scientific reports

Check for updates

OPEN

Diamond-DETR: lightweight real-time quality evaluation algorithm for synthetic diamonds

Xin Yan[1], Saidong Yang[2], Shixiong Zhang[1]✉, Xingchong Li[1] & Ang Li[1]

Synthetic diamonds are prone to subtle defects during the manufacturing process, which can significantly impact their quality grading and market value. Due to the highly symmetrical and complex geometric structures of synthetic diamonds, accurately detecting defects of varying scales in complex backgrounds poses a considerable challenge. To address this issue, we propose a target detection algorithm specifically designed for synthetic diamond quality evaluation, named Diamond-DETR, aiming to improve detection accuracy and model generalization under resource-constrained environments. Diamond-DETR optimizes the backbone network by introducing a lightweight multi-scale feature extraction module and utilizes the RepFasterNet block to reduce computational complexity and enhance inference speed. Additionally, it incorporates an encoder structure based on a high-level screening-feature fusion pyramid network, which employs channel attention mechanisms to filter and fuse features at different scales, thereby enhancing the model's ability to detect defects of various sizes. Furthermore, a cross-stage fusion module is introduced, which leverages dilated convolutions to expand the receptive field without increasing computational cost, improving the model's capacity to perceive long-range dependencies and complex geometric structures. Experimental results demonstrate that Diamond-DETR outperforms the original RT-DETR model in terms of parameter efficiency, inference speed, and detection accuracy, making it particularly well-suited for deployment in resource-constrained inspection scenarios. The model also exhibits competitive performance in a cross-dataset evaluation on an external industrial dataset, indicating potential applicability in related industrial inspection scenarios.

Keywords Diamond-DETR, Synthetic diamond quality evaluation, RepFasterNet block, HSFPN encoder, DRBC3 block

Diamond is the hardest known material in nature and is widely used in industrial applications due to its unique physical properties. Since the emergence and gradual maturation of synthetic diamond technology in the 1960s, synthetic diamonds have almost completely replaced natural ones in industrial use.

The physical properties of diamonds are closely related to their crystal structure. Figure 1 shows images of high-quality (a, b, c) and low-quality (d, e, f, g) diamonds. In this study, diamond quality is defined from an industrial visual inspection perspective, based on surface-visible morphological characteristics observed under optical microscopy. By optimizing ellipticity, roundness, and improving purity and transparency, the impact toughness of diamonds under high temperatures has been significantly enhanced[1]. In addition, impurities and defects in the crystal structure can greatly affect their mechanical properties[2]. Therefore, evaluating the morphological quality of synthetic diamond particles is not only of great technical and economic value but also significantly improves the precision and efficiency of automated sorting.

Traditional methods for sorting synthetic diamonds include vibrating screens, magnetic separation, and heavy liquid separation[3], but all have inherent limitations. Vibrating screens perform poorly for small particles, magnetic separation is ineffective for non-magnetic inclusions, and heavy liquid separation depends strongly on density differences, making these approaches unsuitable for high-precision quality-based sorting of purer diamonds. As a result, quality evaluation still largely relies on manual inspection under optical microscopes, which is time-consuming and susceptible to operator subjectivity and fatigue, limiting efficiency, consistency, and quantitative analysis. With increasing demands for precision and automation, automated visual inspection has become an inevitable trend in industrial diamond processing.

[1]School of Mechanical and Electrical Engineering, Henan University of Technology, Zhengzhou 450000, China. [2]Faculty of Engineering, Huanghe University of Science and Technology, Zhengzhou 450000, China. ✉email: zsx@haut.edu.cn

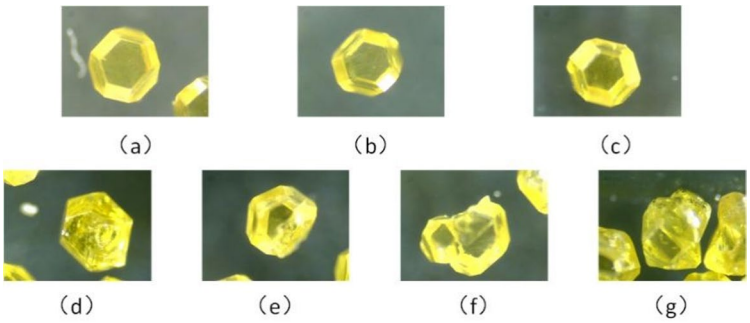

Fig. 1. Images of high-quality (**a**–**c**) and low-quality (**d**–**g**) synthetic diamond particles.

Since the breakthrough of AlexNet in 2012, convolutional neural networks (CNNs) have driven rapid progress in computer vision. Object detection methods such as YOLO[4] and Faster R-CNN[5], as well as instance segmentation approaches such as Mask R-CNN[6], with point-based mask refinement modules like PointRend[7], enable efficient object localization and feature extraction by learning spatial and channel representations within local receptive fields. However, CNN-based methods are inherently limited in modeling global context and long-range dependencies. Recently, lightweight CNN detectors have been applied to synthetic diamond quality inspection, where improved YOLOv8n-based models demonstrated effective extraction of local morphological features under microscopic imaging conditions[8]. While these anchor-based one-stage detectors are well suited for real-time inspection, they mainly rely on local feature aggregation and scale-specific fusion.

In contrast, the multi-head self-attention mechanism in Transformers excels at modeling long-range dependencies but is less effective in extracting local features. Therefore, combining CNNs and Transformers to simultaneously capture local details and global context has become the core idea behind convolutional vision Transformers. DETR[9], the first end-to-end object detection framework based on Transformers, treats detection as a set prediction task, eliminating hand-crafted components like anchor generation and non-maximum suppression, thus simplifying the detection pipeline. However, DETR suffers from slow convergence and complex query optimization.

To address these issues, several improved versions have been proposed. Deformable DETR[10] enhances training speed by optimizing the multi-scale attention mechanism. Conditional DETR[11] introduces conditional queries to simplify optimization and accelerate convergence. Anchor DETR[12] combines queries with anchors, reducing training difficulty. DAB-DETR[13] uses 4D reference points to progressively refine bounding boxes and improve localization accuracy. DN-DETR[14] incorporates a denoising mechanism during training to improve efficiency. DINO[15] further boosts both accuracy and speed. Nonetheless, DETR still faces challenges due to its high computational cost, limiting its practical applications.

Recently, Baidu introduced RT-DETR[16], a real-time end-to-end object detector that effectively eliminates inference delay caused by non-maximum suppression. It achieves notable improvements in both speed and accuracy, showing strong potential for real-time object detection. We briefly follow the standard RT-DETR pipeline[16], which adopts a backbone–encoder–decoder architecture without NMS (Fig. 2), and focus on lightweight, task-oriented modifications described in the following sections.

Although RT-DETR performs well in real-time object detection, it still faces limitations in industrial inspection, particularly in synthetic diamond quality evaluation, where accurate discrimination of small and highly similar objects under multi-scale conditions remains challenging.

Diamond-DETR model and various improvements

To address the limitations of RT-DETR in industrial detection, this paper proposes an optimized version—Diamond-DETR—designed for resource-constrained platforms, with a focus on lightweight design, multi-scale feature extraction, and long-range dependency modeling.

The proposed Diamond-DETR introduces three integrated improvements over RT-DETR[17,18]:

First, a lightweight RepFasterNet backbone is adopted, which integrates partial convolution and re-parameterization to reduce computational complexity while preserving discriminative local features critical for defect recognition.

