symmetry

MDPI

Article

PG-SalDETR: A Method for Detecting Small Defects in Steel Plates Based on Physically Guided Saliency and Asymmetric Perception Network

Xiaodong Zhang [1], Cuiyun Li [1] and Shengye Zhao [2,*]

[1] School of Software, Shenyang University of Technology, Shenyang 110870, China; sutzhxiaodong@163.com (X.Z.); sutlicuiyun@smail.sut.edu.cn (C.L.)
[2] School of Computer Science and Engineering, Hunan Institute of Technology, Hengyang 421002, China
* Correspondence: zhaoshengye@hnit.edu.cn

Abstract

Steel plate defect detection is confronted with problems such as weak features of small defects, disconnection between physical priors and detection tasks, and semantic inconsistency of multi-scale fusion, which can easily lead to the misdetection of small defects. To solve these problems, this paper proposes a detection method named PG-SalDETR. Firstly, this paper proposes a physics-guided saliency perception mechanism (PGSPM), which transforms physical priors into learnable guidance signals and directly embeds them into the detection network for joint optimization. Secondly, this paper proposes the token sequence saliency perception network (TSSP-Net), which is designed to help improve the perception and representation of small defect features through an asymmetric dual-branch architecture, adaptive fusion, and residual fusion. Thirdly, a two-stage query refinement mechanism (TSQRM) is proposed. Through physically guided offset correction and adaptive multi-scale feature aggregation, it optimizes the query while preserving fine-grained defect details. Finally, the dynamic cross-scale fusion module (LCASF) is proposed. Through the dynamic cross-scale fusion strategy, the semantic inconsistency problem of small defect features in multi-scale fusion is alleviated. Experimental results demonstrate improvements. Compared to Salience DETR, PG-SalDETR achieves an AP increase of 3.8% and 2.6%, and an AP_S increase of 2.8% and 3.9% on the NEU-DET and GC10-DET datasets, respectively. These results indicate the effectiveness of the proposed method for small defect detection on steel plate surfaces.

Keywords: transformer; defect detection; small defect detection

Check for updates

Academic Editor: Shuang Xu

Received: 10 May 2026
Revised: 23 June 2026
Accepted: 25 June 2026
Published: 29 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In modern industries, steel plates play a vital role in fields like construction, machinery manufacturing, and automobiles. Surface quality of steel plates directly affects the safety and reliability of the final products. Especially in complex production processes, steel plate surfaces are susceptible to various defects such as cracks, which pose a potential threat to product quality [1]. Manual inspection cannot meet the demands of modern production, so there is an urgent need for automated defect detection [2].

Steel plate surface defect detection is a typical industrial application of computer vision, which has long been confronted with two core technical challenges [3]. First, defect feature degradation is caused by imaging environment interference. Intense reflection, uneven illumination, and low contrast on steel plate surfaces [4] lead to severe image

quality degradation. This further masks the weak feature signals of small defects with background noise, posing great difficulties for feature extraction in the model perception stage. Second, inherent limitations of intra-model perception and multi-scale fusion: the features of small defects with a few pixels are already extremely weak, and they are further subject to semantic submergence and localization distortion during multi-scale feature fusion. Deep high-semantic features directly override shallow weak features, resulting in a sharp drop in the signal-to-noise ratio of small defects. Meanwhile, conflicts between high-level semantic cues and low-level positional details further distort the representational structure of defects. These problems jointly lead to a persistently high missing detection rate. The above challenges jointly constitute the main technical bottlenecks in the current detection of small surface defects on steel plates.

Recently, detection methods represented by deep learning have achieved remarkable progress. From detection frameworks based on convolutional neural networks [5–7] to the recently developed Transformer architectures with global modeling capabilities, the representational and detection performance of models has been continuously improved. Nevertheless, when facing the above industrial-level challenges, existing methods still generally suffer from three systemic problems:

- Disconnection between physical perception and task reasoning: Most methods follow a serial "preprocessing–detection" paradigm, placing physical enhancement modules based on theories such as Retinex [8,9] at the front end. Their optimization objectives are inconsistent with those of the final detection task, resulting in the failure of physical priors to be deeply integrated into the core perception process of detectors in a learnable and guidable manner.
- Coarse multi-scale feature fusion [10] mechanisms: Existing fusion strategies lack refined cross-scale semantic alignment and calibration capabilities for the problems of small defect feature suppression and distortion caused by semantic gaps.
- Insufficient sensitivity to the weak features of small defects: The feature selection or query initialization mechanisms of existing models are slow to respond to extremely weak small defect features. Their filtering and refinement processes are unable to effectively preserve the fine details of small defects while suppressing background noise [11].

To address the above-mentioned problems, this work proposes a method named PG-SalDETR based on physical guidance and saliency enhancement. This method incorporates four innovations:

- To address the disconnection between physical perception and task reasoning, we integrate the PGSPM (physics-guided saliency perception mechanism) into the detection network, which converts the Retinex physical prior into learnable guidance signals. These signals are directly embedded into the detection network for end-to-end optimization, solving the problem of disconnection between physical optimization and detection objectives in traditional methods. This helps improve the discriminative capability for small defect features under complex illumination conditions.
- To address insufficient sensitivity to weak features of small defects, we propose TSSP-Net (token sequence saliency perception network), which alleviates the traditional spatial distance symmetry constraint through three asymmetric designs: asymmetric dual-branch architecture, adaptive fusion strategy, and asymmetric residual fusion. These designs jointly enable TSSP-Net to perform operations, replacing the traditional distance-based saliency prediction. This helps improve the perception and retention ability of fine and asymmetrically distributed small defect features.
- To address the preservation of small defect details, we propose TSQRM (two-stage query refinement mechanism). Through physics-guided offset correction and adaptive

multi-scale feature aggregation, it achieves fine-grained retention of small defect details, effectively improving detection accuracy and recall.

- To address the coarse multi-scale feature fusion mechanisms, we propose LCASF (dynamic cross-scale fusion module). By employing a dynamic cross-scale fusion strategy, LCASF alleviates the problem of small defect features being overwhelmed during multi-scale fusion and enhances the detection performance for multi-scale objects.

The remaining part of this paper is structured as follows: Section 2 reviews the relevant image preprocessing and physical guidance methods, deep learning-based defect detection methods, and Transformer-based detection frameworks. Section 3 elaborates on the PG-SalDETR method proposed in this paper, including the design principles and implementation details of its core modules. Section 4 presents the comparative experiments, ablation analysis, and visualization results on the two datasets. Section 5 is the conclusion, which summarizes the work presented in the entire paper and provides directions for future research.

2. Related Work

2.1. Image Preprocessing and Physics-Guided Methods

Early industrial vision systems relied on image preprocessing to improve defect visibility. For example, contrast-limited adaptive histogram equalization [12,13] can enhance global contrast, but it is prone to introducing noise when processing local highlights or shadows. Methods based on Retinex theory [8,9] process the separated reflectance components, providing a physical basis while enhancing contrast. Their multi-scale variants are widely used to alleviate uneven illumination. However, these traditional methods are usually used as independent front-end preprocessing modules, and their optimization objectives are decoupled from the back-end detection task objectives.

In recent years, research has begun to explore combining physical models with deep learning. Wei et al. [14] were the first to encode Retinex theory into a learnable deep network. In terms of unsupervised physical decomposition, the KinD series of works proposed by Zhang et al. [15] achieved Retinex decomposition without pairwise data training by introducing a consistent prior for the reflectance components. To reduce the dependence on paired data, Zhang et al. [16] further explored weak supervision or self-supervision paradigms. They developed a self-supervised enhancement network that only requires low-light images and constructed a cycle consistency constraint guided by physical information for training. However, the goal of these methods is general image enhancement or enhancing visual image quality. They are not optimized in coordination with specific detection tasks or aligned with the goals of defect detection tasks and fail to achieve end-to-end co-optimization of physical priors and detectors. Recently, in defect detection, Y. Pernando et al. [17] developed an enhancement model based on Retinex for low-light steel surface defect detection. However, this method still adopts a serial architecture of "enhancement first and detection later". The physical prior only acts on the input image and fails to be transformed into a learnable saliency signal that can run through and guide the feature extraction and reasoning process of the entire detection network. How to achieve end-to-end deep fusion of physical perception and semantic understanding (rather than serial preprocessing) is still an urgent direction to be explored.

2.2. Deep Learning-Based Defect Detection Methods

The emergence of deep learning, particularly the rapid development of convolutional neural networks (CNNs), has greatly advanced defect detection technology and completely transformed traditional methods that relied on manual design and shallow features. Current methods are generally categorized into two-stage detectors [5–7] and one-stage

detectors [18–20]. Two-stage detectors tend to achieve higher accuracy but require more computational resources, whereas one-stage detectors are faster but often miss defects that occupy very small pixel areas.

To address the core challenge of misdetection of small defects, researchers have proposed improvement schemes from multiple perspectives, including multi-scale feature fusion, attention mechanisms, feature enhancement, and structural optimization. For example, the feature pyramid network (FPN) [21] has become the standard paradigm for fusing multi-scale features. However, its simple top-down addition operation ignores the semantic differences between layers, which can easily lead to the semantic submergence of small defect details. Specifically, FPN generates deep feature maps through multiple downsampling steps. Consequently, small defects smaller than 32 × 32 pixels correspond to areas smaller than 1 pixel, resulting in the near-complete loss of their location and discriminative features. Although FPN introduces shallow features through lateral connections, the semantic gap between deep-level semantics and shallow-level positional details persists, making it impossible to recover the precise boundaries of small defects through simple linear addition. The improved BiFPN [22] improves efficiency through weighted fusion but still cannot show the modeling and calibration of semantic bias. In its weighting mechanism, large-scale defects dominate the gradient, while the feature contributions of small defects are systematically diluted. Furthermore, BiFPN still relies on fixed-magnification cross-scale scaling operations. The spatial resolution loss of small defects in deep feature maps is irreversible, and the weighted approach cannot compensate for the signal loss caused by sampling. Various attention modules [23] (such as CBAM) have been introduced to focus on potential defect regions. However, these methods are usually used as post-processing modules and cannot fundamentally solve the semantic misalignment problem during feature fusion. In terms of feature enhancement and structural optimization, for example, Luo et al. [24] designed an encoder–decoder network based on cross-attention Transformer to enhance cross-scale interaction. Li et al. [25] proposed to enhance features through dual-path adaptive fusion and detail compensation. The challenge of small target detection is universal and is also prominent in scenarios such as remote sensing and autonomous driving. For instance, Han et al. developed the CGFINet network [26], which significantly improved the recognition accuracy of multi-scale land cover changes by designing cross-scale guidance and high-order feature interaction modules. This work indirectly confirms the importance of explicitly modeling scale relationships and protecting small target features in multi-scale feature fusion. However, most existing methods are limited to local optimization at the feature level and lack a systematic mechanism to calibrate cross-scale semantic bias and protect weak features.

2.3. Transformer-Based Detection Frameworks

Transformers have opened up a new path for object detection with their strong ability to model global information. DETR [27] was the first to approach detection as an end-to-end set prediction problem. However, it faces challenges in small object detection, such as slow training convergence, high computational demands, and less effective performance. To address these issues, Deformable DETR [28] introduced a deformable attention mechanism, greatly improving computational efficiency and small object detection, thereby enabling Transformer models to be applied in industrial settings. To address the specific requirements of industrial defect detection—such as real-time processing and detecting small objects—enhanced models like REDef-DETR [29] and SSR-DETR [30] have been developed. These models mainly focus on balancing accuracy and speed through lightweight architectures and optimized feature interactions. Recently, the Dynamic DETR [31] proposed by Dai et al. can enhance query flexibility, but its generation process does not explicitly consider

the visual saliency of defects, and it does not design a special refining stage for the detailed recovery of small targets. SH-DETR [32] introduces a multi-channel random coding module to efficiently encode diverse features and combines it with the parameter-free UPC-SimAM module to strengthen multi-scale feature fusion, achieving good comprehensive detection performance on datasets such as NEU-DET. Notably, Salience DETR [33] introduced a query filtering mechanism based on saliency guidance to reduce computational redundancy by filtering background queries. However, its significance prediction depends on the spatial distance between the query point and the target center. This implies a distance-based symmetry: query points with equal distances receive equal significance scores regardless of their local context or feature distribution. The closer a point is to the center of the target, the more likely it is to be regarded as a defect. Points at the same distance, regardless of which direction they are in, are treated equally. The features of small defects are usually asymmetrically distributed, which makes distance-based saliency prediction difficult to detect accurately and also vulnerable to interference from annotation noise. In addition, approaches such as the MLP-Mixer proposed by Tolstikhin et al. [34] employed position-agnostic token-mixing methods, in which the token-mixing layer applies globally shared weights to all image patches for mixing, retaining absolute positional information solely through additive positional encoding. However, this design, which is insensitive to spatial position, weakens the ability to model local spatial structures and may result in the loss of fine-grained spatial information that is useful for defect detection. These limitations provide motivation for researching more effective mechanisms for predicting the significance of small defects.

