IR-DETR-Based Algorithm for Detecting Surface Defects on Porcelain Cups

Xiao Mingyang， Deng Xiaojun*， Xiao Shikang*

School of Computer,Hunan University of Technology,Hunan 412007.China

Abstract—To address the limitations of manual inspection, including missed and false detections in porcelain-cup manufacturing, we propose IR-DETR, an improved RT-DETR framework for detecting small-scale surface defects under low computational and parameter budgets. In the backbone, we introduce a wavelet convolution (WTConv) module that enlarges the effective receptive field, thereby improving robustness to defect patterns and promoting shape-oriented rather than texture-oriented responses. In the neck, we incorporate a hierarchical scale-sequence feature fusion (SSFF) module to uniformly fuse multi-scale sequence features, enhancing representation capacity and the detection of diverse defect types. To offset the overhead introduced by SSFF, we further integrate a dynamic upsampling (Dysample) module that reformulates the upsampling process from a point-sampling perspective, preserving feature continuity while reducing complexity and improving sensitivity to tiny defects. Experiments on a self-constructed dataset show competitive performance: compared with representative defect-detection methods, IR-DETR achieves mAP@0.5 = 54.3% with a lightweight design of 14.4M parameters and 48.4 GFLOPs. Overall, IR-DETR maintains high detection accuracy at substantially reduced complexity, achieving a favorable accuracy–efficiency trade-off.

Keywords—Ceramic cup defects; DETR;Multi-scale features; WTconv

I. INTRODUCTION

Porcelain cups are widely used in household and commercial applications, yet kiln firing often induces surface defects such as specks and pinholes. Severe defects impair both functionality and aesthetics, resulting in economic losses. Conventional manual inspection is labor-intensive, inefficient, and error-prone, making it inadequate for high-throughput production. Therefore, developing an automated, high-precision defect detection algorithm—formulated as an object-detection problem for ceramic surfaces—is of practical importance for industrial ceramics[1] manufacturing.

Traditional classical-vision defect detection methods[2]—for example, Zuo Cai et al.[3]—apply image thresholding and the Canny edge detector, followed by Fast Fourier Transform (FFT)-based image registration. Jiang Hongyu et al.[4] used an improved WOA to tune SVM hyperparameters for defect feature segmentation and extraction, thereby improving SVM generalization and enabling classification of different defective steel wires. Wan Jiyao et al.[5] combined an improved Gaussian filter, rectangular masks, and gray-level histograms to detect multiple defect types.

Deep learning-based ceramic surface defect detection methods are generally divided into one-stage detectors, such as the YOLO series, SSD, and DETR-style models, and two-stage detectors, including Fast R-CNN[6], Faster R-CNN[7], and Mask R-CNN[8]. Owing to their favorable speed–accuracy trade-off, YOLO-based methods are widely used in this field. For example, Pan et al.[9] improved YOLOv5 by introducing a global attention module and an enhanced loss function, while Ding et al.[10] further improved detection by adding a small-object detection head and a position-attention-based feature reconstruction module. However, YOLO-based methods remain limited in detecting fine-grained defects on porcelain cups, particularly in terms of small-target sensitivity and computational cost under high-resolution inputs. In contrast, RT-DETR[11], as an end-to-end and NMS-free detector, directly predicts object sets through a Transformer decoder, achieving high accuracy while preserving real-time inference. Considering the small size, morphological diversity, and low contrast of porcelain-cup defects, as well as the strict efficiency requirements of industrial deployment, RT-DETR is selected as the baseline model in this work.

Motivated by these observations, we propose IR-DETR, which further improves both accuracy and throughput for ceramic-cup defect detection.

II. RT-DETT ALGORITHM

RT-DETR is an end-to-end detector for real-time object detection that directly predicts class probabilities and bounding boxes without NMS. As shown in Fig. 1, it consists of a backbone, a neck, and a detection head. Multi-scale features from the last three backbone stages are encoded by a hybrid encoder with AIFI and CCFM, and IoU-aware query selection is used to initialize the decoder queries. The decoder then iteratively refines these queries with the aid of an auxiliary prediction head to produce the final detection results.

