2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

MDTH: A Multi-Scale Deep Learning Network for Steel Surface Defect Detection with Trans-Ham Feature Fusion

Yihong Wu[†1], Yuhao Guo[†2], Chen Yang[2] and Chao Zhang[1(✉)]

[1] School of software, Henan University, KaiFeng 475004, Henan, China
[2] School of Physics and Electronics, Henan University, KaiFeng 475004, Henan, China
2225050331@henu.edu.cn

Abstract. In steel surface defect detection, accurate identification of various defect types is crucial. However, the diverse morphologies of defects and complex backgrounds encountered in real-world industrial production pose significant challenges for existing object detection networks. To address these challenges, this paper proposes MDTH, a deep learning-based network model built upon YOLOv10. MDTH integrates multi-scale deep convolutional feature extraction with Swin Transformer encoding through an enhanced Hybrid attention mechanism (Trans-Ham). Firstly, the Multi-Angle Perception and Depth-wise separable convolution module (MAPD) captures the edges and texture details of steel surfaces, effectively identifying minor defects. Secondly, the Trans-Ham module extracts comprehensive and fine-grained feature information, enabling the model to focus on both local details and global structures simultaneously. Finally, MPDIoU optimizes the overlap and shape matching of bounding boxes, enhancing the accuracy of defect localization. Experimental results on the NEU-DET and PKU-Market-PCB datasets demonstrate that the proposed MDTH model achieves mean average precisions of 78.2% and 95.3% at IoU threshold 0.5, respectively. These results significantly surpass those of commonly used models, highlighting the effectiveness of the added modules and the superior performance of MDTH in defect detection tasks.

Keywords: Defect detection, Hybrid attention mechanism, MPDIoU loss function, Swin Transformer, YOLOv10

1. Introduction

Steel remains a crucial structural material in various industries like construction, automotive manufacturing, and aerospace engineering due to its exceptional mechanical strength and cost efficiency [1]. Despite its benefits, surface imperfections like scratches and cracks frequently arise during production, impacting its longevity [2]. With the increasing need for top-notch steel, ensuring its reliability is paramount. Hence, precise and advanced defect detection techniques are imperative to uphold standards and mitigate potential hazards.

In recent years, deep learning approaches have greatly boosted the ability to detect surface defects. Models such as SSD [3], YOLO [4], and DETR [5] have advanced one-

stage detection, while R-CNN variants [6] have improved two-stage detection. However, achieving high detection accuracy in steel defect detection remains challenging due to diverse defect morphologies and complex backgrounds, which often lead to missed or misclassified defects. To tackle these challenges, the multi-angle perception and depth-wise separable convolution module is incorporated into the YOLOv10 framework. This module is designed to dynamically capture fine-grained texture details across various scales and orientations while maintaining low computational cost. As a result, it facilitates robust feature extraction, effectively distinguishing subtle defect variations even in complex industrial environments.

Transformer architectures [7] overcome the constraints of convolutional networks by capturing global dependencies. Traditional YOLO detectors, on the other hand, underperform due to limited global feature integration. To address this issue, we introduce a novel approach by integrating a Trans-Ham module into the YOLO framework, resulting in a hierarchical feature fusion architecture. The introduced shifted window attention mechanism effectively captures multi-scale contextual patterns, thereby improving detection accuracy by facilitating cross-scale feature interactions.

The work's contribution can be summarized as below:

1. We propose an advanced MAPD module to accurately extract boundary and texture information from steel defect images. This module captures both local structures and global context, enhancing the model's capacity to analyze fine-grained image details and thereby significantly improving defect recognition performance.

2. To effectively extract and emphasize critical positional information, we propose a module integrating Swin Transformer and an enhanced hybrid attention mechanism into YOLOv10. This integration highlights crucial positional key feature regions, enhancing target localization accuracy and the capture of image details. Consequently, it improves recognition accuracy and model robustness, elevating overall performance.

3. The introduction of the MPDIoU loss function enhances the detection of irregular-shaped defects and improves localization accuracy in complex backgrounds. This results in higher prediction precision and significantly boosts the overall detection performance.

4. This paper proposed MDTH model, integrating multi-scale deep convolutional feature extraction with Trans-Ham feature fusion, provides a deeper, more re-fined understanding of the defect images, significantly enhancing image recognition capabilities.

2. Related Work

2.1 Defect Detection on Steel Surface

Traditional defect detection methods on steel often utilize manual operations like surface visual inspection [8] and assess the degree of defects, whose results are easily influenced by human factors, leading to low detection accuracy and speed. Then, machine learning methodologies emerged as powerful tools to substantially elevate defect detection capabilities, drastically reducing manufacturers' workload and bolstering the

efficiency of identifying and classifying surface defects through automation. Notable examples include thresholding methods [9], SVMs [10], and decision trees [11]. However, these techniques are constrained by external factors such as lighting conditions and necessitate manual extraction of feature information, posing challenges to meet the rigorous demands of real-world steel production environments.

