A Multi-scale Defect Detection Network for Wind Turbines Utilizing Margin Aware Features

SI, Yuxin <http://orcid.org/0009-0005-8852-7919>, DING, Yunfei <http://orcid.org/0000-0001-5223-7585>, GE, FuDi <http://orcid.org/0009-0004-6477-5398>, WU, Xingtao <http://orcid.org/0009-0008-1785-3671>, LIU, Jinglin, DING, Dong and ZHANG, Hongwei <http://orcid.org/0000-0002-7718-021X>

Measurement Science
and Technology

ACCEPTED MANUSCRIPT

A Multi-scale Defect Detection Network for Wind Turbines Utilizing Margin Aware Features

To cite this article before publication: Yuxin Si *et al* 2025 *Meas. Sci. Technol.* in press https://doi.org/10.1088/1361-6501/ae08db

Journal XX (XXXX) XXXXXX Author et al

A Multi-scale Defect Detection Network for Wind Turbines Utilizing Margin Aware Features

Yuxin Si[1], Yunfei Ding[1*], Fudi Ge[1], Xingtao Wu[1], Jinglin Liu[2], Dong Ding[1], Hongwei Zhang[3]
[1] School of Electrical Engineering, Shanghai Dianji University, Shanghai 201306, China
[2] School of Energy and Power Engineering, Jiangsu University, Jiangsu 212013, China
[3] Advanced Food Innovation Centre, Sheffield Hallam University, Sheffield S1 1WB, UK
*Correspondence: dingyf@sdju.edu.cn;

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

The long-term operation of wind turbines (WTs) leads to multi-scale surface defects that critically compromise operational reliability. Drone-based defect detection offers a viable approach for real-time assessment of WT operational status. However, the current deployment of UAV-based detection systems struggles to simultaneously achieve both sensitivity and positioning accuracy for such multi-scale defects. To address this limitation, we propose a novel Defect Marginal-aware and Multi-scale Collaborative Attention Network (DMCA-Net). First, we propose a Defect Marginal Detail Transfer backbone (DMDT) to enhance edge information in shallow features, which can be fused with multi-scale features. Second, a Triple-layer Anchor Attention Feature Selection and Fusion Pyramid Network (TAAFSFPN) is introduced to optimize channel-space interactions, which can dynamically balance local details and global features, thereby improving defect localization accuracy. In addition, a Histogram-based Synergistic Attention Head encoder (HSAH) is designed to detect small object defects by co-optimizing frequency-domain split-box attention and cross-box attention to enhance the feature intensity of small object defects. Finally, the Normalized Wasserstein Distance–Inner Distance–IoU (NWD-InnerDIoU) loss is introduced to enhance model generalization and mitigate severe data imbalance, effectively reducing performance fluctuations resulting from interactions among multi-scale targets. Experimental results demonstrate that DMCA-Net achieves state-of-the-art performance with 83.1% mAP50, representing a 3.1% improvement over baseline, while maintaining real-time detection capability at 81.3 FPS on the WT defect dataset. Especially, it outperforms commonly used detection models in terms of detection performance.

Keywords: Fusion pyramid, Real-time detection, Wind Turbine, Edge information, Defect Detection

1. Introduction

Under the dual pressures of global fossil energy scarcity and climate change, human civilization is transitioning from "black" to "green" energy. As core clean energy facilities, WT is expanding rapidly[1]. Critical components such as WT blades, when chronically exposed to extreme weather, salt spray corrosion, and mechanical loads, are prone to surface defects including cracks, damage, and erosion. These defects compromise aerodynamic performance, cause structural failures, and may even lead to fracture incidents. Statistics indicate that operation and maintenance costs account for 25%–30% of a WTs lifecycle expenditure, with surface defects contributing to over 40% of these failures[2].

Conventional defect detection methods include manual inspection, ultrasonic testing, infrared thermography, and robotic inspection. However, manual inspection poses significant risks due to the necessity of working at height and is susceptible to blind spots, whereas non-destructive techniques (e.g., ultrasonic and thermographic testing) require specialized equipment and highly trained personnel. These approaches often lead to slow detection rates, high operational costs, and limited capabilities for real-time monitoring. The deployment of UAVs not only adapts to complex terrains and eliminates risks associated with manual high-altitude inspections but also enables millimeter-level crack identification and comprehensive coverage through mounted high-resolution equipment[3].

Integrating UAVs with deep learning-based defect detection frameworks facilitates holistic monitoring of wind turbine operational status. Consequently, intelligent detection technologies are essential to support efficient wind power operations and maintenance under the "Dual Carbon" goal[4].

In recent years, advancements in deep learning technology have opened a novel path for the intelligent operation and maintenance of WTs. Currently, there are two types of models applied to WT defect detection: one is the convolutional neural network represented by R-CNN [5-7] and YOLO [8-10] series of algorithms. The second is the Transformer-based DETR family of frameworks [11]. Diaz and Davis' team built a detection model based on Cascaded Mask R-CNN and improved Mask R-CNN [12][13], which reduces the model complexity but weakens the feature characterisation capability. Gu et al. proposed WT-YOLO to enhance wind turbine detection accuracy by integrating an Omni-Dimensional Dynamic Network (ODDNet) and a Pyramid Squeeze Attention Network (PSANet). However, this framework overlooks a critical factor in WT blade performance[14]. Zhang et al. proposed GCB-YOLO, a lightweight detection architecture that significantly reduces model size through the integration of GhostNet and a Bidirectional Feature Pyramid Network (BiFPN). Unfortunately, this approach over-emphasizes surface-level features while inadequately extracting defect edge characteristics[15]. Yu et al. introduced the YOLOv8-WTDD algorithm, which enhanced multi-scale defect detection accuracy by incorporating the Diverse Branch Block (DBB) module and Receptive Field Attention Network (RFANet). Nevertheless, its detection capability shows limitations when applied to small-scale defect identification[16]. The Transformer-based Real-Time Detection Transformer (RT-DETR) algorithm integrates the ResNet backbone with an Attention-based Intrascale Feature Interaction (AIFI) hybrid encoder and an enhanced decoder structure. This design achieves significant improvements in inference speed while preserving the global modeling capabilities of DETR. By incorporating the DINO denoising learning strategy to accelerate convergence, this framework constitutes a state-of-the-art detection solution tailored for real-time UAV streaming data processing demands[17]. Zhao et al. proposed WTNet, a lightweight detection model that refined RT-DETR through a Reparameterized Efficient Layer Aggregation Network (RepELANNet) and Sparse Parallel Feature Pyramid (S-FPN). Further improvements are needed to enhance the detection capability for closely spaced minor defects[18].

However, deploying a UAV-based defect identification network in wind farms or natural environments faces significant challenges. Firstly, scale imbalance in defects complicates multi-scale object detection, as corrosion can extend up to 1 m, while cracks may measure only 1 cm[19]. Secondly, image quality is often compromised by cluttered backgrounds, viewpoint variations, and lighting fluctuations. Simultaneously, critical spatial features (crack textures and corrosion patterns, etc.) are obscured, and WT defect edges are blurred. Moreover, conventional detection networks like YOLO and R-CNN, despite effectively extracting global semantic features, exhibit limited localization accuracy due to insufficient attention to fine edge details. These issues are hindering the progress of UAVs in WTs industrial detection.

