

Article

Multi-Scale Deformable Transformer with Iterative Query Refinement for Hot-Rolled Steel Surface Defect Detection

Haoran Wang ^{1,2}, Fan Zhang ² and Rong Yi ^{1,*}¹ College of Mechanical Engineering, University of South China, Hengyang 421001, China; wanghaoran@usc.edu.cn² State Key Laboratory of Precision Manufacturing for Extreme Service Performance, School of Automation, Central South University, 410083 Changsha, China; zhangfan219@csu.edu.cn

* Correspondence: 11930903@mail.sustech.edu.cn

Abstract

Accurate and efficient detection of small and complex surface defects on hot-rolled steel plates remains a significant challenge in industrial quality assurance. Current deep learning detectors often exhibit limitations in detection accuracy and training convergence speed, particularly for small objects, which limits their practical deployment in real-time industrial inspection systems. To overcome these deficiencies, this paper proposes a multi-scale deformable transformer iterative query refinement network (MDT-Net). MDT integrates three key innovations: a Swin Transformer backbone for robust multi-scale feature representation, a deformable attention mechanism to significantly reduce computational complexity and accelerate convergence, and an iterative bounding box refinement strategy for precise localization. Extensive experiments on the NEU-DET dataset demonstrate MDT's superior performance, achieving 82.7% mAP_{50} . Crucially, MDT significantly outperforms other mainstream detectors in small object detection, reaching an $mAP_{50:95}$ of 0.55, and exhibits remarkably faster training convergence. These findings confirm MDT's effectiveness and robustness for accurate and efficient steel surface defect detection, thereby providing a crucial tool for enhancing sensor-based quality control and offering a promising solution for industrial quality management.



Academic Editor: Jiawei Xiang

Received: 2 October 2025

Revised: 27 October 2025

Accepted: 7 November 2025

Published: 11 November 2025

Citation: Wang, H.; Zhang, F.; Yi, R. Multi-Scale Deformable Transformer with Iterative Query Refinement for Hot-Rolled Steel Surface Defect Detection. *Sensors* **2025**, *25*, 6890. <https://doi.org/10.3390/s25226890>

Copyright: © 2025 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 (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: surface defect detection; deformable attention; swin transformer; steel surface defects

1. Introduction

The steel industry is a crucial component of modern industrial production [1]; however, in the actual production process, constrained by manufacturing processes and technology, the surface inevitably develops various defects [2]. For instance, common defects in hot-rolled steel plates include pitting, scratches, holes, scabs, cracks, a pressed-in oxide scale, rust spots, roll marks, flashes, edge cracks, oxidation discoloration, and inclusions [3]. Traditionally, steel surface defect identification [4] was performed using a combination of manual inspection and traditional machine vision methods. However, these methods lack a learning process [5,6], impose stringent requirements on the detection environment, demand the manual selection of high-quality features, and are not suitable for open operating environments. They require specific improvements based on on-site conditions, making generalization difficult [7]. With the widespread application of Convolutional Neural Networks (CNNs) in various industrial image classification and recognition tasks, deep learning-based object detection methods have also gradually found application in a wide

variety of surface defect detection tasks [8–12]. For instance, models like the Single-Shot MultiBox Detector (SSD) [13], You Only Look Once (YOLO) [14], Region-based Convolutional Neural Network (Faster R-CNN) [15], and Mask R-CNN [16] have demonstrated impressive performance within the domain of object recognition. Wang et al. [17] designed a Faster R-CNN algorithm that integrated multi-level features, thereby solving the problem of detecting various random defects on metal plates and strips. Dai et al. [18] proposed an efficient algorithm that employed an enhanced Faster R-CNN. This algorithm addresses the limitations and poor precision of part-surface defect detection, significantly enhancing its performance compared to conventional techniques. Zhang et al. [19] boosted the detection performance for small targets in steel plate surface defect detection tasks by modifying the YOLOv3 network's structure. Nicolas et al. [20] utilized the Detection Transformer (DETR) network structure to achieve end-to-end object detection within the field of computer vision, greatly simplifying the object detection process.

However, these object detection algorithms still result in missed detections and false positives for small targets in surface defect detection tasks. Consequently, we designed a novel method for detecting surface defects using a modified DETR algorithm. By introducing a Swin Transformer as the backbone network, deformable attention, and iterative bounding box refinement, it aims to effectively address the limitations of traditional DETR models' training efficiency, detection accuracy, and particularly their ability to detect small targets, thereby enhancing the detection precision of surface defects.

The MDT-Net contributions in this study are presented as follows:

(1) Utilizing Swin Transformer as the feature extraction backbone network significantly expands the model's capacity to capture multi-scale features, effectively acquiring both local and global contextual information from steel plate images.

(2) Introducing deformable attention, which replaces conventional multi-head attention with sampling-based deformable attention, reduces computational complexity and significantly accelerates convergence speed. Concurrently, iterative bounding box refinement improves the localization precision and accuracy of the model.

(3) Using the NEU-DET dataset, the proposed MDT-Net was tested to verify the network's accuracy in steel plate surface defect detection and to assess the individual and synergistic contributions of Swin Transformer, deformable attention, and iterative bounding box refinement strategies to its performance. The results show an mAP_{50} of up to 82.7%, with a significant improvement in small target detection capability.

The remainder of this paper is structured as follows. Section 2, 'Methods', details the proposed multi-scale deformable transformer iterative query refinement network (MDT-Net) and its core components. Section 3, 'Experiments', presents the experimental setup, results, ablation studies, and a comparative analysis with state-of-the-art object detection algorithms. Finally, Section 4, 'Conclusions', summarizes the key findings and contributions of this work.

2. Method

2.1. Network Architecture Overview

The architecture of the multi-scale deformable transformer iterative query refinement network is shown in Figure 1. Assuming that the input steel plate image has dimensions $H \times W \times 3$ (H for height, W for width), the image was first divided into several 4×4 -pixel patches. Subsequently, these image patches underwent a linear embedding operation, which projects the features of each patch into a high-dimensional space and expands them along the channel direction. At this point, the image size is $\frac{H}{4} \times \frac{W}{4} \times 48$. The main body of the network consists of 4 consecutive stages (Stage 1 to Stage 4). Each stage comprises an operation and several transformer block units that utilize deformable self-

attention computation. Except for Stage 1, which employs a linear embedding operation, all other stages use block combination operations. After the patched image undergoes linear embedding, its size is $\frac{H}{4} \times \frac{W}{4} \times C$, where C is a hyperparameter, and is 96 in the Swin-Tiny architecture. Subsequently, after each stage, the image's width and height are both half of their original size, and the number of channels doubles. This part of the operation is very similar to a convolutional hierarchical structure.