In conclusion, the existing methods still fail to address the three core issues of small defect detection: the disconnection between physical perception and task reasoning, the rough multi-scale feature fusion mechanism, and the insufficient sensitivity to the weak features of small defects. PG-SalDETR specifically proposes four modules: PGSPM, TSSP-Net, TSQRM, and LCASF. It simultaneously addresses the above limitations within a unified framework, which is also the core innovation of this paper.

3. Method

This work develops a steel plate surface defect detection method based on physically guided saliency and an asymmetric perception mechanism (physically guided salience DETR, PG-SalDETR), which aims to address the misdetection issues caused by weak features of small defects and inconsistent multi-scale semantics of traditional methods in complex imaging environments. Illustrated in Figure 1, the entire structure of this method follows the encoder–decoder architecture and is based on the improved Salience DETR [33].

3.1. PGSPM

To tackle the problem of small defect details being obscured in steel plate images because of intense reflections and uneven lighting, we have developed an end-to-end trainable saliency perception mechanism based on physical information. Unlike the existing methods based on Retinex, those methods treat physical priors as independent preprocessing steps (such as image enhancement before detection). In contrast, our PGSPM transforms the Retinex prior into a learnable guidance signal, which is directly embedded within the detection network. This guidance signal is jointly optimized with the detection task in an end-to-end manner, rather than being fixed. This system is designed to produce a saliency guidance map $G \in \mathbb{R}^{H \times W}$ based on imaging physics, which closely relates to defect characteristics and thus offers prior information to support the following detection components.

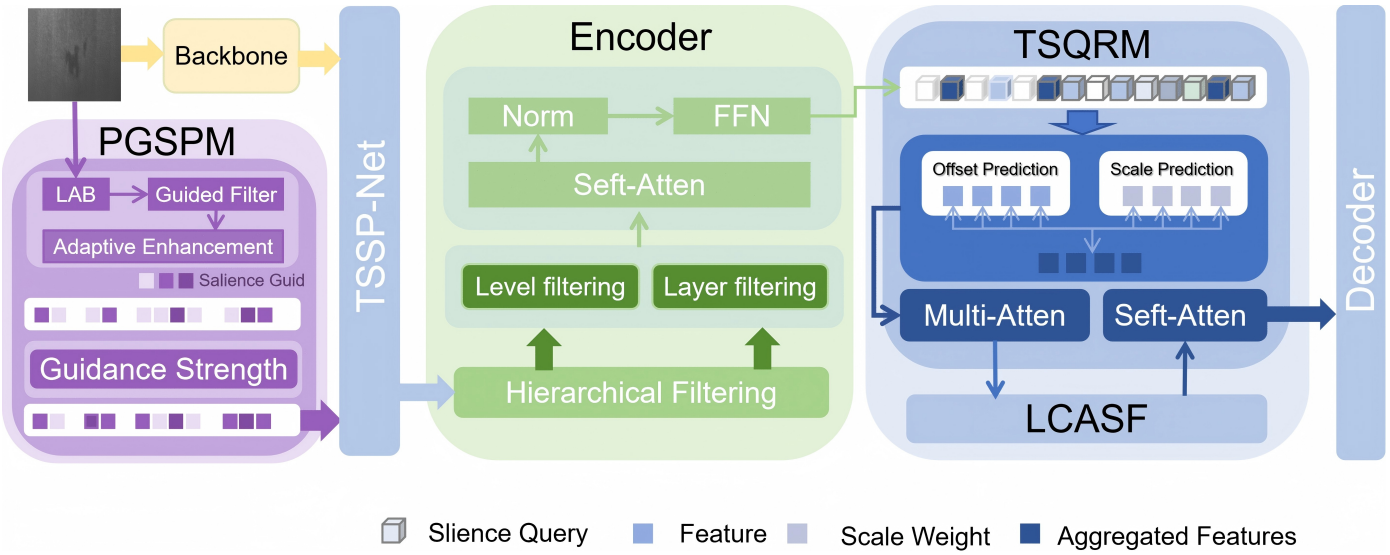


Figure 1. Overview of the proposed PG-SalDETR framework. The framework is built upon a Salience DETR backbone with a 6-layer encoder and a 6-layer decoder. The four novel modules, PGSPM, TSSP-Net, TSQRM, and LCASF, are integrated as plug-and-play components to guide feature extraction, query selection, refinement, and multi-scale fusion respectively, forming a physics-aware and saliency-enhanced detection pipeline. The guidance signal G generated by PGSPM is fed into two core modules: (1) TSSP-Net, where G is used to refine the generated baseline saliency scores; and (2) TSQRM, where it contributes to offset correction and scale selection.

First, the input RGB image $I_{\text{rgb}} \in \mathbb{R}^{3\times H\times W}$ is transformed into the LAB color space, from which the luminance component L_{ab} is extracted.

$$B, D = \text{multi_scale_guided_filter}(L_{ab}). \tag{1}$$

where B is the incident component and D is the reflected component. Through multi-scale guided filtering, it is possible to effectively separate global lighting changes from local texture details while preserving defect edges. This multi-scale guided filter adopts convolution kernels of three different scales, with sizes of 3 × 3, 5 × 5 and 7 × 7 respectively. The 3 × 3 kernel is used to retain fine-grained defect details; the 5 × 5 kernel is used to connect details with the global structure; the 7 × 7 kernel is used to capture global illumination variations. An adaptive gain enhancement strategy is used for the reflection component D. First, the gain coefficient is dynamically adjusted based on the local mean and standard deviation, while a gain map M_{gain} is generated to record the per-pixel gain distribution. A higher gain is applied to the low-contrast area to highlight the weak defect features. Then, noise level estimation is combined to suppress over-enhancement in high-noise regions, preventing noise from being mistakenly amplified as a false defect.

The enhanced reflection component D_{enh} is mapped to the basic saliency map using the Sigmoid function, achieving numerical normalization and initially activating the defect response.

$$S_{\text{base_guid}} = \sigma(D_{\text{enh}} \cdot \gamma). \tag{2}$$

where $\gamma = 8$ is the scaling factor, and σ is the sigmoid function, which can compress the value of D_{enh} to the (0, 1) interval to avoid response saturation caused by extreme values. Dynamic guidance intensity, noise suppression mask, and small target enhancement mask are introduced to finely modulate the basic saliency map, generating the final physical guidance signal:

$$S_{\text{norm}} = \frac{S_{\text{base_guid}} + M_{\text{small}} \cdot \beta_{\text{small}}}{S_{\text{base}}^{\max} + 10^{-6}} \tag{3}$$

$$G = S_{\text{norm}} \cdot g_s \cdot (1 - M_{\text{noise}} \cdot \beta_{\text{noise}}) \tag{4}$$

where g_s is the dynamic guidance intensity, used to globally control the activation intensity of the guidance signal; M_{noise} is the noise mask, estimated through local variance within a 15×15 sliding window and normalized by the Sigmoid function; β_{noise} is the noise penalty factor, which suppresses the response in high-noise regions by $1 - M_{\text{noise}} \cdot \beta_{\text{noise}}$, avoiding noise interference with the detection of small defects; and M_{small} is the small target mask. It is generated by combining multi-scale top-hat transforms (kernel sizes: 3, 5, and 7), Laplace high-frequency components, and 7×7 local contrast, followed by large-region suppression (kernel sizes: 11 and 15) and isolated peak enhancement. The small target enhancement factor, β_{small}, specifically boosts the saliency response of these small defects through $M_{\text{small}} \cdot \beta_{\text{small}}$. Finally, $S_{\text{base}}^{\max} + 10^{-6}$ stands for the global maximum value of the base saliency map and is used for normalization to keep the numerical range of G stable within the interval $(0, 1)$.

3.2. TSSP-Net

Conventional methods for predicting saliency scores depend on the symmetry of spatial distance: the significance of a query point is symmetrically determined by its Euclidean distance from the center of the target bounding box. The coordinates of the query point q are (x_q, y_q), the target center is $c = (x_c, y_c)$, and the Euclidean distance is $d(q, c) = \sqrt{(x_q - x_c)^2 + (y_q - y_c)^2}$. The significance function is Gaussian attenuation $S(q) = \exp(-d(q, c)^2 / 2\sigma^2)$, where σ is proportional to the target scale. If there are two query points q_1 and q_2 such that $d(q_1, c) = d(q_2, c)$, then $S(q_1) = S(q_2)$; this is spatial distance symmetry. This symmetry makes it difficult for them to detect the fine features that are asymmetrically distributed in small defects, and they are vulnerable to inaccurate labeling. To alleviate the negative effects of this symmetry, we propose the token sequence saliency perception network (TSSP-Net). Its architecture is shown in Figure 2.

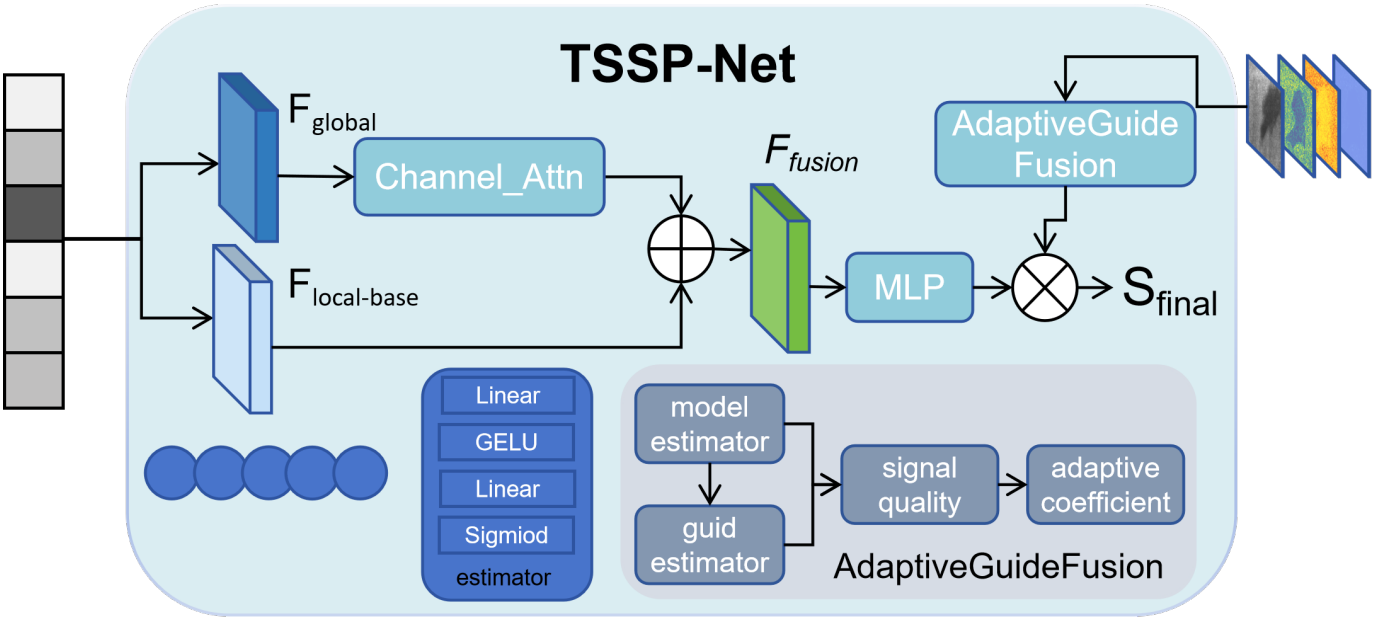


Figure 2. The architecture of TSSP-Net.

TSSP-Net adopts asymmetric structural designs in three key parts. The dual-branch module adopts an asymmetric architecture, where the global branch learns the overall semantics through MLP, while the local branch aggregates adjacent tokens without explicitly relying on any spatial coordinate information. Adaptive guided fusion asymmetrically processes model predictions and physical guidance, dynamically calculating fusion weights based on their respective credibility rather than using fixed or equal contribution weights. Finally, the final score fusion adopts the form of asymmetric residuals: the adaptive term is gated through uncertainty weights so that the physical guidance only takes effect when the model uncertainty is high. These designs alleviate the drawbacks of conventional saliency calculation in the detection of small defects and help improve detection performance.

3.2.1. Dual-Branch Feature Fusion

The token sequence $T \in \mathbb{R}^{B\times N\times D}$ produced by the Transformer encoder is fed into a dual-branch feature extraction module:

Global Branch: A multi-layer perceptron (MLP) is employed to learn a globally consistent semantic representation F_{global} from the tokens.

Local Branch: To capture local relationships among tokens and strengthen the subtle features of small defects, the features of each token's five neighboring tokens are aggregated to form the base local feature F_{local}. This aggregation does not depend on any spatial position information.