Therefore, an innovative DMCA-Net WTs defect detection network is proposed in this study, which is specifically designed for UAV-based aerial inspection scenarios and constructed upon the RT-DETR framework. The specific contributions are listed below:

(1) We propose a novel backbone detection network, termed DMDT, to tackle the challenge of inaccurate defect localization caused by weak edge information in small target defects. The proposed framework incorporates a Hierarchical Edge Pyramid Generator (HEPG) to extract robust and noise-resistant edge features. In addition, a Dynamic Feature Selector (DFS) ensures precise granularity alignment between edge features and backbone features. Furthermore, multi-scale edge information is adaptively integrated into the backbone via a Cross-Channel Edge Integrator (CCEI) through a dynamic feature fusion mechanism.

(2) A novel HSAH encoder is given to adaptively partition and hierarchically extract target features in complex scenes. Subsequently, small defect detection accuracy is significantly boosted in challenging environments.

(3) To solve the problem of weak feature spatial information of small object defects triggered by non-structural factors[20], a novel feature fusion network TAAFSFPN is proposed to replace the Cross-Scale Feature-Fusion Module (CCFM). Dynamic integration of local details and global features is achieved through dual-dimensional channel-space modeling.

(4) To enhance model generalization and mitigate performance degradation from mixed-scale targets, we define a proportionality factor as a ratio, which is conbined with DIoU, Inner-IoU, and NWDLoss to construct a novel NWD-InnerDIoU loss function.

2. Related work

2.1 Backbone networks in detectors

As the central feature extraction unit, the backbone network progressively extracts semantic information from the input image. It can construct multi-scale feature maps to provide discriminative features for the detection head. The evolution of backbone architectures has progressed from manually designed shallow convolutional networks, e.g., AlexNet [21] and VGGNet [22], to automatically searched modern Transformer-based structures, e.g., Swin Transformer [23] and ConvNeXt [24]. As target detection has

matured, backbone networks have diversified. For instance, ResNet-50/101 is commonly employed as the backbone in Faster R-CNN. Within the YOLO series, YOLOv3 utilizes Darknet-53, which consists of 53 convolutional layers [25]. YOLOv5 includes a CSP-optimized variant of Darknet (CSPDarknet), a Focus module and Fast Spatial Pyramid Pooling (SPPF) [26]. YOLOv8 further enhances this architecture by integrating a window attention mechanism adapted from the Swin Transformer into the CSPDarknet framework [27]. To address the limitation of multi-scale modelling in existing architectures, Zhao et al. proposed High- performance GPU Network (HGNetv2) to reduce the model size, which hierarchically aggregates features across receptive fields[28]. A weighted multi-layer feature reconstruction (MLFR) module was proposed for backbone networks in [29]. While traditional network architectures demonstrate effective global semantic feature extraction capabilities in response to increasing accuracy demands in engineering applications, they often exhibit insufficient attention to object edge details. Therefore, it results in constrained localization accuracy, particularly in tasks involving small object defects and low-contrast scenarios, e.g., wind turbine crack identification. To address this issue, we propose a network structure that enhances the extraction of edge information and incorporates multi-scale dynamic feature representation within the backbone architecture.

2.2 Feature Fusion Methods

Feature fusion technology strengthens the model's perception of multi-scale targets by integrating feature information at different levels and scales. The strategic integration of high-resolution shallow features containing detailed visual information with semantically rich deep features significantly improves the performance of small object defects. Concurrently, cross-layer information exchange between hierarchical feature levels strengthens contextual comprehension in complex visual environments. Zhang et al. developed a multi-scale feature pyramid FPN network to enhance small object defects, though its unidirectional propagation mechanism inherently compromises fine-grained detail preservation[30]. Du et al. incorporated bottom-up enhancement pathways into their baseline architecture and developed PANet to facilitate bidirectional information flow, albeit at a substantially increased computational cost[31]. Lu et al. subsequently proposed CSP-BiFPN, which employed cross-scale skip connections and weighted feature fusion to strengthen the efficiency of feature interaction[32]. To achieve an efficient lightweight model, Zhang et al. developed a lightweight dynamic fusion module integrating FFN and DWConv to improve feature representation and reduce both the parameter count and computational overhead[11]. The high-level semantic features are employed in an HSFPN architecture to guide the selection of low-level features and a channel attention mechanism is introduced to dynamically fuse multi-scale information in order to improve cross-scale detection performance in[33]. Chi et al. introduced a lightweight three-path context-guided network (LTCGNet) for sonar frequency-domain applications[34]. However, existing multi-scale pyramid fusion, cross-scale dynamic fusion, and lightweight dynamic fusion methods still struggle to effectively achieve synergistic enhancement between local feature representation and global semantic information. WTs defects exhibit significant morphological variations across different scales. Current feature extraction and fusion architectures struggle to concurrently capture both large-scale deformation patterns and small-scale detail characteristics.

2.3 Comparison with Related Work on Wind Turbine Defect Detection

Although ODDNet[14] dynamically captures local details of blade defects through omni-dimensional dynamic convolution, it imposes stringent computational demands. GhostNet[15] achieves lightweight advantages by replacing regular convolutions with cheap linear operations, but at the expense of fine-grained detail precision. C2f_DBB[16] demonstrates strong multi-scale generalization capability. however, it heavily depends on sufficient training data. RepELAN[18] enhances inference speed via efficient layer aggregation and structural re-parameterization, yet exhibits limited adaptability in dynamic scenarios. In contrast, our proposed DMDT network introduces a unique edge information flow propagation mechanism, which effectively addresses the core challenges of edge blurring and small object defects omission in WT defect detection. Therefore, it can achieve a superior accuracy-efficiency trade-off compared to existing backbone networks.

PSANet[14] employs spatial pyramid pooling to compress global contextual features. However, its fixed-scale pooling operations are less effective in adapting to the slender tubular structures of WTs. Despite its enhanced multi-scale interaction via repeated bidirectional stacking, BiFPN[15] lacks the ability to filter background noise. While SE attention in RFANet[16] enables dynamic receptive field weighting, the architecture fails to establish synergistic perception between global turbine structures and local defects. Moreover, it tends to lose edge continuity in small defects when S-FPN[18] reduces redundant cross-layer connections with sparse parallel pathways. Conversely, the proposed TAAFSFPN synergistically integrates anchor-based attention and channel-wise attention to concurrently capture global and local features, achieving co-optimization of edge preservation and semantic noise suppression through dual-dimensional filtering.

3. Proposed methods

3.1 Introduction to the structure of DMCA-Net

A novel architecture termed DMCA-Net is presented to address the inadequate performance of WT defect detection systems under conditions involving small object defects or low contrast. Therefore, the high-precision defect identification can be obtained in complex operational environments, as illustrated in Fig. 1.

UAV-captured WT surface damage images serve as the input to the DMDT backbone network. HEPG is utilized within the network to extract multi-scale edge information from shallow convolutional features, producing edge feature maps at multiple resolutions. The developed DFS delivers spatially aligned edge features to the CCEI, establishing bidirectional complementary integration between edge characteristics and backbone features. In the neck coding network, a HSAH encoder is designed to reduce the leakage and misdetection rates by capturing the features dynamically and performing the binning operation. An enhanced second-generation TAAFSFPN incorporating a three-layer CAA is applied to replace the CCFM. The output from DMDT is shown in the three-layer fusion feature map $S_3\ S_4\ S_5$, which is used to address the weakening of spatial information of small object defects features triggered by non-structural factors.