Second, the encoder is redesigned with a High-level Screening Feature Fusion Pyramid Network (HSFPN) encoder[19], which combines attention-based feature selection with cross-scale fusion, thereby enhancing the model's robustness to multi-scale variations typical of diamond particles.

Third, a Dilated Re-parameterized C3 block (DRBC3) replaces the RepC3 module, enlarging the receptive field through multi-rate dilated convolutions without additional network depth, strengthening the capacity to capture spatial dependencies in complex geometric structures. The optimized architecture is shown in Fig. 3.

Backbone network improvements

To achieve high detection speed without relying on high-performance hardware, this paper adopts the FasterNet architecture proposed by Chen et al.[18], which replaces standard convolutions with partial channel-wise spatial convolutions (referred to as PConv in FasterNet) to significantly reduce computational cost. However, directly

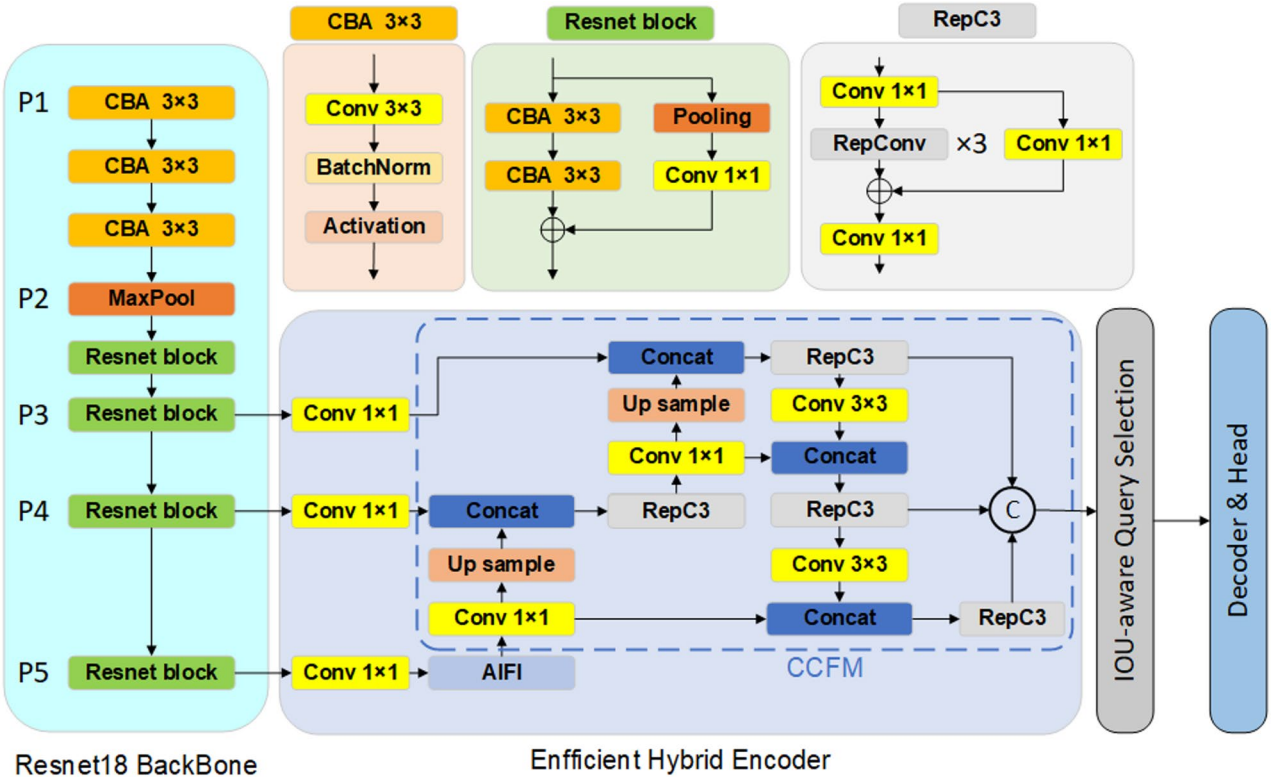

Fig. 2. Architecture of RT-DETR model.

applying the original FasterNet to industrial inspection tasks is insufficient, since the defect patterns in synthetic diamonds are subtle, small-scale, and often embedded in complex geometric backgrounds.

Building upon the partial channel-wise convolution design in FasterNet, we further propose a re-parameterized partial convolution module (RepPConv), which is integrated into a redesigned RepFasterNet block. During training, it introduces a multi-branch structure to enhance feature extraction, and during inference, the branches are fused into a single convolution operation. This design retains the efficiency of partial convolutions while further improving inference speed. Unlike a simple transfer of existing lightweight modules, our RepFasterNet block is specifically adapted to the requirements of synthetic diamond defect detection, where effectively capturing fine-grained local geometric variations—particularly along edges and corners, which are common regions for defect occurrence—plays a critical role under strict computational constraints. The module structure is shown in Fig. 4.

The improved module, RepFasterNet block, replaces the ResNet block[20,21] in the original network. A simple calculation of the FLOPs (floating point operations per second) during the inference phase is conducted to compare the performance with the original module. The FLOPs for a standard convolutional layer can be calculated using the following formula:

$$FLOPs = 2 \times H_{kernel} \times W_{kernel} \times C_{in} \times C_{out} \times H_{out} \times W_{out} \tag{1}$$

Where:

H_{kernel} and W_{kernel} are the height and width of the convolution kernel,

C_{in} is the number of input channels,

C_{out} is the number of output channels,

H_{out} and W_{out} are the height and width of the output feature map.

For example, when both the input and output channels are 64 and the input feature map size is 160 × 160:

In the RepFasterNet block, only 1/4 of the channels (i.e., 16) participate in the 3 × 3 PConv operation. During inference, the PConv branch is fused into a single 3 × 3 convolution. The MLP (Multilayer Perceptron) has a hidden layer with twice the number of output channels and uses two 1 × 1 convolutions to increase the channels to 128 and then reduce them back to 64. Since the input and output channels are equal, the 1 × 1 convolutions for channel transformation are excluded from the FLOPs calculation.

The FLOPs for the RepFasterNet block are calculated as follows:

$$FLOPs_{total} = FLOPs_{CBA} + FLOPs_{RepPconv} + FLOPs_{MLP} \tag{2}$$

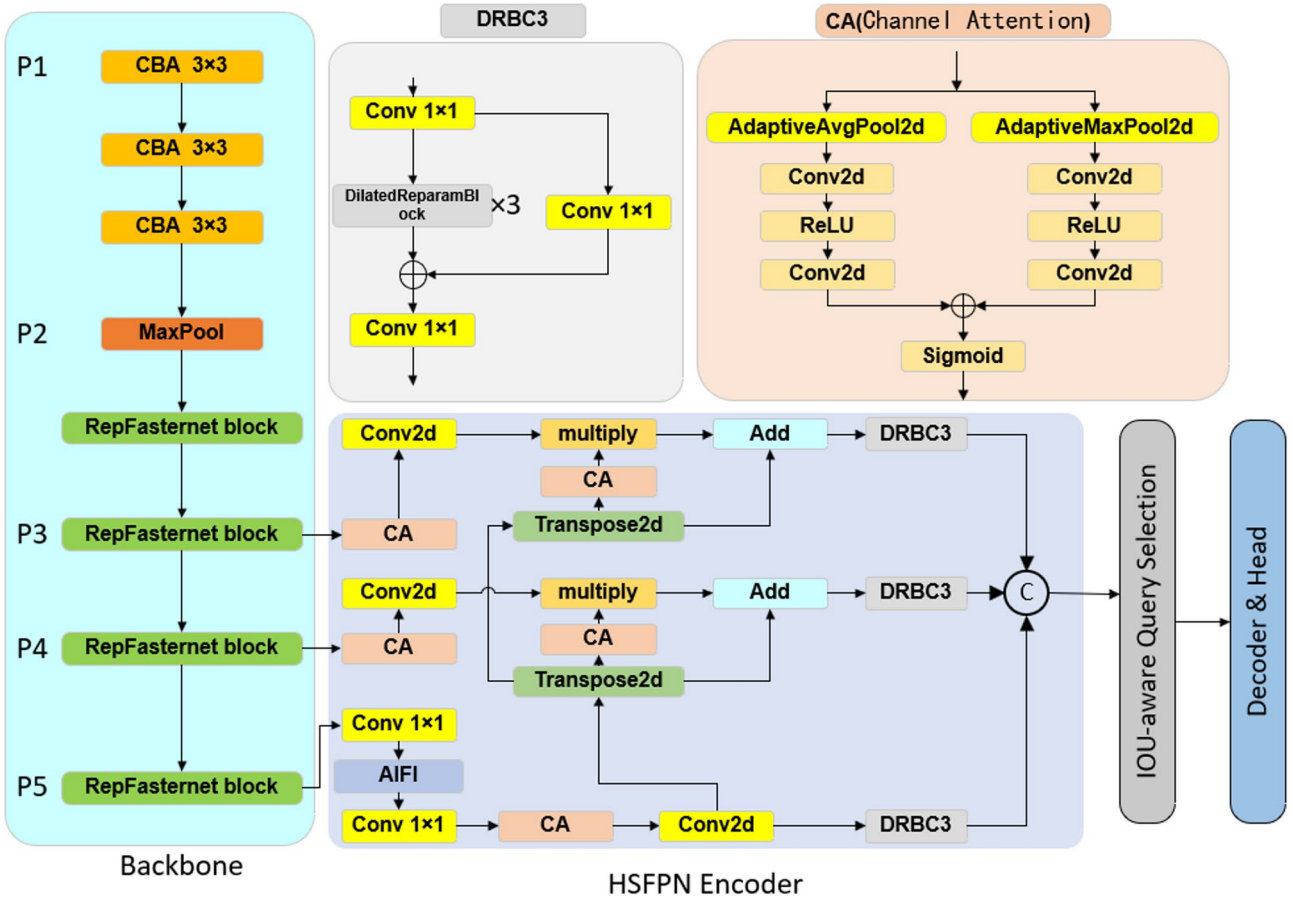

Fig. 3. Architecture of diamond-DETR model.

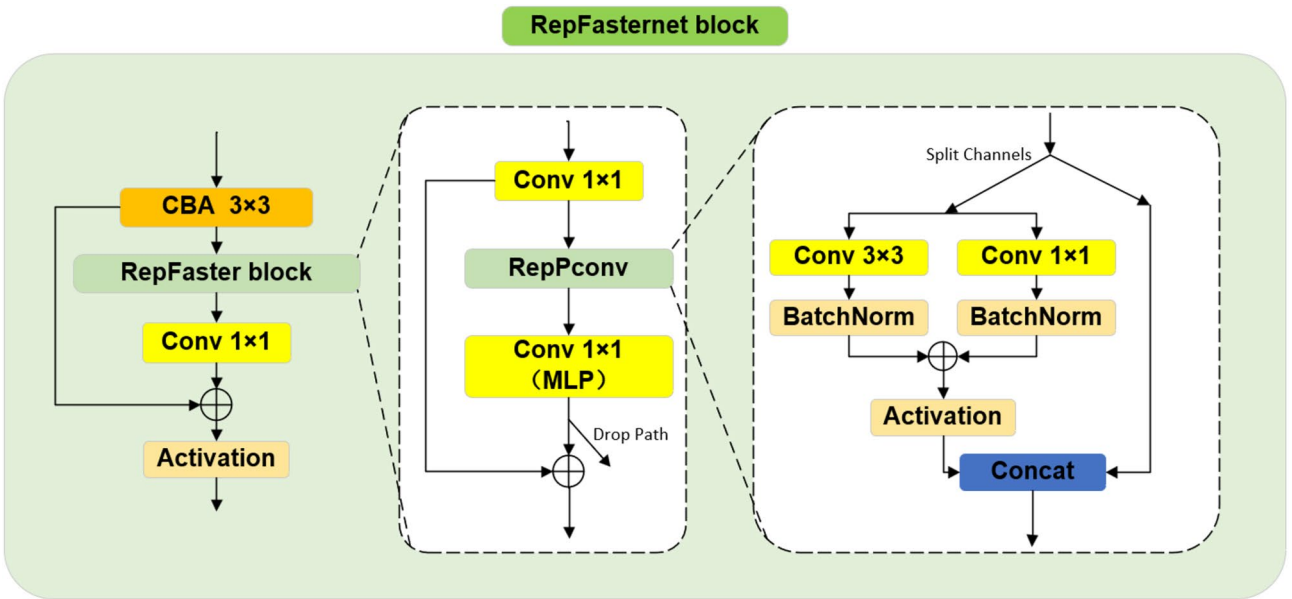

Fig. 4. Architecture of RepFasterNet block.

$$\begin{cases} FLOPs_{CBA} = 2 \times 9 \times 64 \times 64 \times 160 \times 160 = 1887436800 \\ FLOPs_{RepPconv} = 2 \times 9 \times 16 \times 16 \times 160 \times 160 = 117964800 \\ FLOPs_{MLP} = 2 \times 160 \times 160 \times 64 \times 128 \times 2 = 838860800 \\ FLOPs_{total} = 1887436800 + 117964800 + 838860800 = 2844262400 \end{cases} \tag{3}$$

For the ResNet block, two consecutive standard 3×3 convolutional layers are used, each with 64 input and output channels and an output feature map size of 160×160. According to Eq. (1), the FLOPs of a single convolution under this configuration is 2×3×3×64×64×160×160. therefore, the total FLOPs of the ResNet block is obtained by doubling this value, resulting in 3,774,873,600 FLOPs.

It is evident that the RepFasterNet block effectively reduces the FLOPs of the backbone network, thus accelerating the model's inference speed.

In addition to reducing FLOPs and parameters compared with the ResNet backbone, the proposed design enhances the ability to capture local geometric variations that are essential for distinguishing diamond defects. As shown in the ablation study (Sect. 3.4), the RepFasterNet block improves both precision and recall while reducing model size, indicating that its efficiency is effectively integrated with subsequent modules to form a domain-specific adaptation for synthetic diamond quality evaluation.

Encoder improvements

Different batches of synthetic diamonds exhibit varying sizes and shapes under a microscope, and even within the same batch, differences in magnification can result in noticeable scale variations. This makes it difficult to perform detection at a unified scale. Traditional feature extraction methods are poorly adapted to such variations, leading to decreased detection accuracy. Additionally, microscope images of synthetic diamonds often have low resolution and limited features, making it challenging for models to extract sufficient semantic and positional information—especially when feature representations vary across scales, further complicating detection and classification.