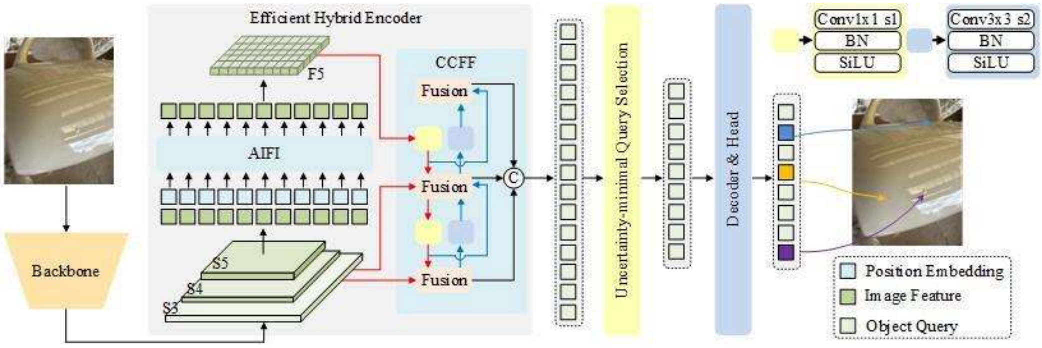

Fig.1. Overall architecture of the RT-DETR framework

III. IR-DETR

To mitigate missed detections and misclassifications on porcelain-cup production lines, we propose IR-DETR.We modify the RT-DETR baseline in three aspects. (i) In the backbone, we replace standard convolutions with the wavelet convolution (WTConv[12]) module. (ii) In the neck, we introduce a Scale-Sequence Feature Fusion (SSFF) module to strengthen multi-scale aggregation. (iii) On top of SSFF, we further refine the upsampling pathway by integrating the DySample dynamic upsampling module. These changes improve defect sensitivity while reducing computational complexity.

A. *Scale sequence feature fusion module*

In porcelain-cup defect detection, flaws such as specks and pinholes are characterized by subtle structures, low contrast, and diverse morphologies. Although feature pyramid networks (FPNs) provide multi-scale representations, their conventional fusion strategies are limited in modeling cross-scale semantic and contextual dependencies, often causing information redundancy and loss of fine details, which leads to missed and false detections of small defects. To address this issue, we propose a Scale-Sequence Feature Fusion (SSFF) module, as illustrated in Fig. 2. Given pyramid features $P3, P4, P5$, we first project $P4$ and $P5$ to 256 channels via 1×1 convolutions to match $P3$. The projected maps are then upsampled by nearest-neighbor interpolation to the spatial resolution of $P3$, achieving spatial alignment. To explicitly model cross-level correlations, we insert a scale axis by unsqueezing each feature from (H, W, C) to $(S = 1, H, W, C)$ and concatenate along the scale dimension to form a unified tensor (S, H, W, C) with $S = 3.$. This tensor is processed by a 3D convolution–BatchNorm–SiLU stack operating over (S, H, W) enabling joint aggregation of cross-scale context while preserving fine details. By leveraging 3D convolutions across the scale axis, SSFF integrates features of different receptive fields, improving robustness to size and shape variation—particularly for minute defects such as specks and pinholes.

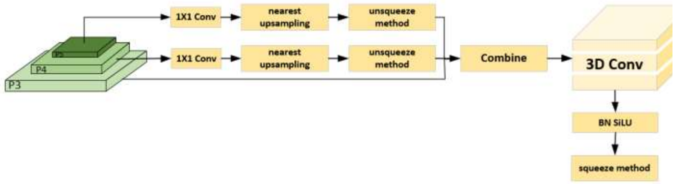

Fig.2. Architecture of the proposed Scale-Sequence Feature Fusion (SSFF) module

B. Dysample

To achieve efficient high-resolution feature reconstruction at low computational cost, DySample is introduced on top of SSFF as a secondary upsampling module. As shown in Fig. 3, it performs content-adaptive resampling by predicting spatial offsets with a lightweight projection and combining them with a regular sampling grid in a bilinearly interpolated continuous feature field. The final reconstruction is implemented via grid_sample, enabling differentiable upsampling. In addition, tailored designs for sampling initialization, offset constraints, and grouped sampling further improve robustness, reduce complexity, and enhance sensitivity to minute defects.

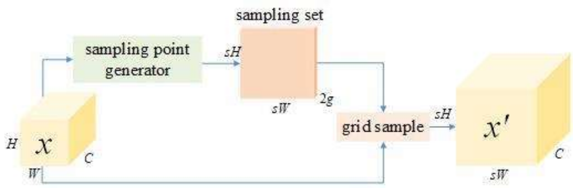

Fig.3. Architecture of the DySample-based dynamic upsampling module

In Fig. 3, sH and sW denote the height and width scaling factors of the upsampling operator, respectively. The symbol g denotes the number of groups used in grouped sampling, and 2g indicates that each group predicts two offset components, corresponding to the horizontal and vertical sampling displacements. Therefore, the offset branch outputs 2g channels before reshaping, where the factor 2 represents the x and y coordinate offsets.

