Measurement 242 (2025) 115753

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

MSIDetector: Detecting Multi-Scenario industrial defects using an adapted visual foundation model and dual thresholding discriminator

Xujie He [a], Jing Jin [a,*], Fujiang Yu [a], She Zhao [a], Duo Chen [a], Xiang Gao [b]

[a] Ubiquitous Computing and Intelligent Systems Research Group, School of Astronautics, Harbin Institute of Technology, Harbin 150001, Heilongjiang Province, China
[b] Institute for Artificial Intelligence, Harbin Institute of Technology, Harbin 150001, Heilongjiang Province, China

ARTICLE INFO

Keywords:
Defect detection
Open-set detector
Foundation model
Industrial feature adapter
Dual thresholding

ABSTRACT

Industrial defect detection is crucial for enhancing product quality. The diverse nature of industrial scenarios has posed challenges for developing a unified defect detection model that can address multiple industrial scenarios. To meet this challenge, we started from a data perspective, developing three self-designed data acquisition apparatuses to capture defect samples from multiple industrial scenarios. Leveraging grounding DINO (GDINO), an open-set detection approach, we propose a unified defect detection method, MSIDetector. By integrating an industrial feature adapter and a context-aware dual-thresholding defect discriminator, we successfully incorporate prior industrial knowledge into the model. The performance of the MSIDetector was evaluated across four industrial scenarios, including a public defect dataset, achieving a total state-of-the-art mAP@50 of 76.2, surpassing the performance of FasterRCNN, YOLO-7X, and the visual foundation model GDINO, with quantitative increases of 13.93, 6.55, and 2.1, respectively. This report represents the first successful attempt to apply foundation models to defect detection.

1. Introduction

Defect detection plays a pivotal role across industries, reducing production costs, and increasing customer satisfaction. For example, in automotive manufacturing technologies such as machine vision systems and X-ray [1] inspection are employed to detect component defects, thereby mitigating accident risks and meeting safety standards. Similarly, in food [2] and pharmaceutical manufacturing [3], techniques such as optical inspection and chromatography identify foreign objects, contaminants, and impurities. Thus, defect detection is a crucial element in ensuring product quality, contributing significantly to the success and sustainable development of enterprises.

Emerging artificial intelligence (AI) technologies are increasingly being integrated into the field of defect detection. Compared with traditional manual inspection, AI-based detection offers advantages such as high consistency, traceability, and cost-effectiveness. However, existing AI-based defect detection methods have yet to form a unified foundational model for defect detection, unlike other computer vision tasks such as object segmentation (SAM) [4], detection [5,6], and tracking [7]. We attribute this to three main factors:

- Scene Diversity Issue: The formation of a unified model for defect detection varies significantly due to the diverse nature of scenes. Unlike generic scenes, defects with similar appearances may have different terminologies in different scenes. We summarize two phenomena, 'one terminology, multiple shapes' and 'one shape, multiple industrial-specific terminologies', as shown in Fig. 1 (a-k), both common in industrial defect detection.
- Constraint of a singular threshold: Current defect detection approaches universally utilize a singular threshold for filtering detection. This approach has inherent limitations for effectively filtering detection boxes leading to the inadvertent exclusion of authentic defects, especially for detection by detectors with a transformer architecture, as depicted in Fig. 1 (l).
- Data Sample Shortage: Currently available defect detection datasets, particularly those in bounding box format, are relatively limited. This scarcity of data strongly hampers the development of a unified defect detection model..

In this study, we propose a unified model for industrial defect detection, MSIDetector. Building upon the open-set GDINO, we

* Corresponding author at: Room 502, Building 1#, Ubiquitous Computing and Intelligent Systems Research Group, School of Astronautics, Harbin Institute of Technology, No. 92, Xidazhi Avenue, Nangang District, Harbin, Heilongjiang Province, China.
E-mail addresses: hexujiee@126.com, hexujie@stu.hit.edu.cn (X. He), jinjinghit@hit.edu.cn (J. Jin), yfujiang@126.com (F. Yu), she_zhao123@163.com (S. Zhao), chenduoo@126.com (D. Chen), gaoxiang_hit@126.com (X. Gao).

https://doi.org/10.1016/j.measurement.2024.115753
Received 25 May 2024; Received in revised form 29 August 2024; Accepted 12 September 2024
Available online 14 September 2024

introduce a learnable industrial adapter to align industrial priors into GDINO, enabling adaptation to diverse defect detection tasks across multiple scenes. To overcome the limitations of existing single-thresholding filtration methods for defects, we employ a dual-thresholding strategy to hierarchically process detection results corresponding to different thresholds. To address the issue of data scarcity, we design three sets of different defect detection apparatuses. Using self-designed detection apparatuses, we collected a considerable number of defect samples and supplemented existing challenging public datasets such as NEU-DET [36], resulting in the Multi-Scenes Defect Detection dataset (MSDD). Extensive experimental results for these 4 industrial scenes demonstrate that the proposed MSIDetector has advanced defect detection capabilities across multiple industrial scenes, achieving a precision of 76.20 for the comprehensive metric of mAP@50, with improvements of 13.93, 6.55, and 2.1 over existing state-of-the-art methods: FasterRCNN integrated with FPN, YOLO7-X and the visual foundation model GDINO, respectively.

To summarize, this work's contributions include:

1) Building upon a foundation visual model, a unified defect detection model named MSIDetector is proposed, designed to address defect detection challenges across multiple industrial scenarios.
2) An industrial feature adapter is proposed to incorporate industrial priors into large visual models.
3) A context-aware dual-thresholding discriminator is proposed to overcome the detection limitations caused by a single threshold.
4) Three large-scale defect detection apparatuses have been independently developed. Valuable datasets suitable for industrial defect detection tasks have been collected and meticulously annotated. We plan to publicly release the collected data to encourage further research.
5) Extensive experiments on 4 industrial scenes were conducted, demonstrating the state-of-the-art performance of the proposed method.

The remainder of this paper is organized as follows. In Section 2, we review the literature on industrial defect detection and current foundation models. In Section 3, we provide the details of the MSIDetector. In Section 4, we report the experimental and comparison results and discuss them in Section 5. Finally, in Section 6, we draw conclusions and suggest possible future work.

2. Related work

The existing intelligent methods for defect detection can be categorized into anomaly detection [8,9] and defect detection based on their detection forms. For scene adaptability, these methods can be divided into single-scene defect localization [8,10,11] and the recently prevailing multiscene foundation methods [4–7].

Defect Detection: Currently, intelligent defect detection methods can be classified, based on the presentation format, into classification-based [12,13], segmentation-based [14,15], and bounding box-based defect detection methods [1,16]. Classification-based defect detection methods [12,13] are primarily used for initial screening of defects in inspected items, as some scenarios only require distinguishing between defect-free and defective products. However, when it is necessary to further locate

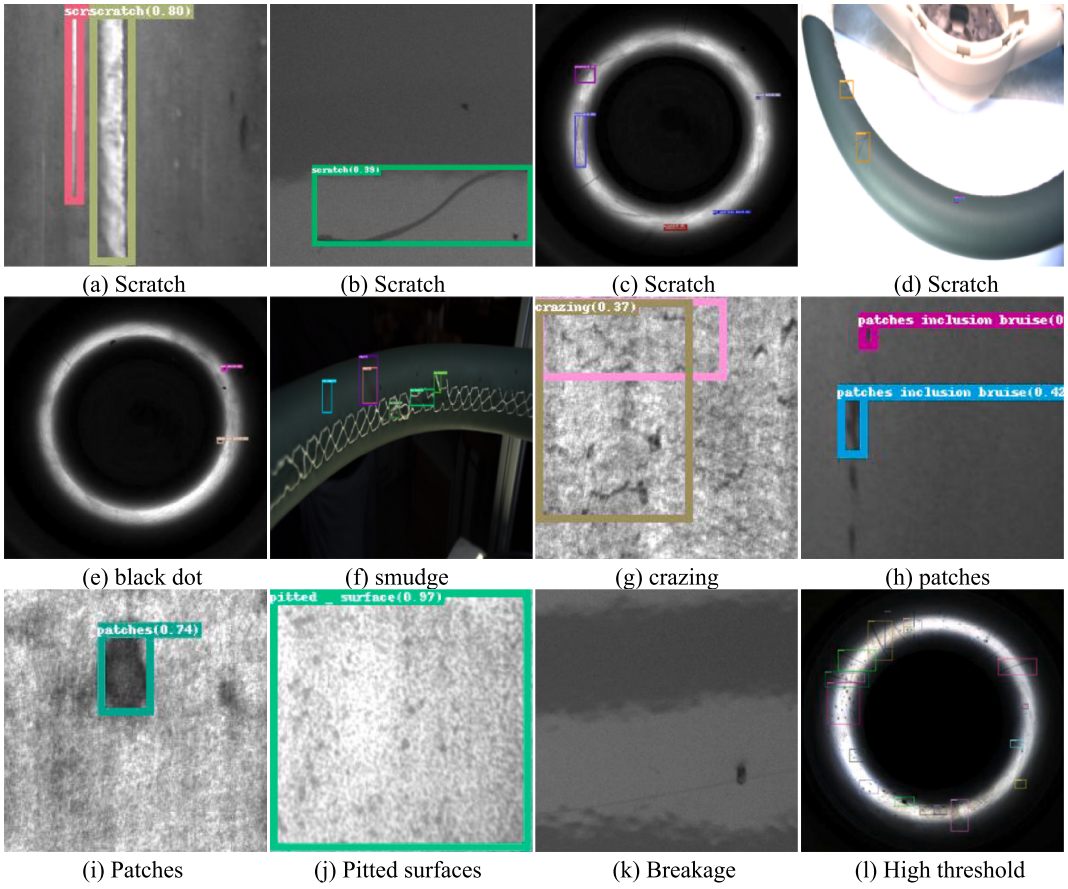


Fig. 1. Illustration of scene diversity issues. The images from (a) to (d) depict "scratches" captured from hot-rolled strips, vehicle bodies, cylinder bores, and steering wheels. This set of four images illustrates the phenomenon of 'one term, multiple shapes', wherein defects of the same category exhibit different forms due to varying scenes. The images from (e) to (k) also originate from the aforementioned four scenes, demonstrating the issue of 'one shape, multiple domain-specific terms' where the same defect appearance is labeled differently in different scenarios.

the specific positions and severities of defects, segmentation-based and bounding box-based defect detection methods should be prioritized. Segmentation-based defect detection methods [14,15] typically represent defects in a segmented form. Segmentation fundamentally involves a pixelwise classification problem, often requiring that the proportion of the defect in the image not be too small to significantly impact the performance. In comparison, bounding box-based detection methods [1,16–18] have a broader range of applications. The defect detection format using bounding boxes has less stringent requirements on defect characteristics than the previous two methods. It also provides information on the defect's location and severity. Commonly used bounding box-based defect detection methods can be further categorized into CNN-based [1,10,12–14], transformer-based [19,20], and, recently, popular hybrid structure defect detection models [21]. Moreover, according to defect searching approaches, methods can be further classified into anchor-based [1,16,21] and anchor-free [19,20] types.

Anomaly detection: Anomaly detection falls under a distinctive category within segmentation-based defect detection. Anomaly detection primarily employs data mining techniques to identify anomalous data regions deviating from the dataset distribution. The prevailing trend in anomaly detection favors the use of defect-free data for model training. Through unsupervised learning [22,23], test samples are reconstructed and compared to defect-free samples, aiming to detect anomalous regions. However, anomaly detection methods segment defects from images, making them suitable for scenarios where the defect area constitutes a relatively large proportion of the image [24] or when detection is performed on locally captured regions.

Visual Foundation Model: The aforementioned defect detection algorithms are all designed for specific scenarios, utilizing datasets tailored to given environments. With gradual advancements in visual foundational models, more researchers are exploring various downstream tasks. For example: 1) In medicine, following SAM [4], MedSAM [25] has been proposed for medical image segmentation. MedSAM is fine-tuned based on SAM using over 1 million (1 M) medical images. Similarly, SonoSAM [26] is used for ultrasound image segmentation, fine-tuned on a SAM with over 2 M ultrasound images and compressed through knowledge distillation. 2) In the remote sensing image domain, RSPrompter [27] extends SAM by integrating an object detection head as a prompter, while keeping SAM's backbone frozen during training to tailor it for remote sensing images. Similarly, SpectralGPT [28] is trained on ~ 1.4 M images based on MAE [29] for remote sensing images. 3) In the object tracking field, SAM-Track [7] uses the GDINO [6] to provide object locations for SAM combined with an existing multi-object tracker [30]. 4) In the anomaly detection domain, AnomalyGPT [31] starts from a large language model and is designed with corresponding image encoders and prompters to retain the generalizability of the language model, and is fine-tuned on specific anomaly detection datasets. Similarly, Myriad [32] utilizes existing anomaly detection algorithms to provide prior knowledge for a large language model (MiniGPT4 [33]) to derive an anomaly detection foundational model.

Leveraging GDINO's strong semantic understanding and multimodal learning capabilities, a truly end-to-end model, we choose to start with GDINO and make improvements accordingly. The novelty of this paper lies in the proposal of the first foundational universal model for defect detection based on GDINO, filling the current gap in universal foundational models for defect detection.