To address these issues, this study replaces the original CNN-based Cross-scale Feature Fusion module (CCFF) with the High-level Screening-feature Fusion Pyramid Network (HSFPN) encoder. By fusing features across different scales, HSFPN encoder captures richer hierarchical information and better adapts to scale variation. High-level semantic features are used as weights to filter and enhance low-level features with precise spatial information, preserving and reinforcing key information during fusion and improving detection performance.

The core mechanism consists of the Feature Selection Module and the Feature Fusion Module, described as follows:

Feature selection module

The feature selection module of HSFPN encoder uses high-level semantic features to guide the filtering of low-level features through a channel attention (CA) mechanism, as illustrated in Fig. 5.

Given a high-level feature map $f_{high} \in \mathbb{R}^{C \times H \times W}$, where C, H, and W denote the number of channels, height, and width, respectively, global average pooling (GAP) and global max pooling (GMP) are first applied along the spatial dimensions to obtain two channel-wise descriptors:

$$\begin{aligned} f_{avg} &= GlobalAvgPool\left(f_{high}\right) \\ f_{max} &= GlobalMaxPool\left(f_{high}\right) \end{aligned} \tag{4}$$

Where $f_{avg}, f_{max} \in \mathbb{R}^{C \times 1 \times 1}$.

These two descriptors are then independently projected by fully connected layers and combined through element-wise addition, followed by a sigmoid activation function to generate the channel attention vector:

$$f_{CA} = \sigma\ \left(W_{avg} f_{avg} + W_{max} f_{max}\right) \tag{5}$$

Where:

$f_{CA} \in \mathbb{R}^{C \times 1 \times 1}$ denotes the channel-wise attention weights,

W_{avg} and W_{max} are learnable weight matrixes,

$\sigma\left(\cdot\right)$ is the sigmoid function.

The obtained attention vector is then applied to the corresponding low-level feature map $f_{low} \in \mathbb{R}^{C \times H_l \times W_l}$ through channel-wise multiplication to produce the filtered low-level feature:

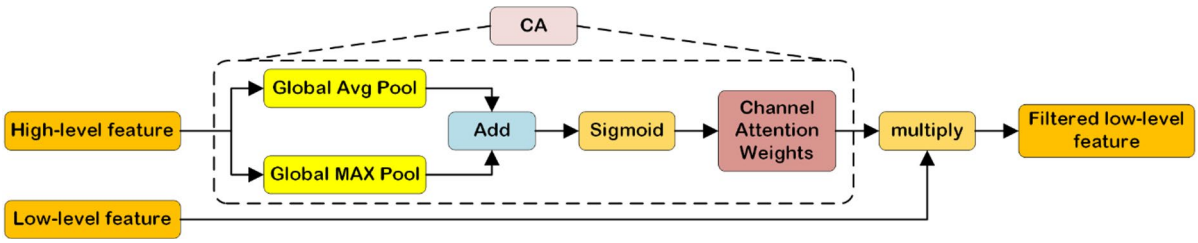

Fig. 5. The process of feature selection in HSFPN encoder.

$$f'_{low} = f_{low} \odot f_{CA} \tag{6}$$

Where f'_{low} denotes the refined low-level feature, $\odot$ denotes element-wise multiplication with broadcasting along the spatial dimensions.

Through this process, high-level semantic information is explicitly used to emphasize informative channels and suppress less relevant responses in low-level features, thereby improving feature discriminability under multi-scale and low-contrast conditions.

Feature fusion module

The feature fusion module fuses attention-modulated low-level features with upsampled high-level features to create a feature map that contains both richer semantic and more precise positional information.

First, the high-level features map f_{high} is upsampled to match the spatial resolution of the low-level feature map f_{low} using transposed convolution:

$$f_{high-up} = \mathrm{Trans}posedConv\left(f_{high}\right) \tag{7}$$

Next, the same channel attention (CA) mechanism described in Sect. 2.2.1 is reused to generate channel-wise attention weights from the upsampled high-level features:

$$f_{att} = CA\left(f_{high-up}\right) \tag{8}$$

Where $f_{att} \in \mathbb{R}^{C\times 1\times 1}$.

It is emphasized that no additional attention structure is introduced at this stage; the CA mechanism is shared and reused to maintain architectural consistency.

The attention weights are then applied to the low-level feature map to obtain an attention-modulated low-level representation:

$$f^{att}_{low} = f_{low} \odot f_{att} \tag{9}$$

Where $\odot$ denotes element-wise multiplication with broadcasting along the spatial dimensions.

Finally, the modulated low-level features are fused with the upsampled high-level features through element-wise addition:

$$f_{fusion} = f^{att}_{low} + f_{high-up} \tag{10}$$

This fusion strategy allows high-level semantic cues to guide the refinement of low-level spatial features while preserving structural information, resulting in a more robust multi-scale representation for subsequent decoding stages.

Cross-stage fusion module improvements

The Cross Stage Partial (CSP) block, also known as the C3 block, is used for feature fusion in the model. In RT-DETR, the RepC3 block integrates reparameterized convolutions (RepConv), which can be simplified into a single efficient convolution layer during inference, achieving both feature fusion and high inference efficiency. Its structure is shown in Fig. 3.

Based on this, we propose an improved version called the DRBC3 block, which replaces the standard convolutions in RepC3 with dilated convolutions. By using multiple dilation rates, this module expands the receptive field without significantly increasing computational cost. This allows the model to capture more spatial information and better model long-range dependencies, thereby improving detection accuracy and robustness. The structure of the DRBC3 block is shown in Fig. 3. The following section introduces the calculation method of dilated convolutions and how to expand the receptive field without increasing the kernel size. For a standard 2D convolution, the output feature map for each element can be represented as:

$$F_{out(i,j)} = \sum_{m=1}^{H_{kernel}} \sum_{n=1}^{W_{kernel}} F_{in(i+m,j+n)} \cdot W_{(m,n)} \tag{11}$$

Where:

$F_{out(i,j)}$ is the value of the output feature map at position (i, j),
$F_{in(i,j)}$ is the value of the input feature map at position (i, j),
$W_{(m,n)}$ is the weight at the corresponding position of the convolution kernel,
H_{kernel} and W_{kernel} are the height and width of the convolution kernel.

In dilated convolution, gaps are introduced between the elements of the convolution kernel, and the output of dilated convolution can be expressed as:

$$F_{out(i,j)} = \sum_{m=1}^{H_{kernel}} \sum_{n=1}^{W_{kernel}} F_{in(i+m\cdot r,j+n\cdot r)} \cdot W_{(m,n)} \tag{12}$$

Where:

r is the dilation rate, which refers to the spacing between the elements of the convolution kernel.

For standard convolution, the receptive field calculation for layer L can be summarized as:

$$R_L = R_{L-1} + (k-1) \times s \tag{13}$$

Where:
R_L is the receptive field size for layer L,
k is the size of the convolution kernel,
s is the stride.
For a dilated convolution with dilation rate r, the receptive field calculation formula is:

$$R_L = R_{L-1} + (k-1) \times s \times r \tag{14}$$

Through the expanded receptive field and enhanced multi-scale feature extraction of the DRBC3 block, the block can better capture information within the image, improve the detection of small objects, and perform more stably when processing input features of various scales. This ultimately enhances the overall detection performance of the network.