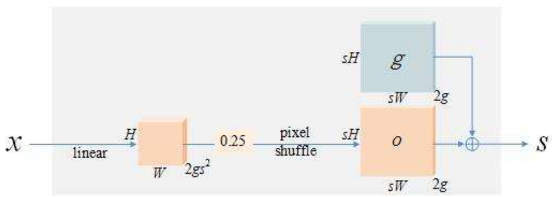

Fig.4. Architecture of the static scope factor in DySample

D. WTConv

To enhance the detection of minute defects in ceramic cup inspection, Wavelet Convolution (WTConv) is introduced, as illustrated in Fig. 7. Unlike conventional convolution, WTConv employs wavelet transform to decompose input features into multiple frequency bands, enabling joint modeling of low-frequency structural information and high-frequency detail information.

The input is first decomposed into low- and high-frequency components via wavelet transform, followed by filtering, downsampling, and depthwise convolution with small kernels on each frequency map. The output is then reconstructed through the inverse wavelet transform, as formulated below:

$$Y = \mathrm{IWT}\left(\mathrm{Conv}\big(W, \mathrm{WT}(X)\big)\right) \qquad (1)$$

In this context, X represents the input tensor, and W denotes the weight tensor for a $k \times k$ deep convolutional kernel, where the input channels of W are four times that of X. This operation not only separates the convolutions for different frequency components, but also enables the smaller kernel to operate over a larger region of the original input, thereby increasing its receptive field relative to the input. As shown in the figure, X is the input tensor, and W is the weight tensor for a $k \times k$ deep convolutional kernel, with input channels that are four times larger than those of X.

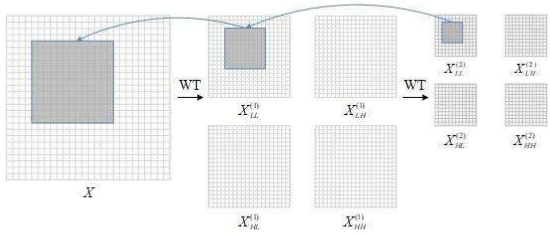

Fig.6. Structure of WTConv

A cascading operation is employed, which is further expanded by applying the same cascading principle. This process is expressed in the following equation.

$$X_{LL}^{(i)}, X_{H}^{(i)} = \mathrm{WT}\left(X_{LL}^{(i-1)}\right) \qquad (2)$$

$$Y_{LL}^{(i)}, Y_{H}^{(i)} = \mathrm{Conv}\left(W^{(i)}, \left(X_{LL}^{(i)}, X_{H}^{(i)}\right)\right) \qquad (3)$$

$X_{LL}^{(0)}$ represents the input to this layer, while $X_{H}^{(i)}$ denotes the three high-frequency maps at level i.

IV. *Experiments*

A. Experimental environment

The experimental environment is summarized in Table I.

TABLE I. EXPERIMENTAL ENVIRONMENT AND IMPLEMENTATION DETAILS

Name	Version
CPU	Intel Xeon Platinum 8362 2.80GHz
GPU	NVIDIAGeForce RTX 3090,24GB
Programming language	Python 3.10
Operating System	ubuntu22.04
Deep learning framework	PyTorch 2.1.0, Cuda 12.1

In addition, Stochastic Gradient Descent (SGD) was used as the optimizer during model training, with the batch size set to 32, the initial learning rate (lr0) set to 0.001, the weight decay factor to 0.0005, and the momentum parameter for SGD set to 0.937.

B. Dataset

The dataset used in this study consists of 362 on-site images of porcelain cup defects collected from a porcelain factory, covering four common defect categories: speckles, mud slag, dissolved holes, and mounting cracks. All defects were manually annotated using Labelme. To enhance model robustness and alleviate overfitting, multiple data augmentation methods, including horizontal and vertical flipping, saturation adjustment, and noise injection, were applied to expand the dataset, as shown in Fig. 8. The augmented dataset contains 3,620 images and is divided into training, validation, and test sets at a ratio of 7:2:1..

