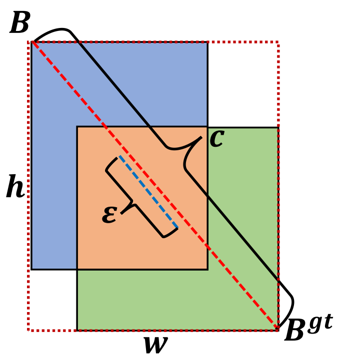


Fig. 4. Predicted box and ground truth box

In Fig. 4, B, B^{gt} represent the center points of the predicted and ground truth boxes, respectively, and ε is the Euclidean distance between these two center points. w, h represent the width and height of the minimum closure region of the predicted and ground truth boxes, respectively, and c represents the diagonal distance of the minimum closure region between the predicted and ground truth boxes. The loss function of YOLOv8 is calculated as follows.

$$\begin{cases} IOU = \dfrac{B \cap B^{gt}}{B \cup B^{gt}} \\ Loss = L(cls) + L(reg) \\ L(reg) = L(CIOU) + L(DFL) \\ L(CIOU) = IOU - \left(\dfrac{\varepsilon^2(B, B^{gt})}{c^2} - \alpha v\right) \\ L(cls) = -\dfrac{1}{n}\sum_{\lambda}[y_\lambda * \log(p_\lambda) + (1 - y_\lambda) * \log(1 - y_\lambda)] \\ v = \dfrac{4}{\pi^2}\left(arctan\dfrac{w^{gt}}{h^{gt}} - arctan\dfrac{w}{h}\right), \alpha = \dfrac{v}{(1 - IOU) + v} \\ DFL(\chi_\theta, \chi_{\theta+1}) = -\big((\psi_{\theta+1} - \psi)\log(\chi_\theta) + (\psi - \psi_\theta)\log(\chi_{\theta+1})\big) \end{cases} \tag{1}$$

where y_λ is the label of λ (positive sample is 1, the negative sample is 0), p_λ is the probability that the prediction is positive, and n is the number of classes. α is the weight coefficient, and v is used to measure the aspect ratio similarity between the predicted box and the ground truth box. The idea of DFL is to transform the regression problem into a classification problem for solving, so the calculation idea of DFL is still the cross-entropy function. Accordingly, $\psi_{\theta}, \psi_{\theta+1}$ represent the sample $\theta, \theta+1$ label (positive sample is 1, the negative sample is 0), $\chi_{\theta}, \chi_{\theta+1}$ represent the sample $\theta, \theta+1$ to predict the probability of positive samples.

Moreover, YOLOv8 adopts the Task-Aligned Assigner dynamic allocation strategy for positive and negative samples. This strategy helps in efficiently allocating samples to the classification and regression tasks based on their relevance and contribution to the overall objective. For the classification prediction score and regression prediction score of all pixels, the K largest positive samples are selected by sorting the weighted scores. The Task-Aligned Assigner policy is weighted as follows:

$$\tau = s^m + u^n \tag{2}$$

where s represents the classification prediction score of all pixels, u represents the regression score between the predicted box and the true box of all pixels, m and n are the weight hyperparameters, and τ is the weighting coefficient.

In summary, due to the excellent detection performance of YOLOv8, this paper employs the improved YOLOv8 algorithm to construct the instance segmentation model for the initial weld position on medium-thickness plate.

3. Methods

3.1 System framework

The BoT-YOLOv8 based instance segmentation system for medium-thickness plate initial weld position includes seven parts: data acquisition, preprocessing, model training, model evaluation, image prediction, post-process and modifying weights. The specific process is shown in Fig. 5.

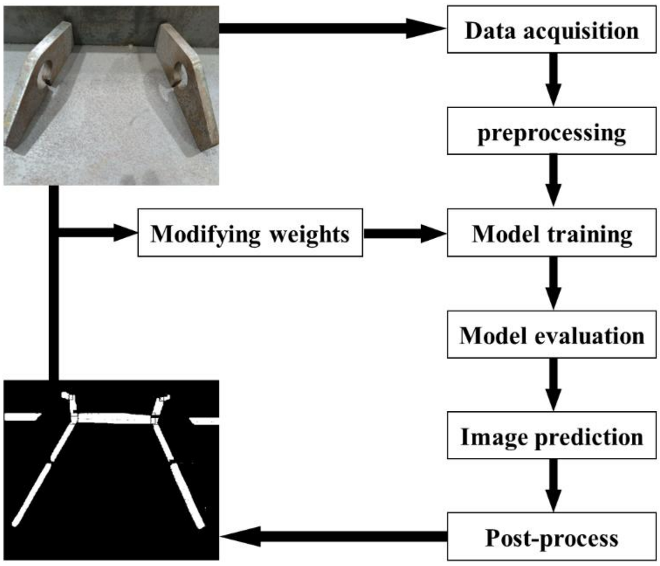


Fig. 5. Flowchart of initial weld position segmentation system for medium-thickness plate

The specific implementation process of the closed-loop system is described as follows:

(1) Weld seam data are collected by a depth camera mounted near the welding gun of the welding robot.

(2) Preprocessing operations such as normalization and data augmentation are performed on the collected data.

(3) The preprocessed images are fed into the model for self-supervised learning.

(4) The model evaluation index is established for error analysis between training results and validation results.

(5) The trained weight file is used to segment and predict the test image, and the initial welding position features are extracted.

(6) The HSV color space region segmentation method is utilized to carry out post-processing operation on the mask image segmented in the prediction step, and the final output is the initial weld position mask binary image of medium-thickness plate.

(7) The weights of the predicted bounding boxes are adjusted based on any discrepancies or calibration issues observed during evaluation. This step ensures better alignment between the predicted positions and the ground truth positions.

3.2 Optimization strategy

Intending to enhance the performance of the model, two optimization strategies are employed after establishing the basic segmentation model for the initial weld position of medium-thickness plate. The specific optimization strategies are as follows:

(1) A new module is obtained by fusing Bottleneck Transformer into C2f and named it C2f_Bottleneck_BoT. In addition, the C2f_Bottleneck_BoT module is used to replace all C2f modules in YOLOv8 neck network.

(2) The Atrous Spatial Pyramid Pooling (ASPP) is used as the spatial pyramid pooling structure connecting the Backbone part and the Neck part. The improved YOLOv8 model is shown in Fig. 6.

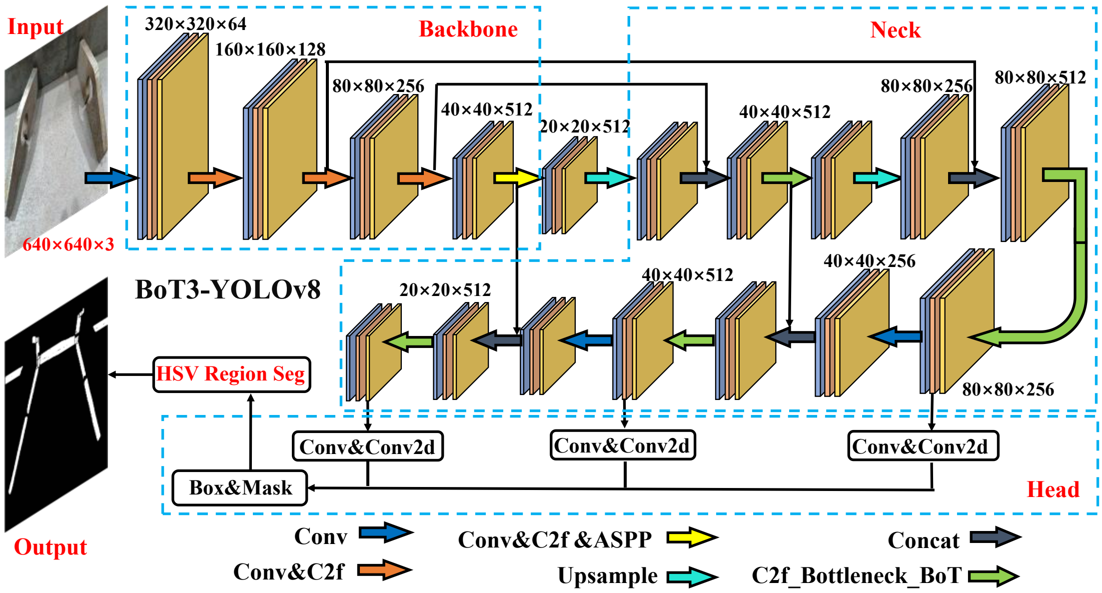


Fig. 6. BoT-YOLOv8 network structure

3.2.1 C2f_Bottleneck_BoT

Bottleneck Transformer model is first presented by Google [22]. Its core idea is to incorporate the Self-Attention Mechanism of Transformer model into the bottleneck structure of ResNet model. Inspired by this design idea, the Transformer Block is integrated into the bottleneck layer of C2f to obtain the C2f_Bottleneck_BoT module in the proposed approach. This integration enhances the local feature extraction ability and global information capture ability of the initial weld position segmentation model of medium-thickness plate. Specifically, the C2f_Bottleneck_BoT module is implemented as follows.