3. Methodology

3.1. Overall

The MSIDetector framework comprises three main components (Fig. 2): the defect initial detection unit (DID), the industrial feature adapter (IFA), and the context-aware dual thresholding defect discriminator (CaDTDD). For a given input image $I \in \mathbb{R}H \times W \times 3$ to be detected, it is separately fed into the image/text encoder of DID and the IFA. In the image/text encoder of DID, the input image $I \in \mathbb{R}H \times W \times 3$ is transformed into an initial feature embedding $FDID \in \mathbb{R}ha \times wa \times ca$. Since DID is trained on generic scenarios, its capability for industrial defect detection is limited. The input image $I \in \mathbb{R}H \times W \times 3$ is simultaneously fed into the IFA, as shown in Fig. 2. The adapter encoder of the IFA extracts industrial domain-specific features $Fa \in \mathbb{R}ha \times wa \times ca$ by encoding the input image $I \in \mathbb{R}H \times W \times 3$ and couples these features containing industrial domain-specific knowledge with the initial feature embedding $FDID \in \mathbb{R}ha \times wa \times ca$ obtained using DID through the aggregation module (AGG). The coupled features are then input into the decoder to decouple into defect locations $Bs, a \in \mathbb{R}n \times 4$ modulated by the IFA. These preliminary detection locations $Bs, a \in \mathbb{R}n \times 4$ are then passed to the CaDTDD. The detection outcomes are refined by adopting dual-threshold machanism to increase defect detection rates while

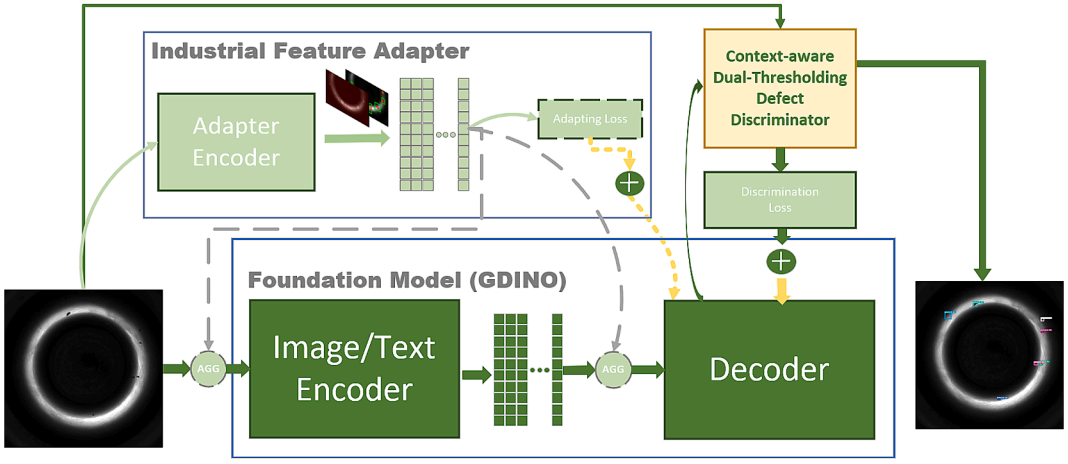


Fig. 2. Deployment of the MSIDetector. For a given industrial image to be inspected $I \in \mathbb{R}H \times W \times 3$, I is first input into the industrial feature adapter (IFA), which extracts corresponding industrial-adapted features $Fa \in \mathbb{R}ha \times wa \times ca$ containing industrial domain-specific knowledge. I is also input into the large visual detection model to obtain the initial feature embedding $FDID \in \mathbb{R}ha \times wa \times ca$. The industrial prior information contained in Fa is then introduced into FDID (at the AGG position in the figure). The defect detection results $Bs, a \in \mathbb{R}n \times 4$ modulated by the industrial adapter are further input into the context-aware dual thresholding defect discriminator (CaDTDD), and a dual-thresholding strategy is applied to filter the detection results while suppressing false-negatives to obtain the final defect detection results. Additionally, all optional parts are indicated with dashed lines in the figure. As illustrated, the aggregation locations (illustrated with AGG) for fusing FDID with Fa are located at the entrances of the image/text encoder and the decoder, and adapting loss is also optional; for more details, please refer to Section 4.4 and Section 4.6.

concurrently mitigating false-positives and false-negatives, ultimately achieving a final defect identification. DID uses an existing high-performance foundational detection model, grounding DINO (GDINO), which follows the encoder-decoder paradigm. A multiscale deformable attention mechanism is employed as the encoder as follows:

$$DeAttn\left(\widetilde{r}_p,\{X^s\}_{s=1}^4\right)=\sum_{h=1}^{H}\left[\sum_{s=1}^{4}\sum_{\partial=1}^{\Re}\mathbf{E}_{ref}\widetilde{W}^h\left(\omega\left(\widetilde{r}_p\right)\Big|_s+\Delta\widetilde{o}_{sh\partial}\right)X^s\right]\cdot W^h \tag{1}$$

where $\widetilde{r}_p\in\mathbb{R}2$ represents the reference point matrix, $\{X^s\}_{s=1}^4$ represents the features extracted from 4 scales after passing through the feature extractor, h denotes the multihead index, and $\mathbf{E}_{ref}$ denotes the attention weights. $\omega(\cdot)$ maps the normalized reference point coordinates to the corresponding layer through a mapping function, where $\Delta\widetilde{o}_{hs\partial}$ represents the offset of the s-th sampled point relative to the $\partial-$th reference point in the h-th attention head, and W is the transformation matrix. The multiscale deformable attention mechanism combines traditional attention mechanisms with multiscale processing and deformable sampling. It processes input features at multiple scales to adequately represent small and large defects. Using learned offsets, it adaptively samples key points from the input features instead of relying on a fixed grid. The model calculates attention weights for the sampled points to indicate the importance of each sampled feature. These weights aggregate the sampled features, generating a context-aware representation for each position in the output, effectively extracting defect features. Furthermore, $\widetilde{r}_p$ and $\mathbf{E}_{ref}$ are obtained through linear projection as follows:

$$\mathbf{E}_{ref}(\boldsymbol{Q},\boldsymbol{P})=Linear(\boldsymbol{Q}+\boldsymbol{P}) \tag{2}$$

where Q is the query matrix and P is the corresponding positional encoding. For the decoder part, a conventional multihead attention mechanism is used:

$$Multihead\_Attn\left(\{X^s\}_{s=1}^4\right)=\sum_{h=1}^{H}\left[\sum_{s=1}^{4}\sum\mathbf{E}_h\widetilde{W}^hX^s\right]\cdot W^h \tag{3}$$

where Eh adheres to:

$$\mathbf{E}_h\propto e\left\{\boldsymbol{Q}_h\cdot\boldsymbol{K}_h^T/\sqrt{d_k}\right\}v_h \tag{4}$$

In the following subsections, we elaborate on the remaining IFA and CaDTDD methods, newly proposed here.

3.2. Industrial feature adapter

To incorporate domain-specific knowledge into the visual foundational model for industrial scenarios, we propose the Industrial Feature Adapter (IFA), as shown in the upper part of Fig. 2. We show the specific structure of the adapter encoder in the Industrial Feature Adapter in Fig. 3. As shown, the adapter encoder primarily comprises a feature extractor, a projection layer, and a multiscale semantic aggregation component. Initially, we extract preliminary features FE-ISA from the input image $I\in\mathbb{R}H\times W\times3$ using the feature extractor. Next, the projection layer is employed to expand these features across different scales through cheap linear transformations, thereby enriching the feature information. Finally, we aggregate these multiscale enriched features using the aggregation layer. In the aggregation layer convolutional kernels with varying receptive fields are utilized to further extract semantic information at different scales from the FE-ISA. The feature with aggregated multiscale semantic information is ultimately upsampled by a factor of 4 to one-fourth the size of $I\in\mathbb{R}H\times W\times3$, resulting in the domain-specific industrial knowledge $Fa\in\mathbb{R}H/4\times W/4$ obtained by the industrial adapter. To adapt to defects of different shapes (such as scratches and rust), we employ deformable convolutions [34] for feature extraction. The deformable convolution is given by:

$$\Phi(i)=\sum_{l\in|\boldsymbol{K}|}\left[\boldsymbol{\kappa}(l)*\phi_f(i+l+\Delta\vartheta_l)\right] \tag{5}$$

where ϕf and Φ denote the input features and the output features, respectively, where i represents the feature map index, $\kappa\in\mathbb{R}k1\times k2$ is the convolution kernel, l denotes the position index of the convolution kernel, and Δϑl represents the positional offset. Deformable convolution is able to flexibly adjust the positions of the convolution kernels by introducing learnable offsets, particularly effective for capturing features of irregularly shaped defects. The offsets are learnable parameters,

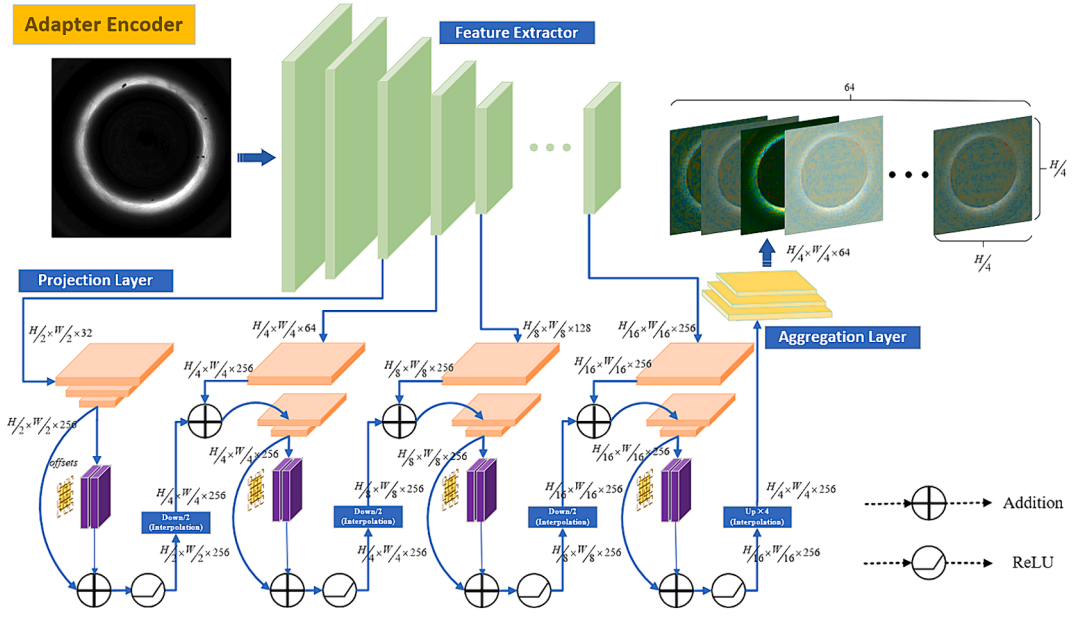


Fig. 3. Deployment of IFA. The input image is initially encoded through a feature extractor, and features extracted at different scales after encoding are projected into high-dimensional space using projection layers with different settings. Ultimately, features projected at different scales are fused via deformable convolution combined with up/downsampling, resulting in industrially adapted features 1/4 the size of the input image.

as shown in the equation representing the displacement of the convolution kernel center on the input feature map. Based on the calculated offsets, the sampling points of the convolution kernel are dynamically adjusted rather than remaining fixed on a regular grid, enhancing feature representation. The details of the IFA are listed in Table 1, and the projection layers in the table all follow the transformations given by

$$\mathscr{P}_j(\boldsymbol{X}) = GroupNorm\left(\widetilde{\mathscr{H}}_{1\times1}(\boldsymbol{X})\right) \tag{6}$$

where X indicates the input features, j indicates the index of the projection layer, and $\widetilde{\mathscr{H}}_{1\times1}$ is the traditional convolution with a kernel size of 1×1. The down/upsampling layers all obey the transformations given by:

$$\mathscr{S}_j(\boldsymbol{X}) = ReLU(BatchNorm(\Phi(\boldsymbol{X})) \tag{7}$$

here, we take features obtained via deformable convolutions as inputs to formalize different shapes of industrial defects.

3.3. Context-aware Dual-Thresholding defect discriminator

The introduction of CaDTDD aims to overcome the limitations of the current single-thresholding approach in defect detection. Single threshold is prone to considerable human bias: selecting a low threshold leads to numerous redundant detections and false-positives, while opting for a higher threshold may result in missing genuine defects. Consequently, an appropriate threshold requires extensive trial and error to suit specific defect detection scenarios. Inspired by [35], we employ a dual-thresholding strategy for defect detection tasks.

By setting dual thresholds, superior detection results can be achieved with minimal experimentation or even in a single attempt. The CaDTDD framework (Fig. 4), outlined in the pseudocode in Algorithm 1, introduces a discrimination branch from the decoder of GDINO to discriminate whether defects truly exist within the detected boxes (patches). The dual thresholds, denoted as $T = \{\tau L, \tau H\}$, consist of a low threshold τL and a high threshold τH. Defects with confidence greater than τH are directly stored in the memory bank (Bh, Ch) without any refinement:

$$(B_h, \boldsymbol{C}_h) = \{(x_1, y_1, x_2, y_2, c_i) | c_i > \tau_H, i = 1, \ldots, n_1\} \tag{8}$$

where (x1, y1) and (x2, y2) represent the top left and bottom right coordinates of the patches, respectively, and n1 is the total number of patches whose confidence is greater than τH. Moreover, defects with confidence falling between τL and τH are temporarily stored in the memory bank (Bl-1, Cl-1):

$$(B_{l-1}, \boldsymbol{C}_{l-1}) = \{(x_1, y_1, x_2, y_2, c_i) | \tau_L < c_i \leqslant \tau_H, i = 1, \ldots, n_2\} \tag{9}$$

where n2 indicates the number of defects with confidence falling between τL and τH. The discrimination operation is imposed on Bl-1, as illustrated in Algorithm 1, and the overall transformation can be expressed as:

Table 1
Details of the IFA structure. C, DC, GN, and BN refer to convolution, deformable convolution, group normalization, and batch normalization, respectively.

Module	Operation	Input Channel	Output Channel	Size
Feature Extractor	—	3	—	—
Projection Layer1	C1-GN	32	256	/2
Projection Layer2	C1-GN	64	256	/4
Projection Layer3	C1-GN	128	256	/8
Projection Layer4	C1-GN	256	256	/16
Downsample Layer1	DC3-BN-R	256	256	/4
Downsample Layer2	DC3-BN-R	256	256	/8
Downsample Layer3	DC3-BN-R	256	256	/16
Upsample Layer4	DC3-BN-R	256	256	/4
Aggregation Layer	DC3-BN-R	256	64	/4

$$\begin{cases} \left\{\ddot{\Upsilon}_i\right\}_{i=1}^{n_2} = \phi_2(\phi_1(\boldsymbol{I}, B_{l-1})) \\ \{\kappa_2^i\}_{i=1}^{n_2} = \sigma_4\left(\sigma_3\left(\left\{\ddot{\Upsilon}_i\right\}_{i=1}^{n_2}\right)\right) \end{cases} \tag{10}$$

where ϕ_1 and ϕ_2 represent ROI_Extractor and Patch_Alignment, respectively, performed at the image level and, where σ_3 and σ_4 represent the feature_Extractor and softmax functionalities, respectively, performed at the feature level. $\ddot{\Upsilon} \in \mathbb{R}n2 \times H1 \times W1 \times 3$ indicates all the defect patches aligned from Bl-1, and κ2 indicates the corresponding defect categories. The final results of CaDTDD are obtained by aggregating the defects belonging to the defect category in κ2 and the defects in Bh:

$$\text{CaDTDD} = (B_h, \boldsymbol{C}_h) \cup \{(b, \kappa_2(b)) | b \in B_{l-1} \; and \; \kappa_2(b) = 'defect'\} \tag{11}$$