Fig.7. Examples of data augmentation applied to the ceramic cup defect dataset

C. Evaluation Metrics

The performance metrics used in this paper are Precision, Recall, mAP@0.5, mAP0.5:0.95, Parameters, Computational Complexity (GFLOPS), and Speed (Frames per Second, FPS).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$mAP = \frac{\sum_{i-1}^{N}\int_0^1 P(\mathrm{R_i})d\mathrm{R_i}}{cN} \quad (6)$$

Where TP is the number of true positive samples correctly detected, FP is the number of negative samples incorrectly detected as positive, and FN is the number of positive samples incorrectly detected as negative. mAP@0.5 represents the mean Average Precision (AP) value at a 50% IoU threshold. mAP@0.5:0.95 represents the mean AP across IoU thresholds ranging from 50% to 95%.

D. Ablation Study

To evaluate the effectiveness of the various improvement modules in IR-DETR, ablation experiments were performed on the self-constructed dataset, with RT-DETR serving as the baseline model. The performance improvements of each component are presented in Table II. Analysis of the results shows that all enhancement methods surpass the baseline model in detection performance.

Under the baseline setting, RT-DETR achieves 58.3 GFLOPs and 20M parameters, with mAP50 and mAP50–95 of 50.3% and 33.6%, respectively. Incorporating SSFF yields gains of 2.0% in mAP50 and 1.1% in mAP50–95 with negligible additional cost, indicating improved feature representation. Further introducing DySample raises mAP50 to 52.9% while maintaining comparable complexity. Replacing standard convolutions with WTConv reduces the model to 41.6 GFLOPs and 13M parameters and improves mAP50 to 53.6%, demonstrating superior lightweight feature extraction. By integrating all three modules, the proposed IR-DETR achieves 48.4 GFLOPs and 14.4M parameters, with mAP50 and mAP50–95 reaching 54.3% and 34.8%, respectively, substantially outperforming the baseline.

TABLE II. ABLATION EXPERIMENTS RESULT

SSFF	Dysample	WTConv	GFLOPS(G)	Params(M)	map@50(%)	map@50:95(%)
			58.3	20	50.3	33.6
√			63.2	20.4	52.3	34.7
		√	**41.6**	**13**	53.6	33.7
√	√		62.9	20.4	52.9	33.5
√	√	√	48.4	14.4	**54.3**	**34.8**

E. Comparative Experiment

To verify the effectiveness and innovation of the proposed algorithm, comparative experiments are conducted with YOLOv8, YOLOv9, YOLOv10, and classical models such as Faster R-CNN, Mask_RCNN, and RT-DETR, as shown in Table III.

According to the experimental results in Table III, the proposed IR-DETR (Ours) demonstrates significant advantages in overall performance, including harmonic mean F1-Score, average precision mAP@0.5, mAP@0.5:0.95, and computational complexity in GFLOPS. In terms of mAP@0.5, IR-DETR (Ours) achieves the highest value (54.3%) among the 10 models, outperforming recent object detection models such as Faster RCNN, Mask RCNN, RTDETR-L, RTDETR-R18, and YOLOv8 by 9.2%, 5.0%, 6.3%, 5.3%, and 3.5%, respectively. For mAP@0.5:0.95, IR-DETR (Ours) outperforms YOLOv11 by 3.7%. In summary, the proposed IR-DETR (Ours) effectively reduces computational complexity while maintaining competitive detection accuracy, demonstrating a favorable balance between efficiency and performance. Owing to its lightweight design and robust detection capability, the model is well suited for deployment in scenarios with limited computational resources. Moreover, its advantages are particularly evident in tasks requiring accurate small-object detection and reliable performance under stringent precision demands.

TABLE III. COMPARATIVE EXPERIMENTAL RESULTS

model	Type	GFLOPS/G	Params/M	F1-Score/%	mAP@0.5/%	Map@0.5:0.95/%
Faster_RCNN	Two-stage	284	60.6	47.3	45.1	23.7
Mask_RCNN	Two-stage	67.1	19.7	52.2	49.3	32.7
RTDETR-L	DETR	66.0	19.6	49.5	48.0	32.7
RTDETR-R18	DETR	58.3	17.2	51.5	49.0	31.7
yolov8m	one-stage	78.8	25.9	53.6	50.8	31.5
yolov9m	one-stage	76.5	20.1	50.4	49.5	30.6
yolov10m	one-stage	61.2	15.8	51.9	50.6	34.1
yolov11m	one-stage	68.1	20.1	54.7	51.2	34.1
IR-DETR(ours)	one-stage	48.4	14.4	**58.8**	**54.3**	**34.8**

F. Experimental comparison based on Visdrone2020 dataset

To further verify the generalization capability of the proposed IR-DETR model, comparative experiments were conducted on the VisDrone2020 dataset, which is a representative benchmark characterized by dense small objects, complex backgrounds, and substantial scale variations. Evaluating on this dataset helps assess whether the proposed method can maintain robust detection performance under more challenging and diverse visual conditions beyond the self-built dataset. The corresponding results are reported in Table IV..