Experiments and discussions

Datasets

Due to the lack of publicly available datasets for synthetic diamond quality evaluation, we constructed an in-house dataset. It contains 627 microscope images, each of which may include multiple synthetic diamond particles, with annotations provided at the instance level. In total, 435 defective instances and 310 high-quality instances are labeled across the dataset.

The task is formulated as a binary classification problem with two categories: high-quality and defective. Defective instances exhibit surface-visible defects such as irregular shape, broken edges, or incomplete appearance, and are consolidated into a single class in accordance with industrial quality evaluation requirements.

All images were acquired under an 80× optical microscope using an MC-D500U high-definition camera and stored in JPG format. The dataset was randomly divided into training, validation, and test sets with a 7:2:1 ratio. Instance annotations were manually performed using the LabelImg tool. The samples were provided by an industrial partner, ensuring practical relevance.

Experimental environment and parameters

All experiments were conducted on a Windows 11 system with an Intel(R) Core (TM) i7-12700 H CPU and an NVIDIA RTX 3070Ti 8G GPU, and all FPS results were measured on the RTX 3070Ti. The number of object queries was set to 300, following the default RT-DETR configuration, and kept fixed for all experiments. Since each image in our dataset contains only a small number of instances (typically 1–3), this query capacity is sufficient and is unlikely to limit detection performance in this low-density detection setting. Model training was performed using PyTorch 1.13.1 with Python 3.9.17 and CUDA 11.7.1. Unless otherwise specified, all input images were resized to 640 × 640 with aspect ratio preserved and normalized to [0,1]. No advanced data augmentation was applied; only random horizontal flipping was used during training. Training parameters are summarized in Table 1. All experiments were conducted using a fixed random seed, and the reported results correspond to a single training run. Performance variability across different random initializations was not explicitly assessed.

Evaluation metrics

The evaluation metrics and their definitions are described as follows. Precision (P) and Recall (R) are defined based on true positives (TP), false positives (FP), and false negatives (FN), measuring the accuracy and completeness of positive sample detection, respectively:

$$P = \frac{TP}{TP+FP} \tag{15}$$

Recall (R) measures the model's ability to identify all positive samples. In actual production, it reflects the algorithm's capability to filter as many high-quality particles as possible, which can enhance economic benefits and reduce resource waste.

$$R = \frac{TP}{TP+FN} \tag{16}$$

Mean Average Precision (mAP) is used to evaluate the overall detection performance across the dataset and is computed as the integral of the precision–recall curve:

$$AP = \int_0^1 P(R)\, dR \tag{17}$$

$$mAP = \frac{1}{N} \sum_{i=0}^{n} AP_i \tag{18}$$

Batch size	Epochs	Initial learning rate	Momentum	Weight decay	Input size	Optimizer
4	150	0.0001	0.8	0.0001	640	SGD

Table 1. Training parameters used in the experiments.

Model Parameters (Param) and Frames Per Second (FPS) reflect the model's efficiency in practical applications. Unless otherwise specified, a confidence threshold of 0.25 was used for inference, and IoU thresholds followed the standard definitions for mAP50 and mAP50–95, which were applied consistently across all models.

Ablation study

To evaluate the contribution of each proposed module, we define Baseline Model (A) as the original RT-DETR architecture without structural modifications. Based on this baseline, a series of variants are constructed: A + B (RepFasterNet block), A + C (HSFPN encoder), A + D (DRBC3 block), A + B + C, and A + B + C + D (Diamond-DETR). By comparing their performance on the test set, the individual and combined effects of each component can be assessed. The results are summarized in Table 2.

The results show that the new feature extraction network effectively reduces the number of parameters while slightly improving both precision and recall. With the introduction of the HSFPN encoder, recall increases by 5.4% and mean average precision (mAP) by 5.2%. The DRBC3 block improves precision by 2.8%, though it slightly impacts recall. Each improvement module contributes to parameter reduction and varying degrees of precision enhancement.

Overall, compared to the baseline model, the optimized network achieves a 3.1% increase in precision, a 2.6% increase in mAP, a 10% improvement in detection speed, and a 29% reduction in parameters. As shown in Fig. 6, after training stabilizes, the improved model consistently outperforms the baseline in both precision and mAP, demonstrating the effectiveness of the proposed enhancements.

Feature map visualization experiments

To visualize the decision-related attention of Diamond-DETR on synthetic diamond images, Grad-CAM (Gradient-weighted Class Activation Mapping)[22] is employed. The resulting attention maps highlight image regions that contribute most to the model's predictions. Representative samples with different quality-related geometric characteristics are selected to compare the attention responses of baseline and ablation models. This visualization provides an intuitive interpretation of the model's focus on geometry-relevant regions, particularly edges and structural details.

As shown in Fig. 7, the baseline model exhibits relatively dispersed attention, especially in structurally complex regions, indicating limited capability to consistently capture geometry-related cues[23]. In contrast, Diamond-DETR shows more concentrated attention on geometry-relevant regions, particularly along edges and vertices, which are critical for distinguishing geometric integrity. This improvement can be attributed to the DRBC3 block, which enhances long-range dependency modeling[24], and the HSFPN encoder, which improves the consistency of attention during multi-scale feature fusion.

Comparative experiments

To further validate the effectiveness of Diamond-DETR, we conducted comparative experiments with representative one-stage detectors from the YOLO family (YOLOv8 and YOLOv10, s/m scales), as well as commonly used two-stage and transformer-based baselines. All baseline models were re-trained on the same dataset split using the unified training parameters in Table 1 and evaluation settings described in Sect. 3.2–3.3, while keeping the official model architectures unchanged. The experimental results are reported in Table 3.

The results of the comparative experiments show that Diamond-DETR outperforms commonly used object detection models across several key metrics. Compared to two-stage models (e.g., Faster R-CNN), it achieves higher precision and recall while effectively reducing parameter count and inference time, supporting its applicability to real-time detection in resource-constrained scenarios. Compared to single-stage models (e.g., YOLO series), Diamond-DETR maintains similar inference speed with improved precision and recall, performing especially well in small object detection and complex backgrounds. Specifically, compared to the original RT-DETR, Diamond-DETR improves precision by 3.1%, mAP by 2.6%, and inference speed by 10%.

Cross-dataset experiments

To verify the applicability and robustness of the Diamond-DETR model, we conducted cross-dataset experiments using metal nuts from the MVTec Anomaly Detection (MVTecAD) dataset as test objects. For clarity, this evaluation is primarily designed as a controlled comparison within the RT-DETR framework, while additionally including a representative CNN-based detector (YOLOv10s) for broader reference rather than conducting a

Model	*P* (%)	*R* (%)	mAP50 (%)	Param/MB	FPS
A	89.7	87.9	88.0	38.6	13.7
A + B	90.3	88.5	88.6	32.8	15.7
A + C	90.4	93.3	90.2	35.0	13.9
A + D	92.5	85.0	87.4	35.0	13.0
A + B + C	91.7	89.2	90.3	29.2	15.6
A + B + D	91.5	88.8	88.3	29.2	15.2
A + C + D	93.9	86.8	89.1	33.2	13.7
A + B + C + D	92.8	88.3	90.6	27.4	15.2

Table 2. Ablation study results.

