

Full length article

A real-time welding defect detection framework based on RT-DETR deep neural network[☆]Gaoyang Liu^{a,b}, Duanrui Yang^a, Jun Ye^{c,*}, Hongjia Lu^d, Zhen Wang^e, Yang Zhao^a^a School of Civil Engineering, Shaoxing University, Shaoxing, China^b Institute of Artificial Intelligence, Shaoxing University, Shaoxing, China^c School of Civil Engineering, University of Leeds, Leeds, UK^d Centre for Innovative Structures and Materials, School of Engineering, RMIT University, Melbourne 3001, Australia^e Department of Civil Engineering, Hangzhou City University, Hangzhou, China

ARTICLE INFO

Keywords:

Weld defect detection

Deep learning

RT-DETR

Data enhancement

Additive manufacturing welds

ABSTRACT

The quality of welds is critical to the safety and reliability of steel structure connections, underscoring the importance of accurate inspection during the welding process. To enhance inspection effectiveness, deep learning methods have gained popularity in weld defect detection for their ability to automatically learn and refine image features. However, the complex multi-stage training and inference process of these methods often fails to meet the requirements of real-time performance and accuracy. To address this problem, a framework based on the Real-Time DEtection TRansformer (RT-DETR) for deep learning-based welding defect detection is proposed. This framework improves the Transformer backbone by eliminating the most time-consuming non-maximum suppression (NMS) step, achieving real-time detection without sacrificing accuracy. A diverse welding dataset with 1,134 images from real-world manufacturing and construction environments was developed for model training and validation. In addition, three data enhancement algorithms were explored to enhance the model's generalization ability. The model achieved detection accuracy scores of mAP@0.5 at 0.996 and mAP@0.5:0.95 at 0.801, with a detection speed of 67 frames per second (FPS). Compared to the previous Faster R-CNN, SSD, YOLOv5, YOLOv11 and DETR models, the proposed RT-DETR model demonstrates superior efficiency and accuracy. The proposed framework was further validated in the on-site inspections of metal additive manufacturing, and the results confirmed that the RT-DETR-based model meets the stringent requirements for real-time inspection in metal additive manufacturing.

1. Introduction

Welding is one of the most popular and widely used methods for connecting steel structures across various industries, including construction, aerospace, and automotive manufacturing[1]. The technology connects metals through high-temperature heating, making it essential to the integrity of numerous applications. The quality of welding is critical, as weld defects can not only affect the appearance of equipment but also compromise its performance and safety. Therefore, timely inspection and thorough assessment of welded joints are crucial for ensuring the safety and reliability of industrial facilities [2–4].

Traditional weld defect inspection involves collecting data using methods such as X-ray, laser scan, ultrasonic, penetration, eddy current,

and magnetic powder, followed by processing the data with computer-aided algorithms[5–7]. Raba et al.[8] developed a Chan-Vese model to segment the X-ray image and detect weld defect boundaries. Domingo et al.[9] employed an edge detection method based on the Laplacian-of-Gauss operator to segment potential defects and extract texture features for weld detection. Cui et al.[10] proposed an image recognition method combining grey gradient co-occurrence matrix (GGCM), cluster analysis and mathematical morphology, effectively identifying various weld conditions such as no defects, cracks and fusion cracks. Li et al.[4] designed a weld contour monitoring and defect detection system based on the line structured light, with an image processing algorithm to extract laser contour feature points and measure the dimensions of multi-layer welds. However, the accuracy of these methods heavily

[☆] This article is part of a special issue entitled: 'AI for the Built World' published in Advanced Engineering Informatics.

* Corresponding author.

E-mail address: j.ye2@leeds.ac.uk (J. Ye).

<https://doi.org/10.1016/j.aei.2025.103318>

Received 15 November 2024; Received in revised form 25 March 2025; Accepted 31 March 2025

Available online 7 April 2025

1474-0346/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

relies on human experience, leading to low efficiency, high subjectivity and unreliable inspection results. Therefore, developing an efficient and accurate end-to-end detection algorithm for weld defects is crucial for practical application in various scenarios[11].

With the advancement of deep convolutional neural networks[12] (CNN), target detection algorithms have gained powerful capabilities in extracting features and learning complex patterns. As a result, CNN has significantly improved the accuracy of image recognition, making it highly applicable to weld detection. Various methods and models have been developed to harness this potential. For example, Asif et al.[13] optimized the Vgg16 model by adding a fully connected flexible maximum activation function and modifying the top layer, improving both accuracy and speed in detecting welding defects. Yang et al.[1] investigated different CNN models and demonstrated that employing yolov5 to check defects in steel pipe welds can improve detection accuracy while meeting real-time detection standards. Paolo[14] trained an artificial neural network model and developed a system for the detection of welding surface defects. Yan et al.[15] combined a CNN with a pre-trained support vector machine classifier (SVM) to extract high-level features from the time–frequency representation of electromagnetic ultrasonic testing signals, developing a deep learning-based ultrasonic pattern recognition method for detecting cracks in pipeline girth welds. Kim et al.[16] introduced a deep learning engine for non-destructive automatic detection of internal defects in welds. Wang et al.[17] shortened the path between the lower and upper layers by adding a feature pyramid structure (FPN)[18], thereby improving detection accuracy and speed. Based on R-CNN, Chen[19] employed a deep residual network, Res2Net, to enhance the original backbone network, improving feature extraction capabilities even for small weld targets. However, despite these advances, many models have complex structures and require multi-stage training, resulting in lower computational efficiency and challenges in achieving real-time detection. In the related research on defect detection framework, Gao et al.[20] proposes a multi-stage active learning framework (MALF) for steel surface defects, combining soft weighted label assignment, evidence branching, and diversity criteria for high-precision detection and localization. Al-Otum[21] utilizes a multi-scale convolutional neural network model, incorporating transfer learning methods (ResNet18 and ShuffleNet) and an independent lightweight deep CNN (CNN-ILD), combined with electroluminescence (EL) imaging data for photovoltaic cell defect detection and classification. Liu et al.[22] proposes a semi-supervised deep learning method based on autoencoders and Assisted Anomaly Localization (AEAL), which trains exclusively on normal samples and enables end-to-end defect detection and localization. Yang et al.[23] proposes a two-stage detection framework based on the YOLO deep learning model for the fast and accurate detection of bolt loosening in wind turbine tower structures.

In complex welding environments with high temperatures and fast speeds, such as metal additive manufacturing, real-time and accurate detection of weld defects is crucial for structural safety and production efficiency. Without real-time detection, defects may go unnoticed due to the overlapping manufacturing layers, leading to hidden flaws. Most real-time detectors rely on CNN-based architectures, such as Faster-RCNN, SSD, and YOLO. However, these detectors often require non-maximal suppression (NMS) for post-processing, which slows down inference and introduces hyperparameters that lead to instability in computational efficiency and effectiveness.

Recently, an end-to-end Transformer-based detector (DETR)[24] has been popular due to its streamlined architecture and reduced manual operations. Transformer introduces a self-attention mechanism that enables parallel computation, which can significantly improve the efficiency of training and reasoning. In addition, Transformer does not rely on data's sequential relationship, making it more effective in capturing global features. However, DETR does not have a preset anchor box, and uses the query vector as a soft anchor box to obtain the location of the target. As a result, the soft anchor box usually converges slowly, leading

to longer training time in the original DETR method. To address this issue, a detection model called Real-Time DETection TRansformer[25] (RT-DETR) is proposed. It employs the Hungarian algorithm to identify the optimal matching pairs, ensuring that each query vector is exclusively responsible for a single target. This approach eliminates the need for complex post-processing steps and enables end-to-end object detection, thereby simplifying the detection pipeline and improving detection efficiency. The architecture of RT-DETR is designed to be flexible, easily extensible, and optimizable, making it suitable for scenarios requiring efficient, real-time, and accurate detection, such as in-situ welding using Wire Arc Additive Manufacturing (WAAM), as illustrated in Fig. 1. Unlike traditional frame-by-frame detection models like YOLO, RT-DETR supports parallel decoding, allowing it to process multiple targets simultaneously, further enhancing real-time performance. Additionally, RT-DETR exhibits strong robustness to noise and complex backgrounds, maintaining high detection accuracy even in challenging environments. It has been widely used in marine ship monitoring[26], microbial detection[27], drone detection[28] and other fields. However, there has been no research on the application of RT-DETR in weld detection. This study proposes a novel welding defect detection framework based on RT-DETR. A dataset of 1,134 weld images was collected, and three data augmentation techniques were applied to enhance the model's generalization ability. The framework leverages deep learning to improve the accuracy and efficiency of defect detection. Additionally, it was validated in the context of in-situ WAAM weld inspection, demonstrating that the RT-DETR model meets the stringent real-time performance requirements of industrial applications, effectively addressing the accuracy and speed limitations of traditional weld defect detection methods.

The paper is structured as follows: Section 2 introduces the model principles and data augmentation methods of DETR and RT-DETR; Section 3 introduces the dataset, including the defect categories of the dataset; Section 4 introduces the experimental part, including the configuration of the experimental environment, the setting of neural network hyperparameters and the operation of data augmentation, and the introduction of evaluation indicators; Section 5 presents the experimental results and discussions, including the comparison of training results after data augmentation and the performance comparison with other detection models; Section 6 employs the WAAM printed weld test to verify the proposed detection framework.

2. Summary of the proposed framework

Welding burrs and dents can result from various factors, such as improper welding current or welding speed. These defects can severely compromise the integrity and safety of the welded structure. The proposed framework based on RT-DETR was trained on a welding defect dataset, which was established in this study. It consists of 1134 welding defect images. Data augmentation was performed to enhance the generalizability of the proposed framework. The performance was validated in the context of weld detection tasks for WAAM. The experimental results demonstrate that the RT-DETR-based weld defect detection framework exhibits significant advantages in both detection accuracy and speed. The 4-stage pipeline of the adopted neural network is shown in Fig. 2.