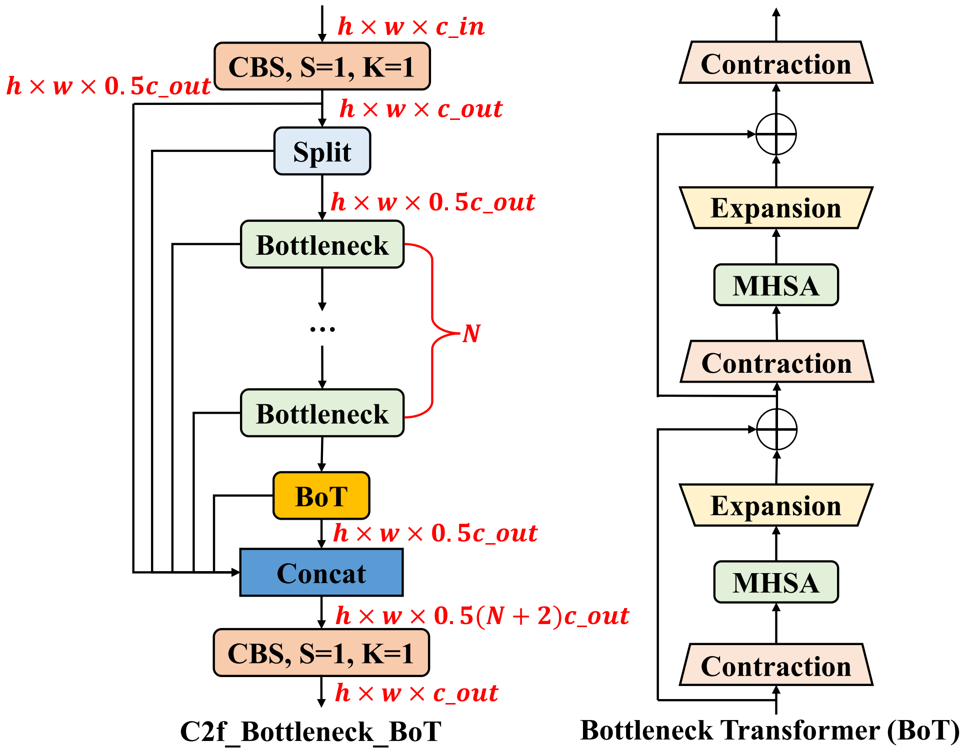


Fig. 7. C2f_Bottleneck_BoT structure

As shown in Fig. 7, the Multi-Head Self-Attention (MHSA) mechanism is the core module in the BoT structure. Compared with the Self-Attention (SA) mechanism, the MHSA adopts the design idea of divide-and-conquer and dimensionality elevation of target information. Leveraging on the parallelism capabilities of graphics cards and employing a space-time tradeoff approach, the MHSA are able to achieve superior model performance. Therefore, Bottleneck Transformer is able to maximize the utilization of graphics card resources and optimize the tradeoff between space complexity and time complexity, resulting in improved overall model performance. The structure of MHSA is shown in Fig. 8.

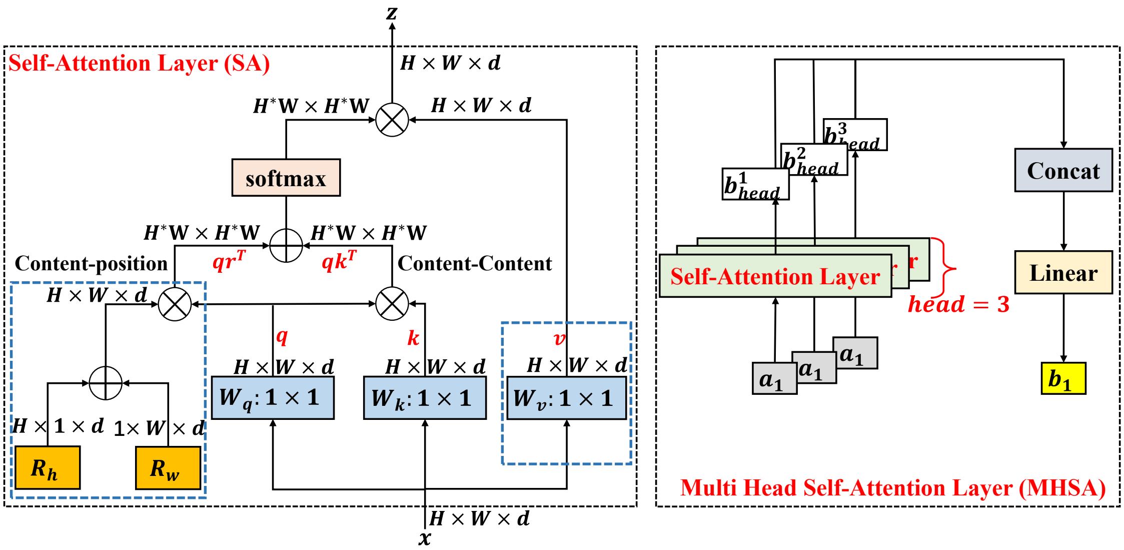


Fig. 8. Schematic diagram of SA and MHSA structures

As illustrated in Fig. 8, MHSA generates multiple output sequences from the input sequence through the multi-head mechanism on the basis of SA, and subsequently integrates these output sequences to obtain a new output sequence. The specific calculation process is as follows:

$$\begin{cases} Q = W^q x, K = W^k x, V = W^v x \\ \hat{\sigma} = Softmax\left(\dfrac{Q^T K}{\sqrt{d_K}}\right) \\ z = \hat{\sigma}^T V = Softmax\left(\dfrac{QK^T}{\sqrt{d_K}}\right) V \end{cases} \tag{3}$$

where x, z represent the input and output respectively, Q is the query, K is the index, V is the content, and $\hat{\sigma}$ represents the attention matrix. The whole computation process of SA is equivalent to generating the attention matrix $\hat{\sigma}$ by weighted summation of the input sequence x and then obtaining the output sequence z.

SA takes the previously calculated Q, K, V as a holistic Head input to the next layer, while MHSA needs multiple groups of W^q, W^k, W^v multiplied with the input x to obtain multiple groups of Q, K, V.

As shown in Fig. 8, taking the diagram on the right as an example, the input sequence a_1 obtains three output vectors $b^1_{head}, b^2_{head},$ b^3_{head} through the multi-head mechanism ($head = 3$). Furthermore, the output sequence b_1 is obtained by concatenation and linear transformation. The calculation process is as follows:

$$\begin{cases} Q_i = W^{q_i}a_1, K_i = W^{k_i}a_1, V_i = W^{v_i}a_1 \\ b_{head}^{i} = \hat{\sigma}^{T}V_i = Softmax\left(\frac{Q_i {K_i}^{T}}{\sqrt{d_{K_i}}}\right)V_i \\ b_1 = Linear\left(Concat(b_{head}^{1}, b_{head}^{2}, b_{head}^{3})\right) \\ i = 1, 2, 3 \end{cases} \tag{4}$$

where $Concat$ denotes the concatenation operation and $Linear$ denotes the linear transformation.

3.2.2 Atrous Spatial Pyramid Pooling (ASPP)

Since the weight matrix of the fully connected layer in the traditional CNN is fixed, the size of the image output of the pooling layer must be consistent, otherwise the network will not be able to be trained. To address this issue, stretching or cropping is commonly used to ensure that the size of the output image of the pooling layer is consistent. However, stretching operation or cropping operation will cause image information loss and image distortion. Hence, the pyramid pooling structure is utilized to solve the previous problem by using the multi-scale information of the image. The pyramid pooling in YOLOv8 model adopts the Spatial Pyramid Pooling Fast (SPPF) structure.