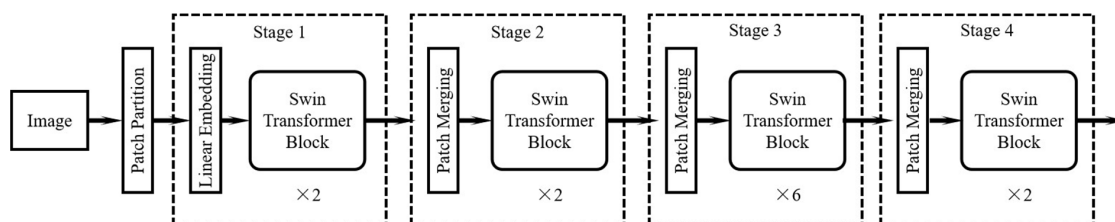


Figure 1. Multi-scale deformable transformer iterative query refinement network.

Figure 2a shows the Swin Transformer unit structure. It is largely consistent with the original transformer unit structure; however, the primary distinction lies in the multi-head attention mechanism, which is replaced by window-based multi-head deformable attention (W-MDA) and shifted window-based multi-head deformable attention (SW-MDA). Furthermore, the Multi-Layer Perceptron, which serves as the feed-forward network, comprises two layers, interspersed with a GELU activation function. Notably, Swin Transformer units are utilized in pairs. In this pairing, the attention mechanism within the first unit is based on fixed-window multi-head attention, while the second unit employs a shifted-window multi-head attention. Additionally, the patch merging operation is functionally similar to pooling, aiming to reduce image resolution for downsampling. However, unlike traditional pooling, the mechanism of patch merging involves combining feature values from the same relative positions within a window to form new feature patches and subsequently concatenating these new patches.

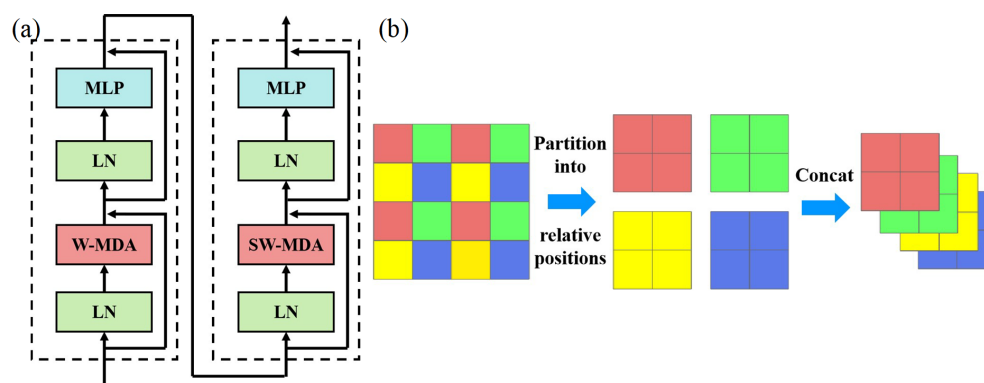


Figure 2. (a) Architecture of two consecutive Swin Transformer blocks; (b) schematic diagram of the window combination process.

2.2. Window Multi-Head Attention

The input image is first divided into a series of non-overlapping fixed-size windows. Then, standard multi-head attention is performed in parallel within each independent window. For instance, assuming that the initial input image size is $hw \times C$, after applying fixed-window multi-head attention, the input dimensions transform to $\left(\frac{hw}{M^2}\right) \times M^2 \times C$.

This means that each individual window has a size of $M^2 \times C$, and there is a total of $\frac{hw}{M^2}$ such windows. The window multi-head attention mechanism is presented as Equation (1).

$$\Omega(W - MSA) = \left(\frac{hw}{M^2}\right) \cdot (4M^2C^2 + 2M^4C) = 4hwC^2 + 2M^2hwC \quad (1)$$

Its time complexity is $O(hw)$, which exhibits an approximately linear relationship with the image size.

2.3. Shifted-Window Multi-Head Deformable Attention

While window multi-head attention reduces computational complexity and overhead, it inherently sacrifices global communication capability. To address this limitation, the shifted-window approach is introduced. Its fundamental process, as depicted in Figure 3, utilizes a Cyclic Shift and Mask Mechanism. This mechanism is crucial for resolving the challenges introduced by sliding windows, specifically the increased number of windows (leading to higher computational complexity) and the inconsistent window sizes (which impede unified parallel computation).

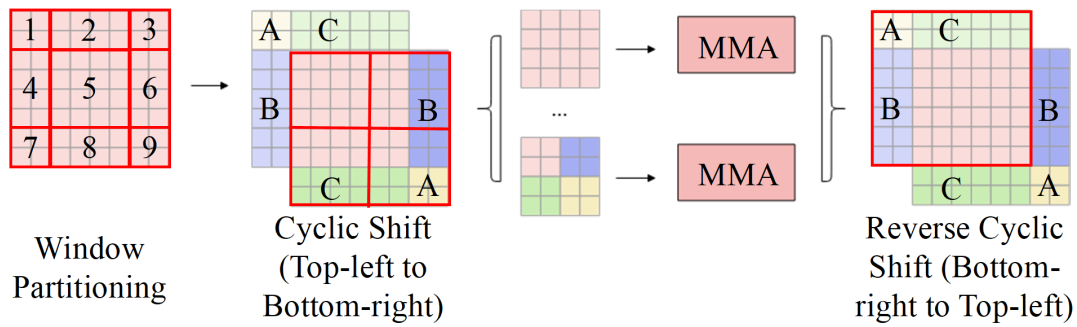


Figure 3. Schematic diagram of Cyclic Shift and Mask Mechanism.

2.4. Deformable Attention Mechanism

The deformable attention mechanism, as depicted in Figure 4, allows each pixel to compute attention solely with a subset of sampled pixels, rather than interacting with every pixel in the entire image. Consequently, this mechanism substantially reduces the computational burden and accelerates the model's convergence speed. Furthermore, deformable attention concurrently incorporates the principle of multi-scale features. Utilizing deformable attention on multi-scale feature maps can also improve small object detection accuracy, eliminating the need for Feature Pyramid Networks (FPNs) [21]. Equation (2) provides the mathematical formulation of deformable attention.

$$\text{DeformAttn}(z_q, p_q, x) = \sum_{m=1}^M W_m \left[\sum_{k=1}^K A_{mqk} \cdot W'_m x(p_q + \Delta p_{mqk}) \right] \quad (2)$$

It should provide a concise and precise description of the experimental results, their interpretation, and the experimental conclusions that can be drawn.

Here, reference points represent the 2D coordinates of z_q within the feature map. Furthermore, Δp_{mqk} denotes the position offset, which is learnable. m denotes the index of the attention head, and W_m, W'_m are learnable weight matrices. Each query computes attention solely with the sampled elements within each attention head. Crucially, both A_{mqk} and Δp_{mqk} are derived from query via a fully connected layer, unlike in a standard transformer where they are typically computed from both query and key.