Stage 1: The backbone network is used to extract key weld features. By incorporating the improved PResNet50 network structure, subtle changes in weld features can be captured more effectively. Multi-layer convolution operation improves the model's ability to perceive weld characteristics, ensuring efficient feature extraction. Additionally, the enhanced residual block design improves the retention of weld information and boosts detection accuracy.

Stage 2: The efficient hybrid encoder improves feature extraction and fusion by combining attention mechanisms with convolution operations. The Attention-based Intra-scale Feature Interaction (AIFI) reduces computational redundancy and focuses on capturing

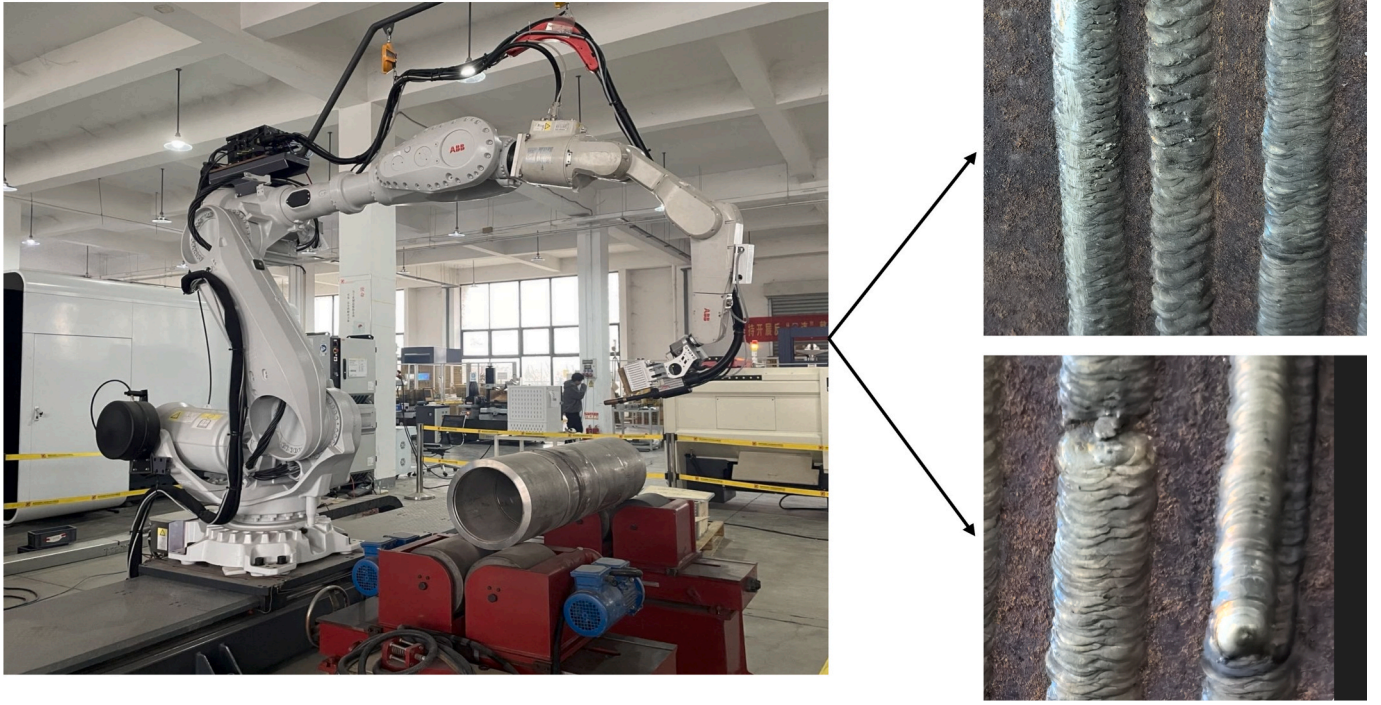


Fig. 1. WAAM on-site welding images.

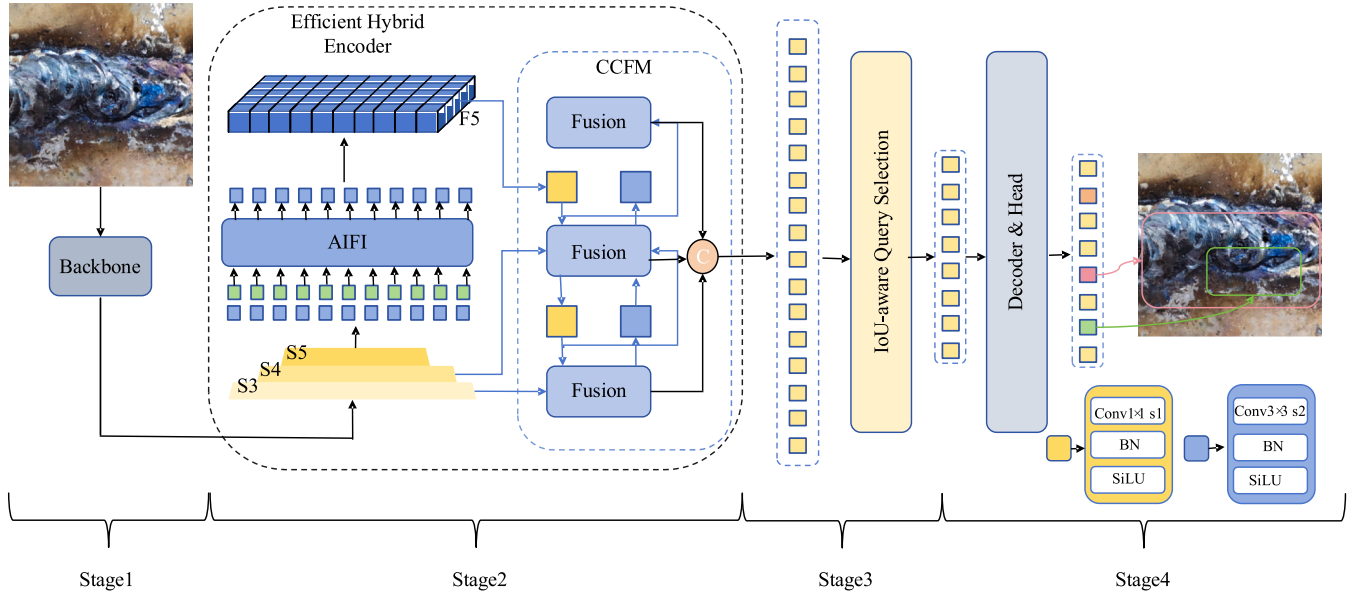


Fig. 2. The weld defect detection framework adopted in this study.

high-level semantic features in welds, aiding in differentiating various weld defects. Meanwhile, the CNN-based Cross-scale Feature-fusion Module (CCFM) fuses multi-scale features through convolution blocks, enriching the overall weld information. This multi-scale feature fusion ultimately improves the detection model's accuracy and efficiency.

Stage 3: IoU-Aware Query Selection constrains model learning by assigning higher classification scores to weld features with high intersection over Union (IoU) while lowering the scores for low IoU features, thereby improving the accuracy of the prediction frame. This method effectively addresses the inconsistency between classification scores and IoU scores and prevents prediction boxes with

high classification scores but inaccurate positions from affecting the detection results. By optimizing object queries, the model is able to select more precise weld features during the decoding phase, ultimately improving the accuracy and robustness of weld defect detection.

Stage 4: The Transformer decoder improves classification and bounding box prediction accuracy through an auxiliary prediction head. Once the decoder optimizes the object query, the auxiliary head maps it into classification confidence and bounding boxes, enabling more precise detection of weld defect characteristics. In addition, the denoising module accelerates the training convergence,

allowing the model to more quickly and accurately learn weld defect locations and classifications.

2.1. Introduction of DETR

In weld defect detection, real-time performance and accuracy are essential, especially in complex steel connections or metal additive manufacturing where multiple layers and beads are involved. The DETR in deep neural networks provides significant advantages in weld defect detection due to its end-to-end nature, which replaces the traditional post-processing operations such as thresholding and non-maximum suppression (NMS) with a single top-k operation on the final predictions, thus significantly reducing the post-processing computational cost. Due to the versatility of the Transformer architecture, DETR demonstrates high performance in computer vision tasks without the need for extensive domain-specific adjustments, highlighting its potential in weld defect detection.

DETR is a transformer-based end-to-end object detector that uses an encoder-decoder structure, with both encoder and decoder comprising multiple superpositions of a self-attention module, capturing long-distance dependencies more effectively. Compared with the traditional recurrent neural network (RNN) model[29], the parallel computing of the multi-head attention mechanism in the Transformer model greatly improves the efficiency of training and inference. In addition, Transformer is applicable to many different types of sequence data, including text, images, and videos.

The model structure of DETR is illustrated in Fig. 3, which consists of five main steps: (1) CNN is used to extract image features, calculate the feature vector corresponding to each patch, and add position encoding; (2) global features are learned through the Transformer Encoder to assist subsequent detection tasks; (3) the encoder output is fed into the Transformer Decoder along with object queries, where the query vector q interacts with the key k and value v from the encoder to generate multiple prediction boxes; (4) the predicted box is matched with the actual box to determine whether the box contains the object, and the loss is calculated in the matched box and backpropagated; (5) the prediction head is established to generate the final prediction result.