Steel surface defect detection has seen a significant improvement in accuracy and speed with the help of deep learning. However, detecting tiny or irregular defects, which are difficult to locate with traditional methods. Therefore, more research is needed to achieve robust and effective defect detection that can handle multiple complex types of defects. To address multi-scale feature extraction in steel surface defect detection, Liu et al. [12] developed MSC-DNet, which employs parallel dilated convolutions with varying dilation rates and an adaptive feature enhancement module to strengthen characteristic defect pattern representation. Demir et al. [13] proposed a dual-path architecture integrating residual connections with attention mechanisms, effectively learning discriminative feature representations for surface defect classification through coordinated spatial-channel interactions. For detecting atypical defect patterns, Sharma et al. [14] designed a multi-task framework incorporating cascaded binary classification with hybrid detection-segmentation modules, demonstrating improved robustness against irregular defect morphologies through hierarchical feature verification.

2.2 Transformer Architecture

Recent work shows that convolutional neural networks employing the encoder-decoder framework have achieved significant advancements in defect detection applications. The Vision Transformer (ViT) proposed by Dosovitskiy et al. [15] was one of the first to treat images as sequences, demonstrating the potential of Transformer architectures in image classification tasks. This work highlighted the ability of Transformer-based models to capture long-range dependencies in images, leading to improved classification performance. Lin et al. [16] introduced a hierarchical structure and sliding window mechanism to enhance the processing of high-resolution images using Transformers. This advancement enabled Transformers to handle larger input sizes more effectively. In the realm of object detection, DETR simplified traditional detection frameworks by applying Transformer architecture. DETR improved detection accuracy and reduced computational complexity by replacing region proposal networks with its self-attention mechanism. Wang et al. [17] proposed Defect Transformer to refine defect localization, enhancing the model's capacity to capture both local and global relationships, thereby improving defect detection precision. ETDNet, a lightweight ViT-based detection network introduced by Zhou et al. [18], decouples local and global feature extraction to enhance defect detection. By integrating local feature representation with global feature aggregation, ETDNet demonstrated a significant performance enhancement. These studies collectively underscore the increasing significance and efficacy of Transformer-based models in defect detection, showcasing their potential to enhance detection accuracy and efficiency by capturing intricate spatial relationships in images.

3. Method

3.1 The overall architecture

This study introduces MDTH, a novel framework developed to improve steel defect detection based on YOLOv10. As illustrated in Figure 1, MDTH analyzes steel defect images using a sequence of purposefully crafted components. Initially, low-level features are extracted from the input images and fed into the backbone that expands the receptive field to encompass extensive high-level features. The enhanced features are then directed to two key modules: Trans-Ham and MAPD. Trans-Ham combines hybrid attention with a transformer architecture to model spatial and contextual relationships, improving defect detection accuracy. The MAPD module focuses on capturing local structures and global information through its Multi-Angle Perception mechanism. This module processes features through parallel branches, using channel compression and spatial feature extraction to enhance detail capture for small defects. Finally, the MPDIoU loss function is employed to further optimize the detection process.

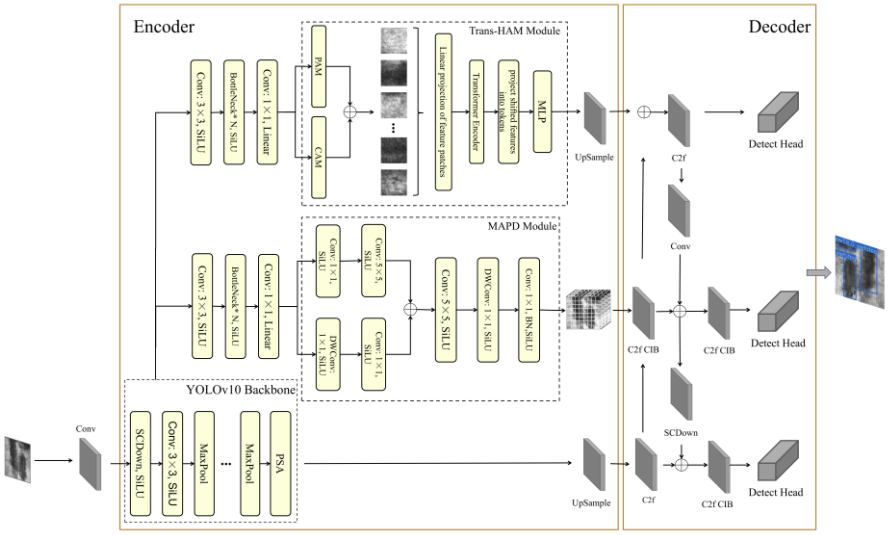


Fig. 1. The diagram illustrates the proposed MDTH structure, which consists of, from left to right, the YOLOv10 backbone, MAPD module, Trans-Ham module, and Detect Head.

3.2 MAPD module

Traditional convolutional neural networks, like the VGG network [19], use fixed kernel sizes and single-path processing. This limits their ability to capture spatial coordinate features and multi-angle information, reducing their effectiveness in precisely detecting small, unevenly distributed objects.

To solve the problem, we propose the MAPD module to capture both local structures and global information. Fig. 2 shows its structure. The Multi-Angle Perception (MAP)

mechanism is introduced to enhance spatial information capture. C is the number of channels, and H and W are the height and width, MAP processes feature map through two parallel branches.