To emphasize the channel features in F_{global} that are important for defect detection, a cross-channel attention mechanism [34] (Channel-Attn) is applied to produce channel weights $W_c \in \mathbb{R}^{B\times N\times D_h}$. These weights are then used to weight and enhance the global features.

$$\begin{aligned} W_c &= \sigma(\text{MLP}_{\text{attn}}(F_{\text{global}})), \\ F_{\text{global-enhanced}} &= F_{\text{global}} \odot W_c. \end{aligned} \tag{5}$$

The improved global features and the local features are combined through direct concatenation, merging the benefits of both feature types to produce the fused feature $F_{\text{fusion}} \in \mathbb{R}^{B\times N\times 2D_h}$:

$$F_{\text{fusion}} = \text{concat}(F_{\text{global-enhanced}}, F_{\text{local}}). \tag{6}$$

The fused feature is then mapped through a multi-layer MLP prediction head to generate the model's own base saliency score $S_{\text{base}} \in \mathbb{R}^{B\times N\times 1}$.

3.2.2. Adaptive Guidance Fusion

Next, the G generated by PGSPM is introduced. First, G is adjusted to match the dimension of the token sequence, resulting in the guidance score S_{guide}. Subsequently, an Adaptive Guidance Fusion (AdaptiveGuideFusion) strategy is proposed to dynamically balance the model's own predictions and physical guidance based on their respective reliabilities. The detailed procedure is as follows:

First, G is adjusted to the same dimension as the token sequence, yielding the guidance score $S_{\text{guide}} \in \mathbb{R}^{B\times N\times 1}$.

Calculate the model confidence (model estimator). Based on the global feature F_{global}, the reliability of the model's own predictions is evaluated through the confidence estimation network. This network outputs confidence scores between 0 and 1, with values nearer to 1 representing greater prediction confidence.

$$C_{\text{conf}} = \text{MLP}_{\text{conf}}(F_{\text{global}}) \in \mathbb{R}^{B\times N\times 1}. \tag{7}$$

Compute the uncertainty weight. The model uncertainty is defined as

$$U = 1 - C_{\text{conf}}. \tag{8}$$

Evaluate the quality of the guidance signal (guidance estimator). Concatenate the global features and the guide scores, compute the reliability of the guidance signal via the guidance quality assessment network, and clip it to the range of $[0.3, 0.8]$ to avoid extreme values. This network evaluates how reliable the guidance signal is at each position, where values closer to 0.8 indicate higher guidance reliability. The uncertainty weight is refined based on the guidance quality. Then multiply this quality score by the uncertainty weight to achieve dynamic adjustment of the uncertainty weight:

$$\begin{aligned} Q_{\text{guide}} &= \text{clamp}(\text{MLP}_{\text{guide}}(F_{\text{global}} \oplus S_{\text{guide}}), 0.3, 0.8), \\ U &= U \odot Q_{\text{guide}}. \end{aligned} \tag{9}$$

To further exploit physical priors, we introduce two auxiliary variables generated by PGSPM, namely the M_{gain} and M_{noise}. The formulas for signal quality and the uncertainty weights after further modulation are as follows (signal quality):

$$\begin{aligned} Q_{\text{signal}} &= M_{\text{gain}} \odot (1 - M_{\text{noise}} \times 0.5), \\ U &= U \odot Q_{\text{signal}}. \end{aligned} \tag{10}$$

Adaptive feature fusion (adaptive coefficient). Compute the fusion weights to adaptively fuse the global features and the guidance features:

$$W_{\text{fusion}} = \sigma(\text{MLP}_{\text{fusion}}(F_{\text{global}} \oplus S_{\text{guide}})). \tag{11}$$

$$F_{\text{global-adaptive}} = (1 - W_{\text{fusion}}) \cdot F_{\text{global}} + W_{\text{fusion}} \cdot \text{MLP}_{\text{trans}}(S_{\text{guide}}). \tag{12}$$

$\text{MLP}_{\text{trans}}$ is a two-layer MLP with GELU activation, which projects the guidance score S_{guide} from a scalar value into a feature space matching the dimension of F_{global}, thereby enabling element-wise adaptive fusion. $W_{\text{fusion}} \in [0, 1]$ is dynamically generated based on the global feature F_{global} and the guide score S_{guide}. It determines the balance between the model's own prediction and physical guidance: when W_{fusion} approaches 1, the guiding features take the lead; when approaching 0, retain the original global features. This allows the model to place greater trust in physical guidance in reliable regions and rely on its own features when guidance is unreliable. The adaptively fused global features are concatenated with the local features again, and the adaptive scores are generated via the prediction head.

Compute the final score. The base score and the adaptive score are weighted and fused:

$$S_{\text{final}} = S_{\text{base}} + \text{MLP}(\text{concat}(F_{\text{global-adaptive}}, F_{\text{local}})) \cdot U. \tag{13}$$

The final output S_{final} serves as the saliency score of TSSP-Net, which is used for subsequent query selection and refinement. Formula (13) suppresses irrelevant features. The core lies in the uncertainty weight $U \in [0, 1]$: when the model's prediction confidence for the current token is relatively high ($U \to 0$), the residual term is suppressed, and the final score is almost solely determined by the base score S_{base}. At this time, the model relies on the discriminative features it has learned. When the model uncertainty is relatively high ($U \to 1$), the residual term is activated, and a physically guided correction signal is introduced. Irrelevant features usually lead to an increase in the uncertainty of model predictions, but the quality of physical guide signals, S_{guide} and S_{signal}, will further modulate U. If the

physical guidance is also unreliable, U will be reduced, thereby suppressing the erroneous correction signal.

3.3. TSQRM

In the Salience DETR detector, the salient queries after hierarchical filtering still suffer from two types of defects: (1) residual background noise, where queries incorrectly attend to non-defect regions, and (2) loss of fine-grained details in small defects, where delicate features are ignored. In contrast, standard deformable attention relies solely on query features to predict offset vectors, which are prone to offset errors for small defects with weak features. Furthermore, its fixed-scale selection strategy cannot be adapted dynamically.

To resolve these issues, this work proposes a two-stage query refinement mechanism (TSQRM). Through physics-guided offset correction and adaptive multi-scale feature aggregation, it gradually enhances the filtered salient query sequence $Q \in \mathbb{R}^{M \times D}$, where M denotes the number of queries and D denotes the feature dimension. The design is shown in Figure 3.

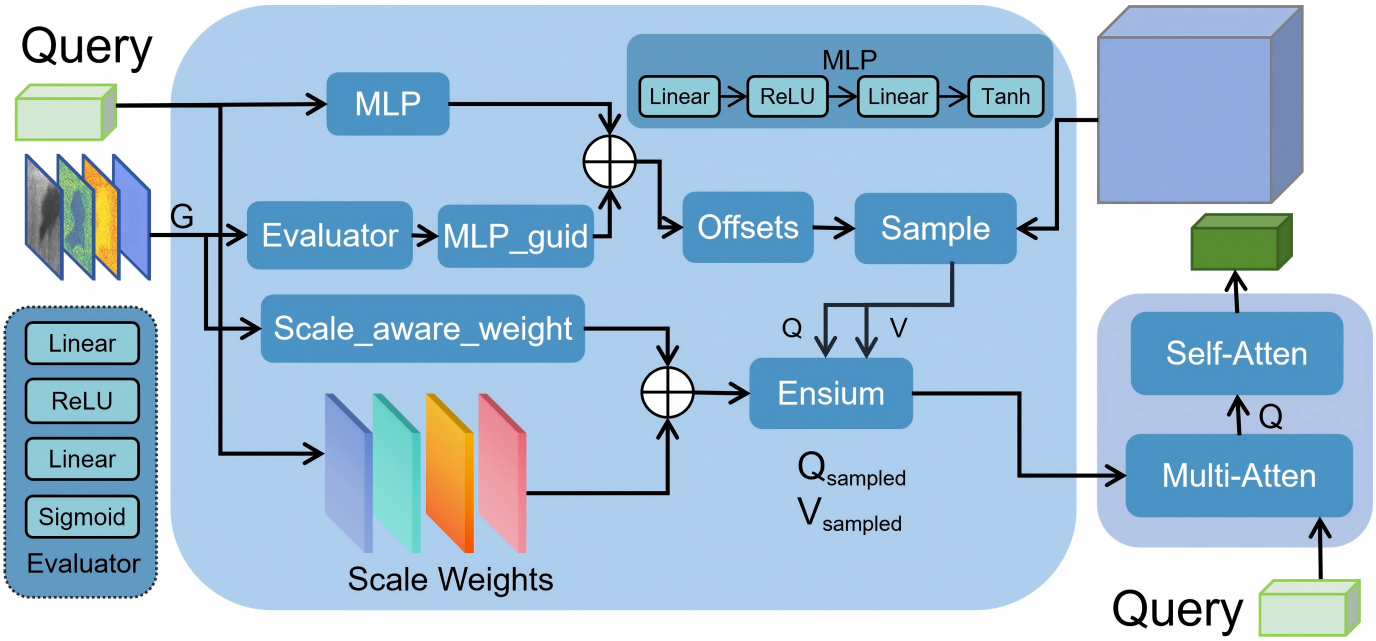


Figure 3. The architecture of TSQRM.

3.3.1. Physics-Guided Offset Correction

The offsets in standard deformable attention are predicted solely from query features. For small defects with weak features, these predictions are often inaccurate, resulting in sampling points falling into background regions. To mitigate this problem, we adopt the G generated by the PGSPM introduced in Section 2.1 to correct the offsets.

Given the input query sequence $Q \in \mathbb{R}^{B \times M \times C}$, where B denotes the batch size and $C = 256$ denotes the feature dimension, the basic sampling offsets are first generated by an offset prediction network composed of a two-layer MLP:

$$\Delta_{\text{base}} = \text{MLP}_{\text{offset}}(Q) \cdot 0.1 \in \mathbb{R}^{B \times M \times N \times 2}. \tag{14}$$

where $N = 4$ denotes the number of sampling points per query, and the coefficient 0.1 is used to constrain the initial magnitude of offsets, preventing sampling positions from deviating excessively from reference points in the early training stage.

To leverage physical prior information, the $G \in \mathbb{R}^{B \times 1 \times H \times W}$ generated by PGSPM is projected into the query space via bilinear interpolation, yielding the guidance intensity fea-

ture $S_{\text{guide}} \in \mathbb{R}^{B\times M\times 1}$ aligned with queries. Meanwhile, a two-layer confidence estimation network evaluates the reliability of model predictions for each query:

$$C_{\text{conf}} = \text{MLP}_{\text{conf}}(Q) \in \mathbb{R}^{B\times M\times 1}. \tag{15}$$

This network outputs confidence scores between 0 and 1, with values nearer to 1 representing greater prediction confidence. For difficult regions with low confidence, we introduce a hard-example enhancement factor $\gamma = 1.2$. The reliability of the guidance signal $Q_{\text{guide}} \in \mathbb{R}^{B\times M\times 1}$ is computed via a guidance quality evaluation network (Evaluator). The final guidance weight is then calculated as follows:

$$W_{\text{guide}} = (1 - C_{\text{conf}}) \cdot Q_{\text{guide}} \cdot S_{\text{guide}} \cdot \gamma. \tag{16}$$

After concatenating the query features and the guidance intensity features, the guidance offset adjustment network generates the offset corrections:

$$\Delta_{\text{guide}} = \tanh\big(\text{MLP}_{\text{guid}}(Q \oplus S_{\text{guide}})\big) \cdot \alpha_{\text{guid}} \in \mathbb{R}^{B\times M\times N\times 2}. \tag{17}$$

where $\oplus$ denotes the feature concatenation operation, the tanh function constrains the output within the range of $[-1, 1]$, and $\alpha_{\text{guid}} = 0.05$ is the guidance offset coefficient, which ensures the correction is a small increment and avoids training instability. The final sampling offset is the sum of the base offset and the guidance correction:

$$\Delta_{\text{final}} = \Delta_{\text{base}} + \Delta_{\text{guide}} \cdot W_{\text{guide}}. \tag{18}$$

3.3.2. Adaptive Multi-Scale Feature Aggregation

After offset correction, features at different scales contribute differently to defect detection: high-resolution features preserve fine details but carry weak semantics, while low-resolution features possess strong semantics but lack detailed information. To tackle this problem, we design a scale selection network, enabling each query to adaptively select the most suitable scale from multi-scale features.

Let the multi-scale feature maps be $F_l \in \mathbb{R}^{B\times D\times H_l\times W_l}$, where $l = 1, 2, 3, 4$ denotes four resolution levels from high to low. The scale selection network first predicts the base scale weights and then refines them using the guidance signal (Scale_aware_weight):

$$L_{\text{scale}} = f_{\text{scale}}(Q) + S_{\text{guide},l} \cdot 0.1. \tag{19}$$

where $f_{\text{scale}}(\cdot)$ denotes the scale prediction network, and $S_{\text{guide},l}$ represents the guidance intensity feature at the l-th scale. The coefficient 0.1 ensures that the guidance modulation only exerts a moderate effect and does not dominate the network's own predictions. This modulation mechanism increases the weights of scales that are primarily affected by defects.