3.4. Loss Penalty

The overall loss comprises three constituent components: the detection box loss, composed of regression and classification losses, and the CaDTDD loss (dis), specified as follows:

$$\begin{aligned} \mathscr{L} &= \lambda_{\text{reg}}\mathscr{L}_{\text{reg}} + \lambda_{\text{cls}}\mathscr{L}_{\text{cls}} + \lambda_{\text{dis}}\mathscr{L}_{\text{dis}} \\ &= \lambda_{\text{reg}}(\mathscr{L}_{GIOU} + \mathscr{L}_{L1}) + \lambda_{\text{cls}}\mathscr{L}_{\text{cls}} + \lambda_{\text{dis}}\mathscr{L}_{\text{dis}} \end{aligned} \tag{12}$$

where λreg, λcls, and λdis represent the balancing coefficients for the bounding box, bounding box category, and discriminator losses, respectively. The regression loss Lreg applied to the bounding boxes integrates the GIOU loss and L1 loss as follows:

$$\begin{aligned} \mathscr{L}_{\text{reg}} &= \sum_{n\in Q}\left[\kappa_{\text{giou}} GIOU\left(\hbar_n, \widetilde{\eta}_n\right) + \kappa_{L1}\left\|\hbar_n - \widetilde{\eta}_n\right\|_{L_1}\right] \\ &= \sum_{n\in Q}\left[\kappa_{\text{giou}}\left(1 - \left(\frac{\hbar_n \cap \widetilde{\eta}_n}{\hbar_n \cup \widetilde{\eta}_n} - \frac{\mathscr{C}\left(\hbar_n, \widetilde{\eta}_n\right) \backslash \hbar_n \cup \widetilde{\eta}_n}{\mathscr{C}\left(\hbar_n, \widetilde{\eta}_n\right)}\right)\right) + \kappa_{L1}\sum_{p=1}^{4}\left|\hbar_n^p - \widetilde{\eta}_n^p\right|\right] \end{aligned} \tag{13}$$

where $Q^{N\times4}$ denotes the set of detection boxes matched with their corresponding annotated boxes $\widetilde{Q}^{N\times4}$ in the current batch. We employ Hungarian matching as the matching criterion, where $\hbar_n \in Q^{N\times4}$ and $\widetilde{\eta}_n \in \widetilde{Q}^{N\times4}$ represent instances corresponding to the detection box and annotated box, respectively, and κgiou and κL1 denote the balancing coefficients for the respective losses. The classification loss Lcls is imposed on categories via binary focal loss and is defined as:

$$\mathscr{L}_{\text{cls}} = \sum_{i\in Q}\left[-\widetilde{\ell}_i(1-\ell_i)^{\upsilon}\log(\ell_i) - \left(1-\widetilde{\ell}_i\right)\ell_i^{\upsilon}\log(1-\ell_i)\right] \tag{14}$$

where $\widetilde{\ell}_i$ represents the category corresponding to the annotated box, ℓ_i represents the category corresponding to the predicted box, i denotes the index, and υ serves as a modulation coefficient. The loss Ldis applied to the CaDTDD module uses cross-entropy loss and is given by:

$$\mathscr{L}_{\text{dis}} = \frac{1}{|Q|}\sum_{i\in Q}\sum_{c=C_T} q_i(c)\log[z_i(c)] \tag{15}$$

where CT represents the set of category numbers corresponding to discrimination, z denotes the confidence of the predicted patches, and q is defined as:

$$q_i(c) = \begin{cases} 1, if \; \text{GT}(q_i) = c \\ 0, otherwise \end{cases} \tag{16}$$

where GT(·) represents the true category corresponding to the extracted patch.

Algorithm 1: CaDTDD

Input: $\boldsymbol{I}\in\mathbb{R}^{H\times W\times 3}$: Current Image to be detected
$\boldsymbol{\mathcal{B}}\in\mathbb{R}^{n\times 4}$: Defect locations to to ensured
$\boldsymbol{\mathcal{C}}\in\mathbb{R}^{n\times 1}$: Box confidence corresponding to $\boldsymbol{\mathcal{B}}$
$\boldsymbol{T}=(\tau_L, \tau_H)$: Dual threshold
δj: Threshold for discrimination
Output: $\boldsymbol{\Theta}\in\mathbb{R}^{k\times 4}$: Discriminated defect locations

Initialization: $\boldsymbol{\mathcal{B}}_h \leftarrow \phi$, $\boldsymbol{\mathcal{C}}_h \leftarrow \phi$, $\boldsymbol{\mathcal{B}}_{l\text{-}1} \leftarrow \phi$, $\boldsymbol{\mathcal{B}}_{l\text{-}2} \leftarrow \phi$, $\boldsymbol{\mathcal{C}}_l \leftarrow \phi$
for i in range(length($\boldsymbol{\mathcal{C}}$)) **do**
  / * classify all detected defects using dual threshold */
  if $\boldsymbol{\mathcal{C}}[i] > \tau_H$ **then**
    / * Directly retain defect boxes above τ_H */
    $\boldsymbol{\mathcal{B}}_h \leftarrow \{\boldsymbol{\mathcal{B}}[i]\} \cup \boldsymbol{\mathcal{B}}_h$
    $\boldsymbol{\mathcal{C}}_h \leftarrow \{\boldsymbol{\mathcal{C}}[i]\} \cup \boldsymbol{\mathcal{C}}_h$
  end if
  elif $\tau_L < \boldsymbol{\mathcal{C}}[i] < \tau_H$ **then**
    / * Temporarily store the boxes with confidences falling between τ_L and τ_H into $\boldsymbol{\mathcal{B}}_{l\text{-}1}$ */
    $\boldsymbol{\mathcal{B}}_{l\text{-}1} \leftarrow \{\boldsymbol{\mathcal{B}}[i]\} \cup \boldsymbol{\mathcal{B}}_{l\text{-}1}$
  end elif
end for
/ * Discrimination on boxes in $\boldsymbol{\mathcal{B}}_{l\text{-}1}$ */
$\ddot{\boldsymbol{Y}} \leftarrow \phi$
for i in range(length($\boldsymbol{\mathcal{B}}_{l\text{-}1}$)) **do**
  $\rho_1 \leftarrow \boldsymbol{\mathcal{B}}_{l\text{-}1}(i)$
  $\rho_2 \leftarrow$ ROI_Extractor($\boldsymbol{I}$, ρ_1)
  $\rho_3 \leftarrow$ Patch_Alignment (ρ_2)
  $\ddot{\boldsymbol{Y}} \leftarrow \ddot{\boldsymbol{Y}} \cup$ {Patch_Augmentation (ρ_3)}
end for
$\kappa_1 \leftarrow$ Feature_Extractor($\ddot{\boldsymbol{Y}}$).flatten()$\in\mathbb{R}^{n2\times 2}$
$\kappa_2 = softmax(\kappa_1).dim(1) = \frac{e^{m(j)}}{\sum_{z=1}^{N} e^{\kappa(z)}}$
for j in range(n_2) **do**
  if $\kappa_2[j] > \delta j$ **then**
    $\boldsymbol{\mathcal{C}}_l \leftarrow \boldsymbol{\mathcal{C}}_l \cup \{\kappa_2[j]\}$
    $\boldsymbol{\mathcal{B}}_{l\text{-}2} \leftarrow \boldsymbol{\mathcal{B}}_{l\text{-}2} \cup \{\boldsymbol{\mathcal{B}}_{l\text{-}1}[j]\}$
  end if
end for
/ * Defect Rectification Unit */
$\boldsymbol{\Theta}_c \leftarrow \phi$, $\boldsymbol{\Theta} \leftarrow \phi$
$\boldsymbol{\Theta}_c \leftarrow \boldsymbol{\mathcal{B}}_h \cup \boldsymbol{\mathcal{B}}_{l\text{-}2}$
$\boldsymbol{\Theta}$ = Non-Maximum_Suppression($\boldsymbol{\Theta}_c$, mode=GIOU)$\in\mathbb{R}^{k\times 4}$
Return: $\boldsymbol{\Theta}$

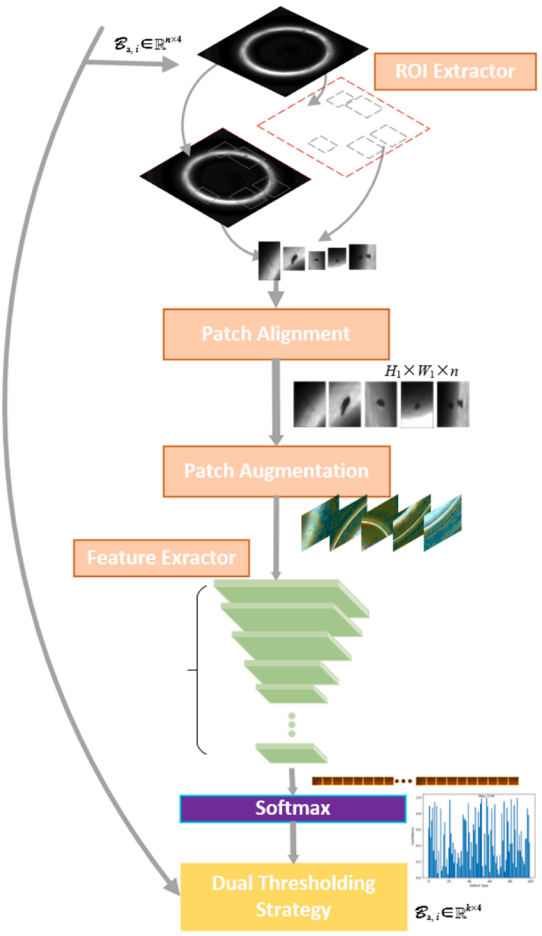


Fig. 4. Illustration of CaDTDD. As illustrated, ROI_Extractor primarily extracts target patches from the image. Patch alignment is mainly responsible for scale normalization of the extracted patches to ensure parallel processing. Patch augmentation involves a series of enhancement operations on the pixel values of the extracted patches to improve their generalizability. The dual thresholding strategy is shown on the left as Algorithm 1.

4. Deployment and experiments

4.1. Dataset

Due to the scarcity of publicly accessible datasets for boundary box-based defect detection, the NEU-DET dataset [36] for hot-rolled strip steel materials released by Northeastern University is commonly used. To address this issue, our team designed and deployed three sets of intelligent quality inspection equipment and collected corresponding data: engine cylinder bore state detection equipment (PX1), steering wheel wrapping area quality inspection equipment (SW1), and car body paint surface quality inspection equipment (CB1). Each set comprises four components: an imaging system, a control system, a central processing system, and a display unit.

The selection of the camera especially for PX1 equipment mainly considers the need to meet the detection requirements for workpiece diameters ranging from approximately φ20 mm. Following a series of selection experiments a camera with a large sensor size was chosen to achieve richer and clearer flat imaging. The main differences among the three sets of equipment are:

1) PX1:

Equipment Introduction: PX1 is designed for capturing images of the inner walls of cylinder bores, placing greater demands on the lighting and camera parts. A customized lens integrated with a light source and a reflector was ultimately selected for PX1, achieving lighting and imaging at the same time and catering to the photography needs of different inner diameter workpieces. A customized LED light source integrated into the lens is utilized for illumination. The customized lens design diagram is unveiled in Fig. 5. Given that the spherical reflector has a uniform curvature, it can reflect light evenly, reducing scattering and distortion. Additionally, spherical mirrors generally have a longer lifespan and lower maintenance requirements, making them more suitable for commercialization. Therefore, we ultimately decided to use a spherical reflector to achieve imaging that reflects a 360° view of the bore interior to the camera. An overall diagram of the PX1 equipment is shown in Fig. 6.

PX1 dataset: Over a span of 6 months, we meticulously gathered nearly 200 diverse instances of engine cylinder bore defects using our PX1, amassing a total of 1,119 images. These images encompass a spectrum of anomalies, including Crack, Sandhole, Bump, Paint flaking, Oil stain, and Uneven surface, totaling 8,379 exemplary defect instances across 6 categories. The data collection originated from common car models on the market, such as Toyota, Volkswagen, Audi, BMW, Mercedes-Benz, Chevrolet, Subaru, and others utilizing cylinder bore types, including inline 4, V6, V8, flat 4, and single cylinders. We named this dataset HIT-EngDV3 (Harbin Institute of Technology Engine Dataset):

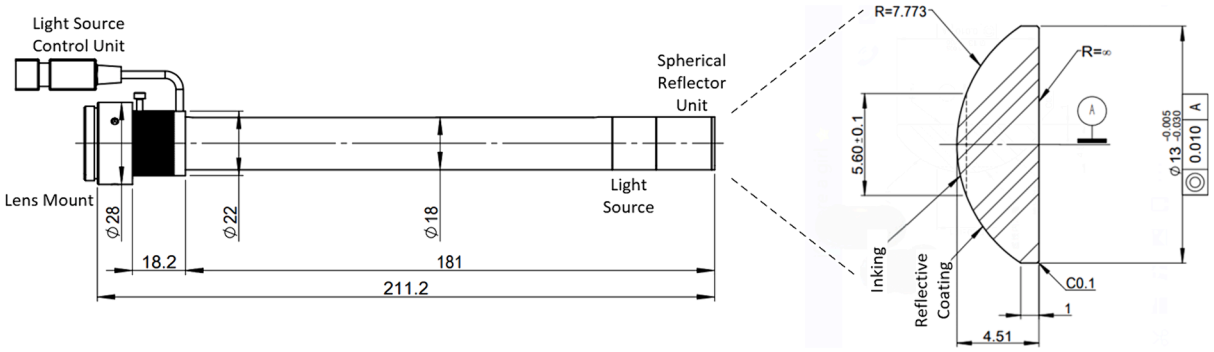


Fig. 5. Lens design schematics.

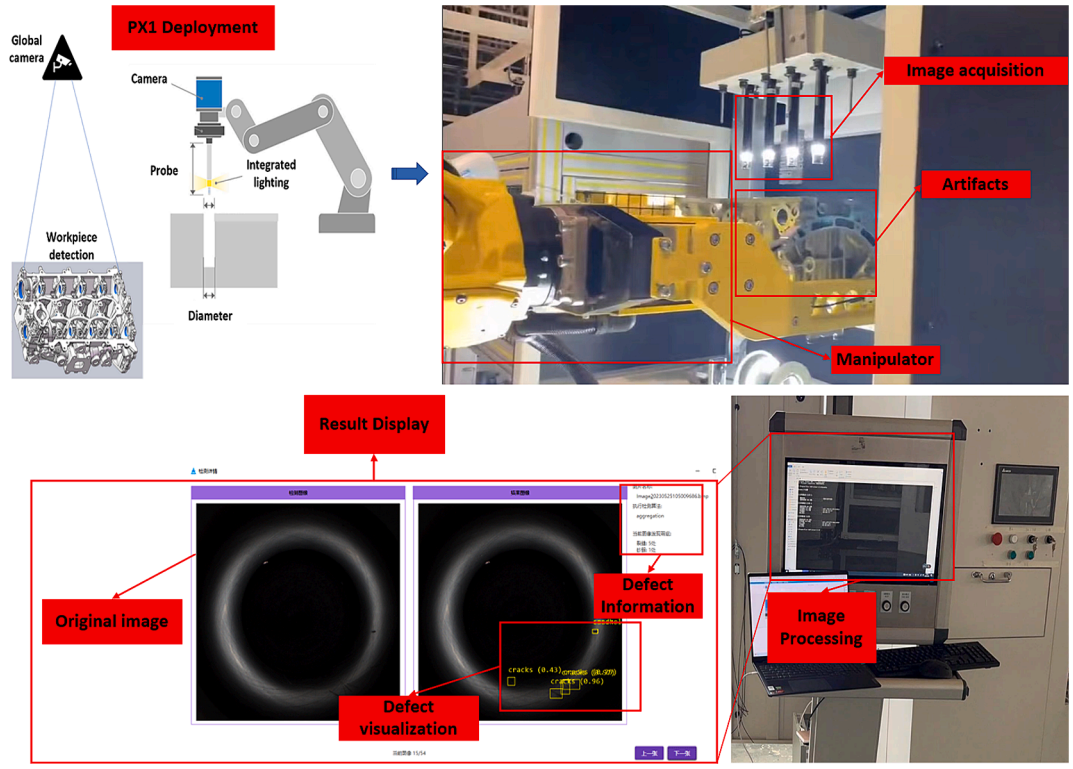


Fig. 6. Deployment of Engine Cylinder Bore State Detection Equipment (PX1).