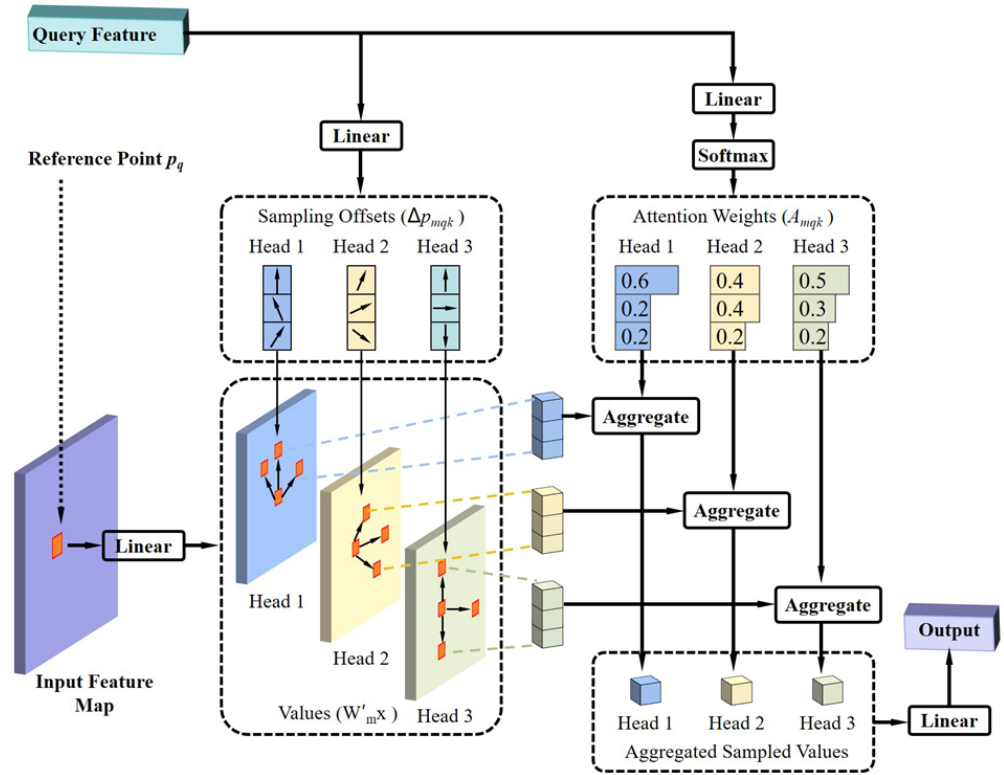


Figure 4. Illustration of the deformable attention module.

Equation (3) presents the mathematical formulation of multi-scale deformable attention.

$$MSDeformAttn\left(z_q, \hat{p}_q, \left\{x^l\right\}_{l=1}^L\right)=\sum_{m=1}^M W_m\left[\sum_{l=1}^L \sum_{k=1}^K A_{mqk} \cdot W'_m x^l\left(\phi_l\left(\hat{p}_q\right)+\Delta p_{mlqk}\right)\right] \quad (3)$$

Here, the l -th multi-scale feature layer is denoted with a total of L layers. Furthermore, $\hat{p}_q \in [0, 1]$ signifies the normalized reference points. Moreover, ϕ_l is a function that maps the normalized coordinates to their respective feature layers. Consequently, each reference point possesses a distinct coordinate within each feature layer, thereby facilitating the computation of sampling point locations across these diverse feature layers. In addition, A_{mqk} represents the normalized attention weight. Specifically, each query samples K points from each of the L feature layers; thus, a total of KL pixels were sampled within each attention head. The normalization of attention weights was also conducted over these KL pixels, meaning that $\sum_{l=1}^L \sum_{k=1}^K A_{mqk} = 1$.

2.5. Iterative Bounding Box Refinement

For each layer of the transformer decoder, predictions are made. These prediction results then serve as new reference points for the subsequent layer, thereby enabling continuous refinement. Ultimately, the final layer's output is taken as the ultimate result. During its computational process, a regression bounding box detection head is first employed to predict the decoded features, which yields offset results. These offsets are then added to the reference point coordinates and subsequently provided to the next layer. Concurrently, gradients are detached before being fed into the next layer, as the reference points serve

as a form of prior information across all decoder layers. For the d -th decoder layer, the corrected regression box is given by Equation (4).

$$\hat{b}_q^d = \left\{ \begin{array}{l} \sigma(\Delta b_{qx}^d + \sigma^{-1}(\hat{b}_{qx}^{d-1})), \sigma(\Delta b_{qy}^d + \sigma^{-1}(\hat{b}_{qy}^{d-1})) \\ \sigma(\Delta b_{qw}^d + \sigma^{-1}(\hat{b}_{qw}^{d-1})), \sigma(\Delta b_{qh}^d + \sigma^{-1}(\hat{b}_{qh}^{d-1})) \end{array} \right\} \quad (4)$$

$$\sigma = \frac{1}{1 + e^{-x}} \quad (5)$$

$$\sigma^{-1} = \log \frac{x}{1 - x} \quad (6)$$

Here, $\Delta b_{q(x,y,w,h)}^d$ represents the prediction result from the regression branch of the d -th decoder layer's output, obtained through the regression bounding box detection head. Furthermore, σ and σ^{-1} denote the sigmoid function and its inverse (logit function), respectively. Their mathematical forms are presented in Equations (5) and (6).

2.6. Loss Function

MDT-Net adopts the Detection Transformer's loss function strategy. The model outputs a fixed number of predictions (e.g., 100), which typically exceeds the actual number of ground truth objects. This creates a class imbalance, with many predictions corresponding to no object or background. To address this, a special label ($\emptyset$) denotes no object or background. When calculating the overall loss ($\mathcal{L}_{Hungarian}$), if a predicted box is matched via bipartite matching (Hungarian algorithm) to a $c_i = \emptyset$ label (i.e., background class), the logarithmic part of its classification loss is down-weighted by dividing it by 10. This effectively reduces the background class's contribution to the total loss, mitigating the class imbalance and allowing the model to focus more on detecting actual targets.

$$\mathcal{L}_{Hungarian}(y, \hat{y}) = \sum_{i=1}^{100} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + 1_{\{c_i \neq \emptyset\}} \mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) \right] \quad (7)$$

Here, y denotes the ground truth, and $\hat{y}$ also represents the 100 predicted boxes. c_i is the class of the i -th ground truth, and b_i denotes the relative values of the bounding box center, width, and height. For the i -th prediction $\hat{\sigma}(i)$, the probability of predicting c_i is defined as $\hat{p}_{\hat{\sigma}(i)}(c_i)$, and the predicted box is $\hat{b}_{\hat{\sigma}(i)}$.