During the training of the image instance segmentation task of the initial weld position of medium-thickness plate, a phenomenon is worth noticing where the average accuracy of the pyramid pooling layer with the atrous convolution structure is higher than that of the SPPF structure. Consequently, this paper selects ASPP [23] structure as the improved pyramid pooling layer of our model. The ASPP structure enhances the model’s ability to capture multi-scale target information. The structure of ASPP is shown in Fig. 9.

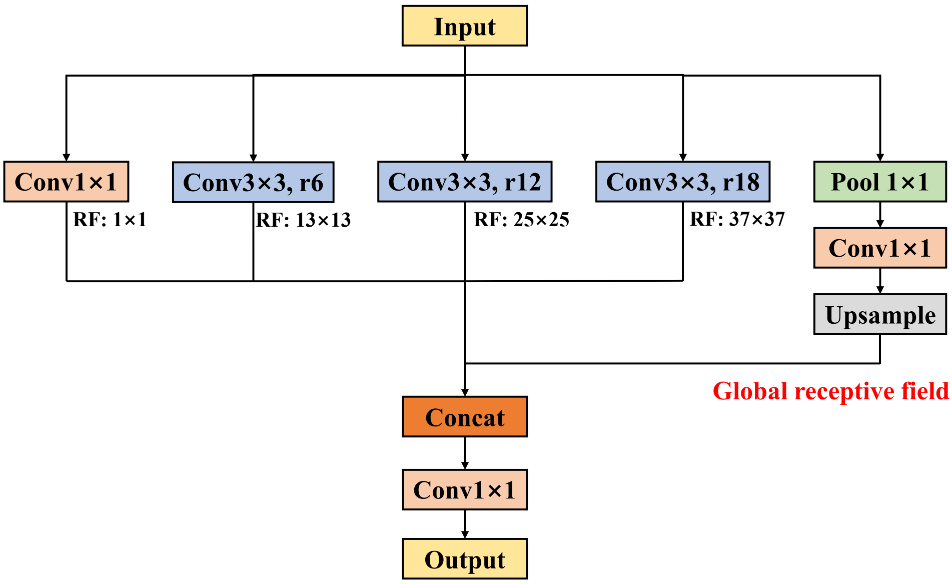


Fig. 9. Schematic representation of ASPP structure

As shown in Fig. 9, for an image input of a given size, the ASPP module uses atrous convolution with different sampling rates to sample the input image. Subsequently, the output results of each dilated convolution are spliced end to end, and the number of channels of the image is expanded. Finally, the number of channels is reduced by 1×1 convolution to output the image.

With the aim to verify the feasibility and effectiveness of the above two optimization strategies, this study conducts ablation experiments on the self-made weld data set. The results show that the improved model achieves superior performance with slightly more parameters and calculation.

3.3 Post-processing

Aiming to obtain the mask binary image of the initial weld position characteristics of medium-thickness plate, the HSV space region segmentation method is adopted to post-process the weld mask image generated by the model prediction. The first step is to determine the color value interval of the weld seam mask using the H component histogram. The distribution of H component histogram is shown in Fig. 10.

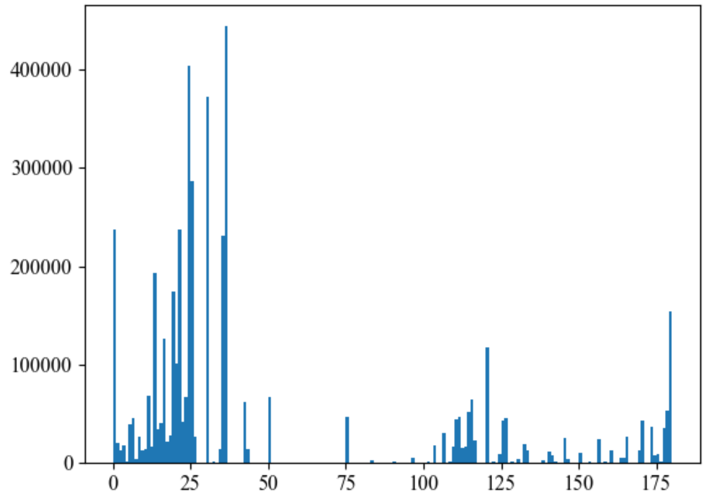
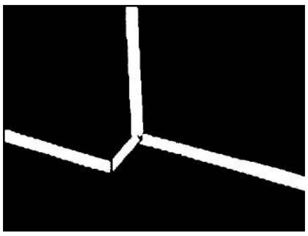


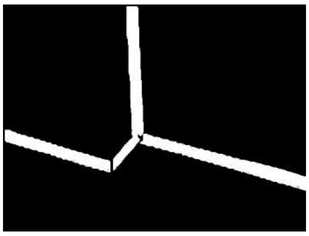
Fig. 10. Distribution of H component histogram

As can be seen from Fig. 10, the majority of the color channels H components in the weld seam mask are distributed between 0 and 25. Utilizing this information, the value range of H component is set to 0~25, and fine-tune the S component and V component. After that, the mask binary image of the initial weld position characteristics of medium-thickness plate is generated according to the HSV components threshold values. Finally, to address the noise in the generated masked binary image, the processing methods of erosion and dilation are utilized to denoise the masked binary image. This post-process step helps refine the accuracy of the segmentation by effectively filtering out unwanted colors and focusing on the desired features. The post-process results are shown in Fig. 11.

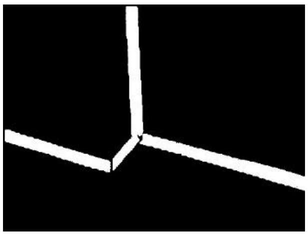
(a) mask image



(b) binary image



(c) eroded image



(d) dilated image

Fig. 11. Post-processing of the mask image

As shown in Fig. 11, compared with the original mask image, the post-processed mask binary image can express the characteristics of the target object in a clearer way. This step is crucial for the subsequent analysis of medium-thickness plate weld seam features and the task of autonomous path planning for welding robots.

3.4 Evaluation indicators

This work uses four indicators: $precision$, $recall$, mean Average Precision (mAP) and frames per second (FPS) to evaluate the performance of the model. The evaluation indicators of the proposed model are calculated based on the confusion matrix, as depicted in Fig. 12.

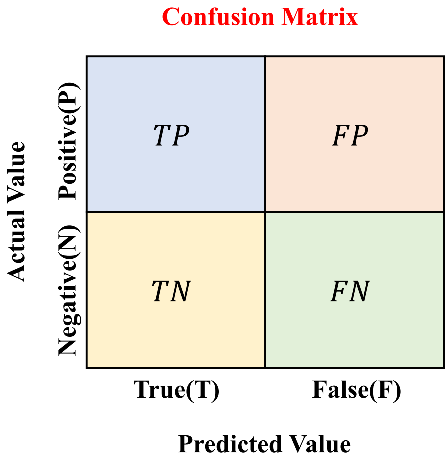


Fig. 12. Diagram of the confusion matrix

where, TP, FP denote the number of positive samples that are correctly and incorrectly classified, respectively. TN, FN represent the number of negative samples that are correctly and incorrectly classified, individually.

The calculation formula of the evaluation indicators of the model are as follows:

$$\begin{cases} Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN} \\ mAP = \dfrac{1}{n}AP = \dfrac{1}{n}\sum_{i=1}^{n} p(i)\Delta r(i) = \dfrac{1}{n}\int_0^1 p(r)dr \\ FPS = \dfrac{1000ms}{Single\ image\ processing\ time} \end{cases} \quad (5)$$

where AP is the area enclosed by the P-R curve and X and Y axis, n is the number of detected classes (there is only one sample class in this article, and $n = 1$ here), p is precision, and r is recall.

As a result, $mAP^{0.5}$ represents the mean average precision of the models when the threshold of IOU was taken as 0.5.