By mapping the problem into a set prediction problem, the architecture greatly simplifies the pipeline of object detection and directly outputs the predicted (object, box) set in an end-to-end manner. In the current popular object detection models, when testing mAP on the validation set, two key steps must be completed first: confidence threshold and non-maximum suppression (NMS)[30]. The former filters out low-scoring detection boxes, and the latter filters out redundant responses to the same object. After completing these two steps, the

remaining detection boxes participate in calculating mAP. Specifically, DETR leverages the multi-head self-attention mechanism within the Transformer to capture long-range dependencies within the image, thereby reducing false positives and duplicate boxes. The decoder of DETR outputs prediction boxes that match the number of objects, and the Hungarian Algorithm is used for one-to-one matching to ensure that each predicted box correctly corresponds to a true object, eliminating the need for any additional post-processing steps.[24,31] This innovative approach introduces new possibilities for computer vision tasks and has significant potential for detecting weld defects. However, complex calculations and large amounts of data still need to be processed during training, leading to longer training time.

2.2. Introduction of RT-DETR

To meet the real-time requirements for weld detection, RT-DETR is employed to eliminate the need for post-processing through key improvements, addressing the high computational cost problem of the DETR model. While retaining the same encoder and decoder architecture as DETR, RT-DETR incorporates several optimization improvements: First, it uses smaller feature maps to reduce computational cost significantly. Second, by reducing the number of attention heads, RT-DETR reduces the number of parameters. In addition, a new group attention mechanism is introduced, further enhancing performance while also effectively shortening training time. RT-DETR consists of three parts: the Backbone, Efficient hybrid encoder, and Transformer decoder with auxiliary prediction heads.

2.2.1. Backbone

In this study, the classic ResNet[32] structure in the backbone part is employed. The specific experiment uses PResNet50, whose network structure is shown in Fig. 4. PResNet50 introduces several improvements over ResNet50. First, the original 7×7 convolution layer in ResNet50 is replaced by three 3×3 convolution layers, which enhances the extraction of weld features while maintaining the same output. Second, in the first residual block of stages 2, 3, and 4, the original 1×1 convolution layer (with a kernel size of 1×1 and a stride of 2) is replaced by a 2×2 average pooling layer (kernel size 2×2 , stride 2) combined with a 1×1 convolution layer. This modification helps better retain the key information of the weld, thereby improving feature extraction performance.

2.2.2. Efficient hybrid encoder

For the neck network, RT-DETR uses a Transformer Encoder, referred to as the Efficient Hybrid Encoder, which consists of two parts: Attention-based Intra-scale Feature Interaction (AIFI) and CNN-based

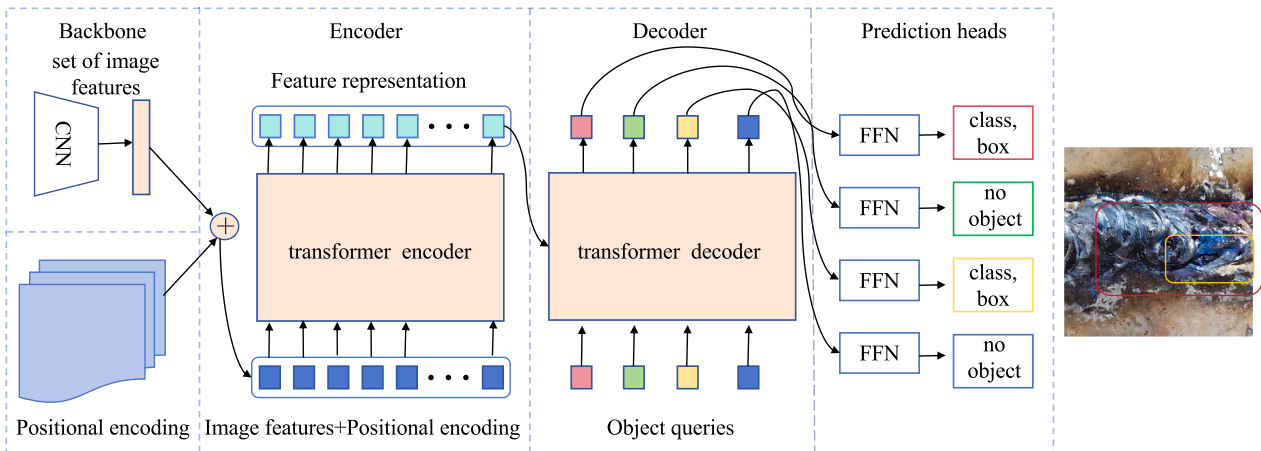


Fig. 3. Architecture of DETR detection neural networks.

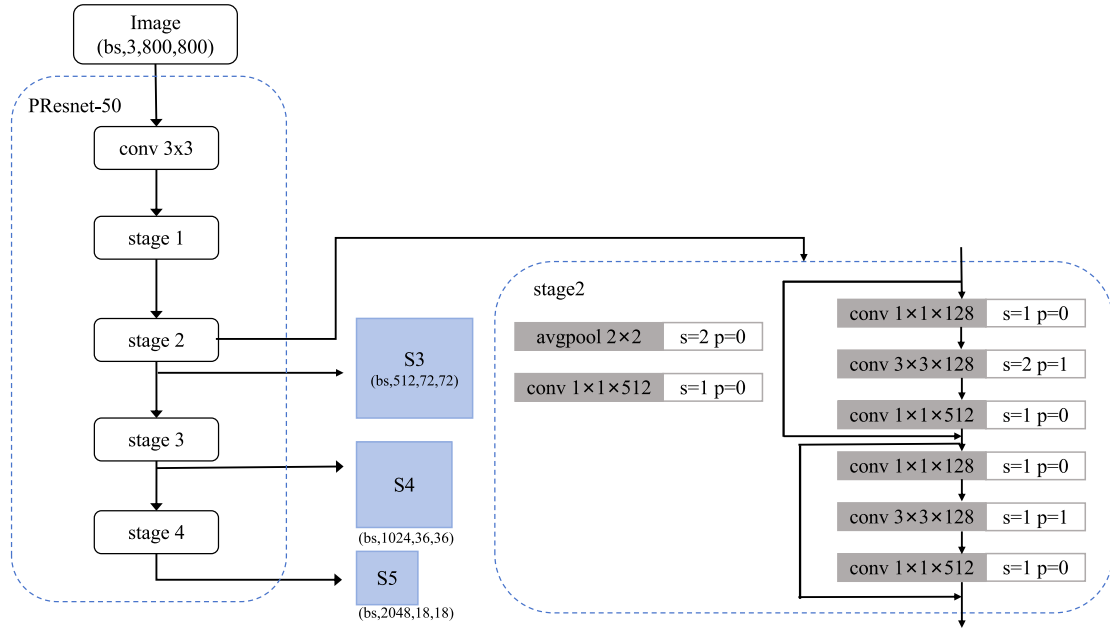


Fig. 4. Detailed architecture of PResNet50.

Cross-scale Feature-fusion Module (CCFM). AIFI reduces computational redundancy, and its structure is shown in Fig. 5. Weld features S3, S4, and S5 are output from stages 2, 3, and 4 of the backbone, respectively. Compared to the shallower S3 and S4 features, the deeper S5 features contain more advanced and richer weld semantic features, which is crucial for the Transformer model in distinguishing between different

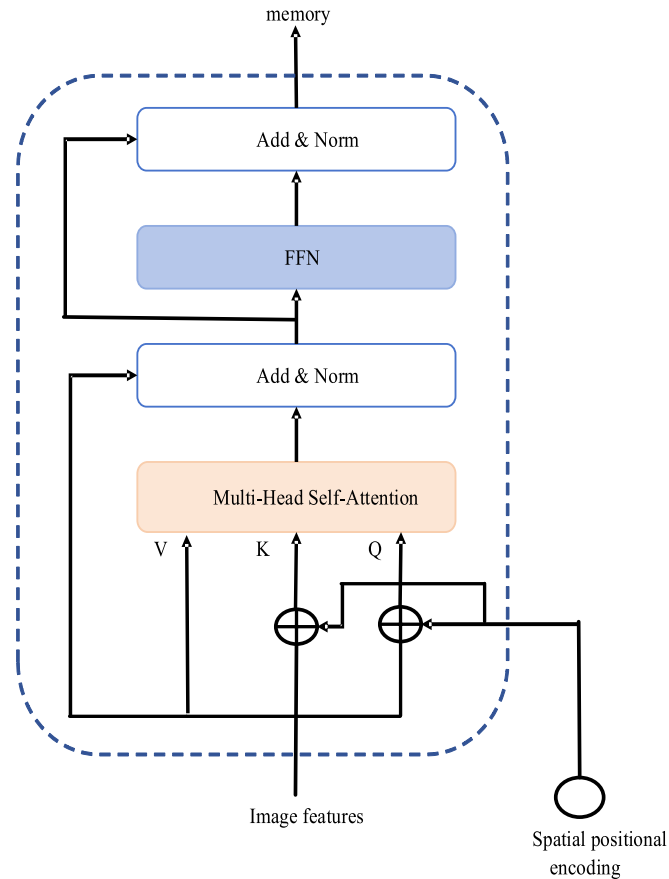


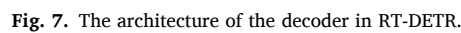
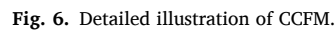
Fig. 5. The structure of AIFI.

weld features. In contrast, since shallow features are less semantic in welds, intra-scale feature interaction is only performed on S5, while S3 and S4 are processed using convolutional layers to extract weld features. Finally, these features are fused across scales through the feature map output by the encoder, enabling multi-scale weld feature acquisition while significantly reducing computational cost. In addition, the CCFM module has been optimized, with its structure shown in Fig. 6. Several convolutional fusion blocks are inserted into the feature fusion path, which merges adjacent feature maps into a new one. The network module adopts a parallel structure to extract features at various degrees from these feature maps, which are then added and dimensionally transformed. As a result, three feature maps of different scales are combined, enhancing the effect of weld feature fusion. At the same time, the three feature maps are stored in inner_outs for use in the next stage.