3. Experimental Section

3.1. Datasets and Evaluation Metrics

This study utilizes the NEU-DET hot-rolled steel strip surface defect dataset [22], an open-source resource from Northeastern University. Specifically, this dataset includes 1800 grayscale images, uniformly sized at 200×200 pixels. Furthermore, it includes six common defect types present on hot-rolled steel surfaces, which are crazing (CZ), inclusion (IS), pitted surface (PS), patches (PC), scratches (ST), and rolled-in scale (RS). Individual defect images and their corresponding bounding boxes are illustrated in Figure 5a. Given that each image may contain multiple defect types, a statistical analysis was performed on all labeled data to ascertain the actual counts for each defect category. The resulting statistical chart is presented in Figure 5b. The original labels within the dataset are stored in YOLO format, where each bounding box is defined by five values: the class ID, normalized width, normalized height, normalized center x-coordinate, and y-coordinate.

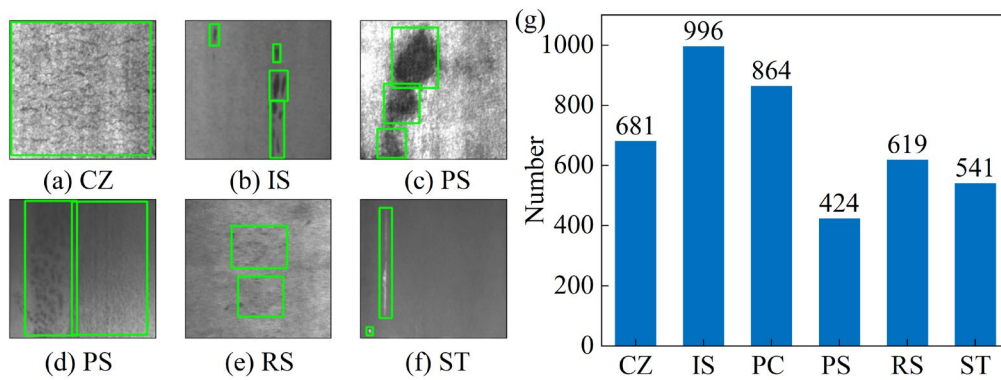


Figure 5. (a–f) Examples of various defect categories from the NEU-DET dataset, and (g) the statistical count of ground truth instances for each defect category.

Given that the surface defect dataset comprises only 1800 images, offline augmentation was initially performed to expand the dataset size. Specifically, this experiment employed random rotation, scaling to 400×400 pixels, and random cropping, thereby increasing the training images to 9000. For training, the dataset was split 9:1 into training and validation sets. The experiments utilized the PyTorch-based deep learning framework MMDetection. The Swin Transformer was chosen as the Swin-Tiny model, and it utilized pre-trained weights from ImageNet-1K. The embedding dimension was set to 96, while the number of Swin Transformer units for each stage amounted to two, two, six, and two, respectively. The multi-head attention heads for each stage were 3, 6, 12, and 24, with a window size of seven and a patch size (pixel) of four. The linear embedding/merging strides were four, two, two, and two.

The main parameters for the deformable attention-based transformer are as follows: 300 object queries, 3 encoder hidden layers, a Dropout probability of 0.1 for the feed-forward network in each encoder layer, 3 decoder hidden layers, a 0.1 Dropout probability for the feed-forward network in each decoder layer, and 4 multi-head attention heads in the decoder. The AdamW (Adam with decoupled Weight decay) optimizer was employed, with a 0.0002 learning rate and a 0.0001 weight decay. In addition, the number of iterations, batch size, and number of workers were 100, 8, and 2, respectively.

This paper adopts the metric of $IoU > 0.5$ as the accuracy indicator for surface defect detection, denoted as mAP_{50} . Mathematically, its meaning is the ratio of the area under the precision (P)–recall (R) curve to the total area. Specifically, precision and recall are defined by Equations (8) and (9) [23].

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

$$AP = \int P(R) dR \quad (10)$$

$$mAP = \frac{\sum_{i=1}^N AP}{TN} \quad (11)$$

Here, TP , FP , and FN represent the number of correctly detected, incorrectly detected, and missed targets for comparative experiments; to validate the enhanced capability of the proposed model for small object detection, the $mAP_{50:95}$ metric was also utilized. Specifically, this metric is derived by considering 10 distinct IoU thresholds (from 0.5 to 0.95 at 0.05 intervals). For each of these IoU thresholds, an individual mean average precision (mAP) value is computed, and $mAP_{50:95}$ is defined as the average of these 10 values.

3.2. Experimental Results

The training process's *loss* curve and mAP_{50} curve are both depicted in Figure 6. Around the 40th epoch, the *loss* curve showed a sudden decrease, while the mAP_{50} curve exhibited a sharp increase. This phenomenon occurred because the training strategy adjusted the learning rate at the 40th epoch, reducing it tenfold from 0.0002 to 0.00002. After 100 iterations, the optimal mAP_{50} value reached 82.7%. The mAP_{50} for each defect category is as follows: crazing, 50.2%; inclusion, 87.4%; patches, 92.7%; pitted surface, 86.0%; rolled-in scale, 83.0%; and scratches, 96.9%. The average value is 82.7%.

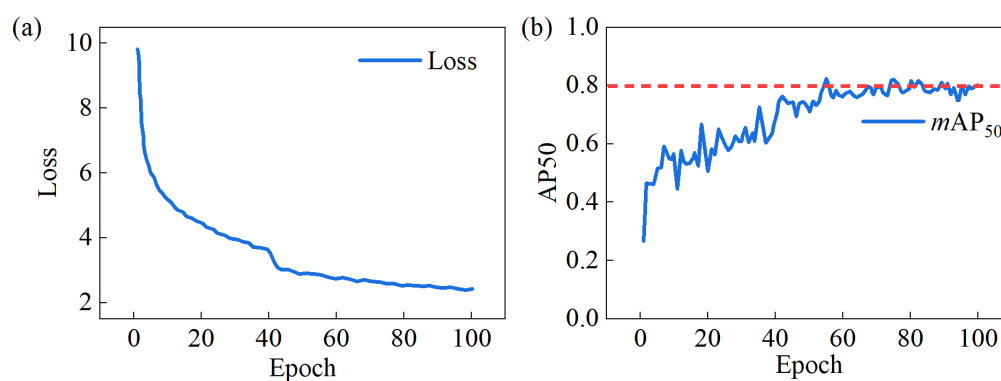


Figure 6. (a) Loss curve and (b) mAP_{50} curve.

From each category of the output results, five images were selected, as shown in Figure 7. Based on the detection images and the numerical results, it is clear that although the MDT-Net model's accuracy for detecting crazing defects is relatively low at only 50.2%, the detection accuracy for other defect categories is above 80%, with the highest accuracy for scratches reaching 96.9%. Therefore, the improved algorithm in this study demonstrates strong performance within the domain of steel strip surface defect detection.