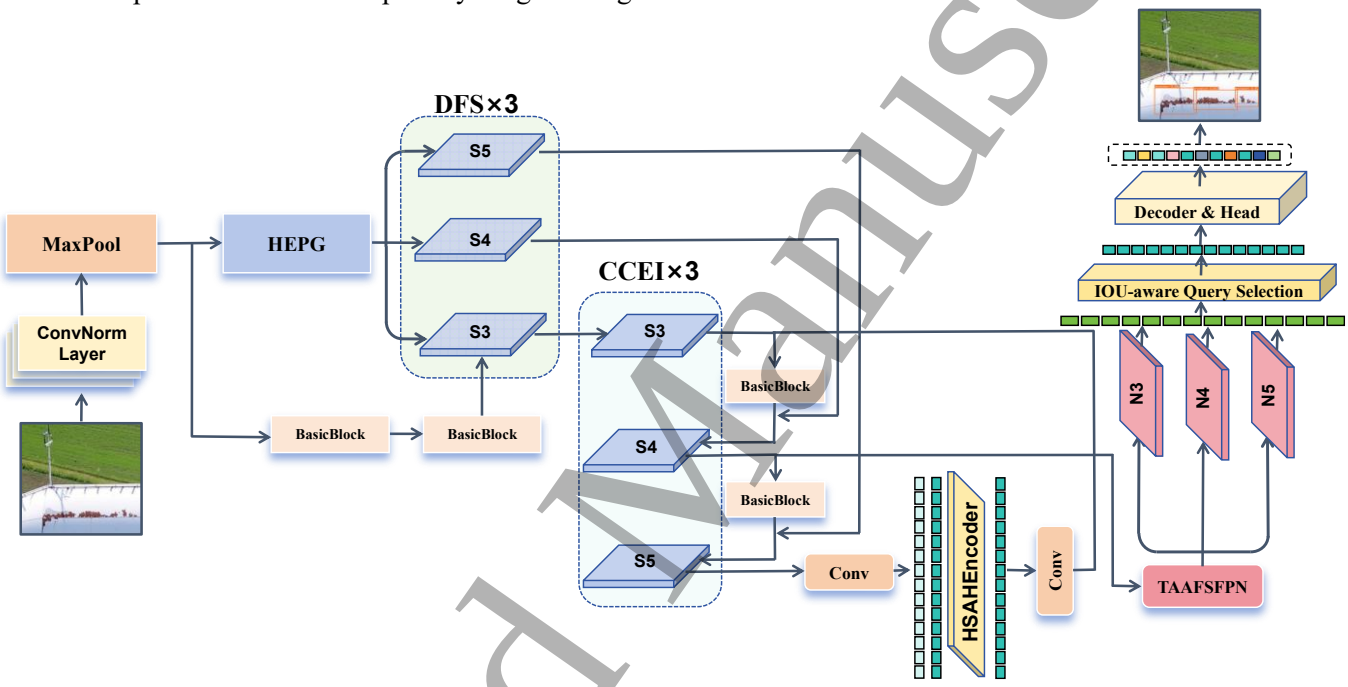

Fig. 1 The overall architecture of our proposed DMCA-Net

3.2 Defect marginal detail transfer backbone network

Traditional small object defect detection methods demonstrate limited accuracy in low-contrast environments due to insufficient attention to boundary details. To address it, we propose a novel architecture, i.e., DMDT, which systematically transfers marginal information from shallow features throughout the backbone network and enables hierarchical multiscale feature integration.

To mitigate background noise and preserves fundamental object contours, we implement shallow feature extraction using three convolutional layers with max-pooling operations, which significantly enhances preliminary edge features. The Hierarchical Edge Preservation Gate (HEPG) resolves the inherent trade-off between detailed feature preservation and limited semantic attention in shallow features by generating multi-scale edge responses. As shown in Fig. 2, the architecture of the spatial marginal pyramids is established to maintain high-resolution information integrity and suppress original image noise interference. The HEPG is designed to extract gradient magnitude features mainly using SobelConv for edge detection. The process is as follows in Eq. (1) and Eq. (2):

$$G_x = \text{Conv3D}(X, K_x), \quad G_y = \text{Conv3D}(X, K_y) \quad (1)$$

$$G = G_x + G_y \quad (2)$$

where $X \in B \times C \times H \times W$ is the input feature map. B is the batch size, C is the number of channels, H is the height of the feature map, and W is the width of the feature map. The horizontal Sobel convolution and vertical Sobel convolution edge feature maps are given by G_x and G_y, respectively. K_x,K_y is Sobel's convolution kernel.

Initial edge features E_k are extracted via the

SobelConv principle depicted in Fig. 2. Maxpool with stride 2 is adopted to realize a multi-stage downsampling operation. Subsequently, $\mathrm{Conv}_{1\times1}$ is used to execute channel adjustment to refine the feature representations.

DFS is developed to achieve feature alignment between output edge features F_k and backbone-derived main features, as illustrated in Fig. 2. The multi-index target feature maps are dynamically selected to provide spatially aligned edge feature through cross-layer integration. The related Eq. (3) is below:

$$F_k = \mathrm{DFS}\left(F[k_1], F[k_2], \ldots, F[k_m]\right) \tag{3}$$

To improve detection of small and morphologically irregular defects in WTs, multi-scale edge information is dynamically integrated into the backbone network to counteract edge blurring caused by downsampling in deep features. As shown in Fig. 2, CCEI is designed to achieve bidirectional complementarity between edge features and backbone features through the residual base block BasicBlock and the CCEI structure. Initially, Conv Channel Fusion[35] is used for feature splicing. $\mathrm{Conv}_{1\times1}$ is used for input channel compression to fuse features across channels. Next, spatial feature extraction is performed by $\mathrm{Conv}_{3\times3}$ to enhance the local details of the edge features. Finally, $\mathrm{Conv}_{1\times1}$ is used to adjust the dimension of the feature map. Specific formulas are described by Eq. (4) and Eq. (5):

$$\begin{aligned} F_{cat} &= \mathrm{Concat}(F_0, F_1, \ldots, F_{N-1}) \\ F_{cat} &\in B \times (\textstyle\sum ouc_k) \times H \times W \end{aligned} \tag{4}$$

$$\begin{aligned} F_{\mathrm{fuse}} &= \mathrm{Conv}_{1\times1} \circ \mathrm{Conv}_{3\times3} \circ \mathrm{Conv}_{1\times1}(F_{\mathrm{cat}}) \\ F_{\mathrm{fuse}} &\in B \times C_{\mathrm{out}} \times H \times W \end{aligned} \tag{5}$$

where F_{cat} is the feature map output after feature channel splicing, and F_{fuse} is the final feature after the fusion of deep semantic features and shallow edge feat ures. By modifying the structure of DMDT1, we derive three variants as illustrated in Fig. 3, and conduct the Backbone Network Comparative Experiment outlined in Section 4.4.1.