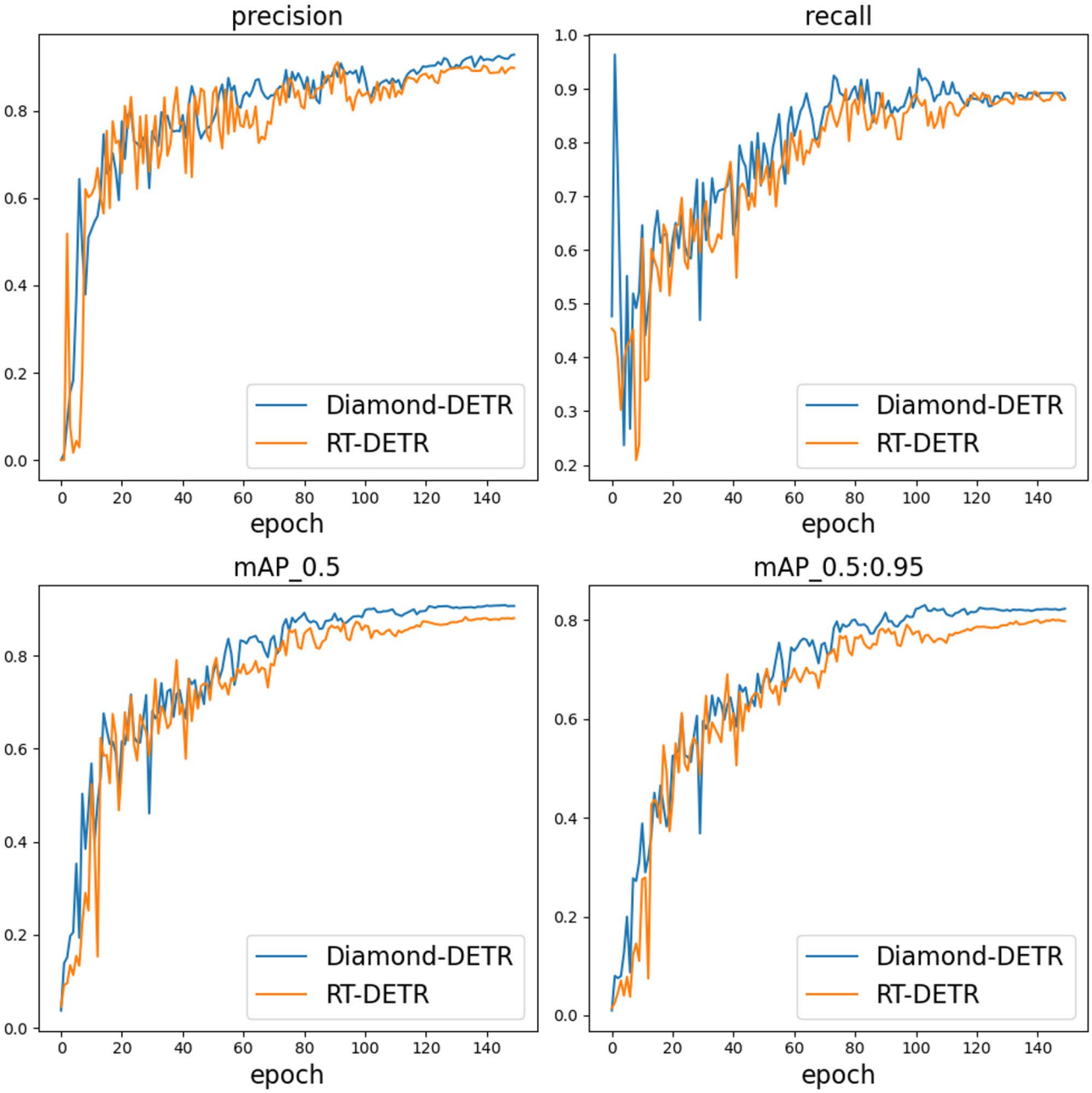

Fig. 6. Comparison of training curves for key evaluation metrics.

comprehensive benchmark against all state-of-the-art detectors. MVTecAD is a standard benchmark for industrial inspection, primarily used for anomaly detection and quality assessment.

Although metal nuts differ in shape from synthetic diamonds, they are also detection targets based on complex geometric features, making them suitable for evaluating cross-dataset transferability of Diamond-DETR across related tasks. The training curves and test results are shown in Fig. 8; Table 4.

All models were trained under the same data split and training settings to ensure fair comparison.

On the MVTecAD dataset, Diamond-DETR delivers strong detection performance, particularly in precision and recall compared with the original RT-DETR, showing consistent improvements over the original RT-DETR under the same training protocol. Although YOLOv10s achieves higher mAP on this small-scale dataset, Diamond-DETR maintains competitive performance while preserving the Transformer-based detection framework.

These results suggest that the proposed architectural design contributes to enhanced cross-dataset robustness of RT-DETR-style detectors. The strong performance of YOLOv10s also indicates that high-capacity CNN-based detectors can quickly saturate on limited-scale industrial datasets. Compared with the original RT-DETR, Diamond-DETR maintains more stable performance when transferred to detection tasks involving complex geometric structures and multi-scale targets. Its lightweight design, optimized multi-scale feature fusion,

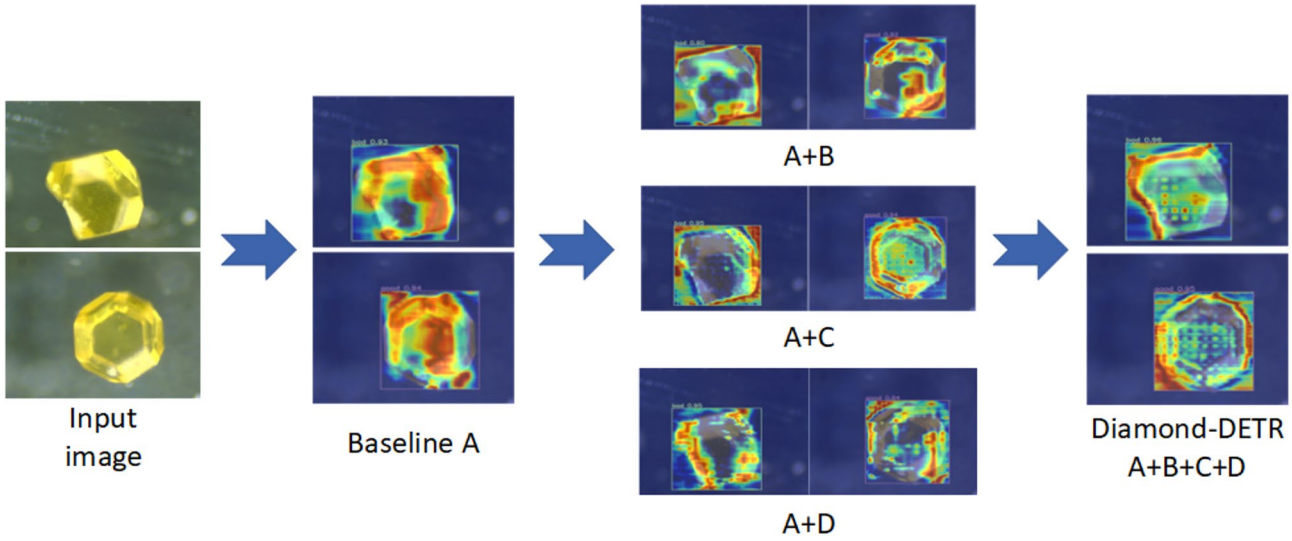

Fig. 7. Grad-CAM attention visualizations of baseline and ablation models.