2.2.3. Transformer decoder with auxiliary prediction heads

In the decoder part, several DETR-related technologies are integrated, as shown in Fig. 7. The Query Selection component selects a fixed number of features from the sequence output by the Encoder to serve as object queries. These queries are a set of learnable embeddings optimized by the Decoder and then mapped to classification scores and bounding boxes through the prediction head. However, the inconsistency between classification scores and location confidence can lead to boxes with high classification scores but poor alignment with the actual target box (GT box). This causes the model to select boxes with high classification scores but low IoU values while discarding those with low classification scores but high IoU values, ultimately reducing detection performance. To address this problem, an IoU-Aware Query Selection method is proposed. This method constrains the model during training, ensuring the consistency between classification and IoU scores. As a result, the object queries contain more encoder features that are accurately classified and precisely located, thereby improving the detector's accuracy. In addition, the Denoising module in the DN-DETR[33] model is introduced. During training, the ground truth boxes are denoised and used as part of the Decoder input. This allows the Decoder's queries to predict specific targets more accurately, as they are generated from noisy targets, speeding up the convergence of the training process. Notably, no denoising is used during the inference phase.

In summary, RT-DETR offers significant advantages over existing real-time object detectors. First, it eliminates the need for post-



processing steps such as NMS, which reduces inference latency and improves detection speed. To efficiently handle multi-scale features and improve computational efficiency, RT-DETR utilizes an efficient hybrid encoder that decouples internal scale interactions and cross-scale fusion. In addition, the IoU-Aware Query Selection mechanism enhances the quality of initial object queries through IoU constraints during training, thereby improving detection performance. RT-DETR also supports adjustable decoder layers, allowing flexible control over inference speed without retraining, ultimately enhancing real-time weld defect detection.

2.3. Data augmentation methods

Data augmentation is a key technique in deep learning, enhancing the model's generalization ability and performance by transforming and expanding the original data. This technique effectively addresses the problems of data imbalance and sample scarcity[34]. Different data augmentation methods offer unique advantages and are suited to specific types of data. For the weld defect detection task, combining pixel-level data enhancement methods can provide more comprehensive and effective training data, which improves the performance and generalization ability of the model. After thoroughly investigating multiple data augmentation methods, we selected Mixup, Mosaic, and Simple Copy-Paste to compare and identify the most suitable data augmentation method.

Mixup[35] is an algorithm used in computer vision to enhance images by mixing different classes, thereby expanding the training data set. Mixup advocates for a linear interpolation of training samples, making the predictions more flexible and reducing the risk of overfitting. Since the decision boundary formed by model learning is smoother, the model will not be very absolute when dealing with uncertainty problems. By building a linear interpolation of samples, the model is relatively less affected by noise samples. In contrast, the model is more likely to learn this linear interpolation from real samples. The expansion of the model's linear interpolation of samples helps to expand its cognitive scope, because the model needs to learn the subtle differences brought about by linear interpolation. This wide cognitive scope improves the model's generalization ability for samples outside the distribution range of training set samples. The construction of new samples based on Mixup is as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_k(1)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_k(2)$$

Where x_i, x_k are raw input vectors; y_i, y_k are one-hot label encodings. (x_i, y_i) and (x_k, y_k) are two randomly drawn samples from the training data, and $\lambda \in [0, 1]$, where $\lambda \sim \beta(\alpha, \alpha)$ and $\alpha \in (0, \infty)$. The Mixup hyperparameter α controls the strength of interpolation between feature-target vectors, and when $\alpha \rightarrow 0$, it reverts to the Empirical Risk Minimization (ERM) principle.

Mosaic[36] is another data augmentation technique commonly used in computer vision tasks. The core idea of Mosaic is to create a new picture by splicing four pictures together by random scaling, cropping, and arrangement. The defects of the four pictures may appear in different positions of the new picture, so the model needs to learn to detect objects at varying positions and scales, making it more adaptable and robust[37].

Simple Copy-Paste[38] data augmentation is a simple yet effective data augmentation technique. Its core idea is generating new training samples by copying and pasting target objects from the original image. Specifically, one or more target objects are selected, copied, and placed in a random new location or background, with optional transformations such as translation, rotation, or scaling. The corresponding labels, including the location and category information, are then generated. This method has been widely used in computer vision tasks such as

target detection, semantic segmentation, and instance segmentation [39].

3. Dataset

This study uses the steel pipe weld target detection dataset as the experimental dataset. The dataset, collected by the author and supplemented with Internet resources, contains a total of 1,134 images, each with a resolution of 800×800 pixels. The weld dataset is labeled as 'good' or 'bad' based on the presence of defects such as burrs and holes. The ground truth labels of the images in the dataset were created using Labelling. The original and the target label images according to the PascalVOC specification are presented in Fig. 8 and Fig. 9, respectively.

Fig. 10 shows different types of target defect images, including those with good quality and defects such as burrs and pits. This dataset comprehensively covers common welding defects, with each type of defect meticulously annotated to ensure both accuracy and diversity. These images provide detailed representations of quality issues that may arise during welding, making the dataset highly representative and practical. For model training and performance evaluation, the dataset is divided into a training set and a validation set in a ratio of 8:2, with the training set containing 908 images and the validation set containing 226 images. This division method ensures sufficient data for training and reliable validation results.

4. Experiments

4.1. Configure the experimental environment

In this study, all deep learning models are written in Python using the PyTorch framework. All models are trained and evaluated using the same software. The hardware environment includes a computer with an Intel(R) Core(TM) i5-13400F processor (CPU), 64G memory, NVIDIA GeForce RTX 4070 graphics processor (GPU), and 12G memory. The operating system is Window10 (64 bits). The specific operating environment is shown in Table 1.

4.2. Neural network hyperparameter settings

The hyperparameter setting of network training is also of vital importance. If the learning rate is too high, the model may eventually oscillate around the optimal solution; if it is too low, more training epochs may be needed, leading to inefficient use of resources. To address this, a piecewise decay schedule and linear warm-up are used to set the learning rate. The base learning rate is set to 0.0001, with a piecewise decay factor of 0.9, meaning that after every 100 epochs, the learning rate will be multiplied by 0.9. A linear warm-up learning rate starts from 0.00001 and gradually increases to the base learning rate of 0.0001 over the first 2000 steps. The optimizer used is AdamW[40], with a weight decay of 0.05 to prevent overfitting from penalizing large weights. In addition, gradient clipping is used during training to limit the gradient values and avoid gradient explosion. The maximum gradient norm is set to 0.1.

Fig. 11 shows the average accuracy mAP@0.5 and mAP@0.5:0.95 of the RT-DETR-based weld defect detection model, demonstrating its high accuracy. Fig. 12 compares the target (red) and the predicted (green) results. The high degree of overlap between the two indicates that the model performs well. The results demonstrate that the current hyperparameter settings result in excellent detection accuracy and are therefore adopted for use in this experiment.

4.3. Data augmentation operations

In order to increase the complexity of the dataset and improve the generalization ability of the model, we used Mixup, Mosaic and Simple Copy-Paste data augmentation methods to process the dataset. The



Fig. 8. Typical images of the experimental dataset.

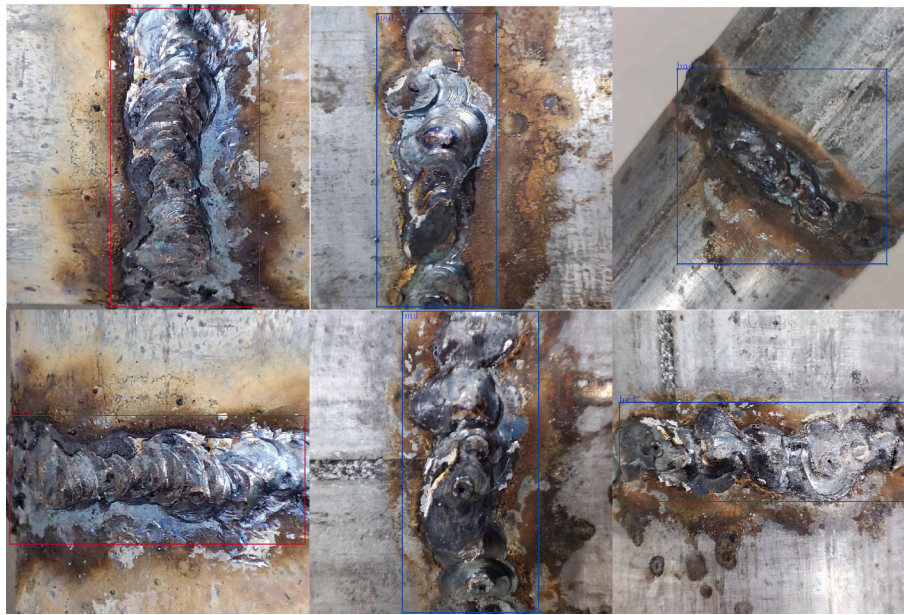


Fig. 9. Typical labeled images of the experimental dataset.

principles of these methods are shown in Section 2.3. Fig. 13 presents examples from the augmented dataset. It can be seen from the observation that the Mosaic data augmentation method starts with four randomly selected pictures and splices them into a new image through random scaling, cropping and arrangement. The Simple Copy-Paste method selects one or more target objects in the original image, placing them in a new position or background using translation, rotation, and scaling. Different from the first two methods, Mixup generates new image samples by mixing two different images in a particular proportion. Each model, including the proposed method, was trained separately with adjusted hyperparameters, and the results were compared to verify the effectiveness of the proposed approach.