In the first branch, a 1×1 convolution compresses channels by a learnable factor of α, reducing the channel dimension to αC while maintaining spatial dimensions. This compression helps to reduce computational complexity and focus on the most relevant channel information. Subsequently, a 5×5 convolution extracts spatial features, helping capture richer details, particularly for minor defects. The second branch utilizes depth-wise separable convolution [20] to separate channel and spatial convolutions, thereby enhancing the model's awareness of spatial dimensions. This multi-path process approach fuses integrates multi-angle information from various angles, boosting the detection of challenging samples. After the parallel branches, another 5×5 convolution deeply extracts features from fused outputs and expands the feature map's channels. This step not only integrates multi-angle information but also balances the feature representation by adjusting the number of channels. Then, depth-wise separable convolution is used again to further extract spatial information and enhance the model's ability to capture more defective features. Finally, a residual connection merges the feature map with the concatenated branch features to prevent feature loss and produce the final feature map. By strategically compressing and expanding the number of channels, the MAPD module effectively balances computational efficiency and feature richness. This design ensures that the module can capture detailed spatial information while maintaining a comprehensive understanding of the feature space, leading to improved detection accuracy for steel defects.

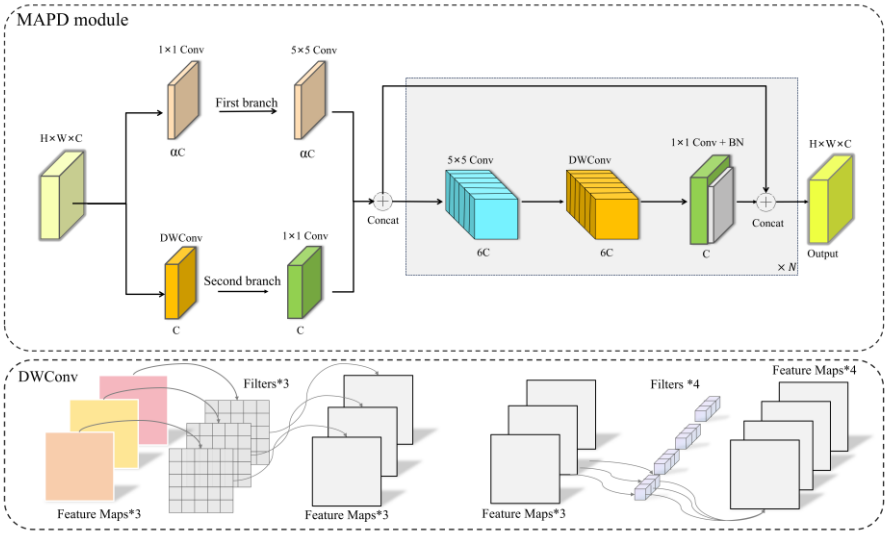


Fig. 2. The workflow of MAPD module

3.3 Trans-Ham Module

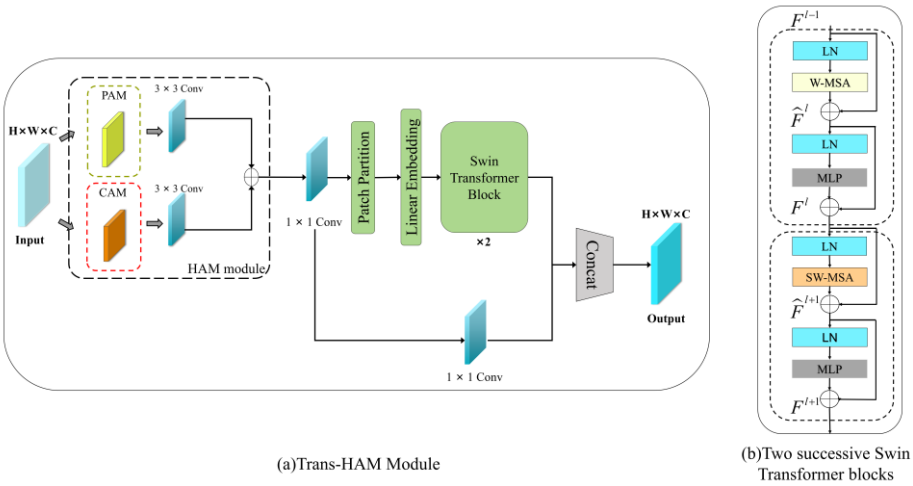
Detecting defects in steel is challenging due to variations in their position, shape, and size, which can hinder accurate localization by models during training. Additionally, the similarity between defect patterns and background textures can lead to missed detections. To address these issues, we incorporated the Swin Transformer and an enhanced Hybrid attention module [21] (Trans-Ham) into our model, enhancing its ability to precisely localize and classify defects. The structure of Trans-Ham is depicted in Fig. 3.