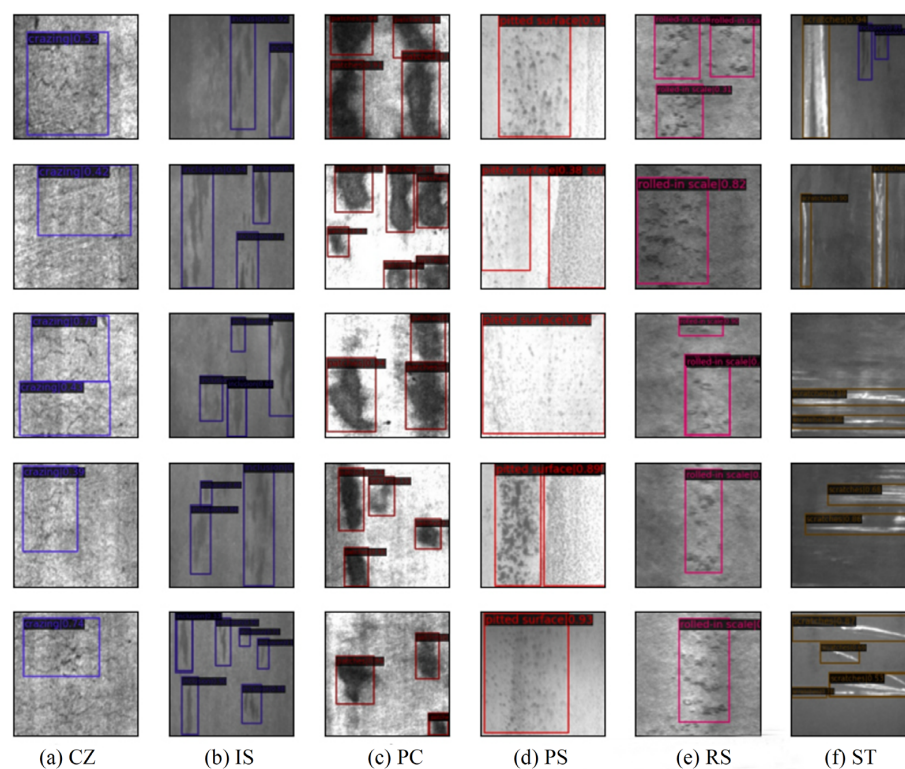


Figure 7. (a–f) Defect detection results for different categories.

3.3. Ablation Experiments

In order to systematically evaluate the effectiveness of the MDT-Net algorithm's key components, we conducted ablation experiments. These experiments aim to quantify the independent and combined contributions of the Swin Transformer backbone network, the deformable attention mechanism, and the iterative bounding box refinement strategy to the steel plate surface defect detection performance. The module configuration information for the five sets of models is shown in Table 1. Experiment 1 serves as the baseline, employing the original DETR algorithm (using ResNet50 as the backbone network). Subsequent experiments progressively introduced or replaced different components.

Table 1. Comparison of data results for ablation experiments.

Schemes	D A	S T	IBBR	mAP_{50}	Parameters	FLOPs	Detection Speed
1				77.0%	32.6 M	15.19 G	28 FPS
2	✓			76.7%	34.49 M	24.67 G	24 FPS
3	✓	✓		77.3%	33.51 M	26.55 G	21 FPS
4	✓		✓	78.6%	34.76 M	24.79 G	23 FPS
5	✓	✓	✓	82.7%	33.78 M	26.67 G	18 FPS

The ablation experiment results are shown in Table 1. It is clear that Experiment 5 (our algorithm) achieved the highest mAP_{50} of 82.7%. Compared to the original DETR baseline (Experiment 1, $mAP_{50} = 77.0\%$), the method presented in this paper achieved a significant improvement of 5.7%. This demonstrates that the proposed combined strategy can effectively increase the accuracy of steel plate surface defect detection.

Comparing Experiment 1 and Experiment 2, it is observed that when deformable attention was introduced in isolation, mAP_{50} slightly decreased from 77.0% to 76.7%. This suggests that, without modifying the backbone network and in the absence of other optimizations, deformable attention might necessitate more meticulous fine-tuning or integration with a more robust feature extractor to fully realize its benefits. However, when combined with iterative bounding box optimization (Experiment 4 vs. Experiment 2) and the Swin Transformer backbone (Experiment 5 vs. Experiment 3), deformable attention exhibited a synergistic effect in performance enhancement. Replacing the backbone network from ResNet50 with Swin Transformer (Experiment 3 vs. Experiment 1) increased mAP_{50} from 77.0% to 77.3%, thereby showcasing the advantage of Swin Transformer in feature extraction. Nevertheless, the Swin Transformer backbone introduced a higher computational load (FLOPs: 26.55 G vs. 15.19 G) and a lower detection speed (21 FPS vs. 28 FPS), which is consistent with the inherent complexity of the Swin Transformer. Iterative bounding box optimization for prediction frames proved crucial for improving performance. Specifically, comparing Experiment 2 and Experiment 4 for the introduction of iterative optimization, in the presence of deformable attention, mAP_{50} was significantly increased from 76.7% to 78.6%. Similarly, under the Swin Transformer backbone network (Experiment 5 vs. Experiment 3), iterative optimization led to a substantial jump in mAP_{50} from 77.3% to 82.7%, making it a key factor in achieving the final high accuracy. This unequivocally confirms the effectiveness of the iterative correction mechanism in progressively refining the predicted bounding boxes.

To further evaluate the performance of the MDT-Net in terms of the training efficiency and detection of different-sized targets, we compared its mAP_{50} convergence curve and detection performance for various target sizes against the original DETR algorithm, as illustrated in Figure 8. It can be observed that the MDT-Net model entered the convergence stage around the 55th epoch, whereas the original DETR algorithm only reached convergence around the 130th epoch. This indicates that the MDT-Net model not only

significantly accelerated the training convergence speed but also achieved higher detection accuracy after convergence. Figure 8b compares the mAP_{50} of the proposed model and the original DETR algorithm for small, medium, and large targets. These results show that the MDT-Net model's detection capability for small targets is significantly stronger than that of the DETR algorithm, while for medium and large targets, the detection performance of both models is not substantially different. This highlights the advantage of the MDT-Net model in handling common small defects on steel plates, which is crucial for practical industrial applications.