- The images are captured at an ultrahigh resolution of 2000 × 2000 pixels. The sizes of the defects vary significantly, ranging from occupying half of the entire image to only a few pixels (<10 pixels).
- Oil stains splattered on the production line and improper storage methods lead to the appearance of dense rust and oil stains in patches.
- Differences in cylinder bore color result in large variations in lighting brightness after reflection, which obscures defect judgment.

Therefore, experiments on HIT-EngDV3 are challenging. Details of the HIT-EngDV3 dataset are provided in Table 2, and sample images are illustrated in Fig. 13 (a).

Table 2
Details of HIT-EngDV3.

Defect Category	Crack	Sandhole	Bump	Paint flaking	Rough area	Oil	All
Amount	4,642	2,282	293	561	24	577	8,379

2) SW1:

Equipment Introduction: For the steering wheel wrapping area defect detection equipment we opted for the highly precise Hikvision MV-CS050-10GC-PRO as the final capturing camera. A programmable logic controller (PLC) serves as the central control unit, regulating the elevation and rotation of the electric cylinder through output signals. SW1 utilizes a total of 4 cameras for data acquisition. These cameras capture the upper side A, lower side B, and inner side C of the steering wheel (Fig. 7). The inner side C is captured by two individual cameras to compensate for any blind spots caused by a single camera setup. The electric cylinder rotates the steering wheel by 60 degrees during each operation, with 6 images captured for each side A and B of the steering wheel and 8 images captured for side C, totaling 20 images per steering wheel. Position limit sensors control the elevation height of the electric cylinder to prevent any out-of-view issues. For illumination, a linear light source is employed to illuminate the darkroom, and a coaxial light source is installed on the camera capturing side B to provide supplementary lighting. The SW1 configuration is illustrated in Fig. 8.

Fig. 7. Explanation of Steering Wheel Positions. The yellow region in (a) indicates the B-side of the steering wheel, the red region in (b) signifies the A-side of the steering wheel, and the blue region in (c) represents the C-side of the steering wheel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

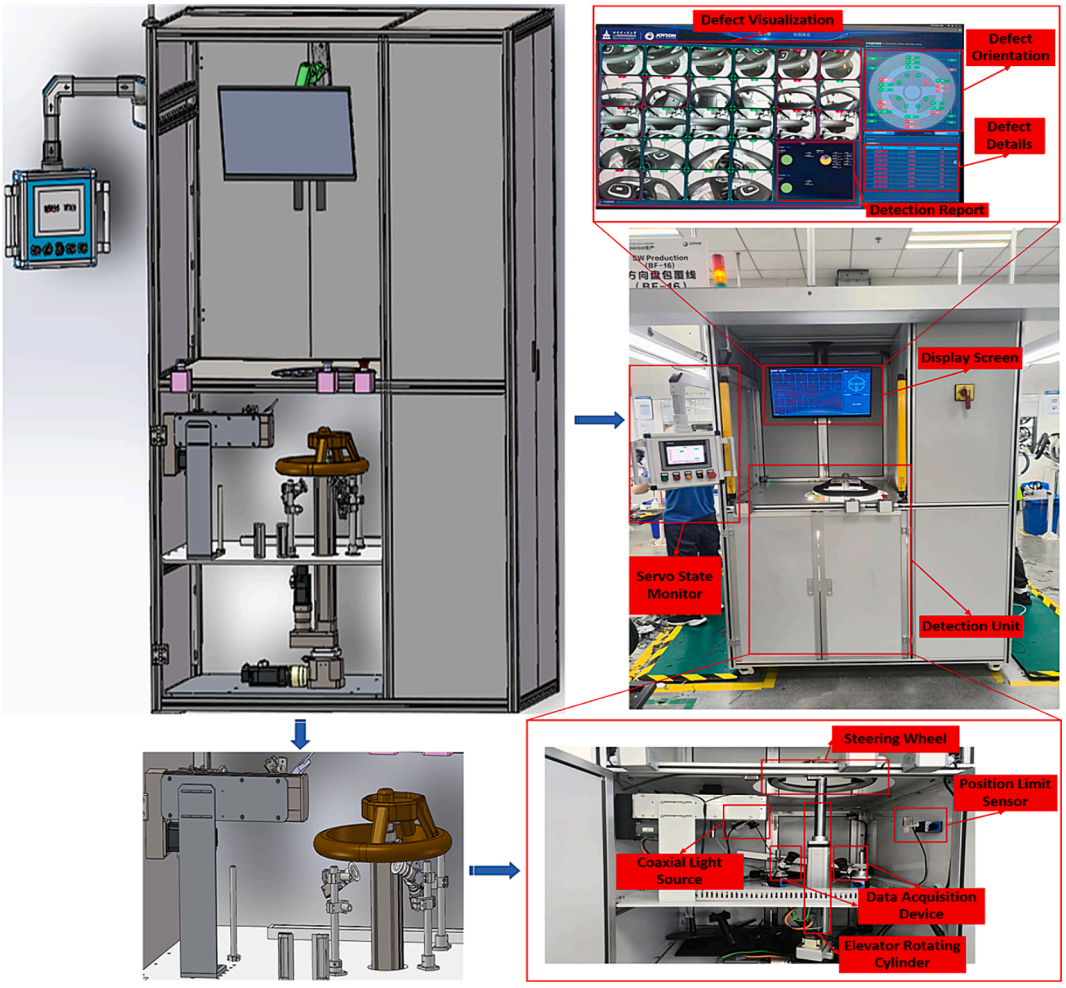


Fig. 8. Deployment of Steering Wheel Wrapping Area Quality Inspection Equipment (SW1).

SW1 dataset: Over a period of 7 months, we collected a total of 1,946 images of steering wheel defects using our SW1 from a steering wheel manufacturer named Joyson.[1] The data were gathered from commonly used car models on the market including BYD, Audi, Great Wall, Volkswagen, and Mercedes-Benz. The collected dataset is named HIT-SWD (Harbin Institute of Technology Steering Wheel Dataset), encompassing 13 typical defect categories including Leather tear, Loose thread, Smudge, Pinhole, Needle exposed, Branding, Leather uneven, Leather wrinkle, Leather dent, Marks, Skipping stitch, Foaming, and Scratch, for a total of 10,024 defect samples. Specifically:

- Defects such as pinholes and surface damage (leather flipping) are relatively small, occupying only a few pixels at ultrahigh resolution (2000 × 2000).
- The scenes are complex and contain a significant amount of background information.
- Steering wheels exhibit various shapes, including circular, elliptical, rectangular, and square shapes.
- There is frequent color variation in the steering wheels, some even appearing as patches (different colors on sides A and B).

Therefore, experiments on the HIT-SWD dataset present significant challenges. Detailed information about the HIT-SWD dataset is provided in Table 3. Sample images from the HIT-SWD dataset are illustrated in

[1] https://www.joysonsafety.com/.

Table 3
Details of the HIT-SWD.

Defect Category	Leather tear	Loose thread	Smudge
Amount	55 + 650	1060	3835
Defect Category	Pinhole	Needle exposed	Branding
Amount	495	406	177
Defect Category	Leather uneven	Leather wrinkle	Leather dent
Amount	24	416	253
Defect Category	Marks	Skipping stitch	Foaming
Amount	47 + 889	808	24
Defect Category	Scratch	All	
Amount	867	10,024	

Fig. 13 (b).

3) CB1

Equipment Introduction: The most notable feature of CB1 is the image acquisition apparatus. In the actual process we employ fixed car positioning and perform data acquisition by controlling different manipulators (Fig. 9). We use four manipulators at locations 1, 5, 7, and 11 on the vehicle, each equipped with an external probe. The external probe (Fig. 9) consists of a light source emitter, 4 lenses, and 4 central processing units. To eliminate noise introduced during image acquisition the light source emitter emits structured light in nine different patterns. Among these, patterns are sinusoidal structured lights with a phase discrepancy of 90 degrees (Fig. 10 (a)), given by:

$$\Xi(x) = \Omega\left[\sin\left(\frac{2\pi}{T^*}x + \Delta\delta_k\right) + 1\right], n = \{1, 2, 3, \ldots\} \in \boldsymbol{N}^* \tag{17}$$

where Ω is set to 128, representing half of the maximum grayscale value, while T* represents the modulation period. The remaining 5 patterns take the form of various gray code structured lights, which adhere to the following:

$$\widetilde{G}_i(x) = \Im_{i-1}(x) \oplus \Im_i(x), \Im = \left\{\widetilde{\Im}_1, \widetilde{\Im}_2, \ldots \widetilde{\Im}_8\right\} \tag{18}$$

using binary codes $\Im_i(x)$ to obtain Gray codes $\widetilde{G}_i(x)$. Finally, the images captured under the nine different light sources at the same position are fused to generate the ultimate image used for subsequent detection.

CB1 dataset: We collected a total of 2,472 images of car body paint defects over a period of 5 months using our CB1 equipment at the Shimada Big Bird[2] Manufacturer in Harbin, China. The data were collected from commonly used car models on the current market such as BYD, the Great Wall, and Volkswagen. The dataset covers a total of 16,758 defect samples including Breakage, Inclusion, Scratch, Condensate, Crater, Run, and Bulge, totaling 7 typical defect categories. We named this dataset HIT-CBD (Harbin Institute of Technology Car Body Dataset). Specifically:

- The defect sizes of the HIT-CBD dataset also vary widely, ranging from thousands of pixels to only a few pixels.
- A car body paint surface is reflective, affecting detection after imaging.
- The contrast between defects and background is relatively low.
- The noise present at the edges can affect the defects.

Detailed information on the HIT-SWD dataset is shown in Table 4, and sample images from the HIT-SWD dataset in Fig. 13 (c).

4) NEU-DET:

To further validate the effectiveness of the proposed method, we introduce a challenging publicly available dataset termed the Northeastern University Detection dataset (NEU-DET) [36]. NEU-DET consists of 6 types of surface defects found in hot-rolled strip steel, namely, rolled scale (RS), patch (Pa), chromium streak (Cr), pitted surface (PS), inclusion (In), and scratch (Sc) defects. The dataset comprises 1,800 grayscale images with 300 samples for each defect category. Each image is 200 × 200 pixels in size. This dataset is currently widely used and presents significant challenges for defect detection. Detailed information about the dataset is provided in Table 5, and sample images are illustrated in Fig. 13 (d).

Fig. 11 (a) further illustrates the geographical distribution associated with the three current quality inspection devices (HIT-PXD, HIT-SWD, and HIT-CBD). As depicted, the development of this equipment and data acquisition are currently being coordinated across 9 provinces and cities in China. This collaborative effort has led to the gradual application of developed equipment, projected to expand further. The dataset collected from the three industrial scenarios comprises a total of 24 distinct defect categories (Fig. 11 (b)). To reduce redundancy and enhance the readability of the paper we amalgamate the data from the four industrial scenarios into a unified dataset, Multiscenes Defect Detection dataset (MSDD). Henceforth, unless specified, all references to the dataset in subsequent sections denote the MSDD.

Fig. 12 further visualizes the data traits of defects in the four datasets mentioned above. Size-related data have been normalized in all figures. Columns one through four in Fig. 12 represent HIT-PXD, HIT-SWD, HIT-CBD, and NEU-DET, respectively. (a) Group visualizes all defects under each industrial scenario in a single graph, with the horizontal and vertical axes representing the coordinates of the defect center (cx, cy). The defective regions in the engine cylinder bores are mainly concentrated around the holes, consistent with actual observations. Defects in steering wheels predominantly appear in the rectangular region of the graph, with fewer defects at the top and bottom, primarily representing background information, aligning with real-world observations. Defects in the car body paint scene are distributed throughout various positions in the collection area. NEU-DET also exhibited defects distributed across various areas in the collection graph. (b) Group visualizes statistical analysis of defect sizes, with the horizontal and vertical axes representing defect height and width, respectively. The defect sizes in our self-collected data (first three columns) are relatively small and are primarily distributed in the bottom-left corner of the graph. Groups (c) to (f) visualize the data in a manner consistent with (a) and (b). Fig. 13 shows different sample instances of data under each industrial scenario.

4.2. Evaluation metrics and training Protocol

Evaluation metric: The evaluation metrics used in our experiments include precision, recall and comprehensive mAP, given by:

$$precision = \frac{TP}{TP + FP} \tag{19}$$

$$recall = \frac{TP}{TP + FN} \tag{20}$$

$$\text{mAP} = \frac{\sum_{i=1}^{N} \text{AP}_i}{N} = \frac{\sum_{i=1}^{N} \int_0^1 P_i(R_i) d(R_i)}{N} \tag{21}$$

where TP, FP, and FN are true positives, false-positives and false-negatives, respectively. R in Eq. (21) denotes the recall, and P(R) indicates the curves formed by the metric precision and recall. The integral intervals are acquired by setting different thresholds using IOU similarity from 0 to 1. N indicates the number of defect classes.

Training protocols: All of the models are fine-tuned under 100 epochs. The learning rate is set to 1e-4 and decays to 1e-5 at 75 epochs. The batch size is set to 32 with GPU memory (Nvidia 3090/24G×8). The models are all pretrained with the COCO dataset and are fine-tuned

[2] https://www.sbi-shimada.com/.