The Trans-Ham module integrates the attention mechanism and the feature expression capabilities of the Transformer to effectively capture both local details and global context in images, thereby enhancing the accuracy of defect localization and classification. This module employs an enhanced HAM module to extract key defect information from the image. The enhanced HAM module preserves the complementary strengths of the Positional Attention Mechanism (PAM) and Channel Attention Mechanism (CAM) [22], facilitating dynamic feature interaction through joint positional and channel attention calculations, followed by merging the outputs through weighted splicing. By incorporating positional attention, which captures inter-position correlations in the feature map, and channel attention, which emphasizes channel importance, the module enables a more comprehensive understanding of image features. The introduction of weighted splicing further enhances the model's feature fusion capabilities, ensuring the retention of detailed information during feature extraction, capturing global semantic information, and providing a richer feature representation for subsequent defect localization and classification tasks. Here is the formulation of the enhanced HAM module:

$$H_{PAM} = PAM(H) = S\text{oftmax}(\frac{Q_P K_P^T}{\sqrt{d_k}})\mathrm{V}_P + H \tag{1}$$

$$H_{CAM} = CAM(H) = S\text{oftmax}(\frac{Q_C K_C^T}{\sqrt{d_K}})\mathrm{V}_C + H \tag{2}$$

$$H_{fuse} = \gamma Conv(H_{PAM}) + (1-\gamma)Conv(H_{CAM}) \tag{3}$$



(a)Trans-HAM Module

(b)Two successive Swin Transformer blocks

Fig. 3. The architecture of Trans-Ham module

Subsequently, the combined feature map undergoes processing within the Swin Transformer framework, where it is partitioned into non-overlapping image blocks of fixed sizes, with each block treated as a 'token'. The mathematical representation of the Swin Transformer block is depicted by the following equations:

$$F^{l} = W-MSA(LN(F^{l-1})) + F^{l-1} \tag{4}$$

$$F^{l} = MLP(LN(F^{l})) + F^{l} \tag{5}$$

$$F^{l+1} = SW-MSA(LN(F^{l})) + F^{l} \tag{6}$$

$$F^{l+1} = MLP(LN(F^{l+1})) + F^{L+1} \tag{7}$$

Where, LN denotes Layer Normalization, which normalizes input features to stabilize the training process. $W-MSA$ represents Window-based Multi-Head Self Attention (MSA), while $SW-MSA$ refers to Shifted Window Multi-Head Self Attention. $W-MSA$ focuses on local context extraction by restricting attention computation within non-overlapping windows, while $SW-MSA$ expands the receptive field through a window-shifting strategy, enabling the capture of broader spatial relationships. This combination enhances the model's ability to understand complex patterns in defect detection tasks. MLP denotes Multi-Layer Perceptron, which consists of two fully connected layers with a $GELU$ activation function [23], facilitating further feature transformation and enhancement. The inclusion of MLP increases the model's expressive power.

The operation of the Shifted Window mechanism is illustrated in Fig. 4. As shown in Fig. 4.(b), the input image is initially partitioned into non-overlapping windows using

$W-MSA$. In the subsequent step, the yellow and red regions which highlighted in Fig. 4.(c). are reassigned to the lower-right position, as depicted in Fig. 4.(d). The strategic re-positioning of the regions enriches the interaction between different parts of the image. This shifting process enables the model to capture a broader range of spatial dependencies.

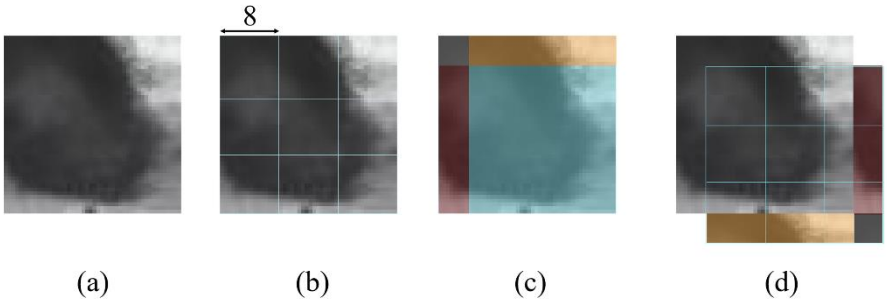


Fig. 4. Flowchart of the operation of the Shifted Window mechanism

3.4 MPDIoU loss function

In real-world industrial production, bounding box prediction is prone to be affected by the irregular shapes and various sizes of different defect categories. To address this issue, we employ MPDIoU [24], a groundbreaking loss function that enhances bounding box regression by considering multiple points on the bounding boxes rather than just the center points. The MPDIoU loss provides a more comprehensive metric for objects with irregular shapes and varying sizes, leading to improved accuracy in defect detection tasks. The MPDIoU loss function is shown below:

$$d_1^2 = (x_{\min}^A - x_{\min}^B)^2 + (y_{\min}^A - y_{\min}^B)^2 \tag{8}$$

$$d_2^2 = (x_{\max}^A - x_{\max}^B)^2 + (y_{\max}^A - y_{\max}^B)^2 \tag{9}$$

$$L_{MPDIoU} = 1 - IoU + \frac{d_1^2 + d_2^2}{h^2 + w^2} \tag{10}$$

For the bounding boxes A and B, the terms d_1^2 and d_2^2 measure the squared Euclidean distances between the corresponding corners of the two bounding boxes. The coordinates $(x_{\min}^A, y_{\min}^A)$ and $(x_{\max}^A, y_{\max}^A)$ represent the top-left and bottom-right corners of bounding box A, respectively. Similarly, the coordinates $(x_{\min}^B, y_{\min}^B)$ and $(x_{\max}^B, y_{\max}^B)$ correspond to the top-left and bottom-right corners of bounding box B. The dimensions w and h denote the width and height of the smallest enclosing box that completely covers both bounding boxes A and B.