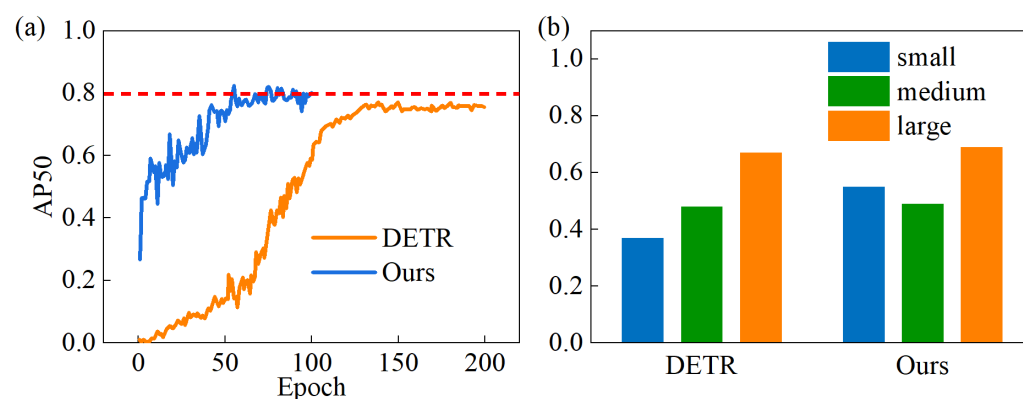


Figure 8. (a) mAP_{50} curves of our and (b) the DETR algorithm.

Figure 9 illustrates a visual comparison of the detection results for different defect types across various experimental groups. As can be seen from the figure, Experiment 5 demonstrated significant advantages in terms of localization accuracy and bounding box tightness in detecting defects. In particular, when detecting small-sized or irregularly shaped defects, Experiment 5 was able to generate more accurate prediction boxes and effectively reduced instances of missed detections and false positives. This is highly consistent with the quantitative results presented in Table 1. For subtle scratches or pitted defects under certain complex backgrounds, the detection performance of Experiment 5 was significantly superior compared to that of other experimental groups. The ablation experiments unequivocally confirmed the effectiveness of the Swin Transformer backbone network, the deformable attention mechanism, and the iterative bounding box refinement strategy for prediction. The synergistic effect of these components significantly improved steel plate surface defect detection accuracy. Furthermore, the performance enhancement was more pronounced after the introduction of iterative bounding box refinement.

Table 2 presents an ablation study on the hyperparameters of our deformable attention-based Transformer, revealing their impact on mAP_{50} . The study identifies an optimal configuration of three encoder layers, three decoder layers, and four attention heads, achieving the highest mAP_{50} of 82.7%. Crucially, increasing the number of attention heads from one to four (while keeping layers at 3,3) leads to a substantial performance gain, improving mAP_{50} from 79.6% to 82.7%, and underscoring the importance of diverse feature learning. Conversely, varying the encoder and decoder hidden layers (with four heads) shows that both shallower (2,2) and deeper (4,4) architectures yield slightly lower performances (80.6% and 81.3%, respectively) compared to the optimal (3,3) configuration, suggesting a sweet spot for network depth. This systematic evaluation highlights the critical role of careful hyperparameter tuning in achieving optimal accuracy for steel surface defect detection.

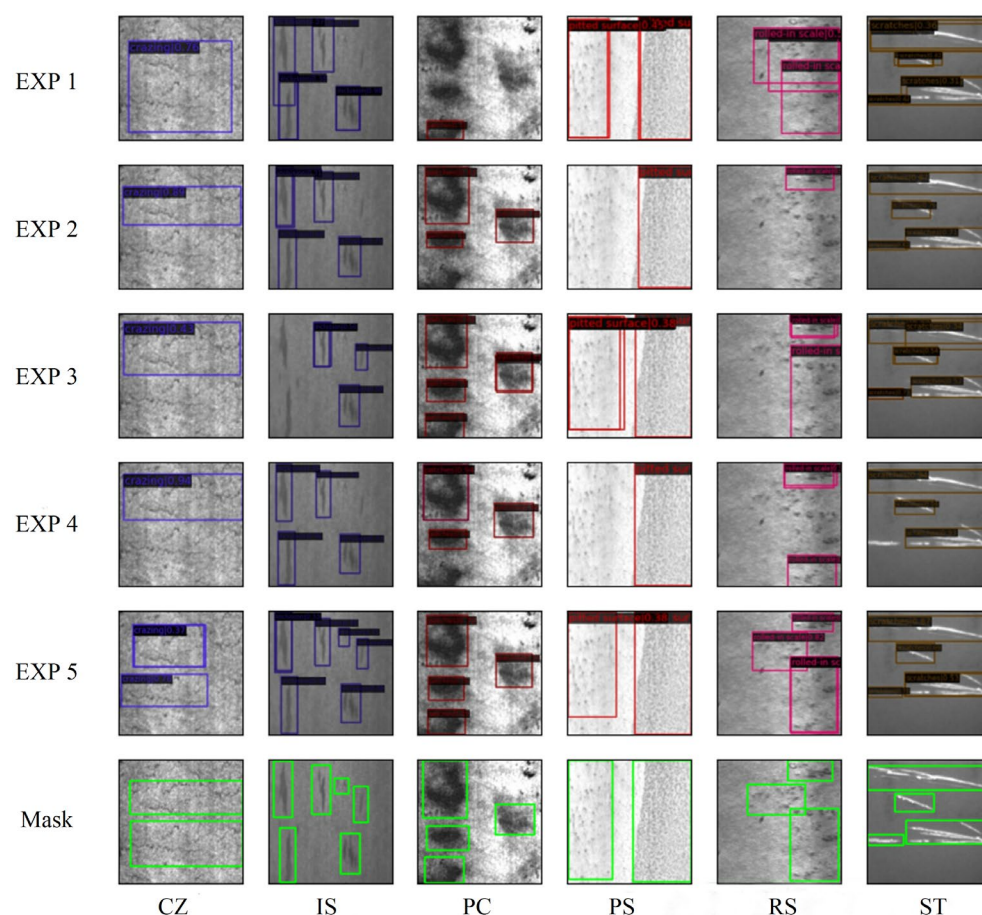


Figure 9. Comparison of steel plate defect detection images from ablation experiment results.

Table 2. Ablation study of the hyperparameters for the deformable attention-based transformer.

Coder Hidden Layers	Decoder Hidden Layers	Head Number	mAP_{50}
3	3	1	79.6%
3	3	2	80.3%
3	3	4	82.7%
4	4	4	81.3%
2	2	4	80.6%

3.4. Algorithm Comparison

To validate the superiority of the MDT-Net model over other models, comparative experiments utilized the NEU-DET dataset. Specifically, YOLOv3, SSD, and Faster R-CNN models were selected for comparison, resulting in a total of four models, including the one proposed in this paper. The mAP_{50} results for the YOLOv3, SSD, Faster R-CNN, and the MDT-Net model were 79.6%, 76.1%, 79.4%, and 82.7%, respectively. This indicates MDT-Net's superior overall performance in steel plate surface defect detection.