Based on the refined offsets Δ_{final}, sparse sampling is performed on the feature maps at the corresponding scales to obtain the sampled feature set for each scale $F_{\text{sample},l} \in \mathbb{R}^{B\times M\times N\times D}$. Linear projection is applied to the sampled features to obtain multi-scale keys K_l and values V_l. Scale weights are obtained by normalizing L_{scale} along the scale dimension using Softmax, based on which multi-scale K and V are aggregated:

$$\begin{aligned} K_{\text{sampled}} &= \sum_{l=1}^{4} \text{Softmax}(L_{\text{scale}})_l \cdot K_l, \\ V_{\text{sampled}} &= \sum_{l=1}^{4} \text{Softmax}(L_{\text{scale}})_l \cdot V_l. \end{aligned} \tag{20}$$

Finally, a multi-head cross-attention mechanism is introduced to establish semantic associations between the query feature Q and the aggregated global features. Meanwhile, an efficient self-attention mechanism is adopted to enhance semantic correlation among queries while keeping computational overhead under control. The final output is the refined query feature Q_{refined}, which is passed to the next layer of the decoder.

3.4. LCASF

Traditional multi-scale feature fusion methods adopt a fixed-range fusion strategy, which leads to a semantic gap between features of different scales and makes it difficult for high-level semantic information to effectively guide low-level features, thus causing the misdetection of small defects. To address this issue, a dynamic cross-scale fusion module (LCASF) is proposed based on the channel-spatial dual attention mechanism. Its core innovative dynamic cross-scale fusion strategy adaptively adjusts the fusion range and scale weights during the training process, which effectively bridges the semantic differences among multi-scale features and enhances the representation capability of small defect features.

The discriminability of features is enhanced via the well-established channel-spatial dual attention mechanism (channel attention learns the weight W_{ca} by fusing average and max pooling, while spatial attention generates the weight W_{sal} through lightweight convolution, yielding the enhanced feature $F_l^{\text{sa}} \in \mathbb{R}^{B\times C\times H_l\times W_l}$).

The fusion range is dynamically adjusted. As the training epoch increases, the fusion range is linearly narrowed from ±2 scales to ±1 scale.

$$\text{current_delta} = 2.0 - 1.0 \times (\text{epoch}/\text{total_epochs}). \tag{21}$$

Only scale differences δ satisfying $|\delta| \leq \text{current_delta}$ are retained for fusion. Different fusion weights are assigned to high-resolution (small-scale) and low-resolution (large-scale) features, which decay smoothly with the training epochs.

$$\begin{aligned} w_{\text{small}} &= 2.0 - 0.5 \times (\text{epoch}/\text{total_epochs}), \\ w_{\text{large}} &= 1.5 - 0.5 \times (\text{epoch}/\text{total_epochs}). \end{aligned} \tag{22}$$

For the current-scale feature F_l and adjacent-scale feature $F_{l+\delta}$ (adjusted to the same resolution via bilinear interpolation), they first undergo a convolutional transformation $\text{Conv}(\cdot)$ as a 3×3 convolution, followed by dynamic Dropout regularization (with the dropout rate gradually increasing from 0.05 to 0.15). This dynamic strategy aims to retain more features during the early stages of training to promote thorough learning and to prevent overfitting in the later stages through stronger regularization. The fusion weights are generated by a gating network from the global feature of the current scale $G_l = \text{AvgPool}(F_l)$:

$$W_{\text{gate}} = \sigma(\text{Linear}(G_l)) \in \mathbb{R}^{B\times C\times 1\times 1}. \tag{23}$$

The final fused feature is given by

$$F_l^{\text{fusion}} = F_l^{\text{sa}} + \begin{cases} w_{\text{small}} \cdot W_{\text{gate}} \cdot \text{Conv}(F_{l+\delta}^{\text{resized}}), & \delta < 0 \\ w_{\text{large}} \cdot W_{\text{gate}} \cdot \text{Conv}(F_{l+\delta}^{\text{resized}}), & \delta > 0 \\ W_{\text{gate}} \cdot \text{Conv}(F_{l+\delta}^{\text{resized}}), & \delta = 0 \end{cases}. \tag{24}$$

Here, G_l stands for the global feature at the current scale, while $F_{l+\delta}^{\text{reg}}$ refers to the feature that will be combined following dynamic dropout regularization.

Finally, the fused features from all scales are aggregated to output features with consistent dimensions.

4. Experiment

4.1. Experimental Setup

Multiple experiments were conducted to thoroughly evaluate our method's performance. We first introduce the datasets used in the experiments and their characteristics, then elaborate on the specific implementation scheme and parameter settings, and finally present the evaluation metrics adopted.

4.1.1. Dataset

To comprehensively evaluate the developed method, two publicly available metal surface defect detection datasets are selected, namely NEU-DET [35] and GC10-DET [36]. In this work, we randomly split the official training set of GC10-DET into a training subset containing 1530 images and a test subset containing 769 images using a fixed random seed (42) to ensure reproducibility. These two datasets are representative and complementary and are employed for quantitative and qualitative experiments, respectively. Detailed information on the datasets is listed in Table 1.

Table 1. Overview of experimental datasets.

Dataset	Ann	Class	Train	Test	Resolution
NEU-DET [35]	4189	6	1200	600	200 × 200
GC10-DET [36]	3194	10	1530	769	2048 × 1000

Ann: number of annotations. Train/Test: number of images in training and testing sets.

Illustrated in Figures 4 and 5, the defect sizes in the NEU-DET are mostly clustered within the medium to small range, making it highly suitable for evaluating the consistency of the model on common-size defects. The defect size distribution of the GC10-DET dataset is wider, and the proportion of extremely small defects, such as pixel-level defects, is significantly higher. This presents a challenge for the model to accurately detect and localize features at these extreme scales, further highlighting the effectiveness and adaptability of our approach under various defect sizes and industrial contexts.

4.1.2. Implementation Details

All experiments were conducted on a compute server equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), an Intel Xeon Gold 5418Y processor (10 cores), and 120 GB of system memory. We used ResNet-50 as the backbone network for initialization, employed the Adam optimizer, set the initial learning rate to 1×10^{-4}, and used cosine annealing for learning rate scheduling. The batch size was set to 4, and the number of data loading workers was set to 4. All comparison models were evaluated under the same hardware and software settings mentioned above to ensure fair comparisons. Because GC10-DET has a larger number of images and is closer to the actual industrial scene, it requires more rounds to converge. NEU-DET and GC10-DET adopt different numbers of training rounds (70 vs. 100). On each dataset, all models use the same number of rounds setting.

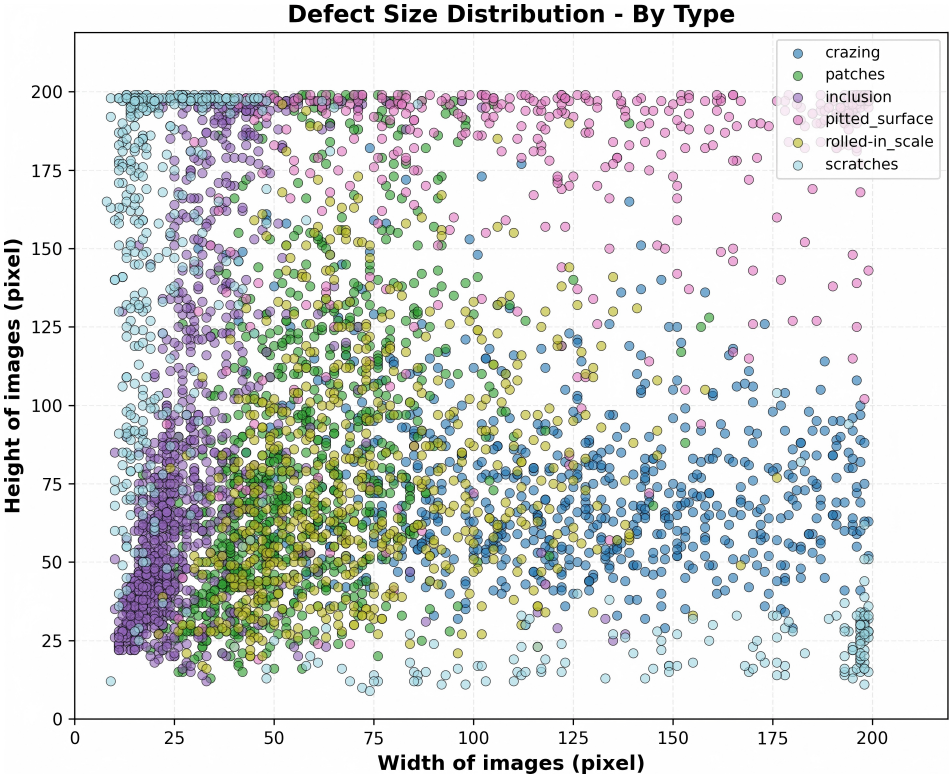


Figure 4. Size distributions of different defect types in the NEU-DET dataset.

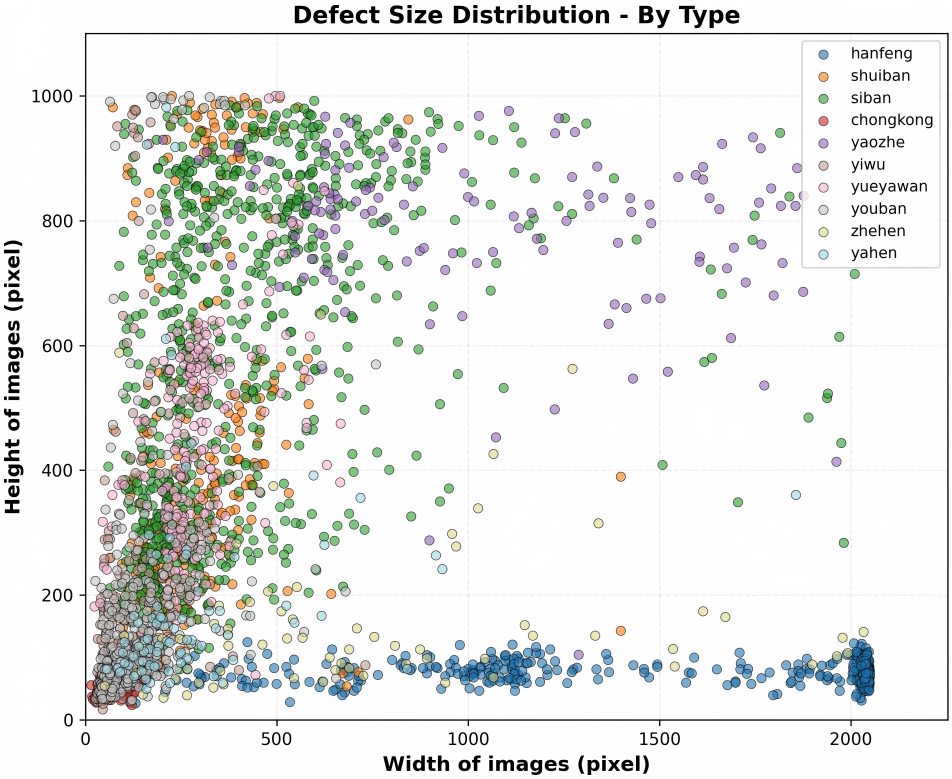


Figure 5. Size distributions of different defect types in the GC10-DET dataset.

4.1.3. Evaluation Criteria

To thoroughly assess the improvement effect of the developed method on the final defect detection performance, we adopted a series of widely accepted evaluation metrics.

AP (Average Precision) is the average precision across IoU thresholds ranging from 0.50 to 0.95 and is the main indicator for measuring overall detection accuracy. AP_{50} and

AP_{75} are the average precision at IoU thresholds of 0.50 and 0.75, respectively, used to evaluate the model's performance under lenient and stringent localization requirements. GFLOPs evaluates a model's computational cost, denoting the billions of floating-point operations it executes per second. The parameter count (Params) reflects a model's size and architectural complexity, thereby revealing its learning capacity. To evaluate the real-time feasibility of our method in industrial deployments, we report the inference speed in terms of FPS (frames/s).

To further analyze the model's sensitivity to defects of different sizes, we adopted the following two scale division criteria:

Scale division based on absolute pixel area:

Considering that the NEU-DET dataset has a fixed image resolution and a clear defect size distribution, we adopt the absolute threshold standard of the COCO dataset: AP_S represents the average precision of small defects with a bounding box pixel area of less than 32^2 pixels; AP_M represents the average precision of medium defects with a bounding box pixel area of greater than or equal to 32^2 pixels but less than 96^2 pixels; and AP_L represents the average precision of large defects with a bounding box pixel area of greater than or equal to 96^2 pixels.

Scale division based on relative image area:

Given the varying image resolutions in the GC10-DET dataset and its closer resemblance to real-world industrial scenarios, we propose an evaluation criterion based on relative area, which has a more clearly defined physical meaning. Relative area refers to the proportion of the defect area to the total image area. AP_S represents the average precision of small defects with a relative area less than 0.5%; AP_M represents the average precision of medium defects with a relative area greater than or equal to 0.5% but less than or equal to 5%; and AP_L represents the average precision of large defects with a relative area greater than or equal to 5%. AR represents the average recall rate when detecting a maximum of 100 bounding boxes per image. The selection of these thresholds is detailed in Table 2. AR_S, AR_M and AR_L represent the average recall rates for targets at different scales.