When the predicted box is distant from the groundtruth box's center without overlapping, as illustrated in Fig. 5(a), the MPDIoU loss accounts for both factors. In this

scenario, the DIoU [25] and CIoU [26] losses are 1.29, whereas the MPDIoU loss notably increases to 1.58. This holistic consideration yields more informative gradients, facilitating quicker and more accurate convergence in bounding box regression. In cases where the predicted box overlaps with the real box, as depicted in Fig. 5(b), the MPDIoU loss integrates IoU overlap and a corner distance penalty. Here, the DIoU and CIoU losses measure 0.97, while the MPDIoU loss is 1.08. Unlike CIoU, which emphasizes differences in center distance and aspect ratio, MPDIoU also incorporates corner distances, making it particularly suitable for targets with substantial spatial misalignment or irregular shapes. This balanced integration of geometric factors enables MPDIoU to effectively capture positional and shape disparities between bounding boxes, enhancing regression precision and furnishing detailed gradient insights for optimization.

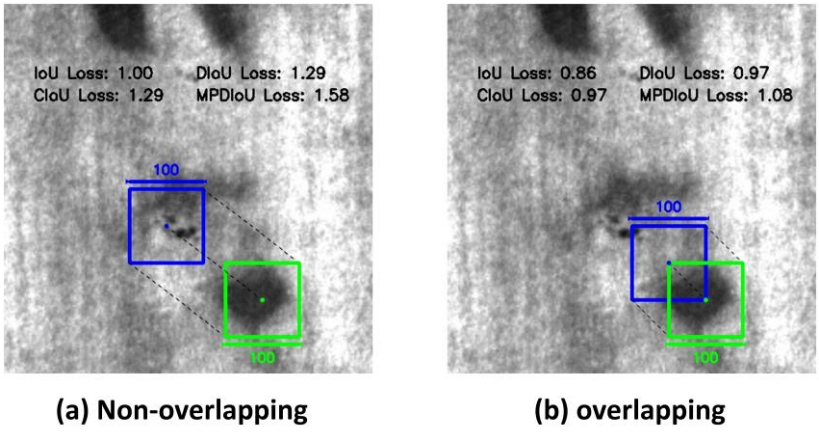


Fig. 5. Two cases with different boxes of regression results. The green boxes represent the groundtruth boxes and blue boxes represent predicted boxes. (a) The result of non-overlapping bounding box; (b) The result of overlapping bounding box.

4. Experiments

4.1 Data Description

To demonstrate the validity of MDTH, we performed experiments on two distinct datasets: NEU-DET dataset [27] and PKU-Market-PCB dataset [28]. The dataset contains 1,800 grayscale images in total, organized into six defect types: Crazing, Inclusion, Patches, Pitted Surface, Rolled-in Scale and Scratches, which is shown in Fig. 6. Each defect category has 300 annotated images, with labels indicating defect type and location. To manage data efficiently, we randomly selected 1620 for the training sets, with 180 images for the test sets. To further assess the generalization ability and robustness of MDTH, we conducted additional experiments on the PKU-Market-PCB dataset. This dataset comprises 693 defect images with an average resolution of 2046×2016 pixels

and is also categorized into six defect types: Missing Hole, Mouse Bite, Open Circuits, Short, Spur, and Spurious Copper. The defect types of PKU-Market-PCB are illustrated in Fig. 7. The dataset was split at an 8:1:1 ratio for training, validation, and testing.