To further analyze the MDT-Net detection capabilities for defects of different scales, we additionally compared the $mAP_{50:95}$ of each algorithm for small, medium, and large targets, and Figure 10 displays the detailed results. The MDT-Net model achieved an $mAP_{50:95}$ of 0.55 for small target detection, which is significantly higher than YOLOv3 (0.47), SSD (0.30), and Faster R-CNN (0.49). This fully demonstrates the remarkable performance of the MDT-Net model in handling small defects commonly found on steel plate surfaces.

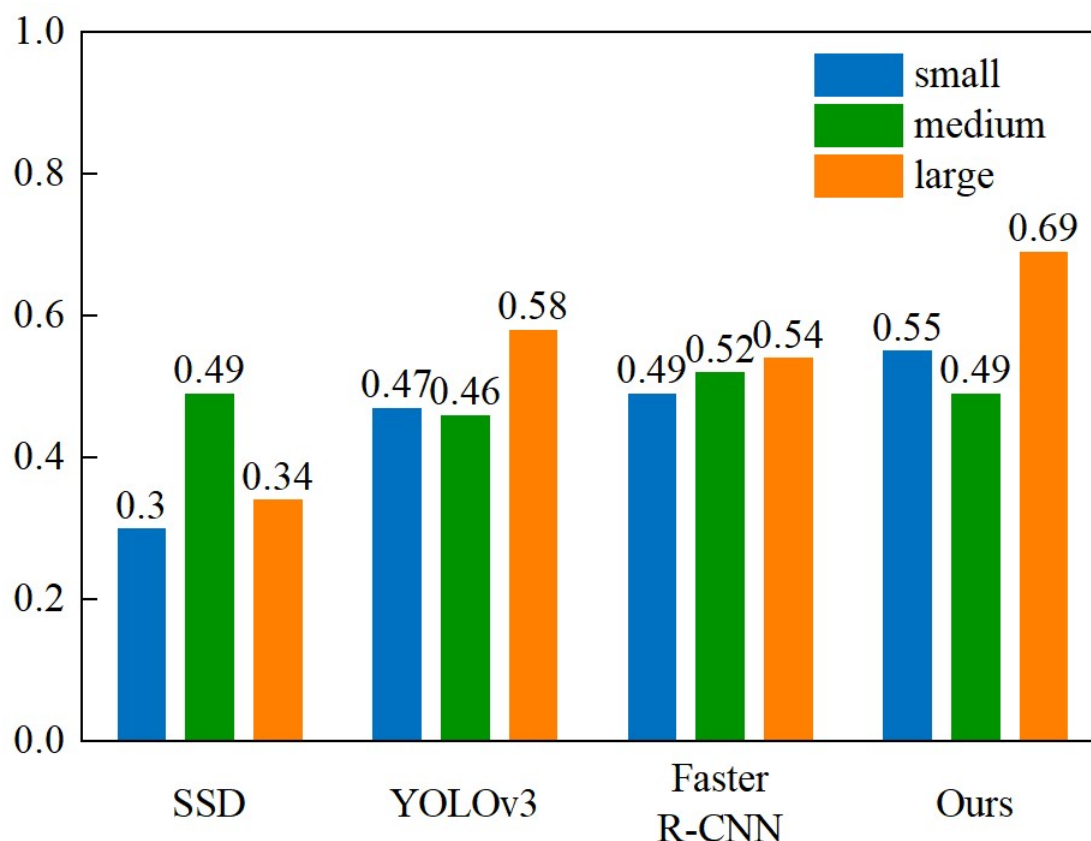


Figure 10. Comparison of $mAP_{50:95}$ for SSD, YOLOv3, Faster R-CNN, and our object detection algorithm across three sizes.

From Figure 11, it is evident that the MDT-Net model exhibits clear advantages in the localization accuracy and bounding box tightness of defects, especially when detecting small-sized or irregularly shaped defects. Particularly in areas with complex backgrounds or dense defects, the proposed model can generate more accurate prediction boxes and effectively reduce missed detections and false positives. This is highly consistent with the significant improvement in small target detection shown in Figure 10.

Table 3, detailing per-category mAP_{50} results across various state-of-the-art methods, confirms MDT-Net's overall superiority in steel plate defect detection, achieving the highest average mAP_{50} of 82.7%. Specifically, MDT-Net demonstrates a leading performance in most defect categories, notably achieving the highest accuracy for inclusion (87.4%), patches (92.7%), rolled-in scale (83.0%), and an impressive 96.9% for scratches, significantly surpassing competitors and highlighting its capability with subtle defects. However, while competitive for pitted surface (86.0%), MDT-Net records the lowest mAP_{50} for crazing defects at 50.2%, which is consistent with earlier observations and suggests challenges with their irregular and subtle characteristics. Despite this specific challenge, MDT-Net's robust and often superior performance across other defect types solidifies its position as a highly effective and practical solution for industrial steel plate surface defect detection.

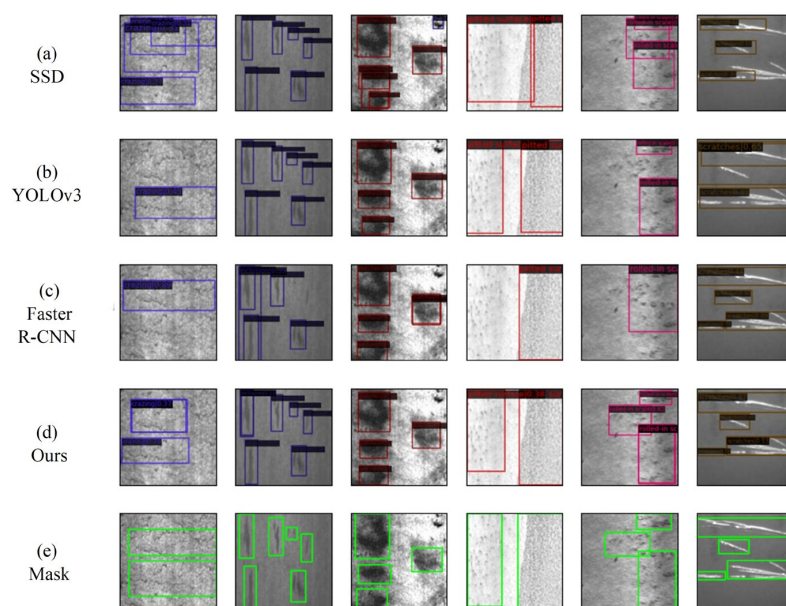


Figure 11. (a) SSD, (b) YOLOv3, (c) Faster R-CNN, (d) our method, and (e) Mask steel plate defect detection results.

Table 3. Comparing the proposed method with the other approaches for each type of defect.