Model	*P* (%)	*R* (%)	mAP50 (%)	Param/MB	FPS
Faster R-CNN	88.5	90.9	86.0	97.7	7.2
SSD	80.3	87.5	77.6	21.1	20.5
YOLOv8s	88.7	87.2	83.2	21.5	29.9
YOLOv8m	82.1	90.9	80.0	49.6	13.7
YOLOv10s	89.2	89.7	87.4	15.7	25.3
YOLOv10m	89.6	87.4	87.9	31.9	10.4
RT-DETR	89.7	87.9	88.0	38.6	13.7
Diamond-DETR	92.8	88.3	90.6	27.4	15.2

Table 3. Comparative results with lightweight CNN-based YOLO detectors and baselines.

and enhanced long-range dependency modeling support stable performance in this cross-dataset setting, demonstrating applicability in related industrial inspection scenarios.

Discussion

The Diamond-DETR model demonstrates solid detection performance and practical potential for deployment in resource-constrained smart manufacturing and automated inspection scenarios. With a model size of 27.4 MB, an inference speed of 15.2 FPS, and an mAP50 of 90.6%, it achieves a balanced trade-off between accuracy and efficiency. Compared with recent CNN-based approaches for synthetic diamond inspection, such as the improved YOLOv8n detector[8], Diamond-DETR follows a different design philosophy. While YOLO-based methods emphasize extreme lightweight design and local feature extraction, Diamond-DETR focuses on end-to-end detection with enhanced global dependency modeling and multi-scale geometric consistency. Although the improved YOLOv8n model in[8] was evaluated on a related but not identical dataset, both studies report similar performance ranges under similar microscopic inspection conditions. This indicates that Diamond-DETR achieves competitive practical performance while adopting a transformer-based end-to-end detection framework. While lightweight CNN backbones have been explored in other domains, this work adapts RepFasterNet specifically for synthetic diamond quality inspection, targeting fine-grained geometric discrimination under strict real-time and resource constraints. Rather than pursuing extreme parameter minimization, the proposed model maintains sufficient feature representation capacity for reliable industrial quality evaluation within RT-DETR-style frameworks[25].

The model also shows practical potential in several use cases: synthetic diamond sorting, where it can partially replace manual microscopic inspection for initial quality classification; inline micro-particle inspection, enabling real-time image analysis on production lines; and multi-magnification particle detection, ensuring adaptability across varying resolutions and magnification levels.

Experiments indicate the model maintains stable performance under challenging conditions, such as occlusion, complex geometries, and densely packed small objects. The DRBC3 block further strengthens long-range dependency modeling, improving the extraction of key features and reducing false positives and missed detections.

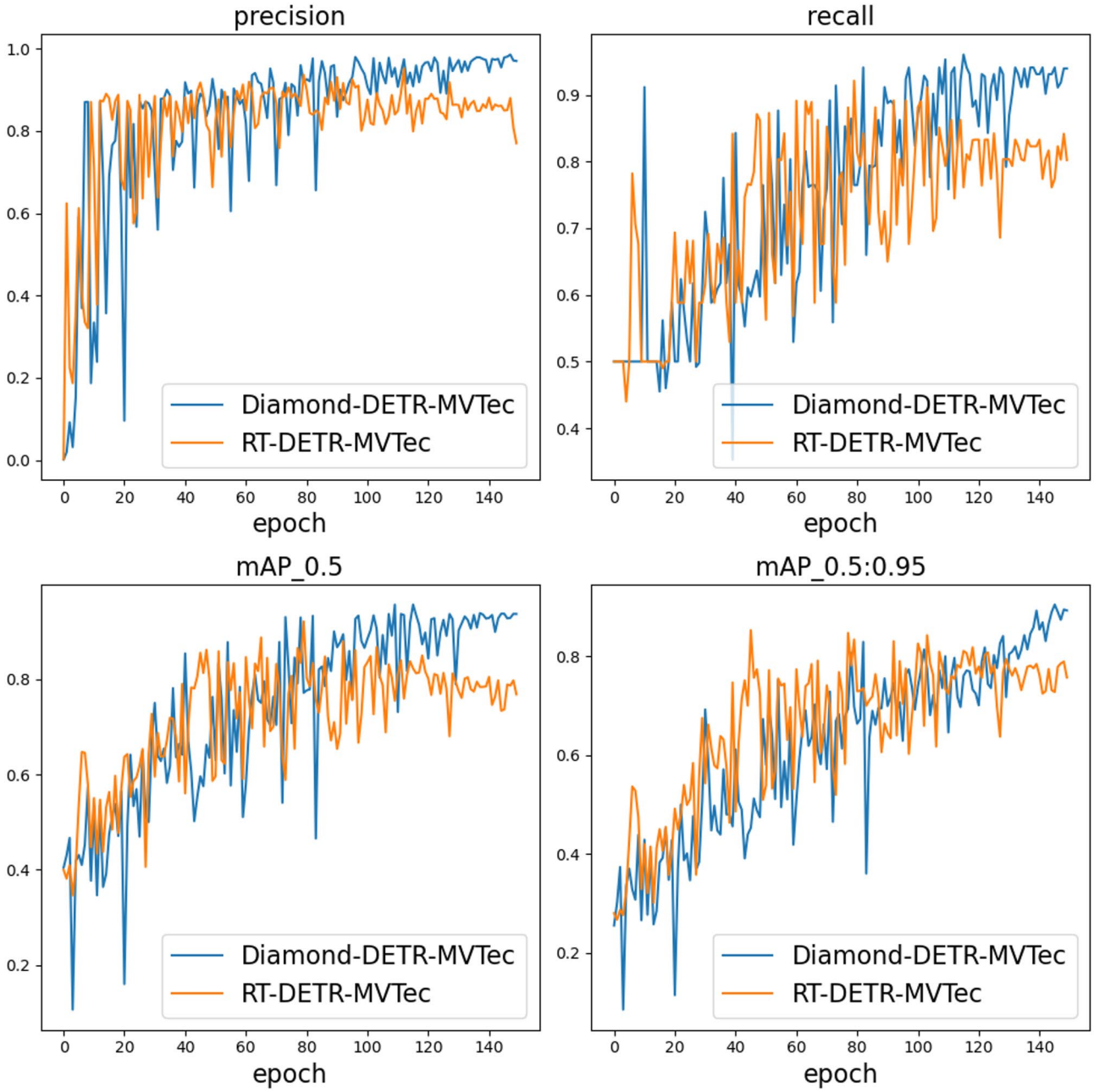

Fig. 8. Training curves of RT-DETR and Diamond-DETR on the MVTec metal nut subset.

Model	*P* (%)	*R* (%)	mAP 50 (%)	mAP 50–95 (%)
RT-DETR	91.8	76.5	85.6	85.3
YOLOv10s	89.0	83.4	94.6	94.6
Diamond-DETR	97.9	94.1	93.8	90.4

Table 4. Cross-dataset performance comparison on the MVTec metal nut subset (test set).