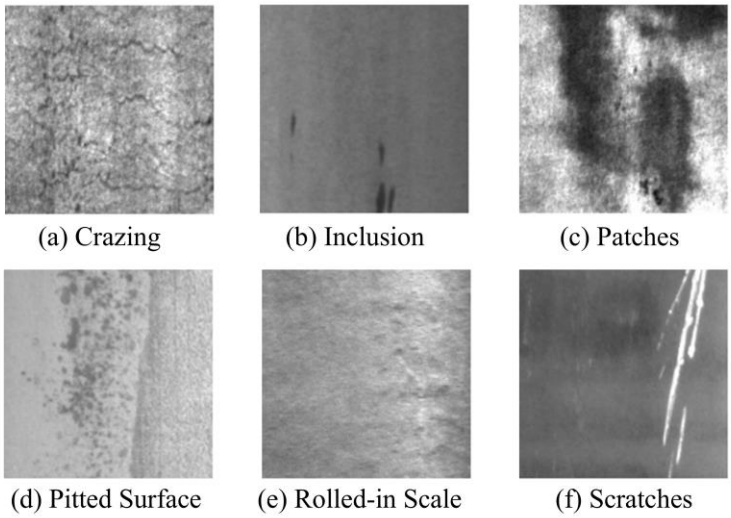


Fig. 6. Defect types of NEU-DET dataset

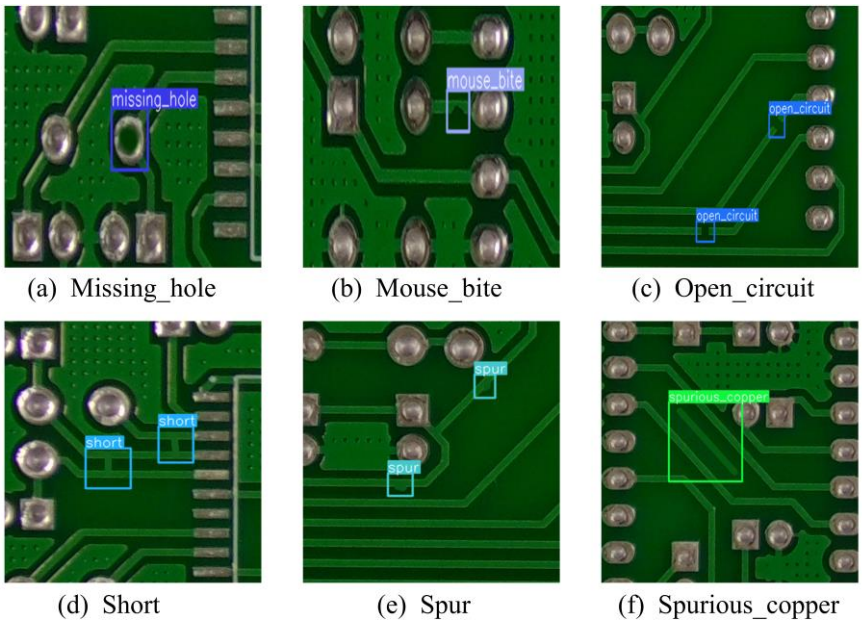


Fig. 7. Defect types of PKU-Market-PCB dataset

4.2 Experimental platform and hyperparameter setting

To validate the effectiveness of the proposed MDTH model, all experiments were conducted on an NVIDIA GeForce RTX 3090 GPU (24GB) in a Pytorch 2.0.1 environment. The hyperparameter settings for the models are presented in Table 1.

Table 1. The experiment settings

Hyperparameters	NEU-DET	PKU-Market-PCB
Optimizer	SGD	SGD
Learning rate	0.01	0.01
Weight decay	0.0005	0.0005
Batch size	16	16
epochs	300	300
Image size	640×640	640×640

4.3 Evaluation metrics

The common and wide used indicators for evaluating defect detection model are the precision (P), recall (R), parameters and mAP@0.5. Mean average precision is often considered a comprehensive metric across all defect types for evaluating defect detection models. Here is definition of P, R and mAP@0.5:

$$precision = \frac{TP}{TP + FP} \tag{11}$$

$$recall = \frac{TP}{TP + FN} \tag{12}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{13}$$

In the formula, TP (True Positive) denotes the number of cases where defects have been accurately detected, FP (False Positive) denotes the number of cases where defects have been inaccurately detected. FN (False Negative) denotes the number of cases that have been falsely judged as non-defective.

4.4 Ablation Experiment

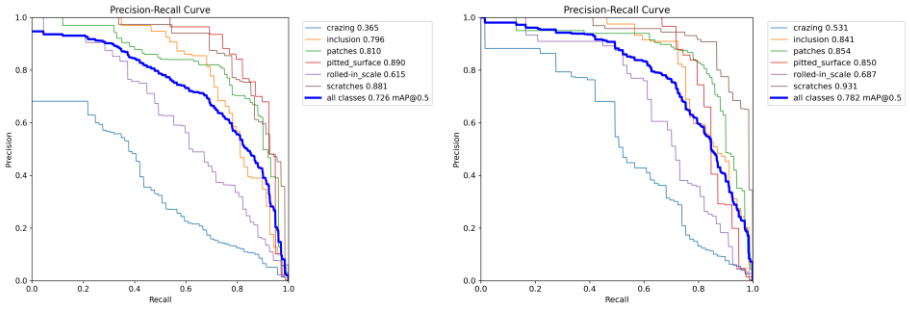
To evaluate the effectiveness of the proposed modules, we conducted a series of ablation experiments on the NEU-DET dataset, systematically comparing the impact of different enhancement strategies on model performance. The detailed evaluation metrics are presented in Table 2. First, we evaluated the precision, recall, and mAP@0.5 metrics of the original YOLOv10n model, with results of 67.0%, 66.9%, and 72.6%, respectively. Next, we introduced the MAPD module, which improved the model's

Precision by 7.2% and mAP@0.5 by 1.9%. We then introduced the Trans-Ham module, which reached 75.6%, 67.3%, and 75.2% on the precision, recall and mAP@0.5 metrics, respectively. In addition, by replacing the original loss function with MPDIoU, the precision, recall and mAP@0.5 metrics were improved by 2.6%, 3.8%, and 1.7% on the original model, respectively. Ultimately, our model improved by 13.9%, 1.6%, and 5.6% on the Precision, Recall, and mAP@0.5 metrics compared to the original YOLOv10 model, respectively. These results fully demonstrate the effectiveness of the modules we added and the superior performance of the baseline model in defect detection tasks.

Table 2. Ablation results on NEU-DET

MAPD	Trans-Ham	MPDIoU	Precision (%)	Recall (%)	mAP@0.5 (%)
			67.0	66.9	72.6
√			74.2	66.3	74.5
	√		75.6	67.3	75.2
		√	69.6	**70.7**	74.3
√	√		78.8	67.6	77.6
√		√	75.1	67.9	75.4
	√	√	78.2	68.2	77.1
√	√	√	**80.9**	68.5	**78.2**

Fig. 8 shows the comparison of the P-R results before and after the improvement, from which it can be seen that the proposed MDTH model has different degrees of improvement for each class, the AP value of the crazing class is improved from 36.5% to 53.1%, which is 16.6 percentage points; the AP value of the inclusion class is improved from 79.6% to 84.1%, which is 4.5 percentage points; the AP value of the patches class AP value from 81.0% to 85.4%, an improvement of 4.4 percentage points, etc. The AP values of many different defective target detections are all improved, reflecting the good generalization ability of the model.