4. Experiment and results

4.1 Data preparation

Intended to capture images of the initial weld position of medium-thickness plate workpiece, two 6-DOF Fanuc M-10iD/8L welding robots equipped with a depth camera are used for data acquisition. As can be seen from Fig. 13, the data acquisition is enabled by the cooperation of four components: welding robot, depth camera, guide rail and gantry truss. During the data acquisition process, the gantry truss mounted on the guide rail moves in the X direction to help the manipulator scan the entire workpiece area. In the process of scanning, the image of the workpiece is taken by the depth camera at different distances and angles.

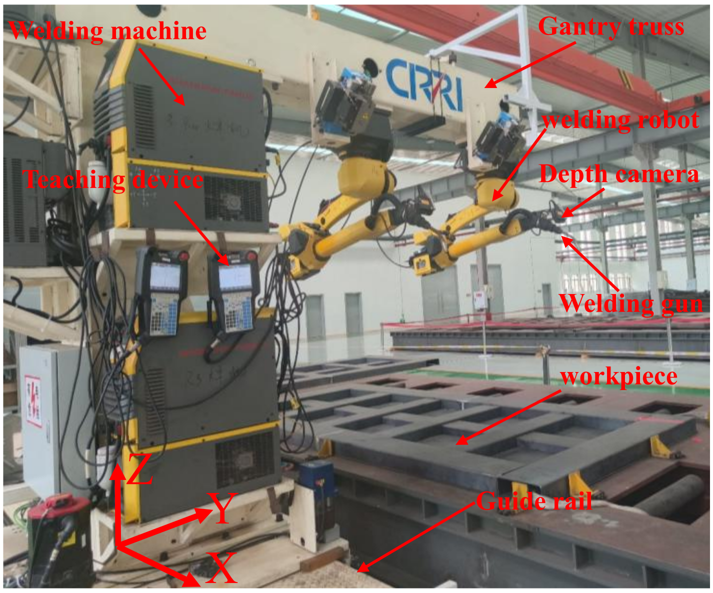


Fig. 13. Data acquisition equipment

To ensure the representativeness of the experimental results, the acquisition of image data is guided by the following principles. The dataset primarily includes images of the baffles and frames of engineering special vehicles, along with images for steel structures. These images depict intricate weld types within the welding environments of medium-thickness plate workpieces. To this end, a total of 500 images of medium-thickness workpiece with complex weld types are collected. Subsequently, the dataset is split into training, validation, and test sets in an 8:1:1 ratio following the COCO format. Ultimately, the Labelme is used to label the training and validation set.

To accurately extract target features and prevent overfitting during the model training phase, this study employs Mosaic data augmentation method. The specific implementation of Mosaic data augmentation is shown in Fig. 14. Initially, reference points for image stitching were randomly selected. Subsequently, four medium-thickness plate weld images are resized and normalized, and the dimensional alteration of each image is mapped onto its corresponding label. The final step involved stitching the images together, with any detection boxes extending beyond the stitched image boundary being appropriately cropped. The advantage of this methodology lies in its capacity to augment the diversity of the dataset, thereby mitigating issues such as reduced model accuracy attributable to data scarcity. Furthermore, it enhances the model's robustness and elevates its detection capabilities for subtle target features, notably the initial weld position of the medium-thickness plate. In the experimental part, this paper also compares and analyzes the specific impact of Mosaic data augmentation method.

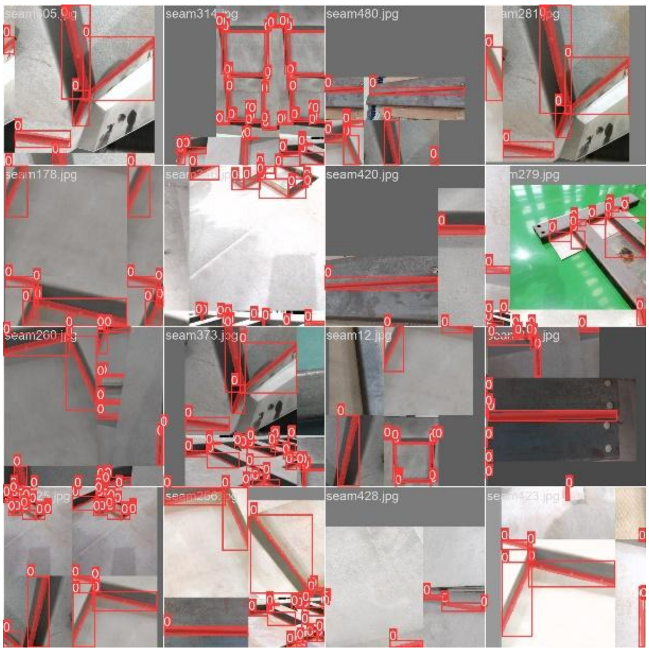


Fig. 14. The results of Mosaic data augmentation

4.2 Model training

To balance computational cost and efficiency, facilitate model deployment in practical production, and ensure transparency, verifiability, and repeatability, the hardware configuration and software environment used in this experiment are detailed below. The experiments in this paper are carried out under the win10 operating system. The hardware configuration of the industrial computer includes an Intel(R) Core(TM) i5-12400F@ 2.50 GHz CPU, a 12GB NVIDIA GeForce RTX 3060 GPU and a 64GB RAM. Furthermore, this study considers the model's deformation calculation requirements and the associated training time costs. The deep learning environment used in the experiment is configured with PyTorch 1.13, Python3.10, CUDA11.7, Ultralytics8.0, and OpenCV-Python 4.6.

The YOLOv8 model offers five distinct compound scaling constants: YOLOV8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. To balance the model's parameter size and accuracy, YOLOv8s is chosen as the compound scaling constant for the model training phase. Subsequently, the pre-trained weights and model configurations associated with YOLOv8s are selected. Furthermore, this paper ensures the learning speed and convergence of the model while fully accounting for the computational efficiency and hardware memory constraints. Table 1 presents the detailed configuration of the model training parameters.

Table 1. Model training parameters

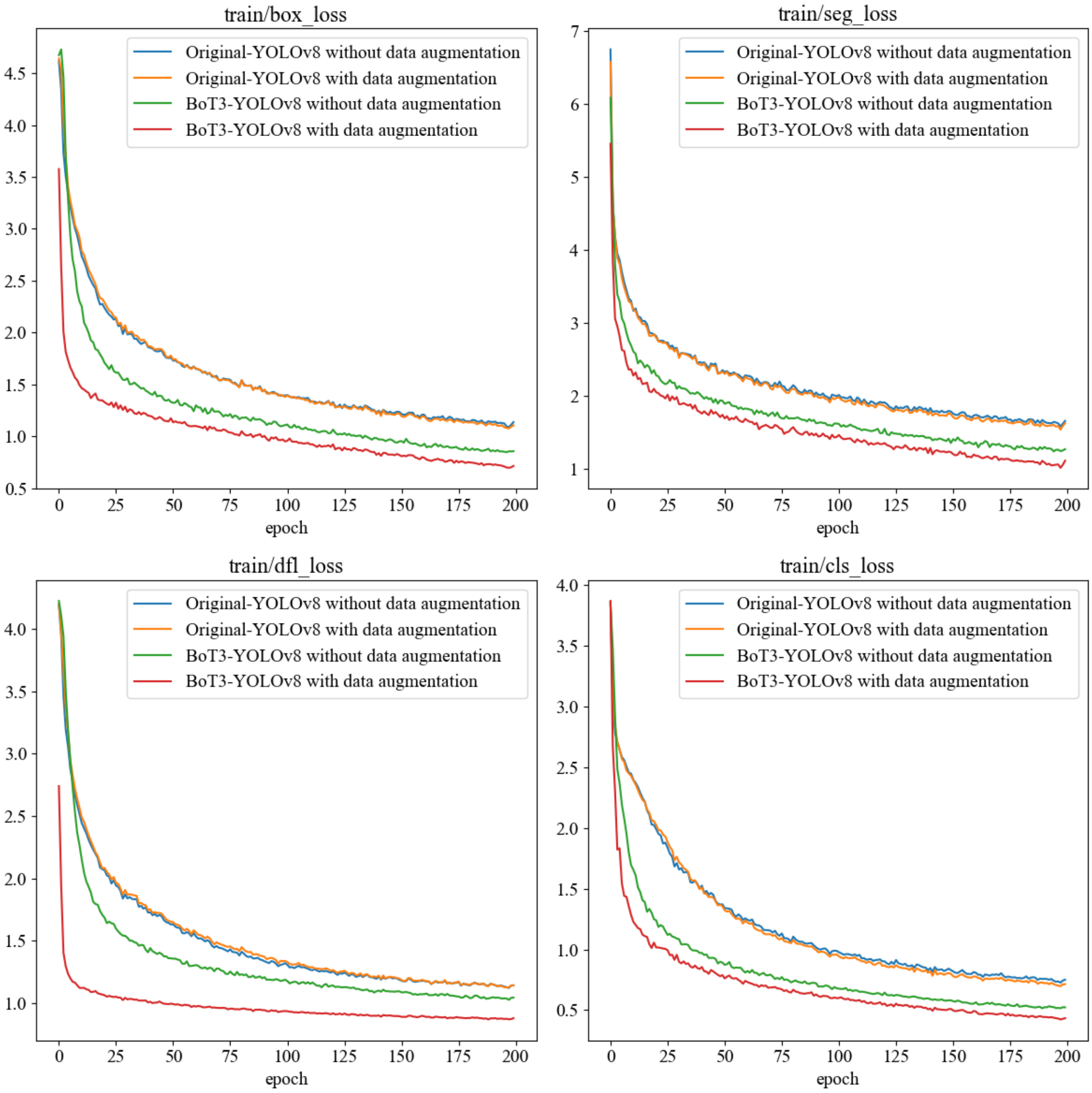
Weights	Model	Epochs	Batch-size	Image-size
YOLOv8s-seg.pt	YOLOv8s-seg.yaml	200	16	640×640
Optimizer	**Learning rate**	**Workers**	**Seed**	**Mask ratio**
SGD	0.01	8	0	4