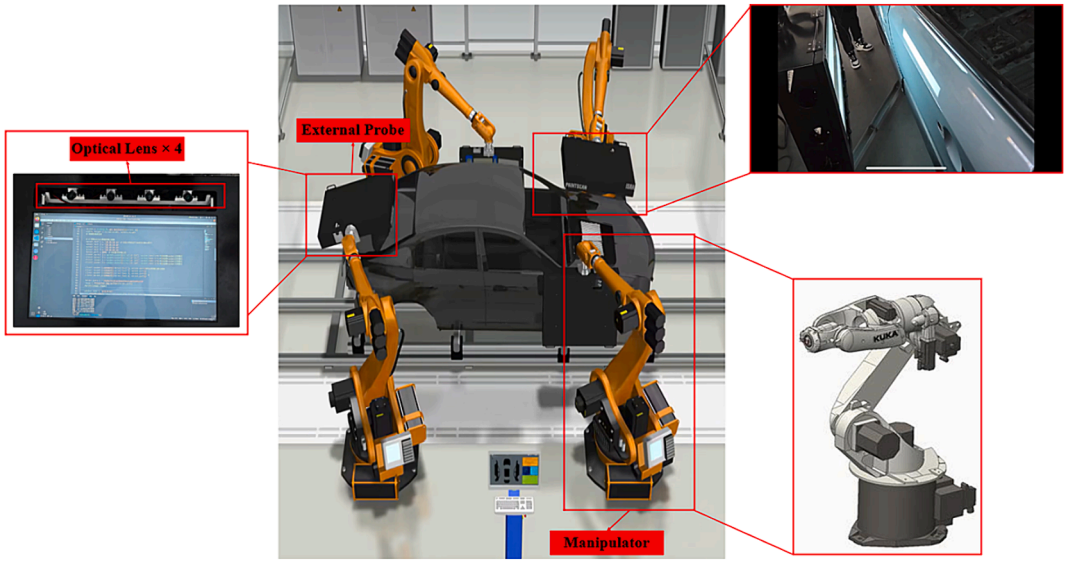


Fig. 9. On-site deployment diagram of the car body paint surface quality inspection equipment (CB1).

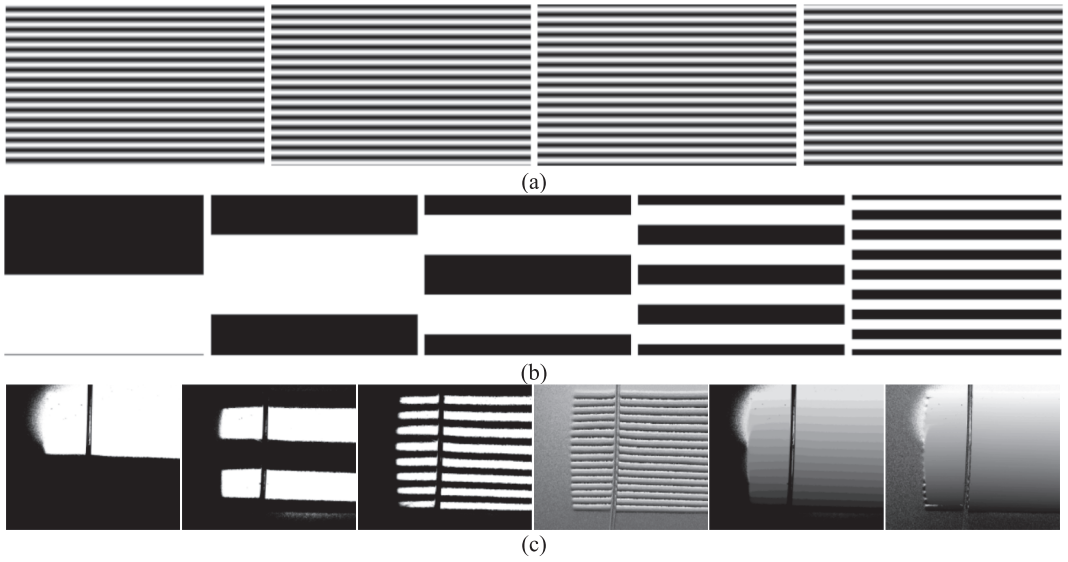


Fig. 10. Explanation of Structured Light and Image Acquisition. (a) Four sinusoidal structured light patterns with a phase difference of 90 degrees. (b) Five different Gray Code structured light patterns used during the experimental data acquisition. (c) Specific captured images where the first three images are samples directly captured after applying gray code structured light. The fourth image is a sample obtained after merging four images captured with sinusoidal structured light, the fifth image is a sample obtained after merging five images captured with gray code, and the last image represents the final sample image used for subsequent defect detection, obtained after merging nine images of different structured light forms.

Table 4
Details of the HIT-CBD dataset.

Defect Category	Breakage	Inclusion	Scratch	Condensate
Amount	4,642	2,282	293	561
Defect Category	Crater	Run	Bulge	All
Amount	24	577	8,379	16,758

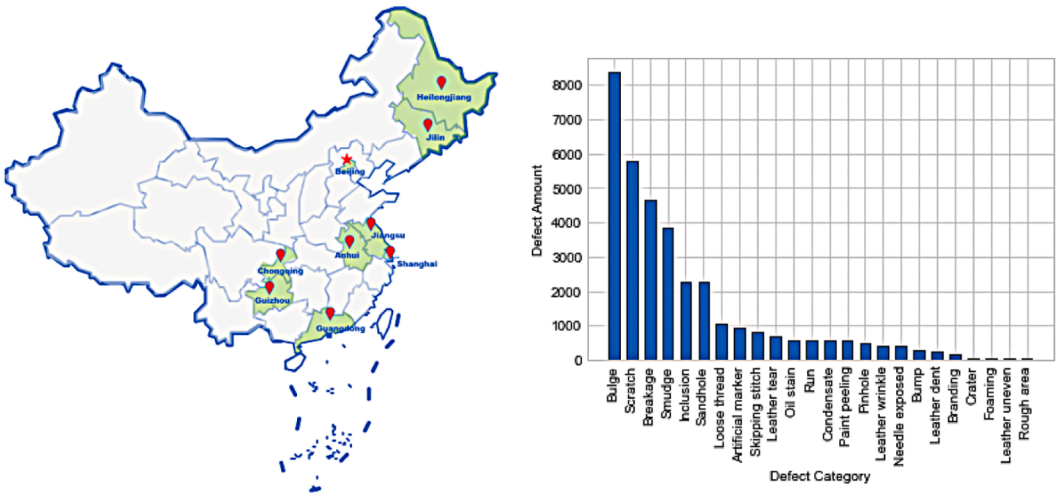
using our 4 industrial datasets via multiscale and warming-up training strategies. Additionally, all defect detectors incorporating large visual models, such as GDINO, were trained by freezing the text module, such as BERT [37].

4.3. Impact of defect terminology on defect detection results

Experimental Design: We designed a set of comparative experiments examining the influence of defect categories on their results. The

Table 5
Details of NEU-DET.

Defect Category	Crazing	Patches	Inclusion	Pitted surface	Rolled in scale	Scratches	All
Amount	689	881	1,011	432	628	548	4,189



(a) Dataset acquisition locations (b) Specific category details of the dataset

Fig. 11. Geographic Distribution of Dataset Collection Sites and Defect Category Information.

experiments were divided into two groups: one group used all current defect categories for experimentation, while the other group conducted experiments under the 'defect' category. All parameter configurations used during the experiment remained consistent with the previous training protocols. To eliminate the influence of data, the models selected for the experiment were all based on the transformer architecture, including DINO, GDINO, and the MSIDetector proposed in this paper. The experimental results are shown in Table 6.

Experimental Results: As shown in Table 6, 1) overall, the newly proposed MSIDetector model is more suitable for defect detection tasks across multiple scenarios, yielding the best comprehensive mAP@50 of 76.20. 2) Defect terminology has a significant impact on the final detection results. Specifically, when the defect terminology is not further subdivided, the comprehensive detection capability of the GDINO model, measured using mAP@50, increases from 57.0 to 74.3, a quantitative increase of 17.3. The stricter metric mAP@50:95 also increased from 27.8 to 33.9, an absolute increase of 6.1. Similarly, the recalls indicated by the ARS and ARM methods increase from 44.3 and 50.8 to 45.5 and 56.7, respectively, increases of 1.2 and 5.9, respectively. 3) The 24C experiment reaffirms GDINO's superior defect recall capability over DINO, with GDINO achieving higher ARS, ARM, and ARL values by 7.2, 7.6, and 8.3, respectively.

4.4. Ablation study of the fusion method

We designed a set of ablation experiments concerning fusion methods, whereby different fusion approaches were applied to the MSIDetector to attain industrial-adapted features consistent with those described in Fig. 2. Fig. 14 provides a detailed expansion of the AGG module in Fig. 2. In this section, we employed four different fusion models in our experiments, represented by ①, ②, ③, and ④ in the figure. These correspond to direct additive fusion (①), additive fusion combined with a feed-forward neural network (FFN) (②), concatenation aggregation combined with an FFN (③), and multisource feature alignment using self-attention combined with a cross-attention mechanism (④). Specifically, we employed elementwise addition for additive fusion. The dashed boxes on both sides of Fig. 14 illustrate the specific deployment details of each fusion mode.

The results are presented in Table 7, where we further supplemented the results under different supervisory modes. The model used in this section is named FU, with the fusion mode occurring at the decoder entry and without the CaDTDD module. The feature extractor used in the industrial adapter is PVT2-Lite [38]. For the supervised mode we used focal loss as the adaptive loss in Fig. 2 to penalize the prior features:

$$L_{fl} = \begin{cases} -(1-\alpha)p_t^{\gamma}\log(1-p_t) & \text{if } y=0 \\ -\alpha(1-p_t)^{\gamma}\log(p_t) & \text{if } y=1 \end{cases} \tag{22}$$

where pt is the probability predicted by the model, α is the balancing factor used to balance the influence of positive and negative samples, γ is the modulation factor used to adjust the loss contribution of easily classified samples, and y is the annotation label. In this experiment, we set α to 1 and γ to 2.

The ADD fusion mode (①) performs optimally in unsupervised mode for domain-specific feature adaptation (Table 7). Overall, the unsupervised mode for industrial-adapted features outperforms the supervised mode across various fusion methods. 1) Without any supervision of the defect-adapted features, when comparing FU1 under u-S and S, the overall detection accuracy, measured by mAP@50, increases from 72.4 under soft supervision to 74.2 without supervision, an increase of 1.8. 2) The stringent comprehensive accuracies indicated by mAP@50:95 and mAP@75 also increase from 32.7 and 24.7 under supervision to 33.9 and 26.4 without supervision, respectively, with increases of 1.2 and 1.7, respectively.

4.5. Ablation study on parameter sharing

To investigate the impact of model capacity on training visual foundation models, we designed a set of experiments regarding model parameters, with two sets of approaches. The first set uses the feature extractor provided by the industrial adapter for feature extraction, while the second set shares the backbone network and the feature extractor of the adapter for feature extraction, directly using the features extracted by the backbone as the input features for the industrial adapter. The feature extractor used in the first set is still the lightweight PVT2-Lite of the transformer paradigm, with experimental parameters consistent with those in the previous sections. The model used in this section is PA, with the fusion of the ADD mode located at the entrance of the decoder without supervision as previously described. The results are shown in Table 8. Notably, when parameter sharing is used, the performance of the model is further improved. Quantitatively, 6 out of the 9 metrics are improved when parameter sharing is employed, especially in terms of recall, where the recalls for small, medium, and large defects are all enhanced after parameter sharing.

4.6. Ablation study of adapting location

Experimental Design: A further set of ablation experiments on

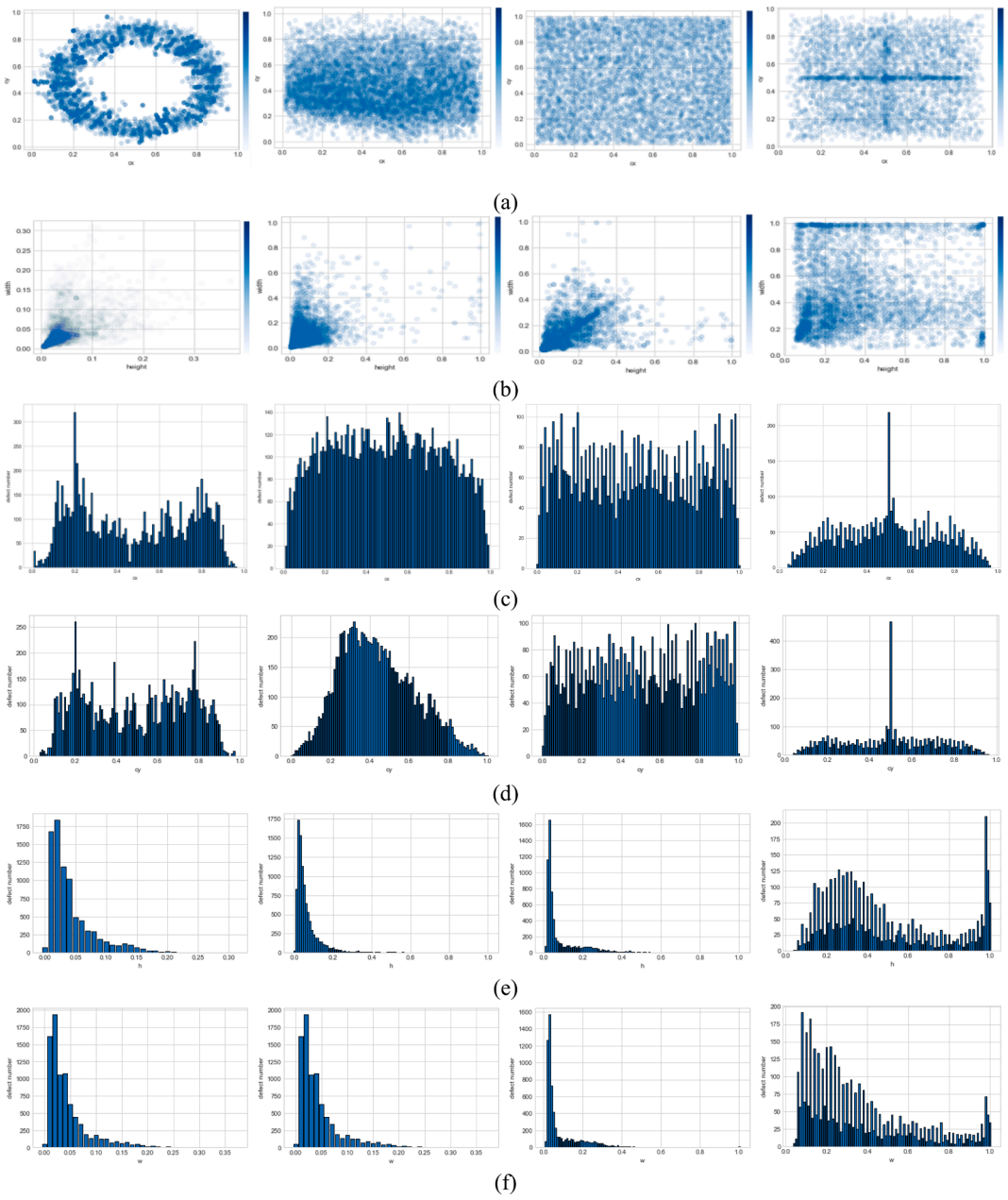


Fig. 12. Defect trait information in different scenes. (a) The distribution of defects in the respective dataset. (b) The number of defects in the dataset. (c) The number of defects contained at each x-coordinate position in the graph. (d) The number of defects contained at each y-coordinate position. (d) The number of defects contained at different height stamps. (f) The number of defects contained at different width stamps.