YOLOv10 on NEU-DET dataset MDTH on NEU-DET dataset

Fig. 8. P-R comparison between YOLOv10 and MDTH on NEU-DET dataset

4.5 Comparison and analysis of different object detection models

To assess the model's generalization capability, we conducted extensive comparative experiments on both the NEU-DET and PKU-Market-PCB datasets. We compared our model with recent single-stage object detection methods, including RT-DETR, YOLOv5, YOLOv8, YOLOv10, and YOLOv11. Table 3 presents a comparative analysis of the proposed method in terms of mAP@0.5 and parameters (params). Our model achieved the highest mAP@0.5 value of 78.2 on NEU-DET dataset and 95.3 on PKU-Market-PCB dataset with an additional 0.16M parameters. This performance surpasses that of other mainstream models, demonstrating a favorable balance between accuracy and complexity for industrial defect detection tasks.

Table 3. Comparison results on NEU-DET dataset and PKU-Market-PCB dataset

Model	NEU-DET		PKU-Market-PCB		Params
	mAP@ 0.5	mAP@ 0.5-0.95	mAP@ 0.5	mAP@ 0.5-0.95	
RT-DETR	72.6%	40.7%	90.5%	44.0%	3.19M
YOLOv5	71.3%	39.4%	90.3%	44.1%	**1.76M**
YOLOv8	75.6%	41.5%	90.7%	43.7%	2.25M
YOLOv10	72.6%	38.5%	90.9%	46.1%	2.69M
YOLOv11	76.2%	41.2%	91.7%	47.8%	2.60M
Our model	**78.2%**	**43.1%**	**95.3%**	**63.3%**	2.85M

Fig. 9 illustrates the visual comparison of detection results between the proposed MDTH algorithm and mainstream methods, including RT-DETR, YOLOv5, YOLOv8, YOLOv10, and YOLOv11, across two different datasets. As observed in the figure, the MDTH model exhibits superior performance in detecting defects under complex backgrounds, demonstrating a clear advantage over existing mainstream algorithms. Furthermore, the MDTH algorithm effectively identifies various defect types from multiple perspectives with high precision. Regardless of scene complexity or defect diversity, the proposed method consistently achieves robust and reliable detection performance, further underscoring its effectiveness in practical defect detection applications.

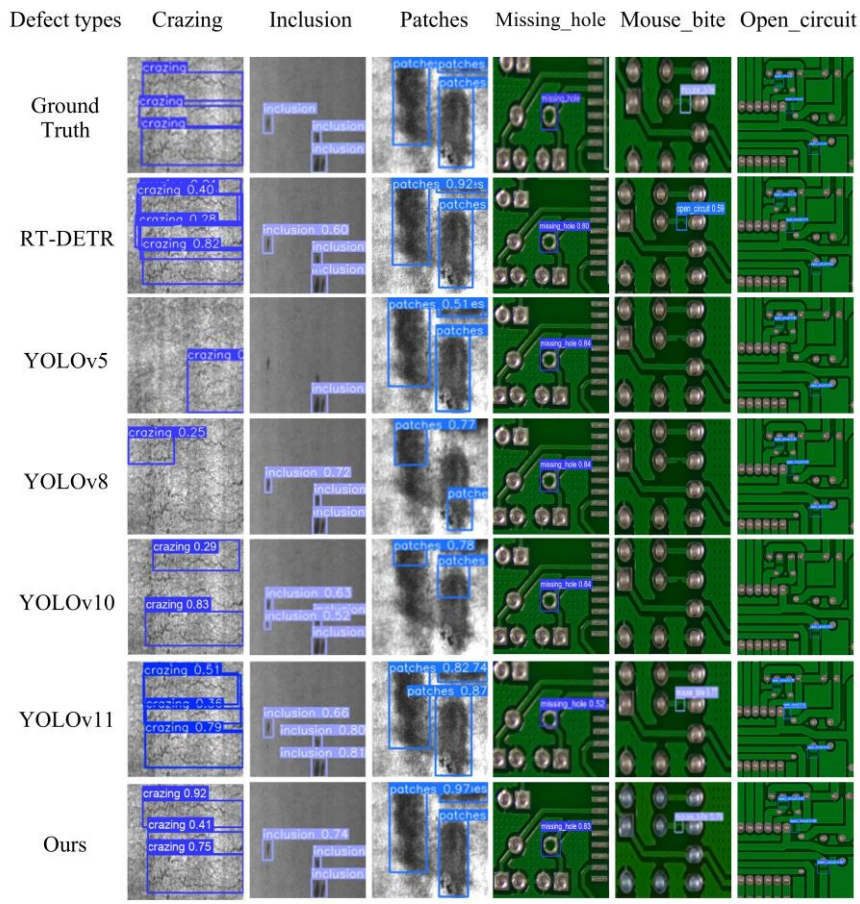


Fig. 9. Comparison of visualization results

5. Conclusion

In this paper, we propose MDTH, a multi-scale deep learning network for steel surface defect detection, which integrates Trans-Ham feature fusion. To enhance the ability to extract fine-grained defect features, we first employ the Multi-Angle Perception and Depth-wise separable convolution (MAPD) module to capture the edge and texture details of the steel surface. Subsequently, the Trans-Ham module is utilized to extract richer and more detailed feature information, allowing the model to simultaneously focus on both local details and global structures. Finally, we introduce MPDIoU to optimize the overlap and shape matching of the bounding boxes, thereby accelerating the model's convergence and reducing the loss, which in turn enhances its robustness.