To alleviate the video memory usage and enhance the inference velocity of the model, specific adjustments to the model's hyperparameters are detailed as follows. Before training, the momentum factor of the model is set to 0.937, the weight decay is set to 0.0005 and the Automatic Mixed Precision (AMP) training mode is enabled. In addition, to prevent model overfitting, the training is terminated early if the model still does not improve after 50 epochs.

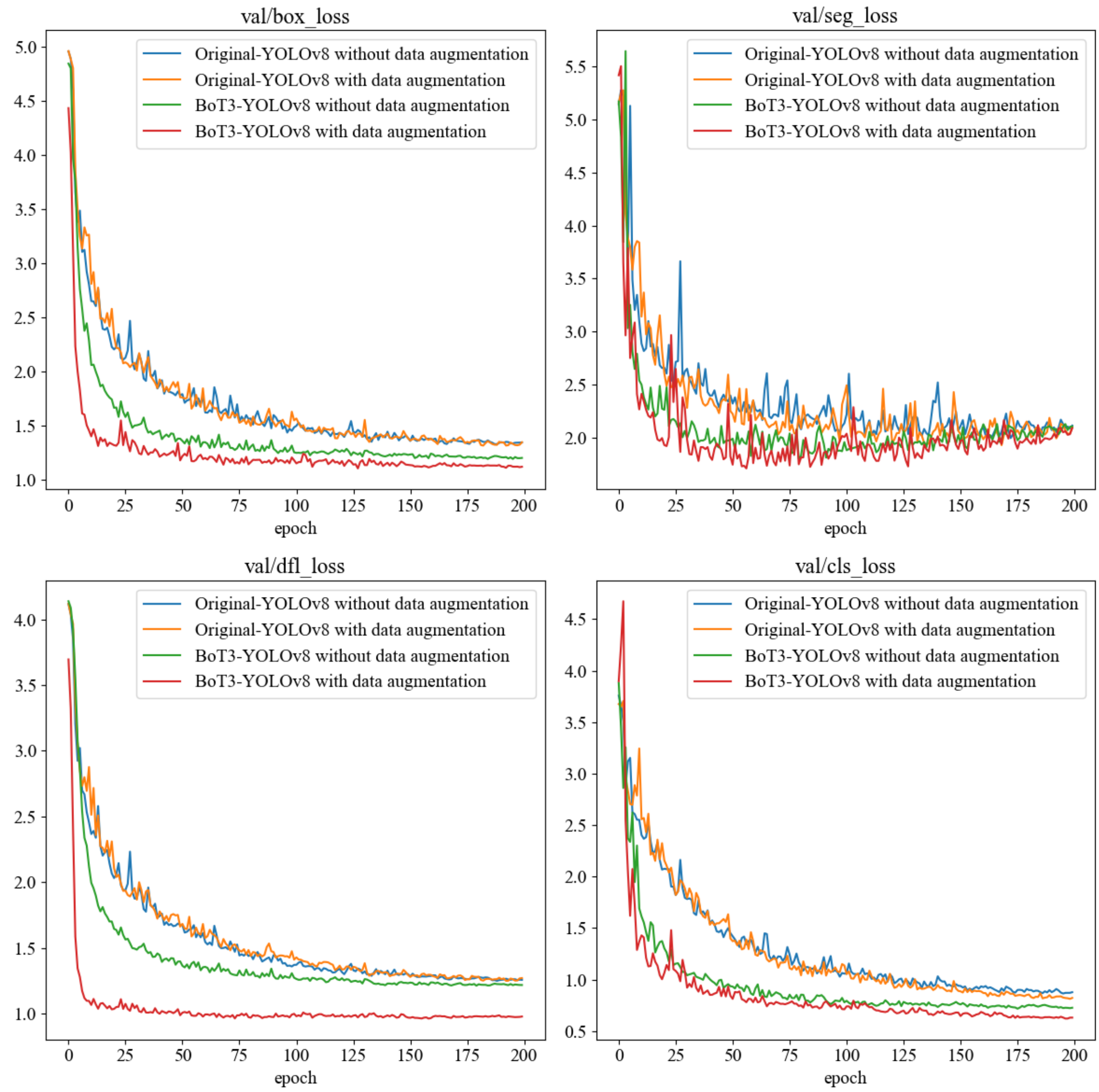
4.3 Benchmarking with existing methods

4.3.1 Analysis of training results

In order to verify the effectiveness and superiority of the improved model, this paper conduct ablation experiments on the Original-YOLOv8 and BoT-YOLOv8 respectively under ensuring the same model training parameters. The training and validation results are shown in Fig. 15 and Fig. 16, respectively. As can be seen, the Original-YOLOv8 model exhibits limitations in feature extraction for specific targets, such as the initial weld position of the medium-thickness plate. It also suffers from slow inference speed and tends to misidentify background features as target features. In practical production scenarios, these shortcomings can lead to incorrect welding and improper vision guidance for robotic manipulation of medium-thickness plate workpieces, potentially resulting in safety hazards such as welding gun collisions. In contrast, the BoT-YOLOv8 model demonstrates enhanced feature extraction capabilities and faster inference speed for identifying the initial weld position of medium-thickness plate, thereby improving the accuracy and efficiency of visual guidance for robotic operations.