Table 6
Impact of Defect Terminology on Defect Detection Results. 'C' denotes 'class', and 24 classes represent the merged categories obtained by consolidating four datasets into one. '1C' signifies experiments conducted after amalgamating all categories into a single 'defect' class. The best results yielded by MSIDetector are reported in red, bold font.

Model		mAP @50:95	mAP @50	mAP @75	APS @50:95	APM @50:95	APL @50:95	ARS @50:95	ARM @50:95	ARL @50:95
24C	DINO	27.5	57.1	22.5	20.5	28.9	35.1	37.1	43.2	53.7
	GDINO	27.8	57.0	22.6	23.1	27.7	34.6	44.3	50.8	62.0
1	GDINO	33.9	74.3	26.1	28.8	38.5	37.1	45.5	56.7	57.4
C	MSIDetector	34.7	76.2	26.7	28.8	39.2	38.8	42.7	53.5	54.9

adapting locations is designed to investigate the impact of the adapting locations on the final performance of defect detection models. Three different locations were identified for introducing industrial-adapted features: the first approach introduces the industrial-adapted features before feature encoding; the second approach introduces them before feature decoding, and the third approach introduces them simultaneously before encoding and decoding. All industrial-adapted features follow an unsupervised approach with parameter sharing settings.

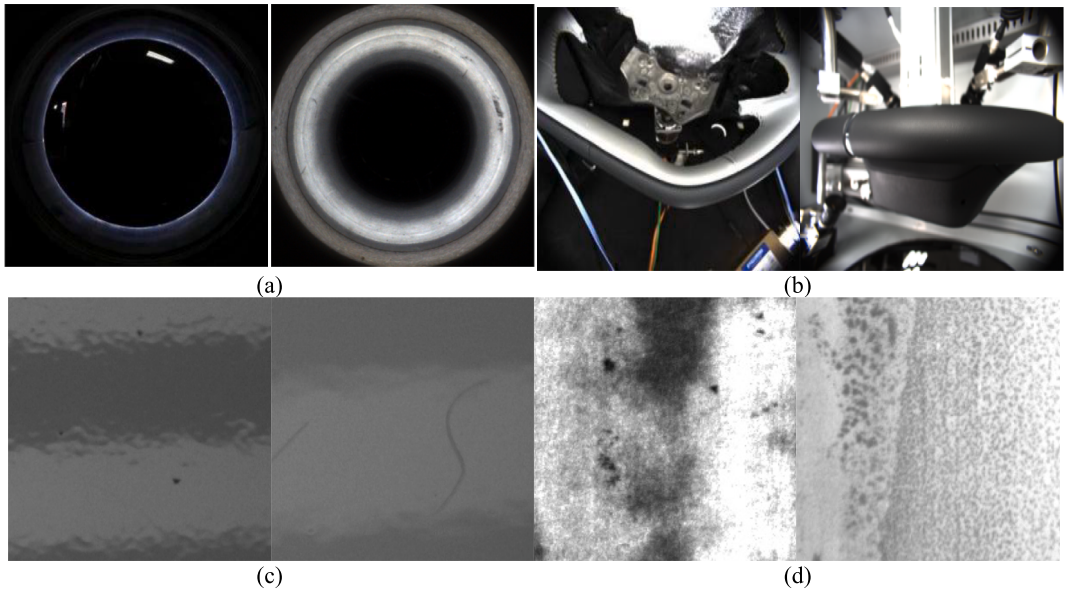


Fig. 13. Mixed dataset samples. (a), (b), (c), and (d) are sample instances from the HIT-EngDV3, HIT-SWD, HIT-CBD, and NEU-DET datasets, respectively.

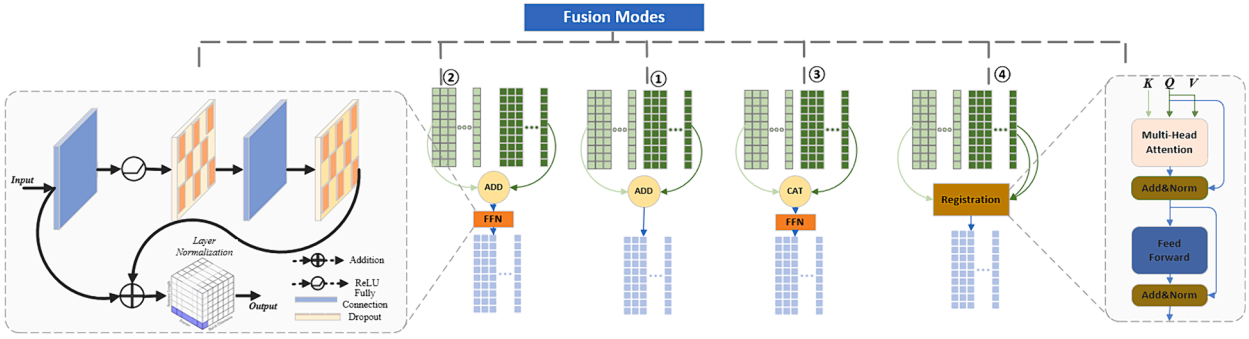


Fig. 14. Detailed diagram of the fusion modes.

Table 7
Experimental Results for the Fusion Method Ablation Study. 'S' and 'u-S' indicate the supervised and unsupervised modes, respectively. 'FU1′, 'FU2′, 'FU3′ and 'FU4′ indicate the fusion modes of AGG, direct additive fusion (①), additive fusion combined with FFN (②), concatenation fusion combined with FFN (③), and multisource feature alignment using self-attention combined with a cross-attention mechanism (④), respectively, as shown in Fig. 14. The best results are reported in red, bold font.

Model		mAP @50:95	mAP @50	mAP @75	APS @50:95	APM @50:95	APL @50:95	ARS @50:95	ARM @50:95	ARL @50:95
u-S	FU1	33.9	74.2	26.4	28.6	38.9	36.3	43.9	55.6	55.2
	FU2	33.9	74.1	26.6	28.0	38.8	37.3	43.6	54.8	55.4
	FU3	32.9	73.5	25.2	27.9	37.8	35.9	44.3	54.7	55.5
	FU4	33.7	74.0	26.5	28.3	38.1	37.3	44.8	56.7	57.2
S	FU1	32.7	72.4	24.7	26.2	37.5	36.4	42.9	53.5	54.8
	FU2	32.7	71.4	25.6	26.7	37.7	35.4	43.3	55.9	56.0
	FU3	33.6	73.4	26.2	28.1	38.2	36.6	43.5	54.9	56.1
	FU4	32.6	72.4	24.6	27.3	36.4	36.2	44.4	55.5	56.4

Table 8
Results of the Ablation Study of Parameter Sharing. The best results are reported in red, bold font. The symbols '√' and '×' denote the use and non-use of parameter sharing approach, respectively.

Model		mAP @50:95	mAP @50	mAP @75	APS @50:95	APM @50:95	APL @50:95	ARS @50:95	ARM @50:95	ARL @50:95
PA	×	33.9	74.2	26.4	28.6	38.9	36.3	43.9	55.6	55.2
PA	√	34.0	74.3	26.4	28.3	38.6	37.2	44.3	56.1	57.2

Fusion modes FU1 and FU4 were employed in this experiment.

Experimental Results: The experimental results are shown in Table 9. 1) Overall, introducing industrial-adapted features at the entrance of the encoder and using the ADD fusion mode achieved optimal defect detection results across multiple industrial scenarios. The comprehensive detection accuracies mAP@(50:95, 50, 75) are all superior to those of the baseline GDINO, with improvements of approximately 0.9, 0.5, and 1.5, respectively. Moreover, as the IOU threshold increases, the improvement in the detection accuracy becomes more significant. 2) Comparing D-FU1 and D-FU4 with E-FU1 and E-FU4 reveals that introducing industrial-adapted features before encoders outperforms introducing them before decoders, with increases of + 1.9 and + 2.7 in terms of mAP@75 and + 1.0 and + 1.8 in terms of mAP@50, respectively. 3) Introducing industrial-adapted features at the entrances of both encoders and decoders does not substantially improve detection accuracy.

4.7. Ablation experiment on CaDTDD

Experimental Design: To investigate the effectiveness of the proposed CaDTDD for defect detection, we conducted ablation experiments with and without CaDTDD. The model configuration employed in this section integrates an industrial adapter, opting for the ADD fusion mode and conducting fusion at the encoder entry while simultaneously employing a parameter sharing strategy, which follows the results obtained in the previous experimental section. The model without CaDTDD is termed MSID-B in this section, and the model with all modules is termed the final MSIDetector. The dual thresholds T = (τL, τH) are empirically set to 0.001 and 0.35, respectively.

Experimental Results: The experimental results are shown in Table 10. Most importantly, the comprehensive accuracy under the quantitative measure of mAP@50 is further improved from 74.8 to 76.2 with the introduction of the CaDTDD on top of the industrial adapter, representing an increase of approximately 2.1 over the original GDINO and approximately 1.4 over the improved version (MSID-B). Fig. 15 illustrates the relationships between the defect detection accuracy and NMS across the current 4 industrial scenarios, indicating that the optimal mAP@50 value of 76.20 was achieved at NMS ≈ 0.55.

4.8. Comparison experiments

Experimental Design: To further compare the proposed defect detection method with current state-of-the-art methods, we conducted a set of comparison experiments. The selected detection methods include both CNN-based and transformer-based architectures as well as single-stage and two-stage structures, varying in anchor initialization with anchor-based and anchor-free methods. Due to the slow convergence speed of DETR [44], we conducted additional training for 400 epochs, resulting in a total of 500 training epochs for DETR. Furthermore, the learning rate decay was implemented at the 400-epoch mark. The dual thresholds selected for the comparative experiments are T = (τL, τH) = (0.001, 0.35), with all other configurations consistent with previous sections.

Experimental Results: The quantitative results (Table 11) reveal that 1) the MSIDetector proposed in this paper achieves the best performance, with a comprehensive detection accuracy of mAP@50 of 76.20, surpassing that of the two-stage FRCNN [39] by 13.93 %. 2) Overall, MSIDetector, DINO, and GDINO, which are based on the transformer architecture exhibit superior performance over those based on CNN architectures. 3) We can also observe that, despite the additional 400 epochs of training, DETR's accuracy, measured as mAP@50, only achieved a value of 7.59.

Table 10
Results of ablation experiments on CaDTDD.

Model	mAP@50:95	mAP@50
GDINO	33.8	74.1
MSID-B	34.8	74.8
MSIDetector	34.7	76.2

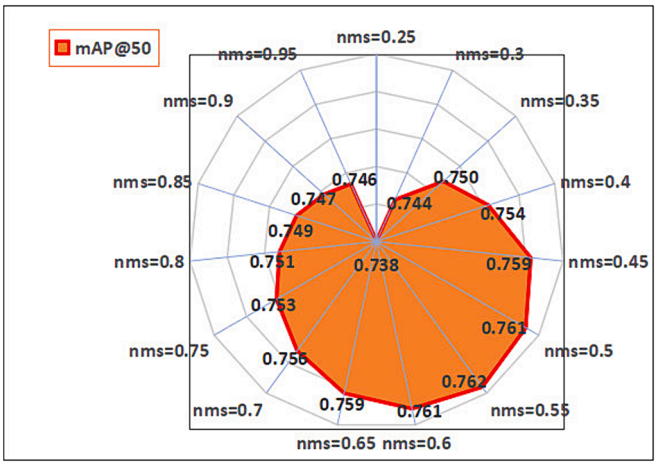


Fig. 15. Experimental results of defect detection under different NMS settings.

Table 11
Comparison of the experimental results obtained with high-performing detectors. The best results are reported in red, bold font.

Model	mAP@50:95	mAP@75	mAP@50
FRCNN+R50 + FPN [39]	28.64	22.84	62.27
SSD [40]	20.29	14.59	48.43
RetinaNet [41]	23.04	16.46	54.35
YOLO5-S3	30.51	23.41	66.82
YOLO5-L3	35.13	30.68	71.17
YOLO7-T [42]	28.40	21.28	63.20
YOLO7-X [42]	34.01	28.92	69.55
DETR [44]	3.00	1.67	7.59
DeformDETR [45]	26.77	17.02	64.96
DINO [43]	32.93	26.04	71.46
GDINO [6]	33.80	25.70	74.10
MSIDetector	34.73	26.66	76.20

Table 9
Results of ablation experiments on prompt location. 'D', 'E', and 'E+D' indicate the introduction of prompt features at the entrances of the encoders, at the entrances of the decoders, and simultaneously at the entrances of both encoders and decoders, respectively. The meaning of 'FU-X' is consistent with that in the previous table. The best results are reported in red, bold font.