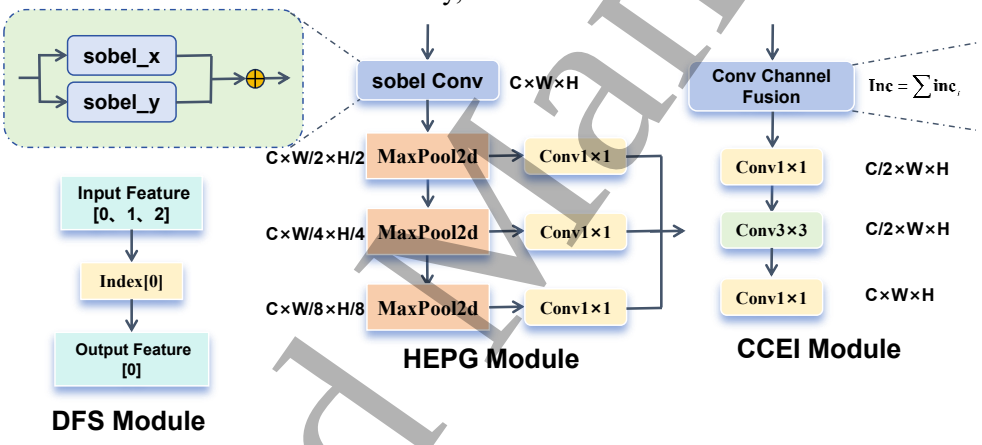

Fig. 2 From left to right in the DMDT structure: DFS, HEPG, CCEI modules

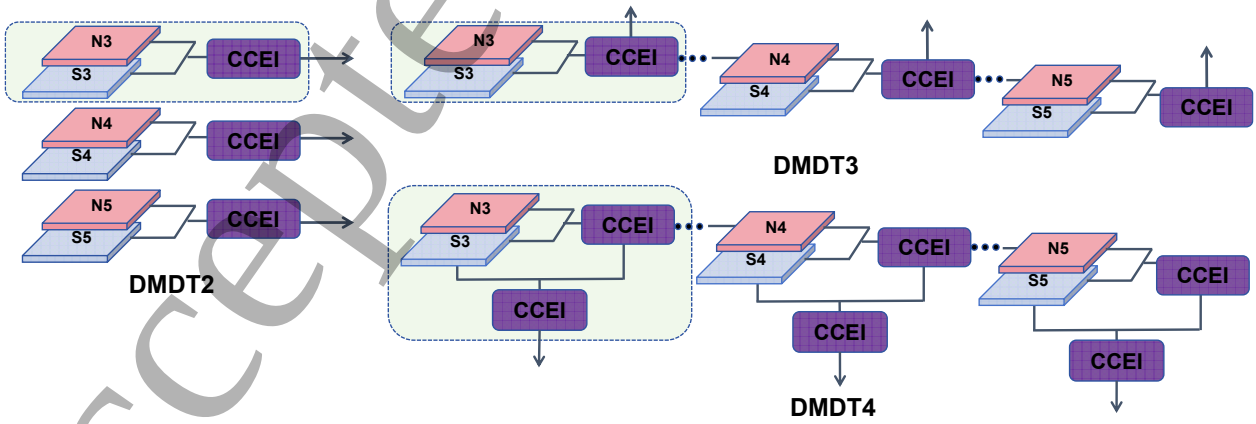

Fig. 3 Network structure of three different marginal information fusion layers (S3, S4, S5) combined with cross-scale feature fusion layers (N3, N4, N5). DMDT2: Parallel fusion is used to fuse S3, S4 and S5 with N3, N4 and N5. DMDT3: A cascade is used to fuse S3, S4, and S5 with N3, N4, and N5. DMDT4: CCEI is re-integrated with S3, S4 and S5 based on DMDT3.

3.3 Histogram-based Synergistic Attention Head

Small defect features often suffer degradation or occlusion under severe weather conditions. Conventional global uniform attention mechanisms struggle to effectively capture these characteristics,

which are dynamically changed. To solve this problem, we propose the HSAH encoder, as illustrated in Fig. 4, which dynamically extracts and adaptively bins target features according to their intensity levels. It significantly enhances detection accuracy for small object defects in challenging environmental conditions.

The input feature $F \in B \times C \times H \times W$ is segmented into two parts using $Split(x)$ along the channel dimension and then $Sort(x)$ is used to rank its intensity along the horizontal and vertical directions, as shown in Eq. (6) and Eq. (7):

$$F_1, F_2 = \text{Split}(F) \tag{6}$$

$$F_{\text{out}} = \text{Conv}^{d}{}_{3\times3}\left(\text{Conv}_{1\times1}\left(\text{Concat}\left(Sort_v\left(Sort_h\left(F_1\right)\right), F_2\right)\right)\right) \tag{7}$$

Frequency-wise paths and Bin-wise paths are Created by Eq. (8):

$$P = \text{Softmax}\left(\frac{Q_B K_B^{\bullet}}{\sqrt{d}}\right)V_B \square \ \text{Softmax}\left(\frac{Q_F K_F^{\bullet}}{\sqrt{d}}\right)V_F \tag{8}$$

Where $Q_B, K_B \in C \times B \times (HW/B)$ and $Q_F, K_F \in C \times B \times (HW/B)$. They are defined as the query key matrixes for Bin-wise paths and Frequency-wise paths. B is the number of bins (to control the global scope), and d is the number of attention heads. $\square$ is the element-by-element multiplication and P is the final attention map obtained.

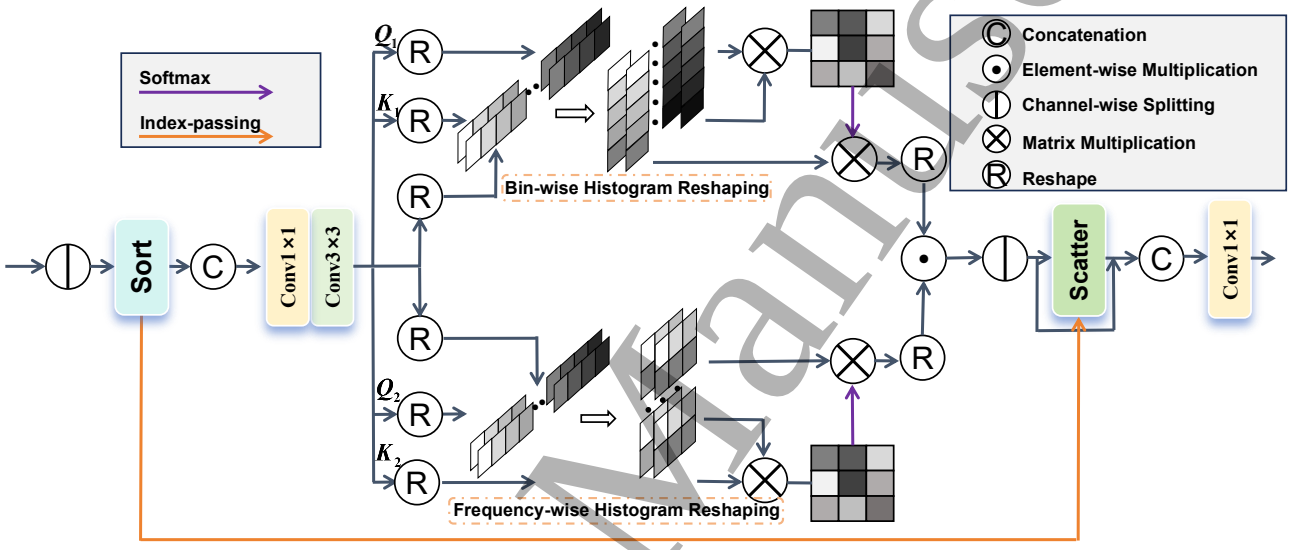

Fig. 4 Detailed structure of HSAH including Dynamic-range convolution (Drconv) and Histogram Self-Attention (HSA). The dynamic-range feature interaction is achieved bin-wise (capturing global features between bins) and frequency-wise (capturing local features within bins), respectively.