(a) Comparison of train/loss functions



(b) Comparison of val/loss functions

Fig. 15. The variation trend of loss functions during training

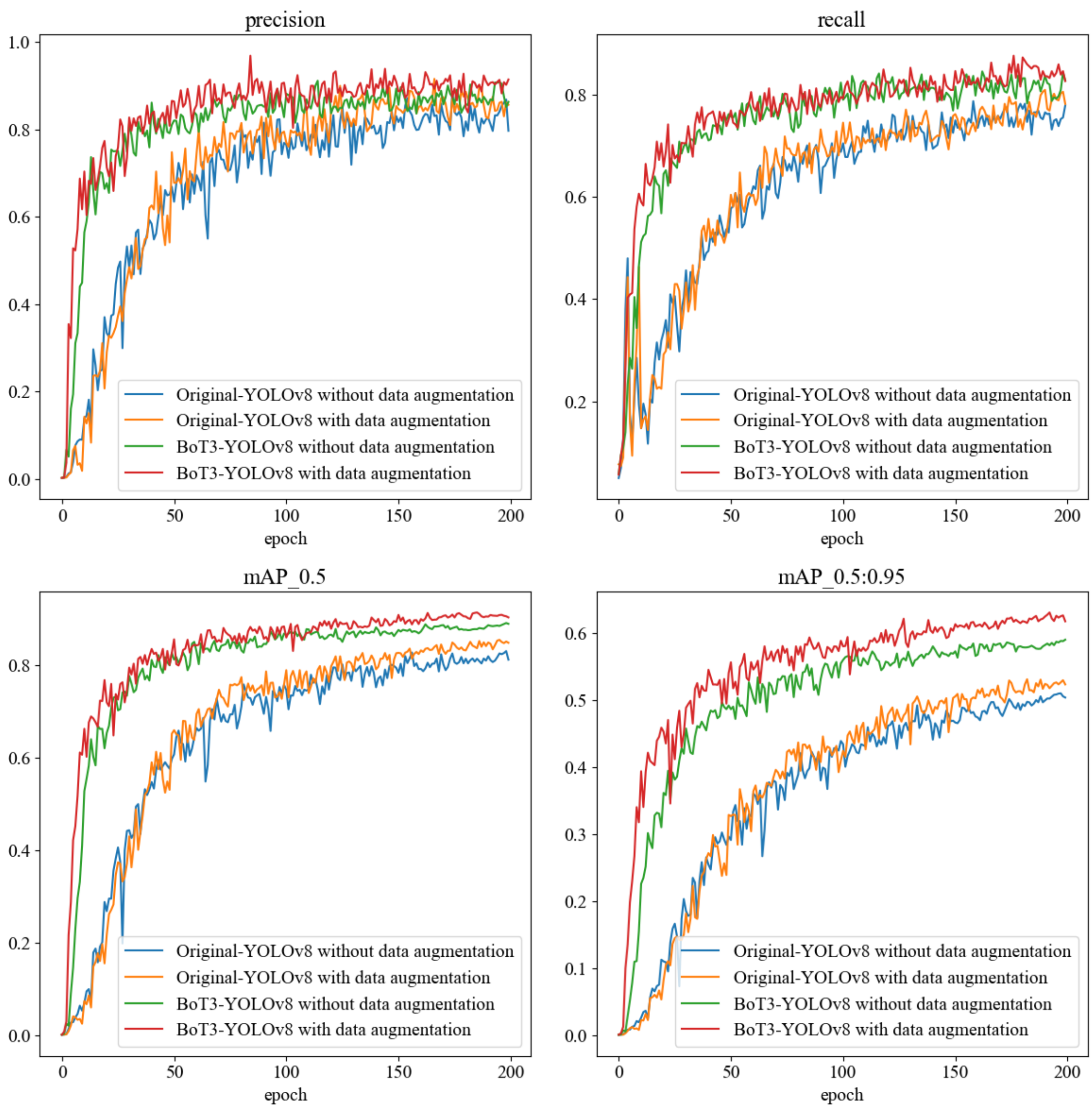


Fig. 16. The variation trend of evaluation indicators during training

By comparing the change of the loss function curve during training and validation, it is obvious that the loss function of the improved model converges faster, and the convergence effect is better. This indicates that the improved model has stronger learning ability for the given samples. According to Fig. 16, the precision, recall rate and mean average precision of the model are significantly improved by fusing the Bottleneck Transformer and replacing the pyramid pooling structure. Furthermore, it can be clearly observed that the Mosaic data augmentation technique enhances both the inference speed and the average precision of the model, and effectively prevents the model from overfitting. The performance parameters before and after model improvement are shown in Table 2.

Table 2. Performance parameters before and after model improvement

Model	$mAP^{0.5}$	$Paramters$/M	$FLOPs$/G	FPS
Original-YOLOv8	87.8%	45.1	42.7	35.2
BoT-YOLOv8	93.1%	56.4	48.9	26.5

It can be seen from Table 2 that compared with the Original-YOLOv8, the mean average precision ($mAP^{0.5}$) of BoT-YOLOv8 is increased by 5.3%, the $paramters$ and $FLOPs$ amount of our model are increased by 11.3M and 6.2G respectively, and the FPS is decreased by 8.7. This indicates that advanced model performance can be obtained by slightly increasing model parameters and network computation, and the accuracy is significantly improved. Furthermore, with the aim of verifying the superiority of the proposed method, three weld feature extraction methods (UNet, SOLOv2, DeepLabV3+) are selected for comparison. The results are shown in Fig. 17.

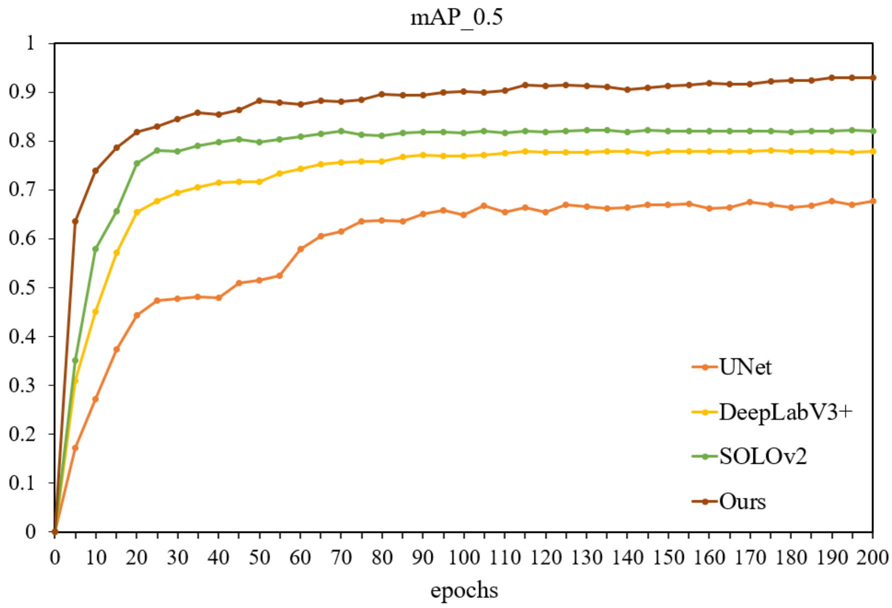


Fig. 17. Comparison results of different methods

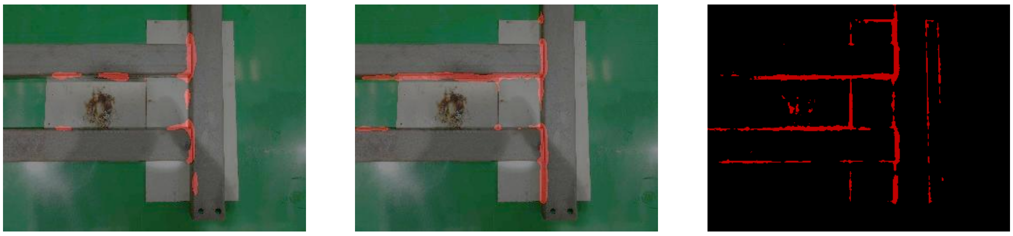
As can be seen in Fig. 17, all methods converged after training for 200 epochs. Among them, the convergence rate of the proposed method is fastest. In terms of model accuracy, the $mAP^{0.5}$ of the propound method basically reaches more than 90%. However, the $mAP^{0.5}$ of the other three methods are significantly lower than that of the proposed method. Specifically, the $mAP^{0.5}$ of SOLOv2 reaches 80%, the $mAP^{0.5}$ of

DeepLabV3+ is around 75%, and the $mAP^{0.5}$ of UNet only reaches 60%. In essence, the propound method demonstrates better theoretical effectiveness in the model training stage.

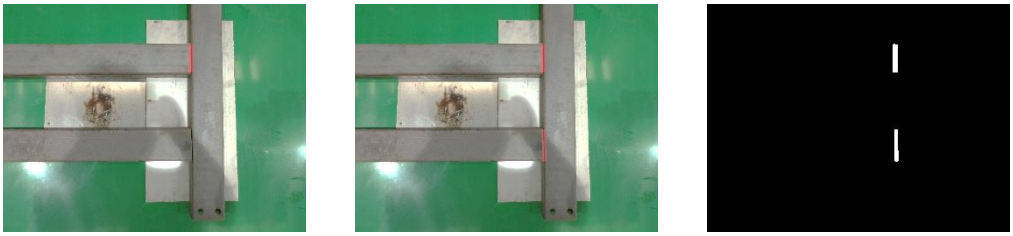
The comparative experimental results indicate that the instance segmentation accuracy for medium-thickness plate weld features achieved by the proposed method significantly outperformed the other three methods. What’s more, the model's training is the fastest. As such, the proposed method has great potential in actual industrial settings to provide robotic vision guidance for medium-thickness plate workpieces with complex structures. Consequently, it is better positioned to satisfy the precision and efficiency requirements essential of robotic welding for medium-thickness plate.