Table 2. Defect distribution under different relative area thresholds.

Small Defects				Medium Defects			
Threshold (%)	**Count**	**Proportion (%)**	**Types**	**Threshold (%)**	**Count**	**Proportion (%)**	**Types**
area < 0.3	420	13.1	8	$0.3 \leq$ area ≤ 3	1045	32.7	9
area < 0.5	758	23.7	9	$0.5 \leq$ area ≤ 5	1082	33.9	10
area < 0.8	901	28.2	9	$0.8 \leq$ area ≤ 8	1407	44.1	10

Note : The 0.5% threshold was chosen because it covers almost all small defects. A stricter 0.3% threshold would exclude some small defects, while a looser 0.8% threshold would begin to include medium defects (e.g., siban). The 5% threshold was chosen because it covers all 10 defect types in the medium category while excluding substantially larger defects.

4.1.4. Statistical Reliability

We compared the performance of PG-SalDETR with that of the baseline Salience DETR on the NEU-DET and GC10-DET datasets. For each dataset, we fixed the training/test split. All experiments were conducted using the same data split. We conducted three independent training runs using different random seeds (42, 123, and 2025) for model initialization and training, while keeping the data split constant. The reported metrics are the mean and standard deviation across the three runs. To assess statistical significance, we performed paired *t*-tests on the results of PG-SalDETR and the baseline. On NEU-DET, AP improved from 0.401 ± 0.005 to 0.439 ± 0.004, and AP_S improved from 0.319 ± 0.006 to 0.347 ± 0.004; on GC10-DET, AP increased from 0.345 ± 0.005 to 0.371 ± 0.003, and AP_S increased from

0.132 ± 0.004 to 0.171 ± 0.003. These improvements are statistically significant (paired t-test, $p < 0.01$). We acknowledge that the number of independent replicate experiments conducted so far is limited (three), and we plan to further validate the robustness of our conclusions through additional runs in the future.

4.2. Comparison Experiment

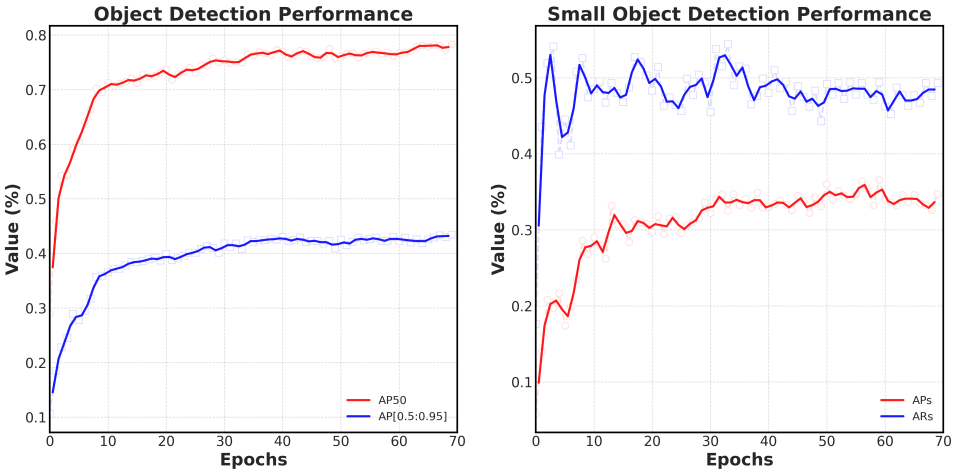
To thoroughly assess the effectiveness of the developed method for detecting surface defects on steel plates, we compared it with seven leading object detection frameworks using the NEU-DET and GC10-DET datasets. As presented in Tables 3–5, a comprehensive comparison was conducted between the proposed method and seven advanced methods. The compared methods cover several mainstream architectures: the one-stage anchor-based YOLOv8 and YOLOv11; the classic two-stage detector Faster R-CNN [37]; and a variety of Transformer-based detectors, including Cascade-DETR [38], Deformable DETR [29], RT-DETR [39], and Salience DETR [12]. We use the COCO standard evaluation metrics, with particular attention to average precision (AP), average precision for small defects (AP_S), and average recall for small defects (AR_S), which are crucial for the accurate location of minute defects in steel plates.

Table 3. Detailed experimental comparison of various models on NEU-DET.

Model	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	AR_S	AR_M	AR_L
YOLOv8	70	0.419	0.748	0.309	0.278	0.516	0.517	0.613	0.397	0.509	0.632
YOLOv11	70	0.434	0.750	0.384	0.320	0.470	0.506	0.632	0.346	0.579	0.655
Faster R-CNN [37]	70	0.384	0.763	0.345	0.311	0.329	0.464	0.559	0.457	0.525	0.589
Cascade-DETR [38]	70	0.244	0.437	0.241	0.176	0.189	0.326	0.683	0.413	0.610	0.753
Deformable DETR [29]	70	0.435	0.750	0.439	0.357	0.381	0.554	0.648	0.485	0.650	0.782
RT-DETR [39]	70	0.361	0.668	0.353	0.037	0.129	0.373	0.654	0.120	0.334	0.669
Salience DETR [12]	70	0.401	0.731	0.398	0.319	0.340	0.476	0.519	0.453	0.512	0.638
Ours	70	0.439	0.780	0.441	0.347	0.372	0.536	0.570	0.493	0.524	0.663

As shown in Table 3 and Figure 6a, our method achieved the highest overall average precision (AP = 0.439), demonstrating that its ability to detect and classify various defects on steel plate surfaces is superior. Compared with the base model, Salience DETR, our proposed model improves the average precision for small defects by 3.8%. Table 3 and Figure 6b illustrate that, on the most challenging small target defect detection (AP_S) metric, our method achieved an excellent performance of 0.347. Figure 7, demonstrates that the PG-SalDETR model effectively enhances the detection precision for small defect carving compared to the base model. Although Deformable DETR achieved the highest AP_S of 0.357, which is slightly higher than our method (0.347), its average recall for small defects (AR_S = 0.485) is lower than that of our method (0.493). This indicates that Deformable DETR's higher precision comes at the cost of missing more small defects. This precision-recall trade-off is critical for industrial applications, as the cost of missing defects (false negatives) is far higher than that of false positives. The root cause lies in the fact that while Deformable DETR's sparse sampling mechanism is efficient, it tends to focus on the most prominent features, potentially overlooking weak signals from minor defects. RT-DETR exhibits more severe limitations: its AP_S is only 0.037 and its AR_S is only 0.120, meaning it detects almost no small defects. This stems from RT-DETR's real-time-oriented architecture, which significantly downsamples features to reduce computational costs, inevitably losing information about small defects at an early stage. In contrast, our method achieves optimal results in both overall average accuracy and localization accuracy, significantly reducing

the defect miss rate, thereby demonstrating the comprehensive performance advantages of PG-SalDETR in the accuracy-recall trade-off.



(**a**) (**b**)

Figure 6. Comprehensive performance comparison on the NEU-DET dataset. (**a**) Visualization of defect detection AP and AP_{50}. (**b**) Visualization of AP_S and AR_S for small defect detection.

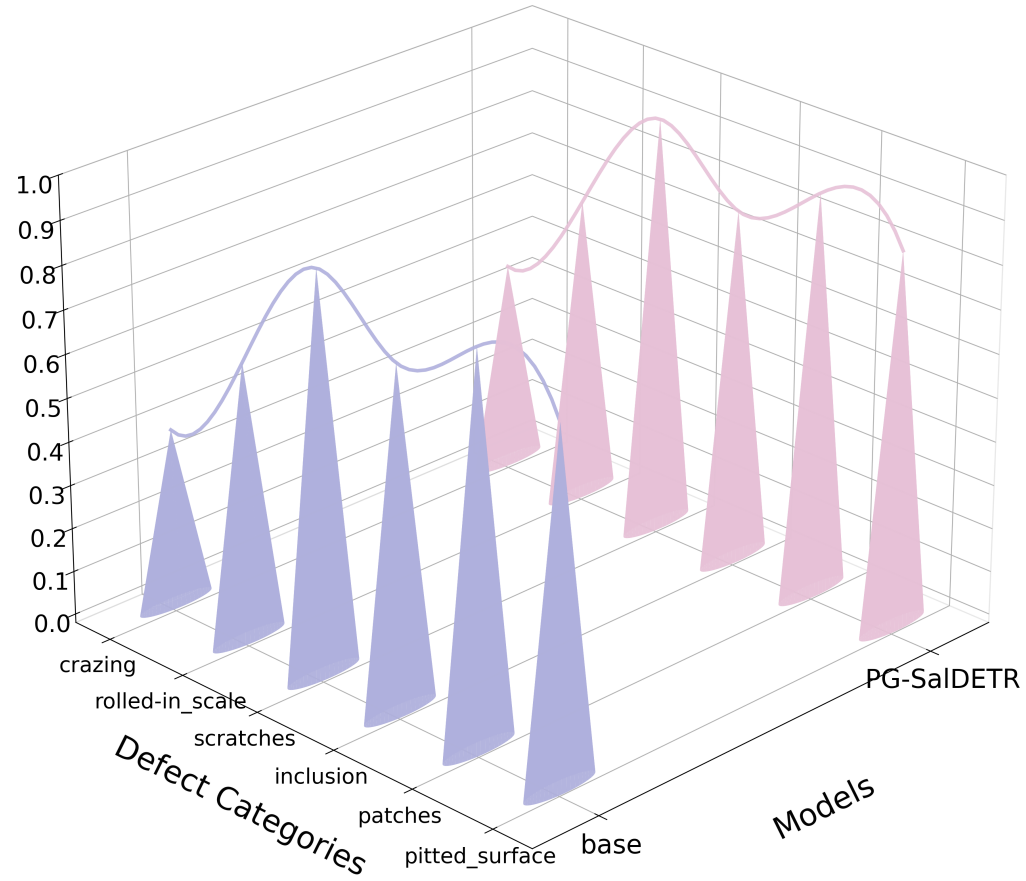


Figure 7. Visualization of detection accuracy for different defect types on the NEU-DET dataset.

Table 4 presents the detection results of different models on six defect categories using the NEU-DET dataset. For the category of small defects, crazing is the most difficult to detect. Its cracks are in the form of fine lines, extremely narrow in width, and occupy very few pixels in the image. The model presented in this paper achieves an AP_{50} of 0.520 for this category, which is significantly higher than that of Deformable DETR (0.442) and YOLOv11

(0.422). This indicates that PG-SalDETR has a stronger feature extraction ability for defects of weak and slender structures. For the "Rolled-in_Scale" defect category, our model also achieved the best result of 0.659, outperforming Faster R-CNN (0.647) and YOLOv11 (0.630), further validating the effectiveness of this method in detecting small-scale defects. For medium-scale defect categories, the model developed in this work also demonstrates stable performance. The AP_{50} for "patches" defects reached 0.919, comparable to YOLOv8 (0.917) and Faster R-CNN (0.915). The defect detection of the "pitted surface" category reached 0.853. Although slightly lower than 0.861 of Faster R-CNN, the gap is small. These results demonstrate that the method developed in this work does not sacrifice its ability to identify defects of other scales to enhance the detection of small defects. In the Scratches category, Faster R-CNN achieved the highest score of 0.970.

Table 4. Comparison of AP_{50} performance across different models for various defect categories on the NEU-DET dataset.

Model	YOLOv8	YOLOv11	Faster R-CNN [37]	Cascade-DETR [38]	Deformable DETR [29]	RT-DETR [39]	Salience DETR [12]	Ours
crazing	0.389	0.422	0.422	0.328	0.442	0.268	0.399	0.520
inclusion	0.800	0.792	0.789	0.454	0.780	0.660	0.782	0.782
patches	0.917	0.903	0.915	0.512	0.881	0.890	0.887	0.919
pitted_surface	0.827	0.823	0.861	0.447	0.843	0.715	0.851	0.853
rolled-in_Scale	0.617	0.630	0.647	0.351	0.631	0.570	0.570	0.659
scratches	0.939	0.931	0.970	0.532	0.923	0.905	0.897	0.947
AP_{50}	0.748	0.750	0.731	0.437	0.750	0.668	0.731	0.780

PG-SalDETR, YOLOv8, and YOLOv11 were 0.947, 0.939, and 0.931, respectively. Although the method developed in this work did not achieve the best performance, it remains at the leading level, with the gap between it and the best methods falling within an acceptable range. Overall, PG-SalDETR achieved the overall optimal AP_{50} on the NEU-DET dataset, particularly making breakthroughs in the two minor defect categories of crazing and roller-in_scale. This method both enhances small target detection and maintains competitiveness in other categories, demonstrating excellent performance balance and generalization ability.

Our method's significant advantage in reducing the misdetection of small defects is directly confirmed by the GC10-DET results (Table 5). Our model achieves a mean accuracy of 0.171 for small defects, significantly outperforming all comparative models. Compared with YOLOv8 and YOLOv11, our method achieves nearly twice the mean accuracy for small defects, and compared to the two variants of DETR, Deformable DETR and Salience DETR, our method achieves performance improvements of up to 5.4% and 3.9%, respectively. This result strongly demonstrates that the proposed method can effectively reduce the misdetection rate of small defects in complex backgrounds. More importantly, regarding the small target recall rate AR_S, which directly impacts the misdetection rate, we achieve an outstanding score of 0.241, marking a significant 2% increase over Salience DETR (0.221). Figure 8 compares the detection accuracy of our proposed model and the baseline model across different defect categories.