3.4 Three-layer Anchor Attention Feature Selection and Fusion Pyramid Network

The cross-layer fusion mechanism in hierarchical scale feature pyramid networks (HSFPN)[36] exhibits limited capacity to correlate global structural patterns of WTs components with localized defects due to restricted receptive fields, which lead to failure in capturing macroscopic deformation features in large-scale damage (e.g., erosion exceeding 50 cm in length). Critical spatial information preservation is diminished in the architectural constraint. Particularly, edge ambiguity artefacts in WT defect detection are introduced via small object defects to reduce crack texture details and erosion defect characteristics. Thus, the TAAFSFPN network is proposed, as shown in Fig. 5. The CAA module[37], illustrated in Fig. 5, mitigates attention dispersion in long-sequence Transformer operations through multi-scale DWconV [38], which explicitly expands receptive fields to encompass meter-scale structural deformations. Simultaneously, the module enhances defect region contrast and suppresses false detections induced by texture ambiguity. Complementarily, a multi-stage screening mechanism is employed in the CA module in Fig. 5 to integrate shallow high-resolution features with deep semantic features and filters irrelevant channels to suppress redundant background interference. The spatial coordinate awareness maintains boundary continuity for irregular large-scale defects. These dual spatial-channel attention mechanisms synergistically complement the HSFPN lightweight architecture through depth-wise separable convolutions, and complete holistic recognition of both microscopic cracks and macroscopic erosion patterns. Finally, a simultaneous leakage detection risk mitigation and parameter reduction can be achieved. As shown in Fig. 5, 'Position' is added to the CAA module one by one to do the

ablation experiment.

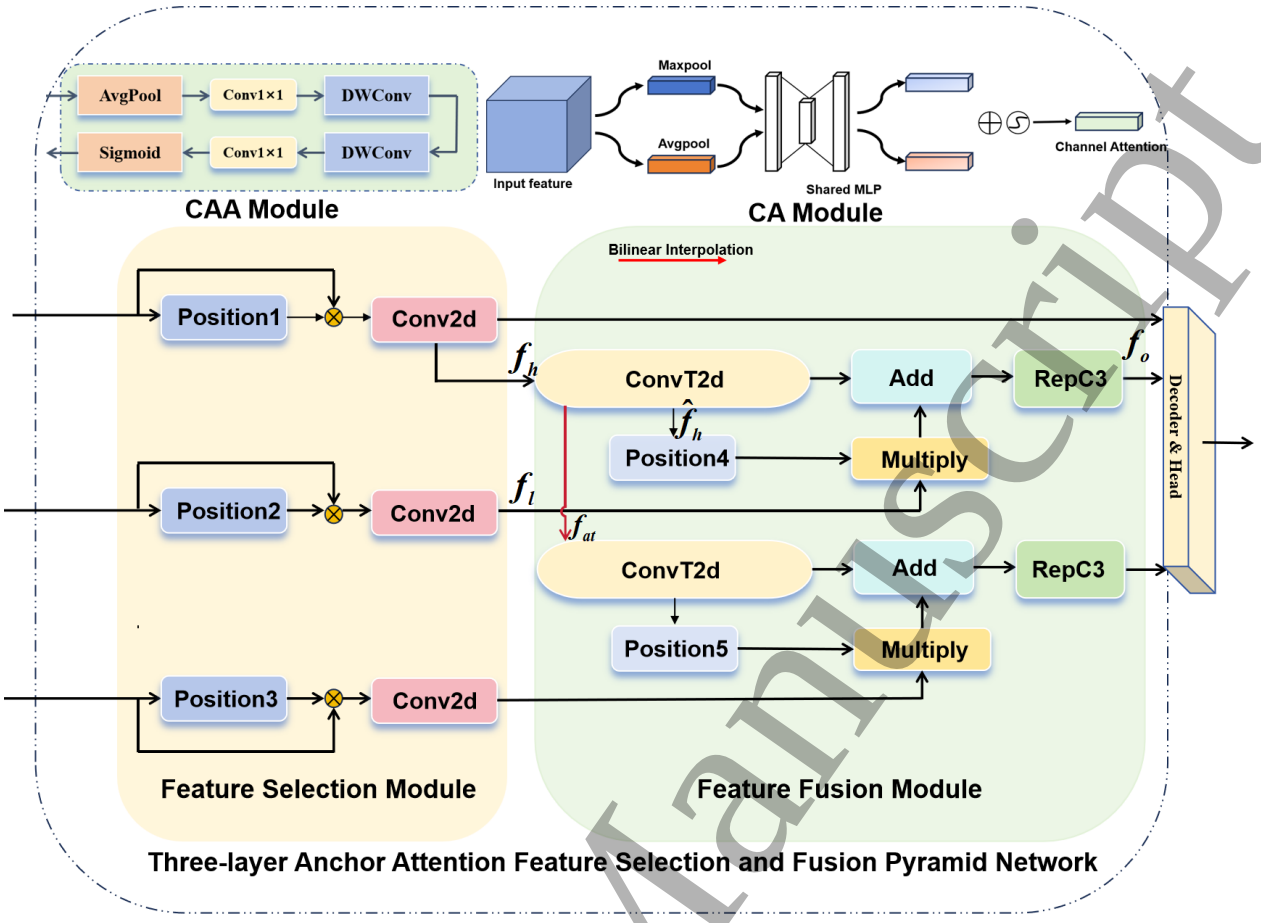

Fig. 5 Three-layer Anchor Attention Feature Selection and Fusion Pyramid Network. Left: Feature Selection Module (FSM). Right: Feature Fusion Module (FFM). Top: the CAA module and the CA module.

3.5 Improvement of the Loss Function

Conventional IoU loss results in frequent leakage and false detection due to insufficient gradient information, particularly for small objects where minor localization errors cause drastic IoU decay. There are unique challenges in WTs defect scenarios that microscopic cracks co-exist with macroscopic erosion damage. Conventional IoU metrics demonstrate critical limitations in such environments, with neither scale adaptability relative to target dimensions nor balanced gradient allocation across multiscale targets.

Enhanced localization is achieved through DIoU's center-distance penalty, which operates in parallel with NWDLoss's distribution alignment. Furthermore, Inner-IoU prevents large erosion targets from dominating the optimization process by adaptively generating dynamic auxiliary boxes based on target scales. In addition, NWDLoss mitigates scale imbalance by modeling bounding boxes as Gaussian distributions and computing similarity via Wasserstein distance. This distribution-based metric ensures equal gradient sensitivity for both small and large targets. The pixel-level vulnerability of IoU is effectively resolved[39-41]. Ultimately, both detection accuracy and generalisation capability are improved. The NWD-InnerDIoU is defined from Eq. (9) to Eq. (13):

$$L_{InnerIoU} = \frac{overlap}{total_{area}} \tag{9}$$

$$L_{\mathrm{DIoU}} = 1 - IoU + \frac{\rho^2(b, b_{gt})}{c^2} \tag{10}$$

$$L_{\mathrm{Inner-DIoU}} = L_{\mathrm{DIoU}} + IoU - L_{InnerIoU} \tag{11}$$

$$\mathrm{NWD}(\mathrm{N}_A, \mathrm{N}_B) = \exp\left(-\frac{\sqrt{W_2^2(\mathrm{N}_A, \mathrm{N}_B)}}{C}\right) \tag{12}$$

$$L_{\mathrm{NWD-InnerDIoU}} = (1 - ratio)L_{\mathrm{NWD}} + ratio L_{Inner-DIoU} \tag{13}$$

where b and b_{gt} are the centres of the prediction and target frames, respectively. ρ is the Euclidean distance between the two centre points. c is the diagonal distance of the smallest rectangle that can cover both the prediction frame and the target frame. $W_2^2(\mathrm{N}_A, \mathrm{N}_B)$ is the Wasserstein distance between the two bounding boxes. C is the number of categories in the dataset.

Gaussian distribution N_A corresponds to the predicted bounding box, while N_B represents the ground-truth bounding box. $NWD(N_A, N_B)$ is the Wasserstein distance normalised by the exponential form.