4.3.2 Analysis of prediction results

After model training, in order to verify the anti-interference ability and generalization ability of the proposed method. Four different scenarios and weld types are set up with the model training weights directly applied. Subsequently, the segmentation effect of each method is analyzed and compared. These four scenarios closely resemble actual welding sites, where baffles and frames welding scenes of engineering vehicles are included, as well as steel structure workpieces. The weld types featured in these medium-thickness plate workpieces are highly representative, encompassing butt weld, lap weld, end weld, corner weld, and T-shape weld. Moreover, these scenarios are selected to ensure that the images contain comprehensive features covering the initial weld position of the medium-thickness plate, thereby enhancing the credibility of the experimental results presented in this paper. The specific results are shown in Figs. 18-21.



(a) UNet (b) SOLOv2 (c) DeepLabV3+



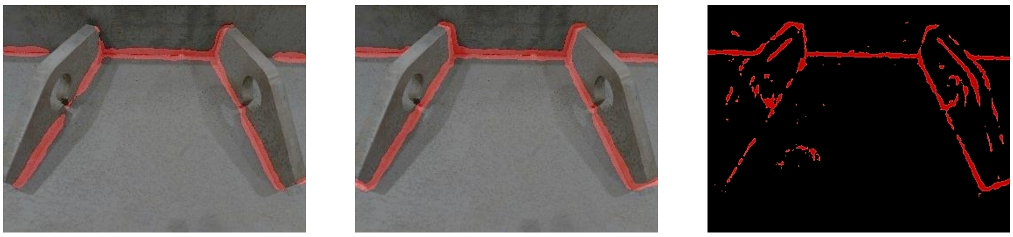
(d) Original-YOLOv8 (e) BoT-YOLOv8 (f) Post-processing

Fig. 18. Butt weld

(a) UNet (b) SOLOv2 (c) DeepLabV3+

(d) Original-YOLOv8 (e) BoT-YOLOv8 (f) Post-processing

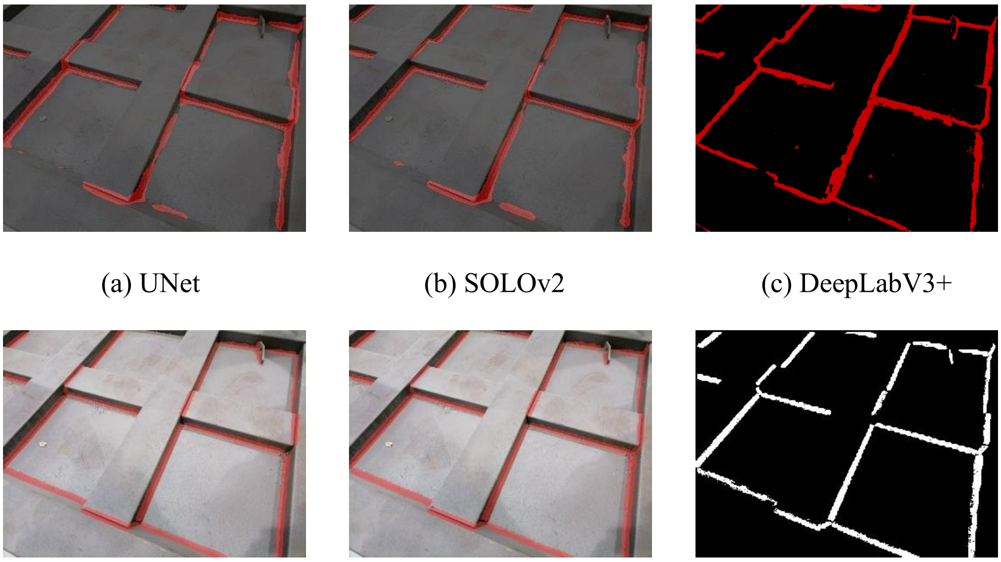
Fig. 19. Lap weld



(a) UNet (b) SOLOv2 (c) DeepLabV3+

(d) Original-YOLOv8 (e) BoT-YOLOv8 (f) Post-processing

Fig. 20. Close-range hybrid weld (corner weld, end weld, T shape)



(a) UNet (b) SOLOv2 (c) DeepLabV3+

(d) Original-YOLOv8 (e) BoT-YOLOv8 (f) Post-processing

Fig. 21. Remote hybrid weld (corner weld, butt weld, lap weld, T shape, end weld)

As can be observed in Figs. 18-21, the proposed method has the best segmentation performance for medium-thickness plate weld images of different scenes and types. Not only target features in each weld image are accurately segmented, but also the mask contour and mask shape of the presented method are more in line with the initial weld position. Furthermore, the HSV region segmentation method is utilized to binarize the image mask, which is convenient for the welding robot to obtain the position information of the target feature and realize the vision guidance before welding. In sharp contrast, Original-YOLOv8, UNet, and SOLOv2 are suffering from false detections, missed detections, and irregular mask shapes. By comparing the prediction results, it is found that the DeepLabV3+ model is more sensitive to the contour of the target object, so DeepLabV3+ is easy to misidentify the contour of medium-thickness plate as a weld.

The occurrence of false detections can be attributed to the classifiers inside these models due to their insufficient discriminative capacity to differentiate between various types of medium-thickness plate welds. Meanwhile the missed detections could be due to the insufficient sensitivity of these methods to identify the characteristics of tiny medium-thickness plate welds within imagery or their inability to distinguish between target features and background elements. With such false detections and missed detections, collisions of the welding gun may happen during the actual robotic welding process, thus compromising weld formation quality and even failing production standards. To this end, the proposed method could effectively address such issues with improved performance, thus providing promising potential for industrial application.

In summary, the proposed method and four existing methods are implemented in this study. Subsequently, the training results of the five methods are analyzed and compared using loss function, evaluation index and performance parameters. Finally, the interference factors are added in the prediction process and compare the prediction effects of these five methods on different types of weld images in different scenes. The experimental results show that.

(1) The loss function of the improved model converges faster, the $mAP^{0.5}$ is highest, and the performance of the model is significantly improved.

(2) Compared with other methods, the proposed method has the highest detection accuracy, the best anti-interference ability and generalization ability.

5. Conclusion

Aimed to address the technical challenge of autonomous vision guidance for welding robots in medium-thickness plate welding scenarios, this paper focuses on the initial weld position features segmentation to assist in path planning for the welding robot. This study enhances the segmentation accuracy of the initial weld position for more precise identification and localization of the welding area in medium-thickness plate. This advancement lays a solid foundation for achieving superior welding quality and uniformity. The improved precision is expected to further bolster the reliability of the automated welding process and augment production efficiency. The main conclusions of this study are summarized as follows:

- A highly accurate and stable initial welding position segmentation method is proposed for medium-thickness plate. The method improvesYOLOv8 with a novel the C2f_Bottleneck_BoT module and utilizes the ASPP module as spatial pyramid pooling structure.

- The HSV region segmentation method is employed as a post-processing technique to extract the weld seam mask binary image from the model's image prediction results. This facilitates the welding robot in acquiring precise position information of the weld seam features.
- The experimental results demonstrate that the BoT-YOLOv8 achieves the best trade-off between accuracy (93.1% $mAP^{0.5}$ for overall) and speed (the inference speed of a single image is 58.3ms, the FPS is 26.5). In comparison with other methods (UNet, DepplabV3+, SOLOv2, Original-YOLOv8), the proposed method exhibits superior reliability and exceptional anti-jamming capabilities.

This research is expected to enhance the precision and automation level of robot welding system with lower error rates, thus providing great potential to reduce labor cost and improve production efficiency. Future work may include integrating the proposed method with embodied artificial intelligence techniques to further advance robotic welding automation. In the subsequent research, we will leverage on the proposed method to enable adaptive welding path planning for medium-thickness plate in robotic systems.

Author contribution Zongmin Liu: Conceptualization, Supervision, Data curation, Writing– review & editing. Jie Li: Methodology, Writing – original draft. Shunlong Zhang: Validation, Experiment. Lei Qin: Validation, Experiment. Changcheng Shi: Data curation. Ning Liu: Conceptualization, Supervision, Writing – review & editing.