Model		mAP @50:95	mAP@50	mAP@75	APS @50:95	APM @50:95	APL @50:95	ARS @50:95	ARM @50:95	ARL @50:95
D	FU1	33.8	73.8	25.7	28.3	38.4	37.0	44.6	56.2	56.3
	FU4	32.9	72.5	24.5	27.0	37.7	36.0	43.1	54.4	55.1
E	FU1	34.8	74.8	27.6	29.0	39.3	38.4	45.0	56.1	57.5
	FU4	34.2	74.3	27.2	29.1	38.7	37.5	44.9	56.4	57.2
E+D	FU1	33.8	74.1	26.0	28.3	38.3	37.4	45.1	56.5	57.2

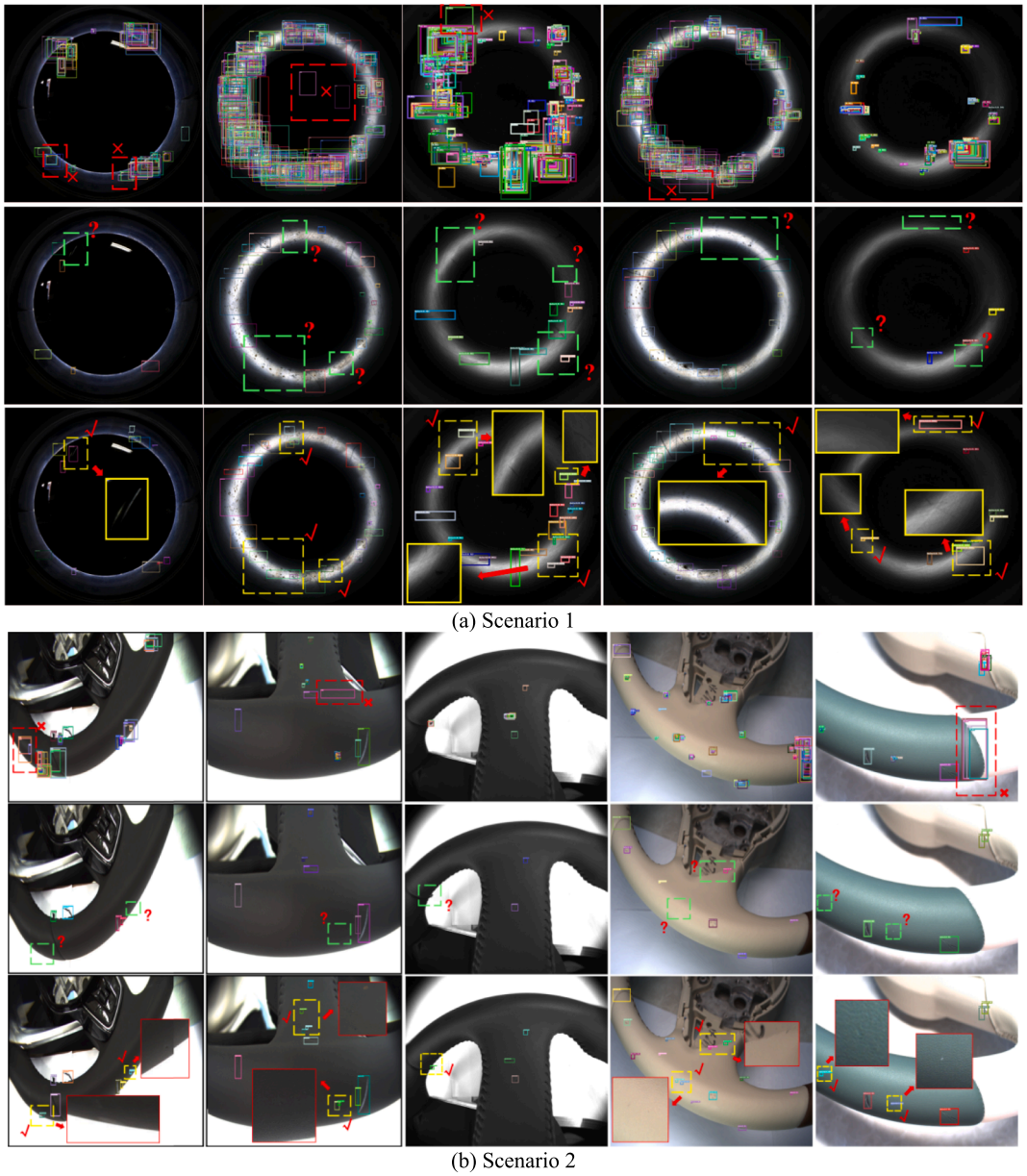
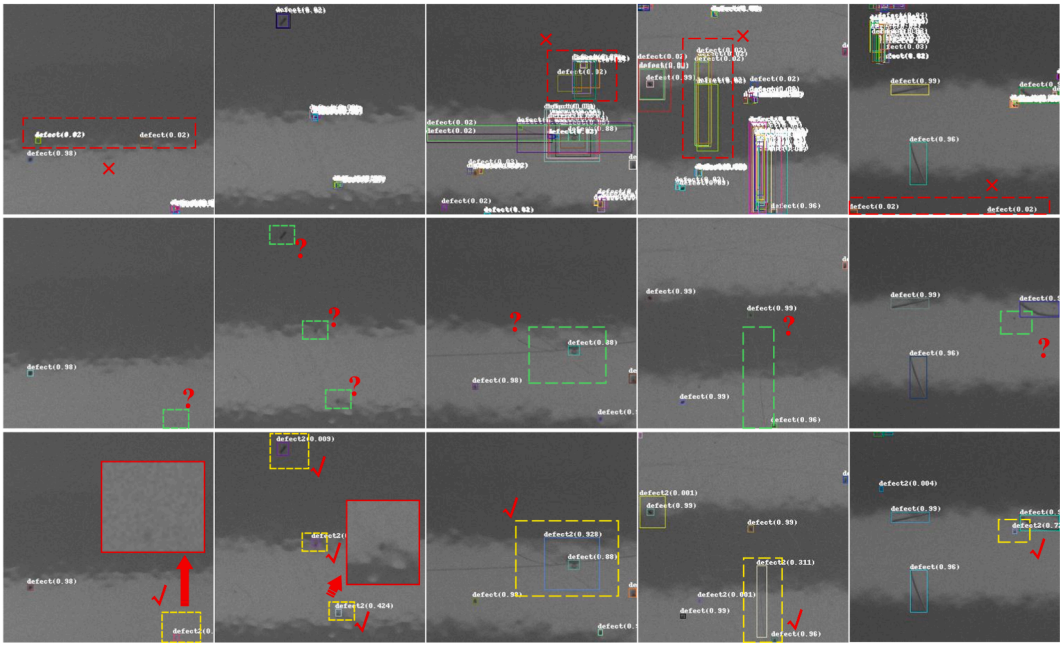


Fig. 16. Visualization of the qualitative experimental results. The data presented above were collected from four industrial defect detection scenarios, where scenarios 1, 2, and 3 were acquired using our self-developed PX1, SW1, and CB1. Within each scenario the three rows of qualitative experimental results are derived from the original GDINO (low threshold = 0.02), the original GDINO (τ_L=0.35), and the proposed MSIDetector model with dual thresholds T=[0.02, 0.35]. The red dashed boxes indicate falsely detected defect results, the green dashed boxes represent missed detections, and the yellow defect boxes signify correctly detected defects (corresponding to the symbols ×,? and √, respectively). In addition, to provide a clearer observation of the regions in the figure we have enlarged and displayed specific local areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

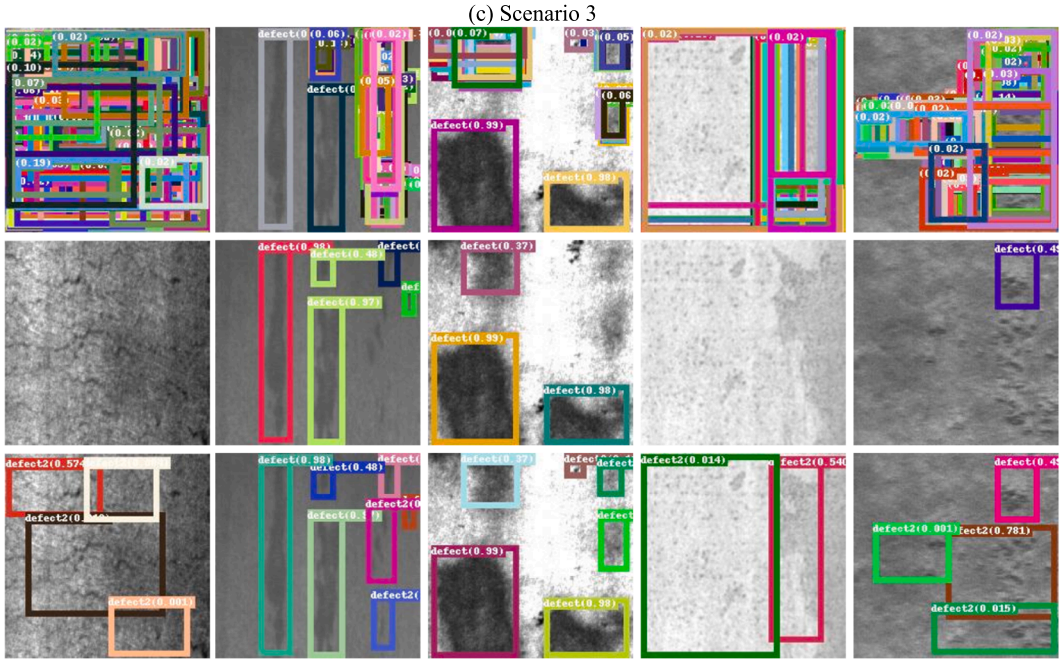
Fig. 16 presents the qualitative results of 5 groups across 4 industrial scenarios. To enable clearer comparisons, we have visualized the qualitative results corresponding to GDINO under both high and low threshold settings. The figure shows that 1) from the first row corresponding to each scenario (low threshold), setting the threshold at a low value results in GDINO generating a plethora of redundant detections alongside numerous false detections, as depicted by the red dashed boxes in the first rows of the 4 scenarios in Fig. 16. The second row corresponds to each scenario (high threshold), and setting the threshold higher leads to many defects being filtered, as illustrated by the green dashed boxes in the second rows of the 4 scenarios in Fig. 16. Ultimately, the MSIDetector proposed in this paper achieves well-balanced mitigation of missed detections, false-positives, and overdetections. 2) Further observation reveals that missed detections caused by high thresholds predominantly occur when the defects are small (spanning only a few pixels) or under conditions of varying illumination, leading to the model generating lower confidence for these defects, resulting in their omission.

5. Discussion

The experimental findings regarding defect terminology underscore its significant impact on the ultimate defect detection outcomes. This finding aligns with our initial premise, which posits that, compared with generic detection scenarios, defect detection scenarios exhibit features of both ‘one term, multiple shapes’ and ‘one shape, multiple terms.’ As



(c) Scenario 3



(d) Scenario 4

Fig. 16. (*continued*).

illustrated in Fig. 1, for example, the same type of scratch may manifest differently in various industrial scenarios, presenting variations in thickness, curvature, and straightness. Moreover, different defect terminologies in distinct industrial scenarios may exhibit similar appearances, such as the presence of dark areas in both the industrial scenario of car body paint and hot-rolled steel strips, named patches and breakages, respectively (Fig. 1). Such nuances pose challenges to the detection task, potentially leading to ambiguity in the model's semantic understanding and thereby impeding convergence. These results further corroborate this contention. Therefore, in the subsequent experimental sections of this paper, unless otherwise specified, we conducted experiments solely within the 'defect' category. However, in practical deployment, adaptive functionalities can be integrated into the backend to align with specific scenarios, for example by further refining categories based on particular defect scenarios (also adopted in our current practical deployment) or directly modifying our discrimination unit to multiple defect categories for practical use.

From the results presented in the CaDTDD table it is observable that two more stringent composite metrics, mAP@(50:95) and mAP@75, and specifically, mAP@(50:95), exhibited no significant variations, while mAP@75 demonstrated a reduction compared with the pre-CaDTDD implementation. We argue that the introduction of dual thresholds may result in subtle positional discrepancies between detections of low confidence and ground truth annotations. Consequently, at higher IOU thresholds such detections may be classified as false-positives (FPs), thereby leading to a decrease in the final mAP@75. Integrating our earlier experimental findings and the qualitative results depicted in Fig. 16, it becomes apparent that detections corresponding to low confidence effectively delineate defects. This observation is particularly pertinent in the context of transformer architectures.

As shown in Fig. 16, in the scenario of engine cylinder bore defect detection, the primary issues are irregular and dense defects caused by splattered oil and rust. For the detection of defects in the steering wheel covering area, the challenge lies in the minimal color contrast between the defects and the leather material, making the defects easily overlooked. When defects in automotive body paint are detected, the diffuse reflection of the paint surface causes significant confusion between the defects and the background structure. In the hot-rolled steel strip scenario, the defects are larger in size than those in the previous three scenarios, but issues arise due to the crazing (first column of scenario 4) and patch (third column of scenario 4) categories, which appear very similar in appearance. Overall, across these four scenarios, selecting a low threshold results in numerous redundant detection results, while opting for a relatively high threshold leads to higher rates of missed detections. Consequently, employing a dual-threshold discrimination strategy enhances detection performance. Additionally, in our experiments regarding the determination of dual thresholds we argue that the selection of the low threshold (τL) should encompass as many detection results from the model as possible. In contrast, the high threshold (τH) should be determined by balancing the detection results and over-detection issues, and this should be established through specific experiments based on practical conditions.

According to the qualitative analysis (Fig. 16), by adhering to the principle of 'existence implies legitimacy', the dual-thresholding approach effectively mitigates the issue of missed defects caused by a single threshold. Moreover, for transformer architectures instances of high rates of missed detections are often associated with defects of smaller sizes or those subjected to varying illumination conditions. These circumstances lead to lower confidence in the detection outcomes by the model, resulting in them being missed under a single-thresholding strategy. Additionally, the presence of CaDTDD further suppresses the drawbacks of excessive redundant and false-positive results induced solely by lowering the threshold. Consequently, our model adapts to hard defect samples without increasing model capacity, offering novel insights for subsequent defect detection model designs. As long as a similar branch, as shown in Fig. 2, is introduced at the output position of the model to complete the dual discrimination of the context within the detection patch, the purpose of dual-threshold discrimination can be achieved. This allows further exploration of the truly existing targets under the low threshold, thereby extending the CaDTDD module to the detection tasks in other scenarios and further improving the detection rate. Therefore, the determination of [τL, τH] does not require extensive consideration of specific scenarios. Instead, as can be seen, we used a commonly adopted threshold for τH and a lower value for τL, which also reduces the complexity of using the CaDTDD.

As described in Section 4.1, the current Multi-Scenes Industrial Defect Detection dataset (MSDD) covers four industrial scenarios: engine cylinder bore defects, car steering wheel covering area defects, car body defects, and hot-rolled steel strip defects. The experiments in this paper demonstrate the effectiveness of the proposed algorithm, MSIDetector, for multiscenario industrial defect detection compared with other advanced algorithms. To enable the model to learn the common and essential features of defects in industrial scenarios more fully and extensively, it is necessary to expand the data volume. Therefore, we promote further expansion of the dataset using our equipment to collect more real data from various scenarios. integrate the currently available public defect detection datasets to form a unified multiscenario industrial defect detection dataset to be shared with researchers.

In our proposed MSIDetector algorithm, guided by the industrial feature adapter, we successfully incorporated industrial priors into the visual foundation model, achieving competitive comprehensive detection accuracy (mAP@50 = 76.20). However, we observed some missed detections in qualitative experiments (Fig. 17). These issues occur mainly in low-light and low-contrast scenarios. The red dashed boxes in the images highlight undetected defects, including a crack in Fig. 17 (a) and a dent in Fig. 17 (b). The common characteristic of these problems is low contrast and illumination leading to less distinct defect features. This results in the geometric shapes or texture features being less prominent visually, making it challenging for the detection algorithm to identify them. To address this issue, we suggest that when the foundation model is deployed in a specific industrial scenario, appropriate adjustments should be made on the basis of the characteristics of the industrial scene. For example, in darker or low-contrast scenarios, input images may need to undergo data augmentation. Furthermore, we have identified another challenge: defect annotation in some industrial scenarios can be quite demanding for annotators. In our qualitative experiments we found that sometimes the detected defects are indeed defects (in most cases, the model is correct, i.e., the highlighted areas do have issues, especially for algorithms based on the transformer structure). However, these defects might be overlooked by annotators because of their subtle nature, as illustrated in Fig. 17 (c). Therefore, we can sometimes utilize a semiautomatic human-machine interaction approach to jointly annotate the data.