Table 5. Detailed experimental comparison of various models on GC10-DET dataset.

Model	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR	AR_S	AR_M	AR_L
YOLOv8	100	0.348	0.648	0.320	0.163	0.331	0.361	0.410	0.216	0.418	0.507
YOLOv11	100	0.361	0.670	0.332	0.157	0.348	0.378	0.404	0.211	0.416	0.509
Faster R-CNN [37]	100	0.286	0.633	0.222	0.118	0.305	0.405	0.423	0.190	0.434	0.526
Deformable DETR [29]	100	0.344	0.661	0.315	0.117	0.326	0.356	0.448	0.235	0.456	0.556
RT-DETR [39]	100	0.361	0.665	0.354	0.146	0.346	0.476	0.539	0.234	0.469	0.658
Salience DETR [12]	100	0.345	0.657	0.332	0.132	0.305	0.422	0.460	0.221	0.427	0.519
Ours	100	0.371	0.704	0.369	0.171	0.353	0.461	0.472	0.241	0.450	0.522

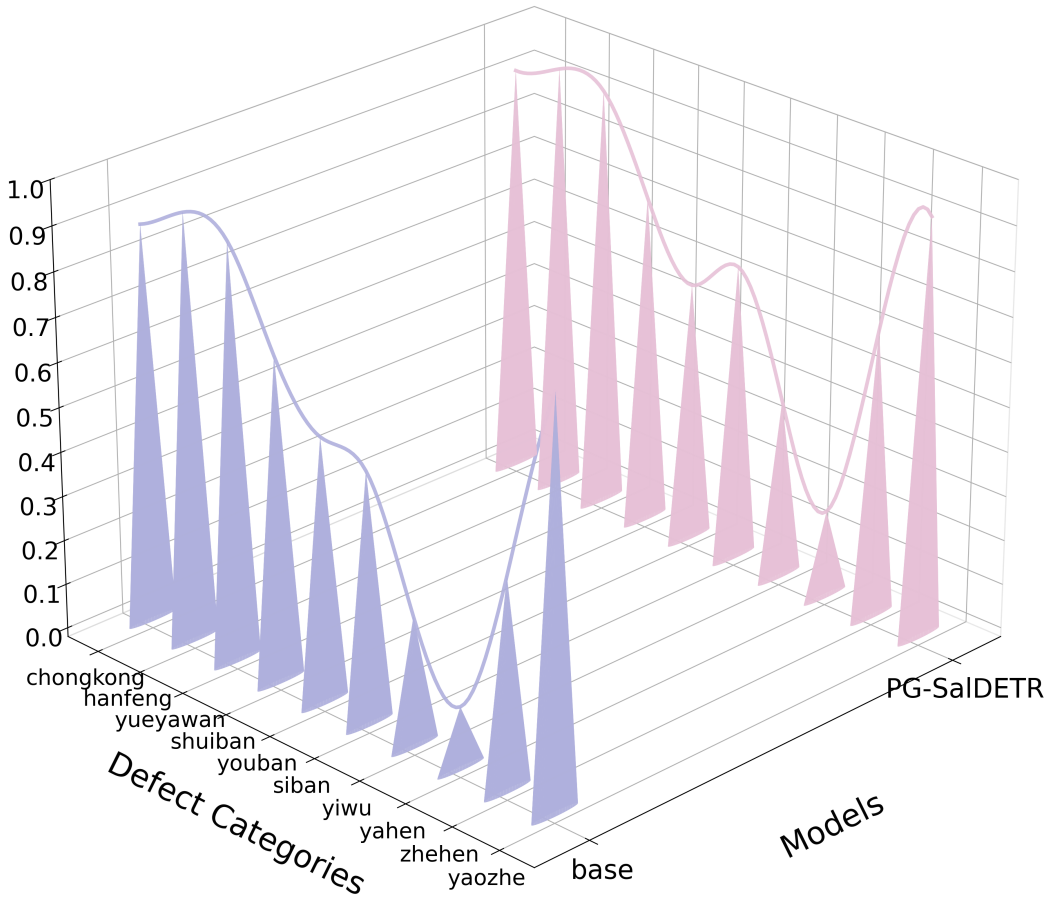


Figure 8. Visualization of detection accuracy for different defect types on the GC10-DET dataset.

Figure 9 presents typical detection results of our PG-SalDETR model on the GC10-DET dataset. As illustrated, the model effectively identifies pixel-level tiny defects that are often missed, such as "chongkong" and "hanfeng", while assigning high confidence scores. This clearly demonstrates why our approach attains a top small target average precision (AP_S) of 0.171 and notably outperforms others in small target recall rate (AR_S), achieving 0.241.

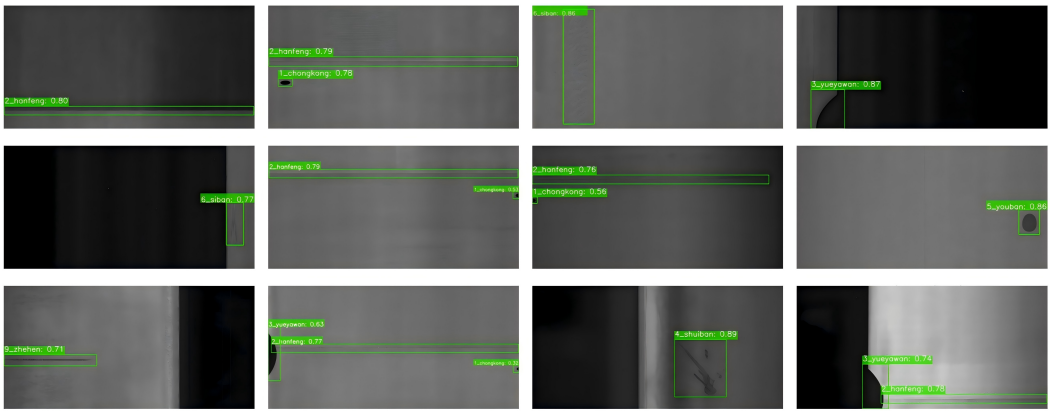


Figure 9. Visualization of detection results.

4.3. Ablation Experiment

To assess the effectiveness of each key component introduced in this paper, we performed a systematic ablation study using the NEU-DET dataset. Beginning with Salience DETR as the baseline model, we progressively incorporated the proposed modules. The outcomes are shown in Table 6.

Table 6. Ablation study results on NEU-DET dataset.

Module Configuration					Performance Metrics								
A	**B**	**C**	**D + C**	**D + A**	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AR**	**AR_S**	**GFLOPs**	**Params (M)**	**FPS (Frame/s)**
✗	✗	✗	✗	✗	0.401	0.731	0.398	0.319	0.519	0.453	254	56.89	16.73
✓	✗	✗	✗	✗	0.412	0.756	0.398	0.311	0.532	0.468	312	57.52	13.28
✗	✓	✗	✗	✗	0.407	0.735	0.398	0.323	0.532	0.454	258	59.52	15.29
✗	✗	✓	✗	✗	0.409	0.745	0.412	0.330	0.530	0.461	259	57.08	15.32
✓	✓	✗	✗	✗	0.418	0.750	0.407	0.316	0.536	0.465	316	60.15	11.74
✗	✓	✓	✗	✗	0.414	0.748	0.414	0.332	0.545	0.459	263	59.71	13.88
✓	✗	✓	✗	✗	0.420	0.767	0.416	0.323	0.543	0.473	317	57.70	11.77
✓	✓	✓	✗	✗	0.425	0.766	0.421	0.329	0.547	0.473	321	60.34	11.10
✓	✓	✓	✓	✗	0.436	0.778	0.435	0.335	0.556	0.485	321	60.42	10.87
✓	✓	✓	✓	✓	0.439	0.780	0.441	0.347	0.570	0.493	321	60.61	10.51

✓: Enabled, ✗: Disabled. A: TSQRM, B: LCASF, C: TSSP-Net, D + C: PGSPM + TSSP-Net, D + A: PGSPM + TSQRM. PGSPM is a guidance signal generator and not a standalone module; it is evaluated in combination with TSSP-Net (D + C) or TSQRM (D + A).

4.3.1. Independent Contributions of Individual Modules

Without our proposed module, the baseline model achieved a comprehensive detection accuracy of 0.401, demonstrating the capabilities of the basic architecture. After adding the TSQRM module, the overall average accuracy (AP) improved to 0.412 (+1.1%), and the average precision under lenient criteria (AP_{50}) improved to 0.756 (+2.5%). This indicates that TSQRM can effectively enhance the model's initial localization capability and classification confidence by optimizing the significance queries obtained through hierarchical filtering in a phased manner. Relative to the baseline model, GFLOPs increased to 312 (+58), and the number of parameters increased to 57.52 M (an increase of 0.63 M).

When the LCASF module was added alone, AP increased to 0.407 (+0.6%), verifying its ability to improve multi-scale object detection through dynamic cross-scale fusion. However, the minor defect metrics have only improved slightly. This is not a design failure

of LCASF but rather because the hierarchical filtering mechanism in the baseline model has filtered out the low-significance regions (including small defects) before the features enter LCASF. LCASF retains details during the multi-scale fusion process, but it cannot restore the information that has been discarded by the filtering mechanism. Therefore, when used alone, its ability to prevent features from being overwhelmed has not been fully exploited.

After adding the TSSP-Net module alone, the AP and AP_{75} of the model reached 0.409 (+0.8%) and 0.412 (+1.4%), respectively, and the AP_S increased to 0.330 (+1.1%), both of which were better than the baseline. This demonstrates that TSSP-Net helps the network better model and identify small defect regions, laying a feature foundation for subsequent processing. This module has initially improved the ability to detect small defects without significantly increasing the computational burden.

PGSPM, as a guiding signal generator in the model, cannot work independently. Therefore, to evaluate its effectiveness, we combined PGSPM with TSQRM or TSSP-Net, respectively. It should be noted that the performance improvements reported below reflect the synergistic effects between PGSPM and the other modules, rather than PGSPM's standalone contribution. As shown in Table 7, when PGSPM is combined with TSQRM, AP_S increases from 0.311 to 0.323 (+1.2%), and AR_S increases from 0.468 to 0.476 (+0.8%); when PGSPM is combined with TSSP-Net, AP_S increased from 0.330 to 0.336 (+0.6%), and AR_S increased from 0.461 to 0.473 (+1.2%). These results demonstrate that PGSPM consistently improves small defect detection performance when used in conjunction with TSQRM or TSSP-Net, indicating that it plays a beneficial role in addressing the disconnect between physical perception and task reasoning. Additionally, it facilitates robust detection of small defects under complex lighting conditions.

Table 7. Ablation study results of PGSPM on NEU-DET dataset.

TSQRM	TSSP-Net	PGSPM	AP	AP_{50}	AP_{75}	AP_S	AR	AR_S
✓	✗	✗	0.412	0.756	0.398	0.311	0.532	0.468
✓	✗	✓	0.415	0.758	0.404	0.323	0.546	0.476
✗	✓	✗	0.409	0.745	0.412	0.330	0.530	0.461
✗	✓	✓	0.420	0.757	0.426	0.336	0.539	0.473

To verify the generalization ability of PGSPM, we conducted ablation experiments on the GC10-DET dataset. Based on TSSP-Net, the performance differences with or without the introduction of PGSPM were compared. As shown in Table 8, after the introduction of PGSPM, AP increased from 0.351 to 0.355, AP_S from 0.140 to 0.151 (+1.1%), and AR_S from 0.224 to 0.229. These results confirm that PGSPM can stably improve the performance of small defect detection on the GC10-DET dataset, indicating its good generalization ability on different datasets.

Table 8. Ablation study results on GC10-DET dataset.

TSSP-Net	PGSPM	AP	AP_{50}	AP_{75}	AP_S	AR	AR_S
✓	✗	0.351	0.664	0.346	0.140	0.463	0.224
✓	✓	0.355	0.676	0.354	0.151	0.465	0.229

4.3.2. Analysis of Module Combination Effects

The combination of TSQRM and TSSP-Net achieved the best pairwise performance (AP = 0.420), representing a 0.8% improvement over TSQRM alone and a 1.1% improvement over TSSP-Net alone. This indicates a synergy between TSQRM and TSSP-Net. TSQRM

enhances global feature representation through a query-response mechanism, while TSSP-Net preserves small defect regions; the two complement each other at the feature extraction level. The combination of LCSAF and TSSP-Net performed best on the AP_S metric (0.332), indicating that this combination has advantages for small object detection. The combination of TSQRM and LCASF improves AP and AP_{75} to 0.418 and 0.407, respectively. Compared to the single-module TSQRM, LCASF effectively enhances boundary localization accuracy (AP_{75}) by strengthening the local spatial correlation in feature maps. The number of GFLOPs increases slightly by 4, and the number of parameters increases by 2.63 M. After incorporating TSSP-Net into the TSQRM and LCSAF framework, the AP, AP_S, and AR_S of the complete model further improved to 0.425 (+0.7%), 0.329 (+1.3%), and 0.473 (+0.8%), respectively. This indicates that TSSP-Net, through its complementary effect with TSQRM and LCSAF, preserves small defect regions that might otherwise have been filtered out, thereby enabling LCSAF to fully leverage its cross-scale fusion capabilities on these retained salient features. The module introduced only 0.19 M additional parameters yet achieved an improvement in AP_S from 0.310 to 0.329. Although GFLOPs increased by 5, this significant improvement in minor defect detection performance was achieved at a low parameter cost, demonstrating excellent parameter efficiency.