Founding This work was supported by grants of the Chongqing Technology Innovation and Application Development Special Key Project (No: cstc2021jscx-gksbX0030), the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No: KJQN202200826) and the Graduate Scientific Research and Innovation Foundation of Chongqing Technology and Business University (No: yjscxx2023-211-52).

Data availability The data that supports the findings of this study are available from the corresponding author upon reasonable request.

Code availability Not applicable

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest Not applicable.

References

1. Yang L, Liu Y, Peng J (2020) Advances techniques of the structured light sensing in intelligent welding robots: A review. Int J Adv Manuf Technol 110(3):1027-1046. https://doi.org/10.1007/s00170-020-05524-2.
2. Xu Y, Wang Z (2021) Visual sensing technologies in robotic welding: recent research developments and future interests. Sens Actuator A Phys 320(1):112551. http://dx.doi.org/10.1016/j.sna.2021.112551.
3. Liu F, Wang Z, Ji Y (2018) Precise initial weld position identification of a fillet weld seam using laser vision technology. Int J Adv Manuf Technol 99(5-8):2059-2068. https://doi.org/10.1007/s00170-018-2574-9.
4. Wang N, Zhong K, Shi X, Zhang X (2020) A robust weld seam recognition method under heavy noise based on structured-light vision. Robot Comput Integr Manuf 61:101821. https://doi.org/10.1016/j.rcim.2019.101821.
5. Li W, Mei F, Hu Z, Gao X, Yu H, Housein A, Wei C (2022) Multiple weld seam laser vision recognition method based on the IPCE algorithm. Opt Laser Technol 155(5-8):108388. https://doi.org/10.1016/j.optlastec.2022.108388.
6. Ma X, Pan S, Li Y, Feng C, Wang A (2019) Intelligent welding robot system based on deep learning. CAC IEEE, China, Hangzhou. https://doi.org/10.1109/CAC48633.2019.8997310.

7. Xiao R, Xu Y, Hou Z, Chen C, Chen S (2019) An adaptive feature extraction algorithm for multiple typical seam tracking based on vision sensor in robotic arc welding. Sens Actuator A Phys 297(9-12):111533. https://doi.org/10.1016/j.sna.2019.111533.

8. Zeng J, Cao G, Peng Y, Huang S (2020) A weld joint type identification method for visual sensor based on image features and SVM. Sensors 20(2):471. https://doi.org/10.3390/s20020471.

9. Tian Y, Liu H, Li L, Wang W, Feng J, Xi F, Yuan G (2020) Robust identification of weld seam based on region of interest operation. Adv Manuf 8(4):473-485. https://doi.org/10.1007/s40436-020-00325-y.

10. Tian Y, Liu H, Li L, Yuan G, Wang W (2021) Automatic identification of multi-type weld seam based on vision sensor with silhouette-mapping. IEEE Sens J 21:5402-5412. https://doi.org/10.1109/JSEN.2020.3034382.

11. Wei H, Zhao H, Shi X, Liang S (2022) Nonlinear identification and control of laser welding based on rbf neural networks. Comput Syst Sci Eng 41(1):51-65. https://doi.org/10.32604/csse.2022.017739.

12. Yang L, Liu Y, Peng J, Liang Z (2020) A novel system for off-line 3d seam extraction and path planning based on point cloud segmentation for arc welding robot. Robot Comput Integr Manuf 64(3):101929. https://doi.org/10.1016/j.rcim.2019.101929.

13. Kim J, Lee J, Chung M, Shin Y (2021) Multiple weld seam extraction from RGB-depth images for automatic robotic welding via point cloud registration. Multimed Tools Appl 80(13):9703-9719. https://doi.org/10.1007/s11042-020-10138-7.

14. Geng Y, Lai M, Tian X, Xu X, Jiang Y, Zhang Y (2022) A novel seam extraction and path planning method for robotic welding of medium-thickness plate structural parts based on 3D vision. Robot Comput Integr Manuf 79(9-12):102433. https://doi.org/10.1016/j.rcim.2022.102433.

15. Ma Y, Fan J, Zhou Z, Zhao S, Jing F, Tan M (2023) WeldNet: A deep learning based method for weld seam type identification and initial point guidance. Expert Syst Appl

238:121700. https://doi.org/10.1016/j.eswa.2023.121700.

16. Zhou F, Liu X, Jia C, Li S, Tian J, Zhou W, Wu C (2023) Unified CNN-LSTM for keyhole status prediction in PAW based on spatial-temporal features. Expert Syst Appl 237:121425. https://doi.org/10.1016/j.eswa.2023.121425.
17. Oskar N, Diyah U, Andi D (2021) Deep learning-based weld spot segmentation using modified UNet with various convolutional blocks. ICIC Express Lett 12(12):1169-1176. https://doi.org/10.24507/icicelb.12.12.1169.
18. Chen B, He S, Liu J, Chen S, Lu E (2023) Weld structured light image segmentation based on lightweight DeepLabv3+ network. Chin J Lasers 50(8):0802105. https://doi.org/10.3788/CJL221398.
19. Zou Y, Zeng G (2023) Light-weight segmentation network based on SOLOv2 for weld seam feature extraction. Measurement 208:112492. https://doi.org/10.1016/j.measurement.2023.112492.
20. Wu Z, Gao P, Han J, Bai L, Lu J, Zhao Z (2022) Real-time segmentation network for accurate weld detection in large weldments. Eng Appl of Artif Intel 117(4):105008. https://doi.org/10.1016/j.engappai.2022.105008.
21. Varghese R, Sambath M (2024) YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. ADICS IEEE, India, Goa. https://doi.org/10.1109/ADICS58448.2024.10533619.
22. Srinivas A, Lin T, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck Transformers for Visual Recognition. CVPR IEEE, USA, Nashville, TN. https://doi.org/10.1109/CVPR46437.2021.01625.
23. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille A (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834-848. https://doi.org/10.1109/TPAMI.2017.2699184.
24. Liu C, Shen J, Hu S, Wu D, Zhang C, Yang H (2022) Seam tracking system based on laser vision and CGAN for robotic multi-layer and multi-pass MAG welding. Eng

Appl of Artif Intel 116:105377. https://doi.org/10.1016/j.engappai.2022.105377.

25. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. CVPR IEEE, USA, Las Vegas, NV. https://doi.org/10.1109/CVPR.2016.90.

26. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: un ified, real-time object detection. CVPR IEEE, USA, Las Vegas, NV. https://doi.org/10.48550/arXiv.1506.02640.

27. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. ICCV IEEE, Italy, Venice. https://doi.org/10.1109/ICCV.2017.322.

28. Ge J, Deng Z, Li Z, Li W, Liu T, Zhang H, Nie J (2022) An efficient syste m based on model segmentation for weld seam grinding robot. Int J of Ad v Manuf Technol 121(11-12):7627-7641. https://doi.org/10.1007/s00170-022-09758-0.

29. Ai Y, Lei C, Yuan P, Cheng J (2022) Analysis of weld seam characteristic parameters identification for laser welding of dissimilar materials based on image segmentation. J Laser Appl 34(4):042050. https://doi.org/10.2351/7.0000734.

30. Chen C, Chen T, Cai Z, Zeng C, Jin X (2023) A hierarchical visual model f or robot automatic arc welding guidance. Ind Robot 50(12):299-313. https://doi.org/10.1108/ir-05-2022-0127.

31. Guo F, Zheng W, Lian G, Yao M (2023) A V-shaped weld seam measuring system for large workpieces based on image recognition. Int J Adv Manuf Tec hnol 124(4):229–243. https://doi.org/10.1007/s00170-022-10507-6.

32. Deng L, Lei T, Wu C, Liu Y, Cao S, Zhao S (2023) A weld seam feature real-time extraction method of three typical welds based on target detection. Measurement 207:112424. https://doi.org/10.1016/j.measurement.2022.112424.

33. Ai Y, Han S, Lei C, Cheng J (2023) The characteristics extraction of weld seam in the laser welding of dissimilar materials by different image segment ation methods. Opt Laser Technol 167:109740. https://doi.org/10.1016/j.optlastec.2023.109740.