4.4. Evaluation metrics

This article uses accuracy (P), recall (R), and mean average precision (mAP) as evaluation metrics. We use the intersection over union (IoU) between the predicted and the actual labeled boxes as the evaluation threshold. The IoU threshold is divided into intervals from 0.50 to 0.95. The 0.50:0.95 average precision (AP) is calculated by sampling every 0.05 within this interval. The final AP is the average value of these individual accuracy calculations. AP50 represents the AP measurement value when the IoU threshold is 0.5. In the target detection model, mAP measures the average maximum precision across categories at varying recall values, balancing precision and recall to provide a single numerical performance indicator. The calculation formulas for P, R, mAP, and

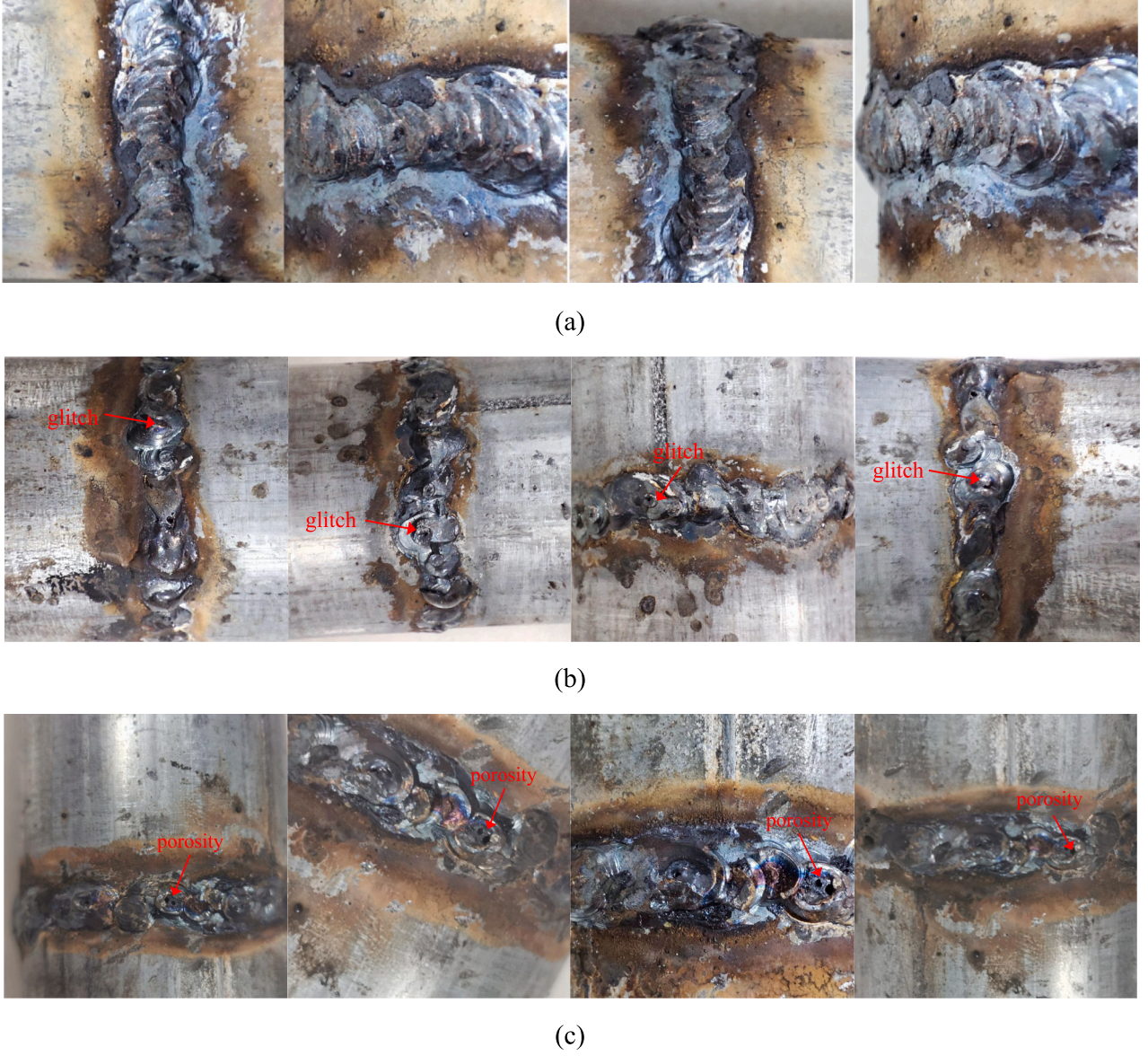


Fig. 10. Images of different categories:(a) Typical images of good quality;(b) Typical images with glitches; (c) Typical images with porosities.

Table 1

The experimental environment configuraion.

Experimental Platform	Environment Configuration
CPU	Intel ^(R) Core ^(TM) i5-13400F
GPU	NVIDIA GeForce RTX 4070 Ti
CUDA	12.2
Python	3.8
Deep learning framework	PyTorch = 2.0.1 + cu118

AP are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$AP = \sum_{k=1}^N \max_{\tilde{k} \geq k} p(\tilde{k}) \Delta R(k) \quad (5)$$

$$mAP = \frac{\sum_{i=1}^n AP_i}{n} \quad (6)$$

where TP , FP , and FN stand for “True Positive”, “False Positive”, and “False Negative”, respectively. They are defined based on the IoU between the predicted bounding box and the ground truth. If the IoU is greater than a threshold, the bounding box is labeled as TP , indicating the number of correctly identified objects; otherwise, it is labeled as FP , meaning the number of incorrectly identified objects.

In the study of detection speed, frames per second (FPS) are used as the main indicator to evaluate detection speed, representing the number of image frames the model can process per second. A higher FPS reflects faster model processing.

5. Results and discussion

5.1. Data augmentation performance comparison

To enhance the model’s generalization ability and improve the accuracy of detecting various weld defect types, various data

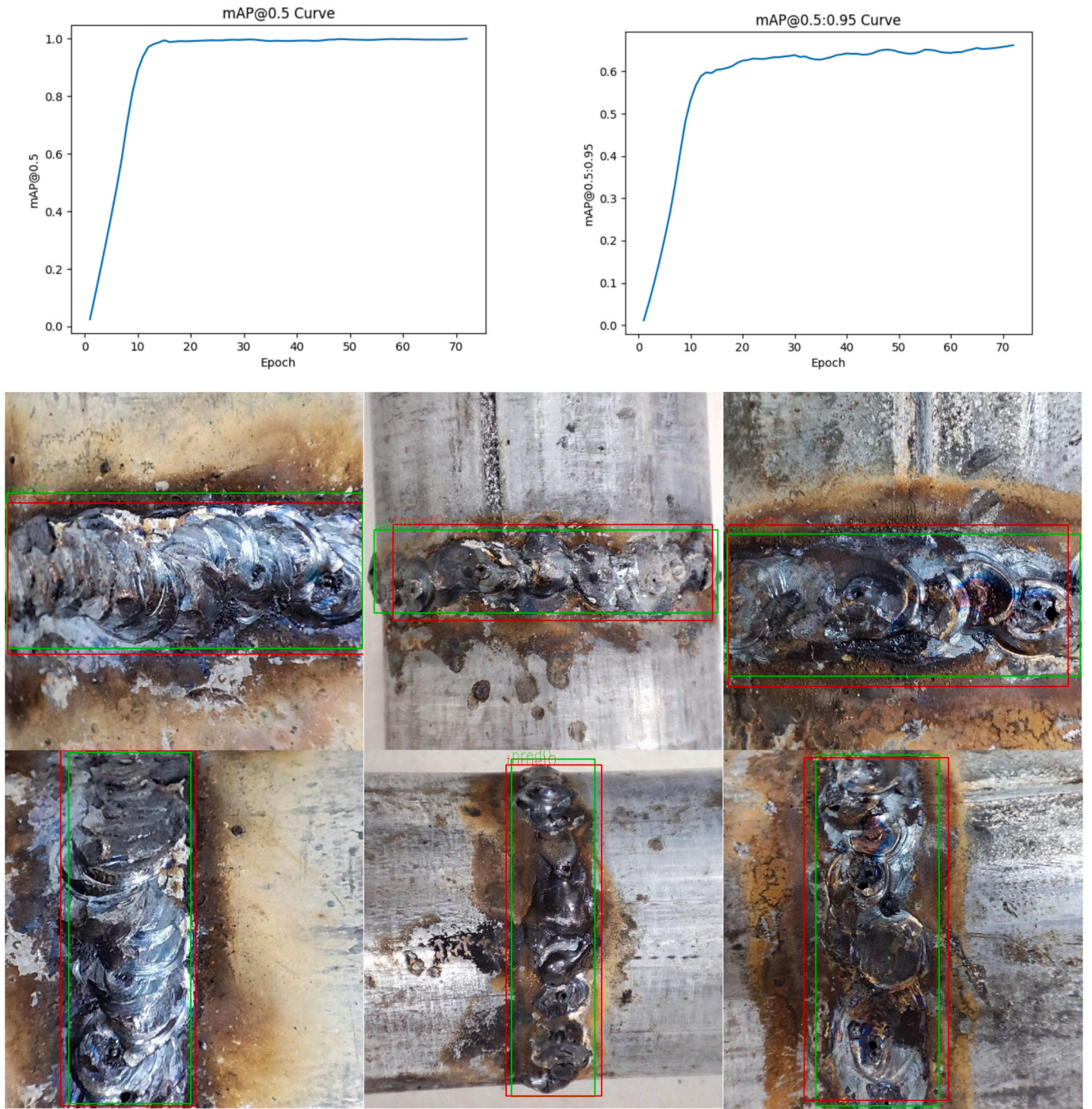


Fig. 11. The mAP evolution during the training process of RT-DETR.