We conduct experiments on the NEU-DET and PKU-Market-PCB datasets to further evaluate the effectiveness of the proposed model. The experimental results demonstrate

that the MDTH model achieves mAP@0.5 scores of 78.2% and 95.3%, respectively, yielding improvements of 5.6% and 4.8% over the baseline model. When faced with minute and irregularly shaped defects, MDTH precisely pinpoints and classifies them, representing a major advancement in defect detection algorithms.

For future research, we plan to focus on three key directions: (1) improving the model to reduce computational complexity, (2) exploring various lightweight modules to identify the most efficient detection algorithms, and (3) extending the MDTH model to accommodate a wider range of application domains.

Acknowledgments. This study was funded by the Henan Provincial Department of Education (grant numbers 202410475105).

References

1. Baddoo, N.R.: Stainless steel in construction: A review of research, applications, challenges and opportunities. J. Constr. Steel Res. 64(11), 1199–1206 (2008).
2. Yu, H.L., Tieu, K., Lu, C., Deng, G.Y., Liu, X.H.: Occurrence of surface defects on strips during hot rolling process by FEM. Int. J. Adv. Manuf. Technol. 67, 1161–1170 (2013).
3. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single Shot MultiBox Detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing, Part I, pp. 21–37 (2016).
4. 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, pp. 779–788 (2016).
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020).
6. Sumit, S.B., Joshi, S., Rana, U.: Comprehensive Review of R-CNN and its Variant Architectures. Int. Res. J. Adv. Eng. Hub 2(4), 959–966 (2024).
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).
8. Czimmermann, T., Ciuti, G., Milazzo, M., Chiurazzi, M., Roccella, S., Oddo, C.M., Dario, P.: Visual-based defect detection and classification approaches for industrial applications—A survey. Sensors 20(5), 1459 (2020).
9. Ng, H.F.: Automatic thresholding for defect detection. Pattern Recognit. Lett. 27(14), 1644–1649 (2006).
10. Gong, R., Wu, C., Chu, M.: Steel surface defect classification using multiple hyper-spheres support vector machine with additional information. Chemom. Intell. Lab. Syst. 172, 109–117 (2018).
11. Aghdam, S.R., Amid, E., Imani, M.F.: A fast method of steel surface defect detection using decision trees applied to LBP based features. In: Proceedings of the 2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA), Singapore, 18–20 July 2012; pp. 1116–1120 (2012).
12. Liu, R., Huang, M., Gao, Z., Cao, Z. & Cao, P.: MSC-DNet: An efficient detector with multi-scale context for defect detection on strip steel surface. Measurement 209, 112467 (2023).

13. Demir, K., Ay, M., Cavas, M., Demir, F.: Automated steel surface defect detection and classification using a new deep learning-based approach. Neural Comput. Appl. 35, 2123–2136 (2023).
14. Sharma, M., Lim, J., Lee, H.: The Amalgamation of the Object Detection and Semantic Segmentation for Steel Surface Defect Detection. Appl. Sci. 12(12), 6004 (2022).
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, (2020).
16. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022, (2021)
17. Wang, J., Xu, G., Yan, F., Wang, J., & Wang, Z.: Defect transformer: An efficient hybrid transformer architecture for surface defect detection. Measurement: Journal of the International Measurement Confederation 211, 112614 (2023).
18. Zhou, H., Yang, R., Hu, R., Shu, C., Tang, X., & Li, X.: ETDNet: Efficient Transformer-Based Detection Network for Surface Defect Detection. IEEE Transactions on Instrumentation and Measurement 72, 3307753 (2023).
19. Simonyan, K., & Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015).
20. Zhang, R., Zhu, F., Liu, J., & Liu, G.: Depth-Wise Separable Convolutions and Multi-Level Pooling for an Efficient Spatial CNN-Based Steganalysis. IEEE Transactions on Information Forensics and Security 15, 3240–3253 (2020).
21. Li, G., Fang, Q., Zha, L., Gao, X., & Zheng, N.: HAM: Hybrid attention module in deep convolutional neural networks for image classification. Pattern Recognition 129, 108785 (2022).
22. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019-June (2019).
23. Hendrycks, Dan, and Kevin Gimpel. "Gaussian error linear units (gelus)." arXiv preprint arXiv:1606.08415 (2016).
24. Ma, Siliang, and Yong Xu. "Mpdiou: a loss for efficient and accurate bounding box regression." arXiv preprint arXiv:2307.07662 (2023).
25. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12993–13000 (2020).
26. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D.: Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans. Cybern. 52, 2255–2266 (2022)
27. Song, K., & Yan, Y.: A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Appl. Surf. Sci. 285, 619–624. (2013)
28. Ding, R., Dai, L., Li, G., & Liu, H. TDD-net: a tiny defect detection network for printed circuit boards. CAAI Transactions on Intelligence Technology, 4(2), 110-116. (2019)