Methods	mAP_{50}	CZ	IS	PC	PS	RS	ST
SSD	76.1%	52.1%	75.3%	86.5%	81.5%	74.1%	87.2%
YOLOv3	79.6%	57.4%	78.4%	88.5%	86.1%	77.5%	89.8%
Faster R-CNN	79.4%	58.6%	76.7%	89.4%	84.7%	76.7%	90.4%
HyperNet	74.8%	54.1%	68.0%	86.5%	87.0%	65.2%	88.1%
Ours	82.7%	50.2%	87.4%	92.7%	86%	83%	96.9%

Based on quantitative results and qualitative analysis, the multi-scale deformable transformer with iterative query refinement proposed in this paper demonstrates superior comprehensive performance compared to mainstream object detection algorithms for steel plate surface defect detection on the NEU-DET dataset; it demonstrates particularly significant advantages in small target detection. This further validates the performance of the Swin Transformer backbone network, the deformable attention mechanism, and the iterative bounding box refinement strategy that we introduced, making it a powerful tool for steel plate surface defect detection.

4. Conclusions

This paper proposes a novel MDT-Net, aimed at enhancing the accuracy of hot-rolled steel plate defect detection. Specifically, (1) the introduction of Swin Transformer as the backbone network significantly improved the network's multi-scale feature representation capability, effectively capturing both local and global contextual information from steel plate images. (2) The integration of a deformable attention mechanism substantially reduced computational complexity and accelerated model convergence. By applying deformable attention to multi-scale feature maps separately, the detection capability for small targets was enhanced. (3) Through iterative bounding box optimization, localization accuracy was improved, leading to a significant increase in both defect localization precision and overall detection accuracy for the steel plates. Experimental results demonstrate that this method can effectively boost steel plate surface defect detection accuracy, achieving an mAP_{50} of 82.7%. Compared to other models, it exhibits significant advantages in small target

detection, thus offering a new approach for steel plate defect detection. Given the NEU-DET dataset's limitations in size and diversity, future work should prioritize acquiring more complex and varied real-world steel surface defect datasets to enable a robust evaluation of MDT-Net's practical industrial performance.

Author Contributions: Conceptualization, H.W.; investigation, F.Z.; data curation, R.Y.; discussion, H.W. and R.Y.; writing—original draft preparation, H.W.; writing—review and editing, H.W., R.Y. and F.Z.; funding acquisition, H.W. and R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Project of State Key Laboratory of Precision Manufacturing for Extreme Service Performance of Central South University (Grant No. Kfkt2022-13) and the National Natural Science Foundation of China (Grant No. 52505490).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors without undue reservation.

Acknowledgments: All authors acknowledge Haoran Wang's part in the discussions.

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

References

1. Cheng, X.; Yu, J. RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2503911. [[CrossRef](#)]
2. Luo, Q.W.; Fang, X.X.; Liu, L.; Yang, C.H.; Sun, Y.C. Automated visualdefect detection for flat steel surface: A survey. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 626–644. [[CrossRef](#)]
3. Yuan, Z.H.; Ning, H.; Tang, X.Y.; Yang, Z.Z. GDCP-YOLO: Enhancing steel surface defect detection using lightweight machine learning approach. *Electronics* **2024**, *13*, 1388. [[CrossRef](#)]
4. He, Y.; Song, K.C.; Meng, Q.G.; Yan, Y.H. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 1493–1504. [[CrossRef](#)]
5. Zhang, F.; Hu, H.L.; Zou, B.Y.; Luo, M.Z. M4Net: Multi-level multi-patch multi-receptive multi-dimensional attention network for infrared small target detection. *Neural Netw.* **2025**, *183*, 107026. [[CrossRef](#)] [[PubMed](#)]
6. Dong, H.W.; Song, K.C.; He, Y.; Xu, J.; Yan, Y.H.; Meng, Q.G. PGA-Net: Pyramid Feature Fusion and Global Context Attention Network for Automated Surface Defect Detection. *IEEE Trans. Ind. Inform.* **2020**, *16*, 7448–7458. [[CrossRef](#)]
7. Wang, C.M.; Yu, Y.; Yu, J.X.; Zhang, Y.; Zhao, Y.; Yuan, Q.W. Microstructure evolution and corrosion behavior of dissimilar 304/430 stainless steel welded joints. *J. Manuf. Process.* **2020**, *50*, 183–191. [[CrossRef](#)]
8. Zhang, F.; Lv, Q.J.; Pan, B.R.; Wang, Y. Boundary semantic interactive aggregation network for scene segmentation. *Expert Syst. Appl.* **2025**, *272*, 126754. [[CrossRef](#)]
9. Yu, Z.L.; Wu, Y.X.; Wei, B.Q.; Ding, Z.K.; Luo, F. A lightweight and efficient model for surface tiny defect detection. *Appl. Intell.* **2023**, *53*, 6344–6353. [[CrossRef](#)]
10. Han, H.; Yang, R.; Li, S.; Hu, R.; Li, X. SSGD: A Smartphone Screen Glass Dataset for Defect Detection. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
11. Zeng, N.Y.; Wu, P.S.; Wang, Z.D.; Li, H.; Liu, W.B.; Liu, X.H. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [[CrossRef](#)]
12. Huang, Q.H.; Zhang, F.; Zhao, Y.Q.; Duan, J.A. Frequency-domain multi-scale Kolmogorov-Arnold representation attention network for mixed-type wafer defect recognition. *Eng. Appl. Artif. Intell.* **2025**, *144*, 110121. [[CrossRef](#)]
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Cheng, Y.F.; Alexander, C.B. SSD: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. pp. 21–37.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]

16. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
17. Wang, H.; Wang, J.; Luo, F. Research on surface defect detection of metal sheet and strip based on multi-level feature Faster R-CNN. *Mech. Sci. Technol. Aerosp. Eng.* **2020**, *40*, 262–269. [[CrossRef](#)]
18. Dai, X.; Chen, H.; Zhu, C. Research on surface defect detection and implementation of metal workpiece based on improved Faster R-CNN. *Surf. Technol.* **2020**, *49*, 362–371.
19. Zhang, N.; Mi, Z.W. Research on surface defect detection algorithm of strip steel based on improved YOLOV3. In Proceedings of the International Conference on Electronic Materials and Information Engineering, EMIE 2021, Xi'an, China, 9–11 April 2021; Volume 1907, p. 012015.
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the 16th European Conference on Computer Vision, ECCV 2020, Glasgow, UK, 23–28 August 2020; pp. 213–229.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Song, K.; Yan, Y.H. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **2013**, *285*, 858–864. [[CrossRef](#)]
23. Li, Z.G.; Wei, X.M.; Hassaballah, M.; Li, Y.H.; Jiang, X.S. A deep learning model for steel surface defect detection. *Complex Intell. Syst.* **2024**, *10*, 885–897. [[CrossRef](#)]

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.