TABLE IV. COMPARATIVE EXPERIMENTS OF DIFFERENT DATASETS

Datasets	Model	mAP50%	mAP50%-95%	F1-Score
Self-built dataset	RT-DETR-R18	49.0	31.7	51.5
	IR-DETR(ours)	**54.3**	**34.8**	**58.8**
Visdrone2020	RT-DETR-R18	45.1	23.7	45.0
	IR-DETR(ours)	51.2	34.1	**54.7**

Table IV compares RT-DETR-R18 and IR-DETR on the self-built dataset and VisDrone2020. IR-DETR consistently achieves better mAP50, mAP50-95, and F1-Score on both datasets, indicating its superior performance and strong generalization capability across different data domains.

V. CONCLUSIONS

To address missed detections and inadequate recognition of subtle defects in finished porcelain cups, this paper presents an improved IR-DETR model incorporating SSFF, WTConv, and DySample. Specifically, WTConv is introduced into the backbone to enhance fine-defect representation with low computational overhead, SSFF strengthens cross-scale contextual modeling, and DySample refines upsampling to improve feature discrimination while reducing complexity. Experiments on the self-constructed dataset show that the proposed model achieves 54.3% mAP50 and 34.8% mAP50–95 with only 14.4M parameters and 48.4 GFLOPs. Compared with classical YOLO- and R-CNN-based detectors, the proposed method attains higher accuracy at lower computational cost, demonstrating a better trade-off among accuracy, efficiency, and complexity for minute porcelain-cup surface defect detection.

ACKNOWLEDGMENT

This research is supported by Hunan Provincial Natural Science Foundation Youth Student Basic Research Project with Grant No. 2025JJ60931.

REFERENCES

[1] N.D., New quality productivity - focus on new breakthroughs in industrial ceramics, China Building Materials (01) (2024) 94-95.

[2] F.B. Zhang, B. Ning, N.L. Kong, J.L. Li, M.M. Zhao, Y.T. Guo, Research on Surface Defect Detection Technology based on Machine Vision, New Technology & New Process (06) (2025) 74-80.

[3] C. Zuo, Y.B. Zhang, Y.S. Qi, X.Y. Li, Y.C. Wang, Detection of Surface Scratch Defects of Printing Products Based on Machine Vision, Printing and Digital Media Technology Study (05) (2023) 42-48.

[4] H.Y. Jiang, Z.S. Dong, Z.J. He, Surface defect detection of wire rope based on machine vision, Journal of Taiyuan University of Science and Technology 44(05) (2023) 434-439+446.

[5] J.R. Wan, T.C. Sun, Y. Lu, K.N. Gu, D.R. Hu, L.J. Pan, Research on surface defect detection method of seaweed based on machine vision, Industrial Control Computer 36(09) (2023) 13-14+17.

[6] R. Girshick, Fast r-cnn. Proceedings of the IEEE international conference on computer vision (2015) 1440-1448.

[7] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6) (2016) 1137-1149.

[8] K. He, G. Gkioxari, P, Dollár, R. Girshick, Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (2017) 2961-2969.

[9] J.J. Pan, C. Zeng, J. Zhang, Z.Y. Li, X.N. Geng, Ceramic Surface Defect Detection Algorithm Based on Improved YOLOv5, Modern Information Technology 8(13) (2024) 70-75.

[10] W.L. Ding, Z.P. Zhang, Z.Q. Lei, P. Sun, Deep learning ceramic surface defect detection algorithm research, Journal of Electronic Measurement and Instrumentation 37(11) (2023) 161-169.

[11] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2024) 16965-16974.

[12] S.E. Finder, R. Amoyal, E. Treister, O. Freifeld, Wavelet convolutions for large receptive fields, European Conference on Computer Vision. Cham: Springer Nature Switzerland (2024) 363-380.