4. Experiment

4.1 Surface damage dataset for WTs

This study addresses the challenge of acquiring high-quality datasets in the wind energy sector. Two complementary data sources are integrated: the DTU-Drone dataset provided by ASM Shihavuddin [42] and a surface damage dataset supplied by the National Jiuquan WT Equipment Quality Supervision and Inspection Centre [43]. The image data was acquired with a Zenmuse P1 camera mounted on a DJI Matrice 300 RTK UAV platform. There are two key specifications that ensure high-quality data for this research: a storage capacity of 128 GB and a maximum photo resolution of 8192 × 5460 pixels. The dataset includes three characteristic defect categories. The first category comprises cracks, specifically linear cracks, bending cracks, and forking cracks. The second category consists of damage types including scratches, depressions, and deformations. The third category involves erosion patterns characterized by irregular wear and material shedding. Multi-angle lighting and multi-view capture strategies are employed under diverse weather conditions to ensure comprehensive scene variability.

We first allocate 1960 images from a dataset of 2986 samples at 640×640-pixel resolution to the training set. The remaining images are divided into a test set (684 samples) and a validation set (342 samples). Subsequently, data augmentation techniques illustrated in Fig. 7 are applied to the training set to enhance model generalization across diverse datasets. Therefore, the training set is expanded to 2392 images, which maintains a 7:2:1 ratio between the training, test, and validation sets.

Fig. 6 Type of defect

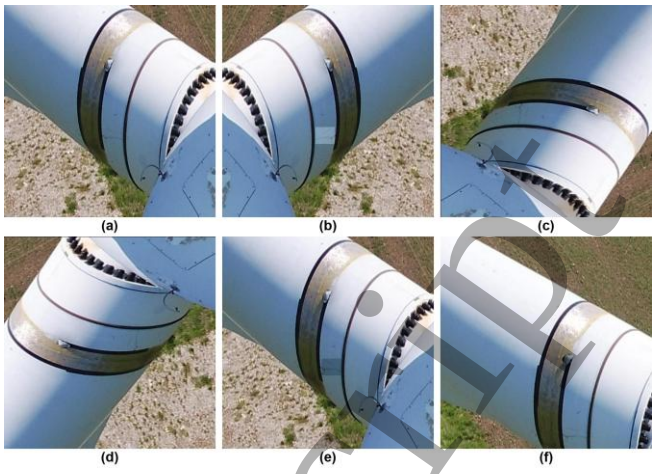

Fig. 7 Data set enhancement methods: (a) original image, (b) mirroring, (c) clockwise rotation 90°, (d) anti-clockwise rotation 90°, (e) mosaic with different brightness, (f) cropping.

Information on the WT defect detection dataset based on drone photography is shown in Fig. 8. The marked box's area is classified as small if $S < 40^2$ pixels, medium if $40^2 < S < 100^2$ pixels, and large if $S > 100^2$ pixels[44]. (a) shows a comparable prevalence of damage and erosion defects, with cracks occurring less frequently, and further provides the size distribution of defects across large, medium, and small categories. (b) visualizes dynamic distributions of bounding boxes, enabling real-time training monitoring and early detection of potential overfitting. (c) reveals uniform annotation distribution, informing loss function selection while exposing dataset characteristics. (d) indicates a predominance of small-scale objects.

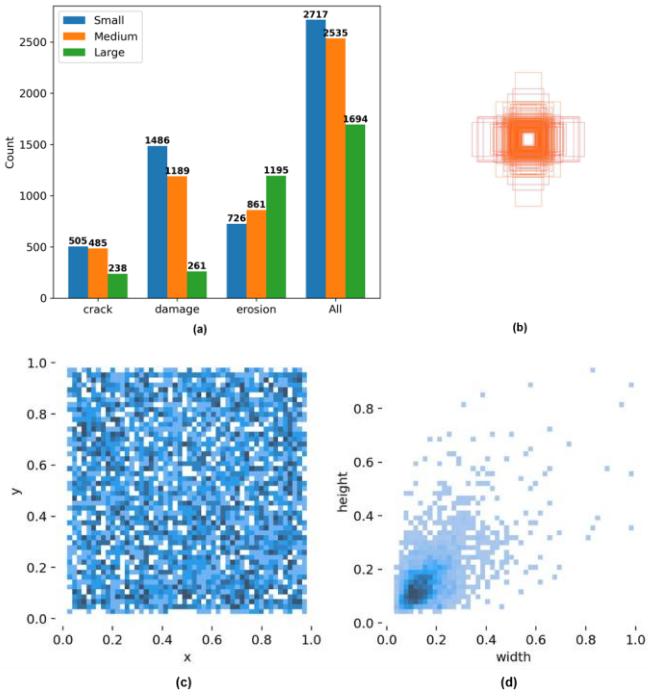

Fig. 8 WT defect dataset information: (a) category infographic, (b) boundary detection box dynamic distribution diagram, (c) statistical distribution of bounding box positions, (d) statistical distribution of bounding box sizes.

4.2 Experimental environment

There are an Intel Core i7-14700HX processor and an NVIDIA RTX 4060 graphics card in the experimental platform. The software environment involves CUDA 11.8, PyTorch 2.0.1, and Python 3.8 to fully support GPU-accelerated model training and inference operations. The iteration cycle is limited to 200 round and the number of samples processed in each batch is set to 4. In addition, the number of worker threads during data loading is set to 8, and the initial learning rate is set to 0.0001. All the other training hyper-parameters are taken as defaults, and all the tests of the model are conducted in the same environment.

4.3 Description of Experimental Indicators

The performance evaluation for WT defect detection employs four established metrics from Eq. (14) and Eq. (18): Precision (P), Recall (R), mean Average Precision (mAP), and Frames-Per-Second (FPS) detection rate.

$$P = TP/(TP + FP) \tag{14}$$

$$R = TP/(TP + FN) \tag{15}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{16}$$

$$mAP = \sum_{n=1}^{N} AP(n) \Big/ N \tag{17}$$

$$FPS = 1000\mathrm{ms}/(T_{preprocess} + T_{inference} + T_{postprocess}) \tag{18}$$

where TP (True Positive) and FP (False Positive) represent the number of correctly or incorrectly detected surface defects on WTs, respectively. FN (False Negative) represents the number of missed defects. AP is the average detection accuracy for each class of defects, and N is the number of defect classes. $T_{preprocess}$, $T_{inference}$, $T_{postprocess}$ denote the image preprocessing time, inference speed, and post-processing time, respectively.

4.4 Experimental results and analyses

4.4.1 Backbone Network Comparative Experiment

The DMDT1 architecture is augmented to evaluate the cross-channel feature fusion benefits of the CCEI module across model locations and scales. Specifically, its marginal information fusion layer is integrated with the deep feature output layer, as illustrated in Fig. 3. There are three variants (DMDT2–DMDT4) generated for comparative evaluation through structural adaptations. The results are shown in Table 1.

The mAP50 is improved by 1.3% when both DMDT1 and DMDT3 are used, but the latter GFLOPs and Parameters are improved by 11% and 9%, respectively. The detection efficiency is seriously affected while the former GFLOPs and Parameters are improved by 4% and 6%, which are relatively small. Comparative analysis reveals that alternative network architectures achieve limited mAP improvement due to the detail degradation and progressive amplification of the shallow noise during deep fusion processes. Fig. 9 demonstrates that DMDT1 achieves faster convergence and higher mAP50 gains than counterparts, potentially attributable to conflicting edge-semantic gradients that impede convergence in comparative models. Consequently, DMDT1 is selected as the optimal architecture in the enhanced backbone network because of superior benefit-cost performance among the four variants.