augmentation strategies are employed in this study. This section discusses the effects of Mixup, Mosaic and Simple Copy-Paste data augmentation methods. Table 2 lists the evaluation indicators of the RT-DETR model trained on the original dataset and the datasets augmented with Mixup, Mosaic and Simple Copy-Paste methods. The table shows that the Mosaic augmentation method leads to the most significant improvement, with the mAP@0.5:0.95 increasing by 0.134 from 0.667 to 0.801. The Simple Copy-Paste data augmentation method increases the mAP@0.5:0.95 to 0.757, while the Mixup data augmentation method causes a slight drop to 0.649. The superior performance of the Mosaic method may be attributed to its ability to stitch multiple different images into a new image, introducing more diversity and complexity and helping the model learn the features across different

scenarios. Although the Simple Copy-Paste method does not offer the same level of diversity as the Mosaic method, it introduces sufficient variation and noise, aiding the model's generalization to new scenarios. In contrast, the Mixup method may cause the mixing of weld defect features, interfere with the learning of the model, and thus reduce the detection performance. The FPS remained consistent across methods, indicating stable operational efficiency and meeting the real-time requirements. Fig. 14 lists the results of the average detection accuracy mAP@0.5 and mAP@0.5:0.95 of the three data augmentation methods. Mosaic consistently showed the highest detection accuracy, demonstrating its effectiveness in improving both the generalization and accuracy of the weld defect detection model, leading to its adoption in this study to enhance the model's overall performance.

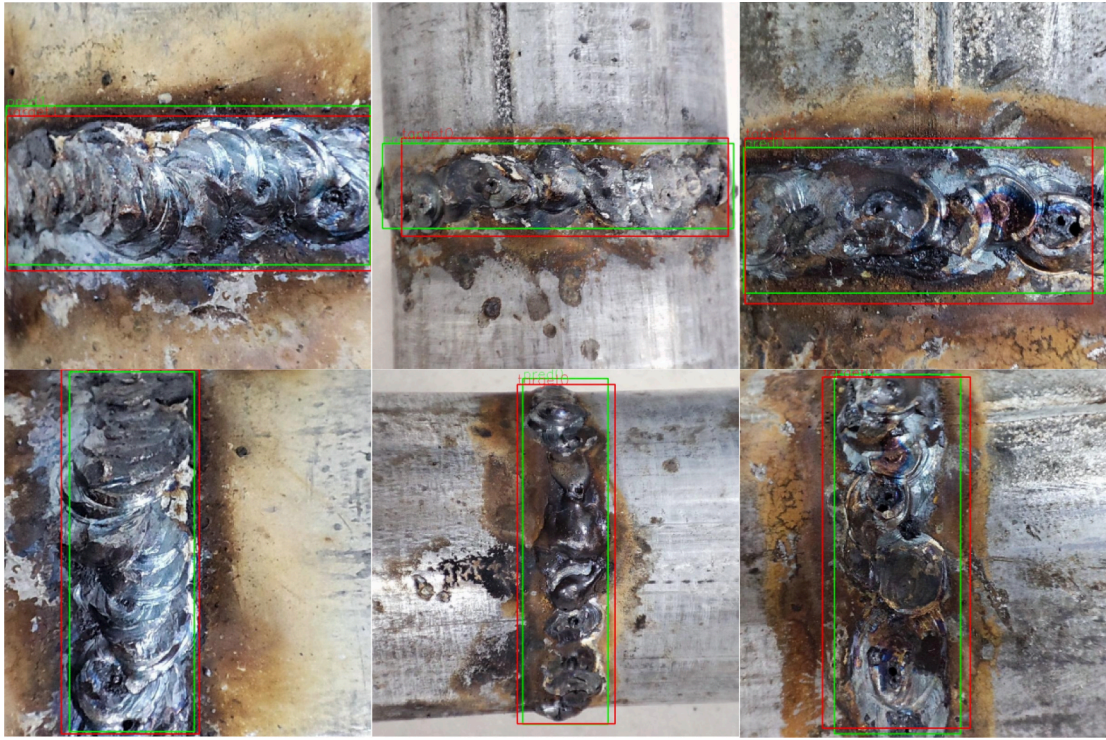


Fig. 12. Detection results of RT-DETR with the optimal hyperparameters.

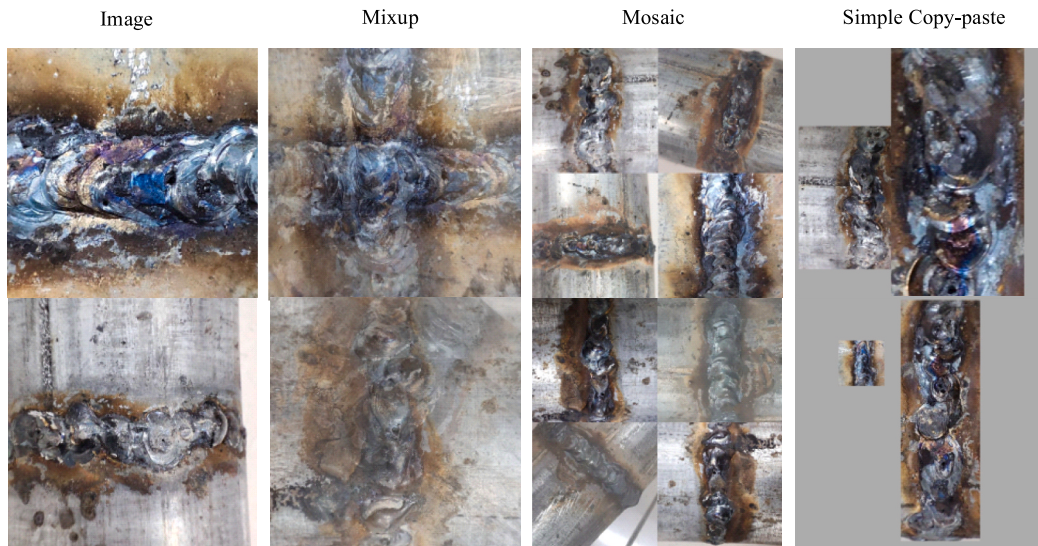


Fig. 13. Typical images of different augmentation methods.

Table 2
The performance comparison of different data augmentation methods.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
Original	0.718	0.956	0.996	0.667	67
Mixup	0.662	0.851	0.992	0.649	65
Mosaic	0.726	0.912	0.998	0.801	68
Simple Copy-Paste	0.686	0.869	0.998	0.757	66

5.2. Comparison of detection performance

In order to verify the effectiveness of the proposed method, this section compares the detection performance of RT-DETR with six other

machine learning models, Faster R-CNN, SSD, YOLOv5, YOLOv11 and DETR, on the weld dataset. To be noted, YOLOv5s and YOLOv11x version are adopted in this study. Table 3 lists the performance of each model in terms of accuracy, recall, and FPS. The results show that RT-DETR achieves the highest mAP scores, with mAP@0.5 reaching 0.996 and mAP@0.5:0.95 reaching 0.801, highlighting the significant advantages of RT-DETR. This is due to the RT-DETR model leveraging the advantages of Transformer and the CNN, allowing it to capture long-distance dependencies and global information, thereby improving detection accuracy. In contrast, the traditional Faster R-CNN, SSD, and YOLO series primarily rely on CNN for feature extraction, which, while effective for local features, falls short of capturing global features. In addition, RT-DETR demonstrates a clear advantage in processing speed,

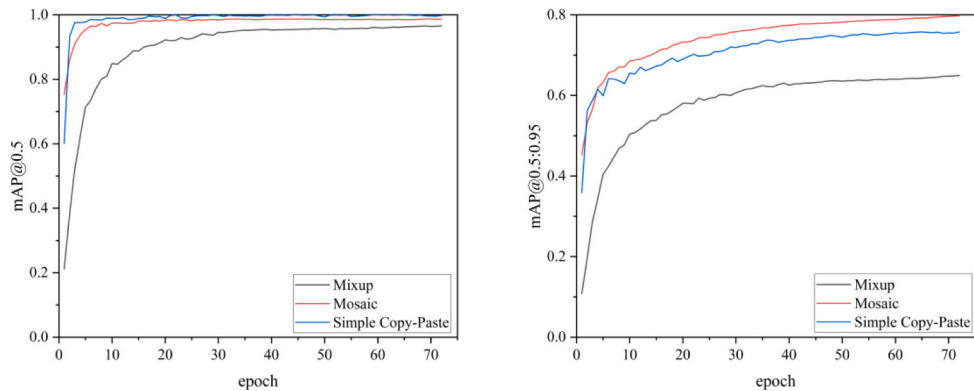


Fig. 14. Map evolution during rt-detr training process after data augmentation.

Table 3
Comparison study of detection performance of various models.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
Faster R-CNN	0.785	0.800	0.973	0.712	16
SSD	0.754	0.803	0.981	0.721	28
YOLOv5*	0.782	0.951	0.987	0.728	48
YOLOv11*	0.802	0.956	0.996	0.768	56
DETR	0.731	0.810	0.986	0.722	26
RT-DETR	0.798	0.963	0.996	0.801	67

*Note: The versions utilized in this experiment are YOLOv5s and YOLOv11x.

delivering 19 more FPS than the widely used real-time detection model YOLOv5. Unlike other models that rely on NMS post-processing, which is often difficult to optimize and slows down detection speed, RT-DETR, using the DETR architecture, eliminates the need for NMS, significantly boosting its speed. It also has certain advantages in detection speed compared to YOLOv11.