A qualitative inspection of test results shows that most confusion occurs between slightly defective particles and high-quality particles with near-regular geometry, particularly when defects are extremely subtle at edges or vertices. Misclassifications are occasionally influenced by uneven illumination or overlapping particles under microscopic imaging. With 640 × 640 RGB inputs, the model operates within moderate runtime memory and bandwidth requirements under the tested GPU configuration, indicating practical feasibility without excessive resource demand.

However, the model has certain limitations. Currently, it only assesses surface morphology and visible quality, and cannot detect internal defects (e.g., cracks, bubbles) or analyze material composition. Comprehensive quality evaluation still requires integration with other sensing technologies such as spectroscopy, X-ray, or CT. Diamond-DETR is better suited as a visual perception frontend.

Additionally, the current dataset comes from a single industrial supplier. Although it includes diverse samples, the limited data source may affect the model's generalizability[26]. Future work will focus on building multi-source datasets and exploring cross-domain adaptability[27].

Conclusion

This paper proposes an optimized algorithm based on RT-DETR, named Diamond-DETR, for automated quality evaluation of synthetic diamonds. By integrating a lightweight multi-scale feature extraction module, a high-level screening feature fusion pyramid (HSFPN encoder), and the DRBC3 block, Diamond-DETR reduces model complexity while improving detection accuracy and inference speed, enhancing overall robustness. Although lightweight backbones have been explored in other domains, this work presents an early task-oriented adaptation of such designs to synthetic diamond quality inspection, driven by practical industrial requirements. Experimental results show that Diamond-DETR outperforms the original RT-DETR in precision, mAP, and inference efficiency, making it suitable for deployment in resource-constrained industrial environments.

Diamond-DETR features a lightweight design that effectively reduces parameter count and computational complexity while maintaining strong detection performance. Its multi-scale feature fusion improves adaptability to targets of various sizes, with reliable performance in detecting complex geometric shapes and high-similarity targets. The DRBC3 block further enhances the model's ability to capture long-range dependencies.

Cross-dataset validation on an external industrial dataset indicates that Diamond-DETR can retain strong detection performance under a related inspection scenario, suggesting its potential applicability to similar industrial tasks.

Future work will focus on further optimizing the model's lightweight architecture and validating its performance in more real-world industrial environments. In addition, exploring its application in other high-precision detection tasks could enhance its adaptability and robustness across diverse scenarios. All performance evaluations in this study are conducted on a desktop GPU platform, and embedded or edge-device deployment is left for future work.

Data availability

The primary code and data supporting the findings of this study are available on GitHub at https://github.com/bigboyliang/DIAMOND. The repository includes the full source code for the model, a portion of the synthetic diamond dataset, the dataset used for cross-dataset validation experiments, and the results from visualization experiments. For access to the complete dataset for scientific research, please contact the corresponding author, providing your identity, affiliation, and intended purpose of use.

Received: 13 December 2025; Accepted: 16 March 2026
Published online: 29 March 2026

References

1. Chen, S. et al. Energy-dependent machining mechanism and process in water-jet guided laser processing single crystal diamond. *Alexandria Eng. J.* **118**, 681–691 (2025). https://doi.org/10.1016/j.aej.2025.01.115
2. Fadhil, A. A. et al. Structural characterization and detecting processes of defects in leaded brass alloy used for gas valves production. *Alexandria Eng. J.* **57**, 1301–1311 (2018). https://doi.org/10.1016/j.aej.2017.04.017
3. Luo, Z. et al. Research on the precision classification and sorting technology of synthetic diamond abrasive. *Diam. Abras. Eng.* **6**, 4–8 (2006). https://doi.org/10.13394/j.cnki.jgszz.2006.06.002
4. Redmon, J. et al. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
5. Ren, S. et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
6. He, K. et al. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2961–2969 (2017).
7. Kirillov, A. et al. PointRend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9796–9805 (2020). https://doi.org/10.1109/CVPR42600.2020.00982
8. Zhang, S. et al. An enhanced YOLOv8n object detector for synthetic diamond quality evaluation. *Sci. Rep.* **14**, 28035 (2024). https://doi.org/10.1038/s41598-024-79549-y
9. Carion, N. et al. End-to-end object detection with transformers. In European Conference on Computer Vision (ECCV), 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13
10. Zhu, X. et al. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv:2010.04159 (2020).
11. Meng, D. et al. Conditional DETR for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3651–3660 (2021). https://doi.org/10.1109/ICCV48922.2021.00363
12. Wang, Y. et al. Anchor DETR: Query design for transformer-based detector. In Proceedings of the AAAI Conference on Artificial Intelligence, 2567–2575 (2022).
13. Liu, S. et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR. arXiv:2201.12329 (2022).
14. Li, F. et al. DN-DETR: Accelerate DETR training by introducing query denoising. IEEE Trans. Pattern Anal. Mach. Intell. 46, 2239–2251 (2024). https://doi.org/10.1109/TPAMI.2023.3335410
15. Zhang, H. et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605 (2022).
16. Zhao, Y. et al. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16965–16974 (2024). https://doi.org/10.1109/CVPR52733.2024.01605
17. Ding, X. et al. RepVGG: Making VGG-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13733–13742 (2021).
18. Chen, J. et al. Run, don't walk: Chasing higher FLOPS for faster neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12021–12031 (2023). https://doi.org/10.1109/CVPR52729.2023.01157

19. Chen, Y. et al. Accurate leukocyte detection based on Deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput. Biol. Med.* 170, 107917 (2024). https://doi.org/10.1016/j.compbiomed.2024.107917
20. Hu, J. et al. Improved detection algorithm of RT-DETR for UAV small target. *J. Comput. Eng. Appl.* **60**, 198–206 (2024). https://doi.org/10.3778/j.issn.1002-8331.2404-0114
21. Yao, T. et al. HGNet: Learning hierarchical geometry from points, edges, and surfaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21846–21855 (2023). https://doi.org/10.1109/CVPR52729.2023.02092
22. Zhou, B. et al. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319
23. Zhang, L. et al. LR-DETR: A lightweight real-time traffic sign detection model based on improved RT-DETR. *J. Real-Time Image Proc.* 22, 77 (2025). https://doi.org/10.1007/s11554-025-01659-1
24. Cui, S. & Deng, H. PMG-DETR: Fast convergence of DETR with position-sensitive multi-scale attention and grouped queries. *Pattern Anal. Appl.* 27, 58 (2024). https://doi.org/10.1007/s10044-024-01281-0
25. Hu, J. et al. RT-DETR-EVD: An emergency vehicle detection method based on improved RT-DETR. *Sensors* **25**, 3327 (2025). https://doi.org/10.3390/s25113327
26. Wang, N. et al. MRA-YOLOv8: A network enhancing feature extraction ability for photovoltaic cell defects. *Sensors* **25**, 1542 (2025). https://doi.org/10.3390/s25051542
27. Liu, Y. & Yang, Z. DDH-YOLO: A dual-head YOLOv8 model for small object detection in UAV aerial images. *Signal. Image Video Process.* **19**, 816 (2025). https://doi.org/10.1007/s11760-025-04429-5

Author contributions
Formal analysis, Xin Yan; Methodology, Xin Yan; Writing – original draft, Xin Yan; Experimental design and Validation, Saidong Yang; Writing – review & editing, Saidong Yang; Resources, Shixiong Zhang; Supervision, Shixiong Zhang; Writing – review & editing, Shixiong Zhang; Investigation, Xingchong Li; Methodology support, Ang Li.

Funding
No funding was received to assist with the preparation of this manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.