In order to validate the superior feature extraction capability of the proposed DMDT backbone network for WT defects, comprehensive comparative experiments with state-of-the-art backbones are conducted. As shown in Table 2, DMDT1 achieves the highest precision among all compared models while maintaining a competitive mAP50 of 0.813. It approaches the top-performing ODDsNet with 10.1% lower computational cost and 7.8% fewer parameters. Obviously, the precision–recall trade-off gap observed in RepELANNet is reduced in DMDT1 through optimized edge-aware feature transfer to enhance crack localization accuracy. GhostNet exhibits lower computational complexity. However, its mAP50 lags significantly behind DMDT1 by 1.5%, indicating critical limitations in micro-defect boundary preservation. Additionally, DMDT1 achieves a 1.3% higher mAP50 than C2f_DBB when 0.4% fewer parameters are used, proving exceptional efficiency in suppressing redundant feature interactions. It is confirmed that the HEPG and DFS play a synergistic role in robust defect characterization under complex operational conditions.

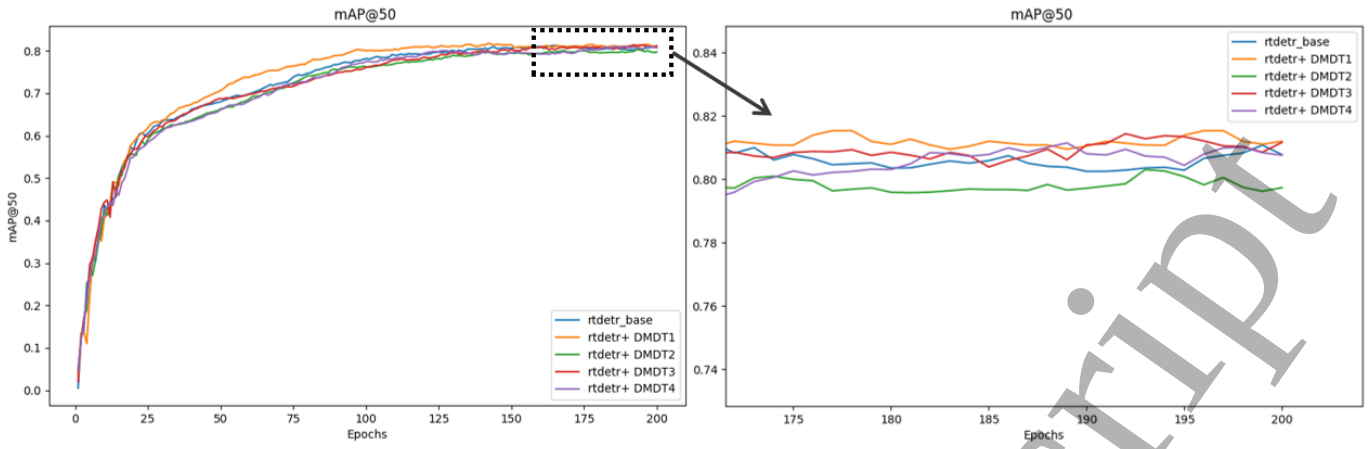

Fig. 9 mAP50 curve comparison chart

Table 1 Comparison of model performance by improved the backbone network.

Model	mAP50	P	R	GFLOPs	Parameters
RT-DETR	0.800	0.817	0.782	56.9	19.87
DMDT1	0.813	0.820	0.786	59.4	21.11
DMDT2	0.802	0.808	0.781	63.2	21.80
DMDT3	0.811	0.817	0.774	63.2	21.80
DMDT4	0.810	0.810	0.796	65.8	22.24

Table 2 Comparing experimental results of backbone networks.

Model	mAP50	P	R	GFLOPs	Parameters
ODDsNet	0.814	0.816	0.796	66.1	22.90
GhostNet	0.798	0.823	0.798	49.1	16.32
C2f_DBB	0.800	0.805	0.758	58.4	21.2
RepELANNet	0.804	0.815	0.806	53.0	20.12
DMDT1	0.813	0.830	0.786	59.4	21.11

4.4.2 Experimental analysis of enhancing model performance using the CAA attention mechanism

To evaluate the combined efficacy of CAA and HSFPN in global context capture, we sequentially integrate the channel attention assembly module at five strategic positions within the hierarchical scale feature pyramid network, as shown in Fig. 5. This systematic ablation study, documented in Table 3, precisely isolates each component's contribution. ("√" indicates that the CAA module is added to that position, and the"-" indicates that the CA module is retained at that position). As evidenced by the defect scale distribution in Fig. 8(a), erosion defects (43.0%, n=1195) predominantly constitute large-scale manifestations, while damage (50.6%, n=1486) displays small-scale anomalies. Precision trends in Table 3 reveal dual-attention specialization. CAA modules progressively enhance large-scale erosion detection accuracy, and the retained CA modules preserve small-scale damage recognition. The optimal multi-scale balance is achieved by integrating three CAA layers and two CA modules into the FSM. Consequently, a 1.7% improvement in mAP50 is obtained over the baseline. The parameters and GFLOPs are reduced to 4.5% and 7%, respectively. Although full CAA integration improves erosion detection, damage accuracy is reduced by 4.5% due to diminished edge sensitivity, while a 4% computational overhead is incurred.

The Grad-CAM++ algorithm is employed in this study to generate infrared heat maps, as illustrated in Fig. 10, enabling visual verification for multi-scale defect recognition. The colour intensity in the heat map directly correlates with the model's defect detection sensitivity, where deeper red hues signify stronger defect recognition. The enhanced model utilizes CA-driven local attention to minimize background interference for small-scale cracks and maintains critical edge details from small object defects. The false identification issues inherent in the baseline model can be effectively resolved by the proposed approach. Regarding large-scale erosion, the CAA module expands hotspot coverage across defects, enabling a holistic perception of metric-scale deformations. The TAAFSFPN network improves large-area erosion detection via a dual attention framework that combines global context modeling with localized attention mechanisms. By prioritizing attention weights on defect core regions rather than peripheral edges, the accuracy of defective area recognition is significantly enhanced in the synergistic system. Quantitative and visual evidence collectively confirms that the three-layer CAA configuration optimizes synergy between global and local features.

As shown in Table 4, comparative experiments are conducted to rigorously evaluate the feature fusion

efficacy of TAAFSFPN. The results show that the second-highest precision (0.835) is achieved among state-of-the-art networks, being slightly outperformed only by RFANet (0.838), while using 9.8% fewer computational resources. Notably, TAAFSFPN obtains a leading mAP50 of 0.817 that matches top-performing BiFPN and RFANet. However, there are 8.2% fewer GFLOPs and 7.5% reduction parameters than BiFPN. Crucially, the proposed method overcomes the critical recall deficiency of S-FPN and mitigates PSANet's precision-recall imbalance through adaptive spatial-channel filtering.