Fig. 15 lists the detection effect images of the six algorithms Faster R-CNN, SSD, YOLOv5, YOLOv11, DETR and RT-DETR on the test set. Of the six models, RT-DETR performs the best, accurately locating defects and providing precise bounding boxes. It is particularly effective on uneven surfaces, where it significantly outperforms the other models.

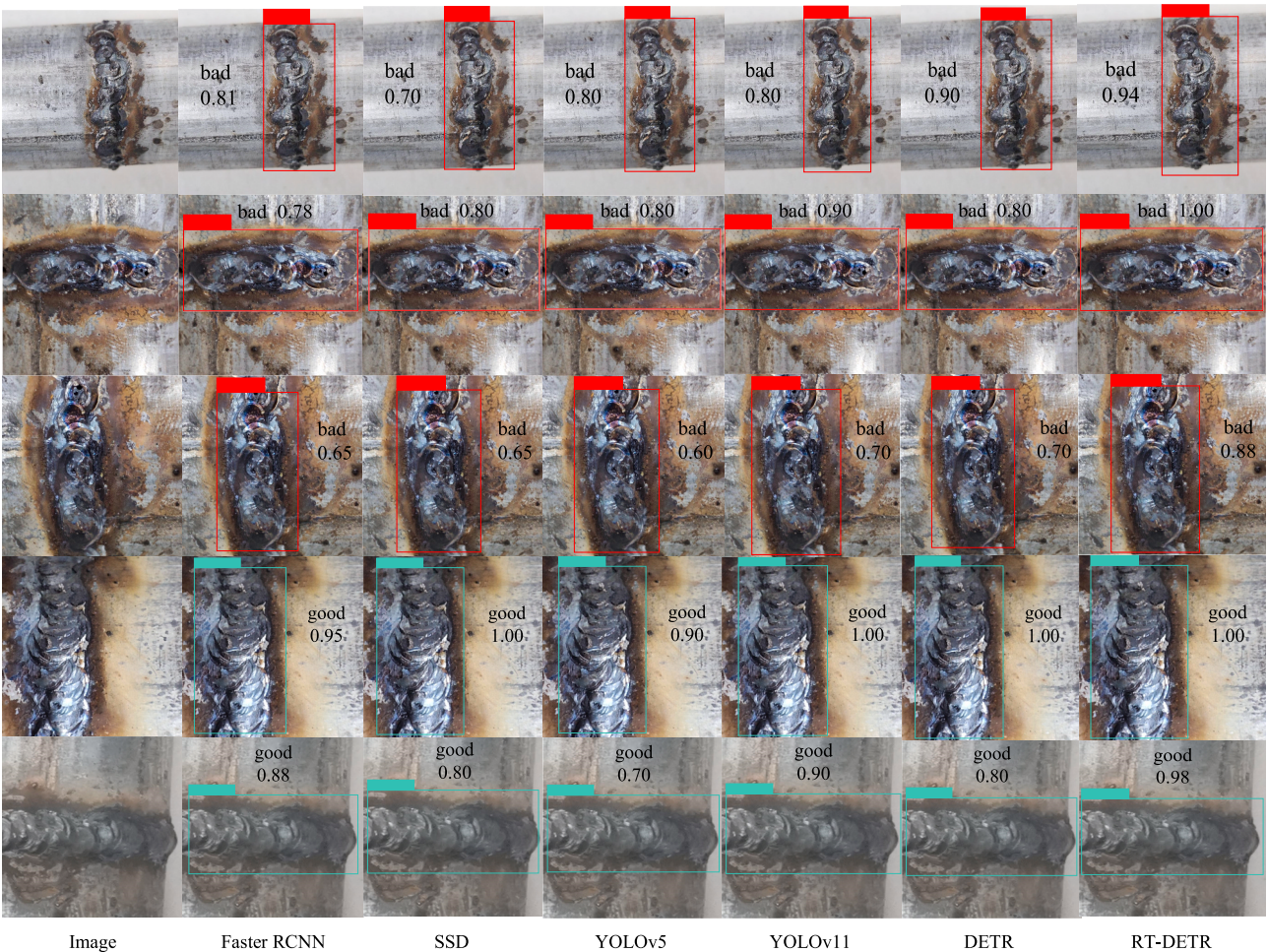


Fig. 15. Detective results comparison study of different neural networks.

Compared with Faster R-CNN and SSD, RT-DETR demonstrates more precise defect localization, especially in complex backgrounds. By utilizing the self-attention mechanism, it captures long-range contextual information, which helps in more accurately determining the defect locations. Compared with Faster R-CNN and SSD, RT-DETR demonstrates more precise defect localization, especially in complex backgrounds. By utilizing the self-attention mechanism, it captures long-range contextual information, which helps in more accurately determining the defect locations. Compared with the YOLO series (e.g., YOLOv5, YOLOv11), RT-DETR shows superior performance in bounding box accuracy. Additionally, CNN based model YOLOv11 may be affected in complex scenes, especially when dealing with overlapping and small objects. RT-DETR generates more accurate bounding boxes based on a global understanding of the context by leveraging the transformer architecture. In contrast to DETR, RT-DETR retains the end-to-end learning capability and the advantage of not requiring NMS post-processing, while improving feature extraction performance. This enhancement allows RT-DETR to offer improved localization accuracy and bounding box regression, particularly in dense and complex defect detection tasks. Thus, RT-DETR not only provides accurate defect localization but also delivers more precise bounding box predictions. When handling complex images, it demonstrates stronger robustness and accuracy compared to Faster R-CNN, SSD, the YOLO series, and DETR.

Therefore, from the above experimental results, it can be concluded that weld defect detection based on RT-DETR not only has a leading advantage in target detection accuracy, but also performs well in real-time application scenarios, providing an ideal solution for tasks such as weld defect detection, where both accuracy and speed are crucial.

6. WAAM weld on-site inspection

In order to verify the generalization performance of the proposed method, it is tested in metal Additive Manufacturing (AM) scenarios. 3D printed welds are a new technology that combines AM technology with the metal welding process, commonly used for repairing and manufacturing complex metal parts or conducting welding operations in high-risk and high-voltage scenarios. Among the available techniques in this area, Wire Arc Additive Manufacturing (WAAM)[41] integrates traditional welding technology with modern robots to produce large-scale components. This technique has great potential for applications in the field of intelligent construction of metal structures. In the WAAM

printing process, the quality and geometry of the weld are crucial to the structural strength and overall quality of the printed parts. However, manual inspection in high-voltage environments poses significant safety hazards and is prone to subjective errors. The detection method proposed in this study aims to achieve automatic real-time detection of weld defects in an end-to-end manner, significantly improving detection accuracy and efficiency.

In this study, WAAM is conducted using the ArcMan S1 Adv, a mini-integrated intelligent arc additive manufacturing system. The setup includes a Fronius brand TPS 4000 CMT Adv welding machine and an ABB IRB 1200-5/0.9 robot arm. Guided by the IungoPNT software, the system utilizes a flexible six-axis robot and an ultra-stable fuse power supply to print a series of welds using stainless steel wire. The arc additive manufacturing system is shown in Fig. 16.

Fig. 17 shows the consolidated weld bead formed during the WAAM printing process. The original image was cropped into an image with a resolution of 800×800 pixels, and then labeled according to the previous requirements, with defect label examples shown in Fig. 18. As can be seen from the figure, there are significant variations in weld quality, indicating that various defects and inconsistencies may arise during the welding process. This underscores the importance of real-time weld detection, which allows for timely identification and correction of issues, ensuring that weld quality meets standards. Real-time detection also reduces rework and material waste, improving both production efficiency and safety.

During the WAAM printing process, the proposed method was used to detect weld defects. The original input image (first row) and the detection results (second row) are shown in Fig. 19. The detection details are shown in Table 4. The proposed method achieved a $mAP@0.5:0.95$ of 0.801, the highest accuracy compared with other models. Additionally, it outperformed the others in FPS, reflecting its superiority in accuracy and speed. The predictive results from the WAAM tests demonstrate that the proposed method in this study is both effective and robust, fully meeting the requirements for WAAM weld detection and highlighting the clear advantages of the RT-DETR-based model in AM applications.

7. Conclusion

This study proposes a real-time weld defect detection based on a Real-Time DEtection Transformer (RT-DETR). The model improves the Transformer architecture by eliminating time-consuming NMS



Fig. 16. Illustration of Arc Additive Manufacturing System ArcMan S1 Adv.



Fig. 17. Typical in-situ weld images of WAAM.



Fig. 18. WAAM welding labeled images.

procedures, achieving accurate real-time detection. Besides, data augmentation methods are adopted to further improve the model's generalizability. The proposed model was validated on a collected dataset of 1,134 weld images and tested on WAAM on-site images. The main conclusions are as follows:

Among the three data augmentation methods, Mixup, Mosaic, and Simple Copy-Paste, the Mosaic and Simple Copy-Paste techniques

achieved mAP@0.5:0.95 accuracy rates of 0.801 and 0.757, improving by 0.131 and 0.09 compared to the original dataset. In contrast, the Mixup method reduced accuracy to 0.649. Therefore, Mosaic is the most suitable augmentation method for weld defect detection, significantly enhancing the model's performance and generalization capability.

The RT-DETR-based welding defect detection model demonstrates exceptional detection capabilities. Compared to existing object



Fig. 19. Illustration of weld defect detection results in WAAM.

Table 4

The comparison of detection performance of various models.