6. Conclusion

In this study, we unify an industrial feature adapter and context-aware dual-thresholding defect discriminator into a large visual detection model, introducing priors of industrial domain-specific knowledge. This approach overcomes the limitations of the single thresholding approach in defect filtration, resulting in a new, higher-precision defect detector termed MSIDetector. Additionally, we independently designed and developed three quality inspection devices and collected a large

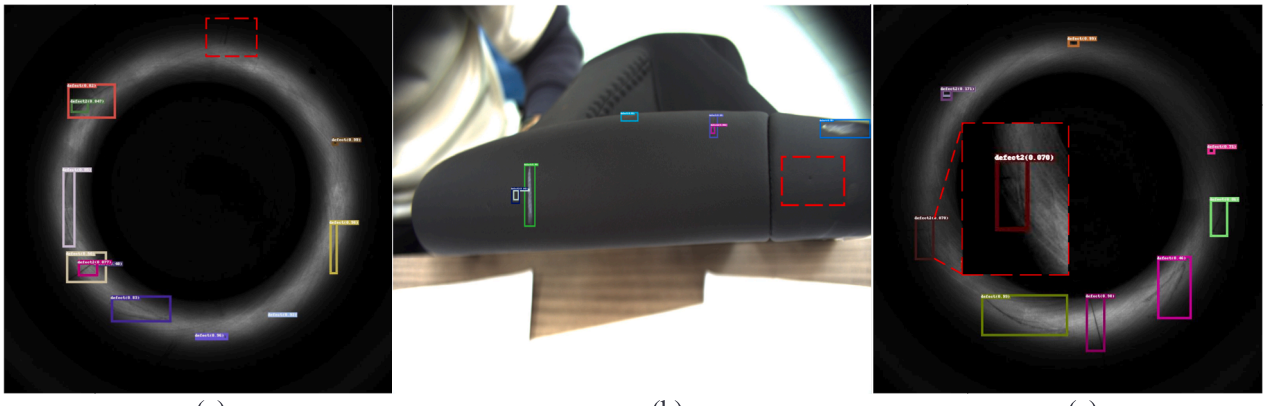


Fig. 17. Issue presentation.

amount of real defect data. Experimental evaluations conducted across four industrial scenarios demonstrate superior detection performance, with an mAP@50 of 76.20, outperforming the SOTA detection methods FasterRCNN incorporating FPN, YOLO7X, and the visual foundation model GDINO by 13.93, 6.55, and 2.1, respectively. We hope that our paradigm could provide a new direction for constructing upcoming industrial defect detectors.

As a member of the visual foundation model, MSIDetector is subject to efficiency issues (currently ~ 1.4 s/image on A100-40G). Therefore, future work will focus primarily on optimizing model efficiency, such as by employing model distillation techniques to compress a large model into a smaller capacity without significantly sacrificing accuracy. Data collection efforts are ongoing, and we will expand our detection equipment and corresponding data to cover a wide range of scenarios in the industry.

CRediT authorship contribution statement

Xujie He: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Jing Jin:** Writing – review & editing, Supervision, Funding acquisition. **Fujiang Yu:** Validation. **She Zhao:** Validation, Data curation. **Duo Chen:** Data curation. **Xiang Gao:** Project administration.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Research and Development Project of Visual Defect Detection Imaging Systems and Software Tools for Complex Workpieces of Heilongjiang Province, China [grant number 2023ZXJ01A01] and in part by Heilongjiang Province's "Million and Ten Million" Major Project in Science and Technology, China [grant number 2021ZX10A01].

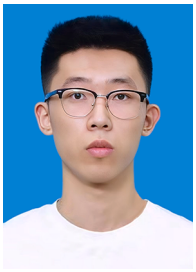
References

[1] M. Liu, Y. Chen, J. Xie, L. He, Y. Zhang, LF-YOLO: A lighter and faster YOLO for weld defect detection of X-Ray image, IEEE Sens. J. 23 (2023) 7430–7439, https://doi.org/10.1109/JSEN.2023.3247006.
[2] Q. Thien Pham, N.S. Liou, The development of on-line surface defect detection system for jujubes based on hyperspectral images, Comput. Electron. Agric. 194 (2022) 106743, https://doi.org/10.1016/j.compag.2022.106743.
[3] Ranebennur1 R, Thirumaleshwar1 S, Somareddy1 HK. Development of Automated Quality Assurance Systems for Pharmaceutical Manufacturing: A Review. J Coast Life Med 2023;11:1855–64.
[4] A. Kirillov, E. Mintun, N. Ravi, S. Whitehead, A.C. Berg, P. Doll, Segment Anything, Proc IEEE/CVF Int Conf Comput vis (2023) 4015–4026.
[5] Jiang Q, Li F, Zeng Z, Ren T, Liu S, Zhang L. T-Rex2: Towards Generic Object Detection via Text-Visual Prompt Synergy 2024.http://arxiv.org/abs/2403.14610.
[6] Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. ArXiv E-Prints 2023: arXiv:2303.05499. 10.48550/arXiv.2303.05499.
[7] Cheng Y, Li L, Xu Y, Li X, Yang Z, Wang W, et al. Segment and Track Anything. ArXiv E-Prints 2023:arXiv:2305.06558. 10.48550/arXiv.2305.06558.
[8] P.I. Gomez, M.E.L. Gajardo, N. Mijatovic, T. Dragicevic, A Self-Commissioning Edge computing method for data-driven anomaly detection in power electronic systems, IEEE Trans. Ind. Electron. (2024) 1–12, https://doi.org/10.1109/TIE.2023.3347839.
[9] A. Origlia, S. Di Martino, E. Battista, Rail anomalies detection: a comparative analysis of three self-supervised models on real data, Comput. Ind. 148 (2023) 103909, https://doi.org/10.1016/j.compind.2023.103909.
[10] X. Shen, J. Liu, L. Jiang, X. Liu, H. Zhang, A Novel weld defect detection method for intelligent magnetic flux leakage detection system via contextual relation network, IEEE Trans. Ind. Electron. 71 (2024) 6304–6314, https://doi.org/10.1109/TIE.2023.3294578.
[11] Z. Xu, Y. Lin, D. Chen, M. Yuan, Y. Zhu, Z. Ai, et al., Wood broken defect detection with laser profilometer based on Bi-LSTM network, Expert Syst. Appl. 242 (2024) 122789, https://doi.org/10.1016/j.eswa.2023.122789.
[12] H. Mewada, I.M. Pires, P. Engineer, A.V. Patel, Fabric surface defect classification and systematic analysis using a cuckoo search optimized deep residual network, Eng Sci Technol an Int J 53 (2024) 101681, https://doi.org/10.1016/j.jestch.2024.101681.
[13] Z. Zhao, J. Wang, Q. Tao, A. Li, Y. Chen, An unknown wafer surface defect detection approach based on Incremental Learning for reliability analysis, Reliab. Eng. Syst. Saf. 244 (2024) 109966, https://doi.org/10.1016/j.ress.2024.109966.
[14] Li P, Li F, Liu M, Bai H, Wei Y, Wang A, et al. Aggregation for CdZnTe Defect Segmentation. IEEE Trans Ind Informatics 2024;PP:1–11. 10.1109/TII.2024.3384517.
[15] X. Zhou, S. Zhou, Y. Zhang, Z. Ren, Z. Jiang, H. Luo, GDALR: Global Dual Attention and Local Representations in transformer for surface defect detection, Meas J Int Meas Confed 229 (2024) 114398, https://doi.org/10.1016/j.measurement.2024.114398.
[16] T. Sun, Z. Li, X. Xiao, Z. Guo, W. Ning, T. Ding, Cascaded detection method for surface defects of lead frame based on high-resolution detection images, J. Manuf. Syst. 72 (2024) 180–195, https://doi.org/10.1016/j.jmsy.2023.11.017.
[17] S. Zhao, R.Y. Zhong, J. Wang, C. Xu, J. Zhang, Unsupervised fabric defects detection based on spatial domain saliency and features clustering, Comput. Ind. Eng. 185 (2023) 109681, https://doi.org/10.1016/j.cie.2023.109681.
[18] C. Zhao, X. Shu, X. Yan, X. Zuo, F. Zhu, RDD-YOLO: A modified YOLO for detection of steel surface defects, Meas J Int Meas Confed 214 (2023) 112776, https://doi.org/10.1016/j.measurement.2023.112776.
[19] F. Guo, J. Liu, Y. Qian, Q. Xie, Rail surface defect detection using a transformer-based network, J. Ind. Inf. Integr. 38 (2024) 100584, https://doi.org/10.1016/j.jii.2024.100584.
[20] H. Shang, C. Sun, J. Liu, X. Chen, R. Yan, Defect-aware transformer network for intelligent visual surface defect detection, Adv Eng Informatics 55 (2023) 101882, https://doi.org/10.1016/j.aei.2023.101882.
[21] X. He, J. Jin, D. Chen, Y. Feng, An integrated defect detection method based on context encoder and perception-enhanced aggregation for cylinder bores, J. Manuf. Process. 114 (2024) 196–212, https://doi.org/10.1016/j.jmapro.2024.02.006.
[22] X. Tao, D. Zhang, W. Ma, Z. Hou, Z.F. Lu, C. Adak, Unsupervised Anomaly Detection for Surface Defects With Dual-Siamese Network, IEEE Trans Ind Informatics 18 (2022) 7707–7717, https://doi.org/10.1109/TII.2022.3142326.
[23] X. Zhang, N. Li, J. Li, T. Dai, Y. Jiang, S.T. Xia, Unsupervised Surface Anomaly Detection with Diffusion Probabilistic Model, Proc IEEE Int Conf Comput vis (2023) 6759–6768, https://doi.org/10.1109/ICCV51070.2023.00624.
[24] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, C. Steger, The MVTec Anomaly DETECTION DATASET: A COMPREHENSIVE REAL-WORLD DATASET FOR UNSUPERVISED ANOMALY DETECTIon, Int. J. Comput. vis. 129 (2021) 1038–1059, https://doi.org/10.1007/s11263-020-01400-4.
[25] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (2024) 1–9, https://doi.org/10.1038/s41467-024-44824-z.
[26] Ravishankar H, Patil R, Melapudi V, Annangi P. SonoSAM - Segment Anything on Ultrasound Images. In: Kainz B, Noble A, Schnabel J, Khanal B, Müller JP, Day T, editors. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 14337 LNCS, Cham: Springer Nature Switzerland; 2023, p. 23–33. 10.1007/978-3-031-44521-7_3.
[27] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, et al., RSPrompter: learning to Prompt for remote sensing instance segmentation based on visual foundation model, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–17, https://doi.org/10.1109/TGRS.2024.3356074.
[28] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, et al., SpectralGPT: spectral remote sensing foundation model, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–15, https://doi.org/10.1109/TPAMI.2024.3362475.
[29] He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2022; 2022-June:15979–88. 10.1109/CVPR52688.2022.01553.
[30] Z. Yang, Y. Yang, Decoupling features in hierarchical propagation for video object segmentation, Adv Neural Inf Process Syst 35 (2022) 1–13.
[31] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, J. Wang, AnomalyGPT: detecting industrial anomalies using large vision-language models, Proc AAAI Conf Artif Intell 38 (2024) 1932–1940, https://doi.org/10.1609/aaai.v38i3.27963.
[32] Li Y, Wang H, Yuan S, Liu M, Zhao D, Guo Y, et al. Myriad: Large Multimodal Model by Applying Vision Experts for Industrial Anomaly Detection. ArXiv E-Prints 2023:arXiv:2310.19070. 10.48550/arXiv.2310.19070.
[33] Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models 2023:1–15.
[34] Xiong Y, Li Z, Chen Y, Wang F, Zhu X, Luo J, et al. Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications 2024;3.
[35] Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Comput. Vis. – ECCV 2022, Cham: Springer Nature Switzerland; 2022, p. 1–21.
[36] Y. He, K. Song, Q. Meng, Y. Yan, An End-to-End steel surface defect detection approach via fusing multiple hierarchical features, IEEE Trans. Instrum. Meas. 69 (2020) 1493–1504, https://doi.org/10.1109/TIM.2019.2915404.
[37] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv E-Prints 2018: arXiv:1810.04805. 10.48550/arXiv.1810.04805.

[38] W. Wang, E. Xie, X. Li, D.P. Fan, K. Song, D. Liang, et al., PVT v2: improved baselines with pyramid vision transformer, Comput vis Media 8 (2022) 415–424, https://doi.org/10.1007/s41095-022-0274-8.
[39] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
[40] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Comput. Vis. – ECCV 2016, Cham: Springer International Publishing; 2016, p. 21–37.
[41] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proc. IEEE Int. Conf. Comput. Vis., 2017, p. 2980–8.
[42] Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors 2023:7464–75. 10.1109/cvpr52729.2023.00721.
[43] Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, et al. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection 2022. http://arxiv.org/abs/2203.03605.
[44] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 12346 LNCS, 213–229. 10.1007/978-3-030-58452-8_13.
[45] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2021. Deformable Detr: Deformable Transformers for End-To-End Object Detection. ICLR 2021 - 9th Int. Conf. Learn. Represent. 1–16.

Xujie He received an M.Eng. degree from Harbin Engineering University, Harbin, China, in 2021, and a B.Eng. degree from Shenyang Aerospace University, Shenyang, China, in 2018. He is currently pursuing a Ph.D. degree in Control Science and Engineering at Harbin Institute of Technology, China. His research interests primarily focus on computer vision, including defect segmentation and detection, multiple object tracking, and visual foundation models.

Jing Jin received her Ph.D. from Harbin Institute of Technology, Harbin, China, in 2008. She is currently a professor at the School of Astronautics, Harbin Institute of Technology, China. She has received research funding from the Chinese Natural Science Foundation and has been involved in several intelligent astronautics projects. Her research interests include computer vision, specifically intelligent detection and diagnosis for medical images; and navigation and localization technologies, including spacecraft pulsar navigation, SLAM, and smart curling robots.



Fujiang Yu received an B.Eng degree from Harbin Institute of Technology, Harbin, China, in 2022. He is currently pursuing a M.Eng degree in Control Science and Engineering at Harbin Institute of Technology, China. His research interests primarily focus on computer vision, including Image Segmentation and General Vision Model.

She Zhao received a B.Eng. degree from YanShan University , Qinghuangdao,China, in 2023. He is currently pursuing a M. Eng degree in Control Science and Engineering at Harbin Institute of Technology,China. His research primarily interests focus on computer vision, including defect detection, Image fusion, and visual-language models.

Duo Chen received an B.Eng degree from Harbin Institute of Technology, Harbin, China, in 2022. He is currently pursuing a M.Eng degree in Control Science and Engineering at Harbin Institute of Technology, China. His research interests primarily focus on computer vision, including object detection and General Vision Model.

Xiang Gao, a PMP-certified project management professional and product expert, currently serves as the Project Management Director at Harbin Institute of Technology AI Research Institute Co., Ltd. Previously, he worked at the publicly listed company Beijing Si-Tech Information Technology Co., Ltd., where he held the roles of Senior Engineer and Product Manager, responsible for the planning and operation of products with annual revenues exceeding 30 million RMB. He possesses extensive project experience and knowledge in industrial IoT platforms, smart manufacturing, smart education, and digital information system development.