As shown in Table 6, the baseline model achieves an FPS of 16.73, while our full model achieves an FPS of 10.51. The full model incurs additional computational overhead due to the introduction of the proposed module. The reduction in inference speed represents a trade-off between improved accuracy and real-time performance.

4.3.3. Improvement in Small Defect Detection Performance

As shown in Table 6, when PGSPM was added on the basis of integrating the above modules, the AP_S rose from 0.329 to 0.347, with an absolute increase of up to 1.8%. AR_S was raised from 0.473 to 0.485 (+1.2%). At the same time, comprehensive indicators such as AP, AP_{50} and AP_{75} also increased effectively. This result proves that PGSPM is the core key to solving the problem of the misdetection of small defects. It enhances the model's ability to focus on and represent the fine and low-significance defect areas in the image. When all modules work together, the model achieves optimal performance, with an AP_S of 0.347 and a higher AR_S of 0.493, the highest among all configurations.

This ultimately confirms the effectiveness of the synergy between the modules. PG-SalDETR not only pushes the overall accuracy to a high level of 0.439 (+3.8) but also achieves a relative improvement of nearly 8.8% in recall for small defects, thus fulfilling the core objective of this paper in addressing the misdetection problem. After integration, the PGSPM module achieved sustained improvements in AP_S and AR_S at the minimal cost of adding just 0.27 M, demonstrating favorable parameter efficiency. Although the overall computational load increased, this was necessary to enhance the performance of small defect detection.

4.3.4. Discussion on Architectural Complexity and Performance Gains

When used alone, LCASF and TSSP-Net do indeed yield limited improvements (AP increases of +0.6% and +0.8%, respectively). This is because TSSP-Net is designed to work in conjunction with PGSPM. The former is responsible for perceiving asymmetric features, while the latter provides physical guidance signals. Therefore, when TSSP-Net is used alone, its capabilities cannot be fully realized due to the lack of input from physical guidance signals. Nevertheless, it still delivers a moderate improvement when used alone, demonstrating its fundamental effectiveness. As shown in Table 7, when the guidance signal generated by PGSPM is introduced, the performance of TSSP-Net further improves:

AP increases from 0.409 to 0.420 (+1.1%), and AP_S increases from 0.330 to 0.336 (+0.6%). This validates the synergistic effect between TSSP-Net and PGSPM.

The limited improvement when using LCASF alone is due to the fact that the baseline model's hierarchical filtering mechanism has already filtered out regions with low-significance small defects prior to its processing, and LCASF itself cannot recover information that has already been discarded. However, it is worth noting that LCASF adds only 4 GFLOPs, imposing virtually no computational burden. Given its extremely low computational cost, this trade-off is acceptable even if the standalone improvement is limited.

It should be noted that the primary increase in computational load comes from the TSQRM module (+58 GFLOPs). LCASF, TSSP-Net, and PGSPM add almost no computational cost. The full model's AP improved by 3.8%, and AP_S improved by 2.8%. This indicates that the overall performance improvement arises from synergies among the modules, rather than from any single module. Although the individual contributions of each module are limited, their integration leads to a substantial overall improvement, which we believe justifies the architectural complexity. In the future, we will explore lightweight designs for PG-SalDETR to reduce computational overhead.

4.4. Parameter Analysis

The sensitivity analysis of γ is shown in Table 9. When γ = 6, the threshold setting is relatively strict. The model is overly sensitive to the response signal of small defects, resulting in the suppression of effective features. Consequently, the AP_S is only 0.298. When γ increases to 8, the threshold tends to be moderate. The model can effectively suppress background noise while retaining the details of small defects, and AP_S increases to 0.336, achieving the optimal performance. This indicates that a moderate γ value can better balance the retention of small defect features and noise suppression. However, when γ further increases to 10, the threshold is too loose, the model introduces more false detections, and at the same time may overly amplify the response of the background region, resulting in AP_S dropping to 0.303. Based on the above analysis, γ = 8 achieves the best balance between detection accuracy and robustness, so we set it as the default parameter.

Table 9. Sensitivity analysis of the PGSPM parameter γ on the NEU-DET dataset.

γ	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AR**	**AR_S**
6	0.412	0.742	0.409	0.298	0.528	0.437
8	0.420	0.757	0.426	0.336	0.539	0.473
10	0.409	0.732	0.397	0.303	0.535	0.416

The sensitivity analysis of α_{guid} is shown in Table 10. This parameter directly controls the maximum magnitude of the sampling offset. When α_{guid} = 0.04, the offset range is limited within ±0.04. This range is too conservative to effectively correct the feature shift of small defects, resulting in the model having difficulty locating the defect area, with an AP_S of only 0.294. When α_{guid} further increases to 0.06, the offset range expands to ±0.06. For small defects that occupy only a few pixels, this margin is too large. This can easily cause the sampling point to move outside the defect area, thereby introducing background noise or causing the target to be missed. Consequently, the model's performance deteriorates, and the AP_S drops to 0.281. Based on the above analysis, α_{guid} = 0.05 achieves the best balance between detection accuracy and robustness, so we set it as the default parameter.

Table 10. Sensitivity analysis of the TSQRM parameter α_{guid} on the NEU-DET dataset.

α_{guid}	AP	AP_{50}	AP_{75}	AP_S	AR	AR_S
0.04	0.408	0.741	0.382	0.294	0.536	0.447
0.05	0.415	0.758	0.404	0.323	0.546	0.476
0.06	0.406	0.740	0.385	0.281	0.534	0.423

The factor γ in Formula (16) is mainly used to enhance the influence of physical guidance on difficult samples, making the model rely more on physical priors when its own judgment is unreliable. We evaluated its impact by changing γ from 1.1 to 1.3 on the NEU-DET dataset, as shown in Table 11. When γ = 1.1, the enhancement intensity is insufficient, the guiding signal for difficult samples is weak, and the performance slightly declines. When γ = 1.3, if the enhancement is too strong, it may introduce noise or over-correct, which also leads to a slight decline in performance. This analysis indicates that γ = 1.2 achieves the best balance between enhancing difficult samples and maintaining stability, and the selection of γ is robust within a reasonable parameter range.

Table 11. Sensitivity analysis of the TSQRM parameter γ on NEU-DET dataset.

γ	AP	AP_{50}	AP_{75}	AP_S	AR	AR_S
1.1	0.414	0.746	0.408	0.310	0.550	0.462
1.2	0.415	0.758	0.404	0.323	0.546	0.476
1.3	0.411	0.757	0.399	0.314	0.532	0.465

As shown in Tables 2 and 12, when 0.3% is adopted as the threshold for small defects, there are only 420 small defects. Some objectively existing industrial micro-defects have been forcibly classified into the medium defect set. The AP_S has dropped from 0.171 to 0.120, while the AP_M has decreased to 0.338. This indicates that the 0.3% classification standard is too strict, resulting in the loss of a large number of real small defects, which does not conform to the inherent sample distribution of the dataset. When the threshold was relaxed to 0.8%, the number of small defects increased to 901, accounting for 28.2%, with only 143 new samples added compared to 0.5%. The types of defects did not continue to increase. The increase in AP_S, AP_M and AP_L is relatively small. Continuing to expand the range of small defects will only mix in some medium defects, and the inspection revenue will be low. Therefore, a 0.5% boundary can not only cover the vast majority of small defect samples and all nine types of fine defects but also will not overly include medium-scale defects.

Table 12. Sensitivity analysis of relative area thresholds on GC10-DET dataset.

Small Defect Definition	APS	APM	APL
area $<$ 0.3%	0.120	0.338	0.443
area $<$ 0.5%	0.171	0.353	0.461
area $<$ 0.8%	0.179	0.358	0.467

To validate the effectiveness of the dynamic fusion decay strategy, we conducted ablation experiments on the GC10-DET dataset. As shown in Table 13, we compared the performance of a fixed current_delta value with that of the dynamic decay strategy. Specifically, the strong constraint imposed by a fixed current_delta of 2 suppressed the differences between small defect and large defect features, leading to excessive averaging of small defect details. The weak constraint of a fixed current_delta of 1 failed to provide

sufficient protection for small-scale features, causing them to be overwhelmed by large-scale features early in the feature-learning process. In contrast, the dynamic attenuation strategy protects small-scale features from suppression through strong constraints during the early training phase and releases their expressive power through weak constraints during the later training phase, thereby achieving superior performance.

Table 13. Ablation study of the dynamic fusion range strategy in LCASF on GC10-DET dataset.

Current_delta	AP	AP_{50}	AP_{75}	AP_S	AR	AR_S
1.0	0.342	0.669	0.325	0.130	0.448	0.222
2.0	0.335	0.656	0.307	0.124	0.475	0.217
(2.0 → 1.0)	0.349	0.669	0.334	0.137	0.464	0.225

Selection of the noise mask window. The 15 × 15 window used for noise mask estimation is selected based on the principle of scale separation. In industrial surface inspection, noise typically manifests as small-scale local gray-level fluctuations, while small defects cover relatively large areas. The 15 × 15 window is significantly larger than the characteristic scale of noise, enabling a stable estimate of local variance while remaining small enough relative to the feature map resolution to avoid variance peaks caused by the suppression of defect signals. A window that is too small may be overly sensitive to isolated noise responses, while a window that is too large may smooth out signals from small defects.

Selection of multi-scale top-hat transform kernels (3 × 3, 5 × 5, 7 × 7). The 3 × 3 kernel captures fine-grained defect details, the 5 × 5 kernel matches the most common defect sizes, and the 7 × 7 kernel covers broader defect structures. Smaller kernels are overly sensitive to noise and cannot capture complete defect structures. Larger kernels treat small defects as background noise and suppress them.

Selection of large-region suppression kernels (11 × 11, 15 × 15). These kernels are used to suppress large contiguous background regions while preserving isolated defect peaks. Their size exceeds that of the maximum top-hat kernel (7 × 7) to ensure that background structures larger than typical defects are filtered out. If these kernels are too small (e.g., close to the size of the top-hat kernel), they cannot effectively suppress background structures; if they are too large, they may suppress genuine defect regions.

Clamp limits [0.3, 0.8]. The clamp limits for the guidance quality evaluator are empirically set to [0.3, 0.8]. Values below 0.3 are considered low-quality guidance and are suppressed to avoid introducing unreliable corrections; values above 0.8 are clipped to prevent saturation, ensuring that the guidance signal retains its discriminatory power. If the lower bound is set too low, unreliable guidance may influence the prediction and potentially introduce false positives. If the upper bound is set too high, the guidance signal may saturate, reducing its ability to distinguish between regions with high and moderate significance.

In LCASF, the dropout rate is dynamically increased from 0.05 to 0.15 during training. This dynamic strategy is designed to preserve more features in the early stages of training to promote thorough learning and to prevent overfitting in the later stages through stronger regularization. A dropout rate below 0.05 does not provide sufficient regularization and increases the risk of overfitting; conversely, a dropout rate above 0.15 may lead to underfitting.

4.5. Visualization Results

As shown in the visualization comparison in Figure 10, our developed method exhibits stronger visual attention and higher localization reliability when detecting minute defects.

For example, in the fine scratch recognition shown in Figure 10d,e, the heatmaps of YOLOv8 and YOLOv11 only show low-attention green areas, indicating that the features of the defect areas are not only weak but also scattered. In contrast, our proposed method generates strong and concentrated red responses in these defect areas, demonstrating that PG-SalDETR has a stronger semantic focusing ability for small defects. In the low-contrast detection shown in Figure 10f,c, the heatmap of Deformable DETR mainly presents light green to yellow tones, and the activation intensity of the defect areas is not only scattered but also low. Our proposed method generates continuous and highly prominent red responses in all defect areas, indicating that PG-SalDETR has a stronger perception ability for small defects. Furthermore, for the defect detection in Figure 10h, the Faster R-CNN heatmap only shows a faint green at the edges, reflecting its limited sensitivity to defect details; our proposed method produces concentrated, high-confidence red activation in the same area, further validating its more focused and accurate detection performance for small defects in complex industrial scenarios.