Model	mAP@0.5	mAP@0.5:0.95	FPS
Faster RCNN	0.971	0.712	16
SSD	0.976	0.656	28
YOLOv5	0.985	0.731	48
YOLO11	0.992	0.768	56
DETR	0.982	0.722	26
RT-DETR	0.992	0.801	67

detection models such as Faster R-CNN, SSD, YOLOv5, YOLOv11, and DETR, the RT-DETR model achieves the highest detection performance, with an mAP@0.5:0.95 accuracy of 0.801. Additionally, it attains the fastest frame rate of 67 FPS, showcasing a notable advantage in detection speed even when compared to YOLOv11. The RT-DETR-based welding defect detection model not only maintains high precision but also delivers superior frame rates per second, demonstrating its feasibility for real-time welding defect detection. The RT-DETR-based welding defect detection model demonstrates high feasibility for detecting weld production in 3D printing of steel structures. The proposed method achieved the highest detection accuracy, with a mAP@0.5:0.95 of 0.801, outperforming other models. Additionally, the detection frame rate reaches 67 FPS, which also shows the best performance. The model's accuracy and real-time detection capabilities make it a reliable solution for the quality control of 3D printing welds.

In summary, the weld defect detection model based on RT-DETR not only achieves high accuracy in weld defect detection but also meets the real-time detection requirements of industrial applications. Additionally, the augmented dataset serves as a high-quality benchmark for model training and evaluation, contributing to improved weld defect detection. This approach enhances detection accuracy and processing efficiency, addressing key challenges in traditional weld defect detection methods. The successful application of this framework in WAAM field inspection further demonstrates its feasibility and provides a reliable solution for weld defect detection. In future research, the

generalizability of the model can be improved by training on welding defect images under diverse environmental and engineering conditions. This will contribute to the development of a more robust welding defect detection model applicable to various real-world scenarios. Additionally, techniques such as pruning and knowledge distillation will be explored to enhance the model's efficiency and reduce the computational requirements. With these methods it would be more suitable for the deployment on various devices, particularly on edge devices, thereby enabling real-time welding defect detection in industrial environments.

CRedit authorship contribution statement

Gaoyang Liu: Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition. **Duanrui Yang:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Jun Ye:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Hongjia Lu:** Writing – review & editing, Conceptualization. **Zhen Wang:** Writing – review & editing, Conceptualization. **Yang Zhao:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge the financial support to this project by the Shaoxing Science and Technology Planning Project (No. 2024B13004).

Data Availability

The code and dataset used in this study will be made available at the following link: <https://github.com/klmyt-ai/Detection>.

References

- [1] D. Yang, Y. Cui, Z. Yu, et al., Deep learning based steel pipe weld defect detection [J], *Appl. Artif. Intell.* 35 (15) (2021) 1237–1249.
- [2] J.M. Antonini, Health effects of welding [J], *Crit. Rev. Toxicol.* 33 (1) (2003) 61–103.
- [3] Zhang H, Chen Z, Zhang C, et al. Weld defect detection based on deep learning method; proceedings of the 2019 IEEE 15th international conference on automation science and engineering (CASE), F, 2019 [C]. IEEE.
- [4] Y. Li, Y.F. Li, Q.L. Wang, et al., Measurement and defect detection of the weld bead based on online vision inspection [J], *IEEE Trans. Instrum. Meas.* 59 (7) (2009) 1841–1849.
- [5] B. Raj, C. Subramanian, T. Jayakumar, Non-destructive testing of welds, [J]. (2000).
- [6] W. Hou, D. Zhang, Y. Wei, et al., Review on computer aided weld defect detection from radiography images [J], *Appl. Sci.* 10 (5) (2020) 1878.
- [7] M. Shaloo, M. Schnall, T. Klein, et al., A review of non-destructive testing (NDT) techniques for defect detection: application to fusion welding and future wire arc additive manufacturing processes [J], *Materials* 15 (10) (2022) 3697.
- [8] R. Abdelkader, N. Ramou, M. Khorchef, et al., Segmentation of x-ray image for welding defects detection using an improved Chan-Vese model [J], *Mater. Today Proc.* 42 (2021) 2963–2967.
- [9] D. Mery, M.A. Berti, Automatic detection of welding defects using texture features [J], *Insight-Non-Destructive Testing and Condition Monitoring* 45 (10) (2003) 676–681.
- [10] W. Cui, K. Wang, Q. Zhang, et al., A recognition algorithm to detect pipe weld defects [J], *Tehnicki Vjesnik-Technical Gazette* 24 (6) (2017) 1969–1975.
- [11] R. Vilar, J. Zapata, R. Ruiz, An automatic system of classification of weld defects in radiographic images [J], *NDT and E Int.* 42 (5) (2009) 467–476.
- [12] R. Yamashita, M. Nishio, R.K.G. Do, et al., Convolutional neural networks: an overview and application in radiology [J], *Insights into Imaging* 9 (2018) 611–629.
- [13] ASIF H, KUMAR S. Welding Defect Inspection Using Deep Learning; proceedings of the International Conference on Emerging Trends and Technologies on Intelligent Systems (ETTIS), Electr Network, F Mar 04-05, 2021 [C]. 2022.
- [14] P. Sassi, P. Tripicchio, C.A. Avizzano, A smart monitoring system for automatic welding defect detection [J], *IEEE Trans. Ind. Electron.* 66 (2019) 9641–9650.
- [15] Y. Yan, D. Liu, B. Gao, et al., A Deep Learning-Based Ultrasonic Pattern Recognition Method for Inspecting Girth Weld Cracking of Gas Pipeline [J], *IEEE Sens. J.* 20 (14) (2020) 7997–8006.
- [16] K. Kim, K.S. Kim, H.J. Park, Multi-Branch Deep Fusion Network-Based Automatic Detection of Weld Defects Using Non-Destructive Ultrasonic Test [J], *IEEE Access* 11 (2023) 114489–114496.
- [17] J. Wang, C. Mu, S. Mu, et al., Welding seam detection and location: Deep learning network-based approach [J], *Int. J. Press. Vessel. Pip.* 202 (2023) 104893.
- [18] LIN T-Y, DOLLAR P, Girshick R, et al. Feature pyramid networks for object detection; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].
- [19] Y. Chen, J. Wang, G. Wang, Intelligent welding defect detection model on improved R-CNN [J], *IETE J. Res.* (2022) 1–10.
- [20] S. Gao, Y. Jiang, T. Xia, et al., A multi-stage active learning framework with an instance-based sample selection algorithm for steel surface defect [J], *Adv. Eng. Inf.* 64 (2025) 103080.
- [21] H.M. Al-Otun, Deep learning-based automated defect classification in electroluminescence images of solar panels [J], *Adv. Eng. Inf.* 58 (2023) 102147.
- [22] Q. Liu, D. He, Y. Shen, et al., Surface defect detection of stay cable sheath based on autoencoder and auxiliary anomaly location [J], *Adv. Eng. Inf.* 62 (2024) 102759.
- [23] X. Yang, Y. Gao, C. Fang, et al., Deep learning-based bolt loosening detection for wind turbine towers [J], *Struct. Control Health Monit.* 29 (6) (2022) e2943.
- [24] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers; proceedings of the European conference on computer vision, F, 2020 [C]. Springer.
- [25] Lv W, Xu S, Zhao Y, et al. Detsr beat yolos on real-time object detection [J]. arXiv preprint arXiv:230408069, 2023.
- [26] E. Guemas, B. Routier, T. Ghelfenstein-Ferreira, et al., Automatic patient-level recognition of four *Plasmodium* species on thin blood smear by a real-time detection transformer (RT-DETR) object detection algorithm: a proof-of-concept and evaluation [J], *Microbiol. Spectrum* 12 (2) (2024).
- [27] S. Li, X.H. Cao, Z.G. Zhou, Research on inshore ship detection under nighttime low-visibility environment for maritime surveillance [J], *Comput. Electr. Eng.* 118 (2024).
- [28] M. Zhu, E. Kong, Multi-Scale Fusion Uncrewed Aerial Vehicle Detection Based on RT-DETR [J], *Electronics* 13 (8) (2024) 1489.
- [29] Salehinejad H, Sankar S, Barfett J, et al. Recent advances in recurrent neural networks [J]. arXiv preprint arXiv:180101078, 2017.
- [30] Neubeck A, Van gool L. Efficient non-maximum suppression; proceedings of the 18th international conference on pattern recognition (ICPR'06), F, 2006 [C]. IEEE.
- [31] WANG A, CHEN H, LIU L, et al. Yolov10: Real-time end-to-end object detection [J]. arXiv preprint arXiv:240514458, 2024.
- [32] TARG S, ALMEIDA D, LYMAN K. Resnet in resnet: Generalizing residual architectures [J]. arXiv preprint arXiv:160308029, 2016.
- [33] LI F, ZHANG H, LIU S, et al. Dn-detr: Accelerate detr training by introducing query denoising; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2022 [C].
- [34] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning [J], *Journal of Big Data* 6 (1) (2019) 1–48.
- [35] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization [J]. arXiv preprint arXiv:171009412, 2017.
- [36] HAO W, ZHILI S. Improved mosaic: Algorithms for more complex images; proceedings of the Journal of Physics: Conference Series, F, 2020 [C]. IOP Publishing.
- [37] Zeng G, Yu W, Wang R, et al. Research on mosaic image data enhancement for overlapping ship targets [J]. arXiv preprint arXiv:210505090, 2021.
- [38] Ghiasi G, Cui Y, Srinivas A, et al. Simple copy-paste is a strong data augmentation method for instance segmentation; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2021 [C].
- [39] BAI Y, CHEN D, LI Q, et al. Bidirectional copy-paste for semi-supervised medical image segmentation; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2023 [C].
- [40] Loshchilov I. Decoupled weight decay regularization [J]. arXiv preprint arXiv:171105101, 2017.
- [41] K. Treutler, V. Wesling, The current state of research of wire arc additive manufacturing (WAAM): a review [J], *Appl. Sci.* 11 (18) (2021) 8619.