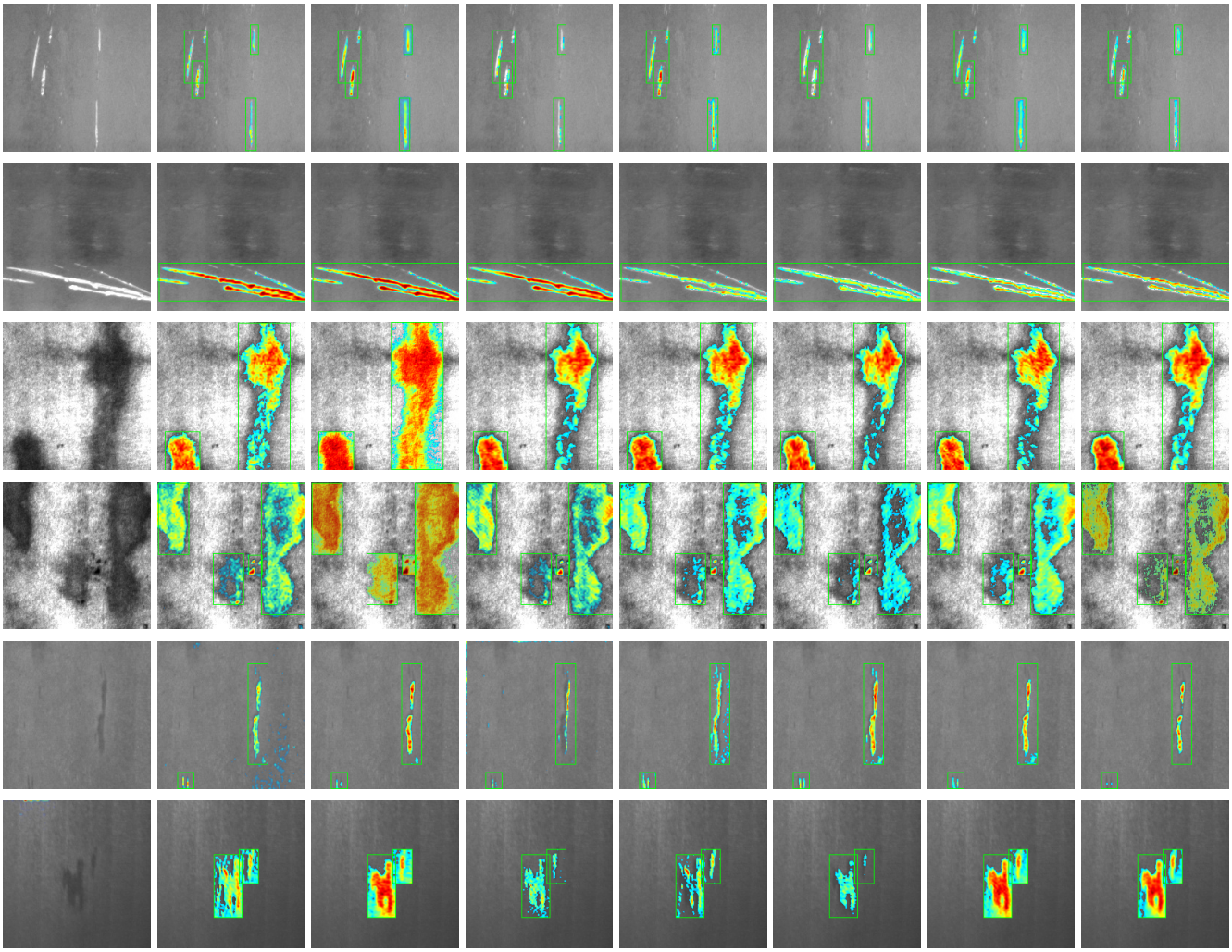


Figure 10. Qualitative comparisons of different models. (**a**) Image. (**b**) Salience DETR. (**c**) Ours. (**d**) YOLOv8. (**e**) YOLOv11. (**f**) Deformable DETR. (**g**) RT-DETR. (**h**) Faster R-CNN.

5. Conclusions

To address the challenge of detecting small defects on steel plate surfaces, this paper proposes a framework named PG-SalDETR. We conducted a comprehensive experimental validation of the PG-SalDETR model using two steel plate defect datasets. Experimental

results show that the PG-SalDETR framework achieves performance improvements on both datasets compared to advanced methods. Particularly in the detection of small defects, both mean precision and mean recall are enhanced. The PG-SalDETR framework not only provides an end-to-end solution to the practical challenge of detecting small defects in industrial surface quality inspection but also offers new research ideas and technical references for the application of physically guided saliency modeling methods in computer vision inspection tasks.

Limitations and Future Work. Despite its performance, PG-SalDETR has the following limitations that need further study. On the one hand, the reasoning speed of the model has declined, which limits its application in high-speed real-time detection scenarios, and the reasoning efficiency needs to be improved. On the other hand, the model has a relatively high complexity, and its deployment on resource-constrained embedded platforms remains challenging. Future work will focus on lightweight network design and model acceleration to enhance inference efficiency and lower deployment thresholds. In addition, although a detailed theoretical basis has been provided for the selection of key hyperparameters, a comprehensive experimental sensitivity analysis covering all parameters is still lacking. We plan to address this limitation in future work through more systematic experimental investigations.

Author Contributions: Conceptualization, X.Z. and S.Z.; Methodology, C.L.; Software, X.Z.; Validation, X.Z., C.L. and S.Z.; Formal analysis, X.Z. and C.L.; Investigation, X.Z.; Resources, X.Z.; Data curation, S.Z.; Writing—original draft, C.L. and S.Z.; Visualization, X.Z.; Supervision, S.Z.; Project administration, C.L.; Funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Joint Fund General Project of the Liaoning Provincial Department of Science and Technology under Grant 2024-MSLH-348, the Basic Research General Project of Liaoning Provincial Department of Education under Grant LJ212410142148, and in part by research funding from the Zhurong Institute of Cybersecurity, Hunan Institute of Technology and was supervised by Shengye Zhao.

Data Availability Statement: The dataset is available at http://faculty.neu.edu.cn/songkechen/zh_CN/zdylm/263270/list/index.htm and https://github.com/lvxiaoming2019/GC10-DET-Metallic-Surface-Defect-Datasets, accessed on 24 June 2026.

Acknowledgments: The authors acknowledge the use of DeepSeek-V3.1 for layout suggestions during the manual creation of Figures 1–3. All figures were manually drawn and refined by the authors using WPS.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, S.; Fan, X.; Zhao, Z.; Qiao, Z.; Wang, N.; Qiu, Y.; Jia, X. Research on steel surface defect detection system based on YOLOv5s-SE-CA model and BEMD image enhancement. *Nondestruct. Test. Eval.* **2025**, *40*, 3080–3099.
2. Yu, S.; Liu, Z.; Zhang, L.; Zhang, X.; Wang, J. FasterNet-YOLO for real-time detection of steel surface defects algorithm. *PLoS ONE* **2025**, *20*, e0323248. [PubMed]
3. Liu, Y.; Guo, Y.; Fan, Y.; Zhou, J.; Li, Z.; Xiao, S. Optical imaging technology application in transmission line insulator monitoring: A review. *IEEE Trans. Dielectr. Electr. Insul.* **2024**, *31*, 3120–3132. [CrossRef]
4. Luo, Q.; Li, B.; Su, J.; Yang, C.; Gui, W.; Silvén, O. CDDNet: Camouflaged defect detection network for steel surface. *IEEE Trans. Instrum. Meas.* **2023**, *73*, 5000313. [CrossRef]
5. Shi, X.; Zhou, S.; Tai, Y.; Wang, J.; Wu, S.; Liu, J. An improved faster R-CNN for steel surface defect detection. In *Proceedings of the 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), Shanghai, China, 26–28 September 2022*; IEEE: New York, NY, USA, 2022; pp. 1–5.
6. Wang, Y.; Wang, X.; Hao, R.; Lu, B.; Huang, B. Metal surface defect detection method based on improved cascade R-CNN. *J. Comput. Inf. Sci. Eng.* **2024**, *24*, 041002.

7. Jiyang, Q.; Yufan, W. Strip steel surface defect detection algorithm based on improved Faster R-CNN. *China Weld.* **2024**, *33*, 11–22.
8. Hari, U.; Ruhan Bevi, A.; Ramachandran, B. Performance analysis of Retinex based algorithms for enhancement of low light images. *J. Phys. Conf. Ser.* **2021**, *1964*, 062046. [CrossRef]
9. Wang, Z.; Zhu, B.; Wang, S.; Gao, W. Image enhancement algorithm for cold cathode weld defect based on improved Retinex theory. In Proceedings of the 2024 6th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP), Xi'an, China, 29 November–1 December 2024; pp. 661–664.
10. Zhang, L.; Li, X.; Sun, Y.; Feng, Y.; Guo, H. Multiscale Feature Fusion for Salient Object Detection of Strip Steel Surface Defects. *IEEE Access* **2025**, *13*, 42689–42702. [CrossRef]
11. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
12. Dhal, K.G.; Das, A.; Ray, S.; Gálvez, J.; Das, S. Histogram equalization variants as optimization problems: A review. *Arch. Comput. Methods Eng.* **2021**, *28*, 1471–1496.
13. Xiong, J.; Yu, D.; Wang, Q.; Xu, J.; He, X. Application of histogram equalization for image enhancement in corrosion areas. *Shock Vib.* **2021**, *2021*, 8883571. [CrossRef]
14. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. *arXiv* **2018**, arXiv:1808.04560.
15. Zhang, Y.; Zhang, J.; Guo, X. Kindling the darkness: A practical low-light image enhancer. In Proceedings of the 27th ACM International Conference on Multimedia (ACM MM), Nice, France, 21–25 October 2019; pp. 1632–1640.
16. Zhang, Y.; Di, X.; Zhang, B.; Wang, C. Self-supervised image enhancement network: Training with low light images only. *arXiv* **2020**, arXiv:2002.11300.
17. Pernando, Y.; Andryani, N.A.C.; Chowanda, A.; Budiharto, W. Low-light welding defect detection using Retinex-enhanced YOLOv11 for steel welding inspection. In Proceedings of the 2025 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Padang, Indonesia, 11–13 December 2025; pp. 77–83.
18. Kong, H.; You, C. Improved steel surface defect detection algorithm based on YOLOv8. In Proceedings of the 2024 7th International Conference on Computer Information Science and Application Technology (CISAT), Hangzhou, China, 12–14 July 2024; pp. 80–83.
19. Zhao, C.; Shu, X.; Yan, X.; Jiang, Y.; Jiang, D. RDD-YOLO: A modified YOLO for detection of steel surface defects. *Measurement* **2023**, *214*, 112776. [CrossRef]
20. Huang, X.; Zhu, J.; Huo, Y. SSA-YOLO: An improved YOLO for hot-rolled strip steel surface defect detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5040017.
21. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19.
24. Luo, Q.; Su, J.; Yang, C.; Gui, W.; Silvén, O.; Liu, L. CAT-EDNet: Cross-attention transformer-based encoder–decoder network for salient defect detection of strip steel surface. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5009813. [CrossRef]
25. Li, L.; Hu, S.; Hou, Y.; Hou, Y.; Li, W. ADMA-YOLO: A lightweight strip steel surface defect detection network via adaptive dual-path and multi-branch aggregation. *Meas. Sci. Technol.* **2025**, *36*, 105410.
26. Han, Q.; Zhi, X.; Hu, J.; Huang, Y.; Chen, W.; Jiang, S. CGFINet: Cross-scale guided high-order feature interaction change detection network for remote sensing image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 14614–14629.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
28. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
29. Li, D.; Jiang, C.; Liang, T. REDef-DETR: Real-time and efficient DETR for industrial surface defect detection. *Meas. Sci. Technol.* **2024**, *35*, 105411.
30. Zhou, Z.; Lu, Y.; Lv, L. A real-time defect detection algorithm for steel plate surfaces: The StarNet-GSConv-RetC3 detection transformer (SSR-DETR). *Ann. N. Y. Acad. Sci.* **2025**, *1548*, 277–288. [PubMed]
31. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-end object detection with dynamic attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 2988–2997.
32. Wu, S.; Yang, H.; Liao, L.; Liu, Z.; Zhang, J. SH-DETR: Enhancing steel surface defect detection and classification with an improved transformer architecture. *PLoS ONE* **2025**, *20*, e0334048. [CrossRef] [PubMed]

33. Hou, X.; Liu, M.; Zhang, S.; Wei, P.; Chen, B. Salience DETR: Enhancing detection transformer with hierarchical salience filtering refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 17574–17583.
34. Guo, M.-H.; Xu, T.-H.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
35. Zheming, Y.; Terekhov, V.I.; He, J. Research of different neural network architectures for real-time metal surface defect detection. In Proceedings of the 2025 7th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), Moscow, Russia, 8–10 April 2025; pp. 1–5.
36. Lv, X.; Duan, F.; Jiang, J.-J.; Fu, X.; Gan, L. Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network. *Sensors* **2020**, *20*, 1562. [CrossRef] [PubMed]
37. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
38. Ye, M.; Ke, L.; Li, S.; Tai, Y.; Tang, C.; Danelljan, M. Cascade-DETR: Delving into high-quality universal object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6704–6714.
39. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.