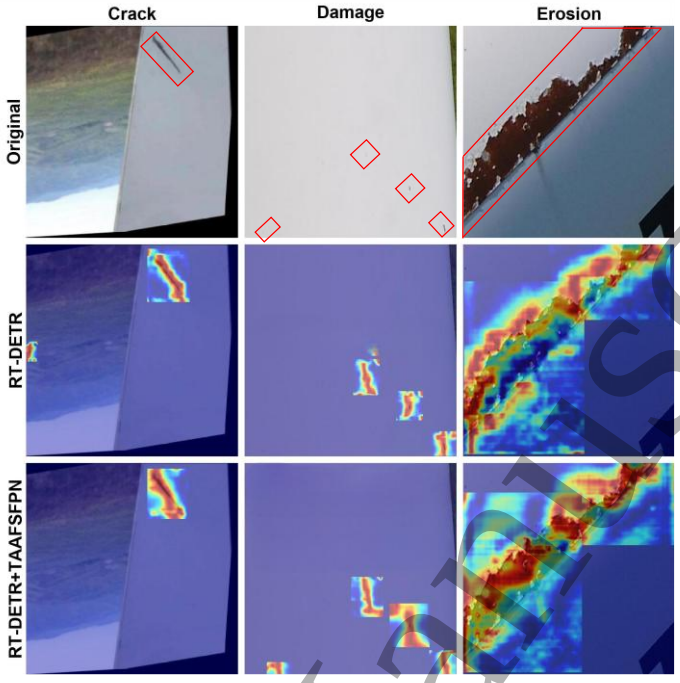

Fig. 10 Heat map comparison

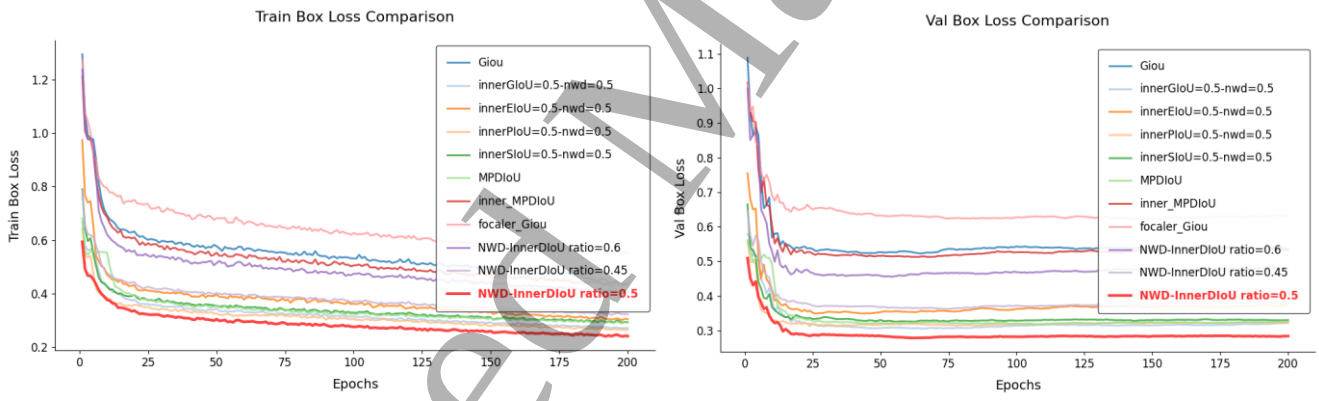

Fig. 11 (a) Train Box Loss curve (b) Val Box Loss curve

Table 3 Ablation experiments by adding CAA modules at each of the five positions of the feature fusion network.

Position1	Position2	Position3	Position4	Position5	mAP50	mAP_{crack}	mAP_{damage}	$mAP_{erosion}$	mAP50:95	GFLOPs	Parameters
-	-	-	-	-	0.806	0.883	0.767	0.786	0.443	53.3	18.11
√	-	-	-	-	0.810	0.849	0.762	0.792	0.438	53.5	18.36
√	√	-	-	-	0.812	0.867	0.753	0.800	0.435	53.9	18.39
√	√	√	-	-	0.817	0.896	0.747	0.810	0.449	54.3	18.43
√	√	√	√	-	0.814	0.869	0.730	0.823	0.432	55.9	18.46
√	√	√	√	√	0.817	0.822	0.722	0.877	0.439	56.5	18.56

Table 4 Comparing experimental results of feature fusion network

Model	mAP50	P	R	GFLOPs	Parameters
PSANet	0.811	0.815	0.806	50.4	17.62
BiFPN	0.817	0.830	0.806	59.3	19.94
RFANet	0.816	0.838	0.802	60.1	20.11
S-FPN	0.812	0.824	0.798	52.0	18.11
TAAFSFPN	0.817	0.835	0.801	54.3	18.43

4.4.3 Ablation experiment

To further investigate the impact of DMDT1, TAAFSFPN and HSAH on the detection model, we perform a careful ablation study on the WTs defect target detection dataset. The effect of each structure on the performance of the RT-DETR baseline model is measured separately. Table 5 depicts the results of the ablation study. The symbol "√" indicates that the structure is used, while "-" indicates that it is not used.

The DMDT1 backbone network contributes a selective 1.3% improvement in mAP50 while maintaining consistent mAP50:95 performance. It indicates the effectiveness in coarse-grained defect identification and a constrained impact on fine-grained localization tasks. The standalone integration of TAAFSFPN produces a 1.4% mAP50:95 improvement with a 4.6% GFLOPs reduction. It demonstrates the efficient feature selection capabilities. The HSAH encoder resulted in a 1.7% improvement in mAP50, which shows the effectiveness of its channel-space dual-dimensional attentional synergy. In the Group 5 experiment, mAP50 is improved by 2.1%, while FPS is decreased by 8.1%. GFLOPs are basically flat. The experimental results show complementary performance characteristics between DMDT1 and TAAFSFPN, with GFLOPs remaining largely unchanged. DMDT1 and TAAFSFPN are combined to jointly enhance classification precision since the former delivers robust edge feature representation, and the latter uses attention mechanisms to suppress background interference. Nevertheless, the computational cost associated with edge feature transmission constrains further localization accuracy enhancement. Group 7 experimental results demonstrate synergistic optimization of the multi-attention mechanism. The 0.823 mAP50 and 0.444 mAP50-95 are obtained to confirm the effectiveness for complex background target detection. Group 8 results show further improvement to 0.828 mAP50 and 0.445 mAP50-95, while a 7.3% FPS reduction to 81.3 compared to baseline. The precision and recall are increased by 2.4% and 2.0% respectively to indicate significant inter-module synergy.

As shown in Fig. 12 and Fig. 13, Group 8 maintains a consistently higher mAP50 and Precision across all training epochs and demonstrates the fastest convergence speed among all experimental models.

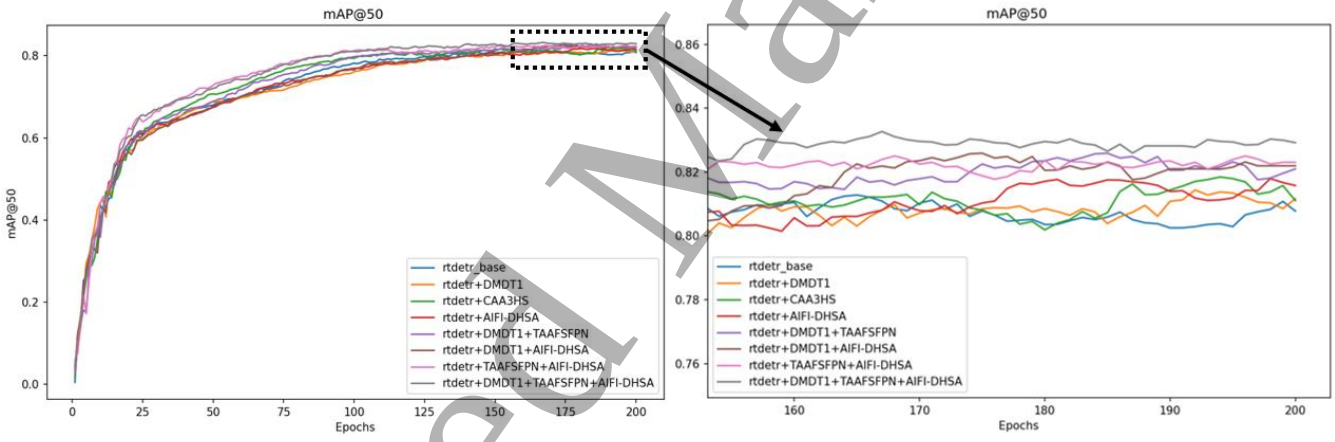

Fig. 12 mAP50 comparison curves

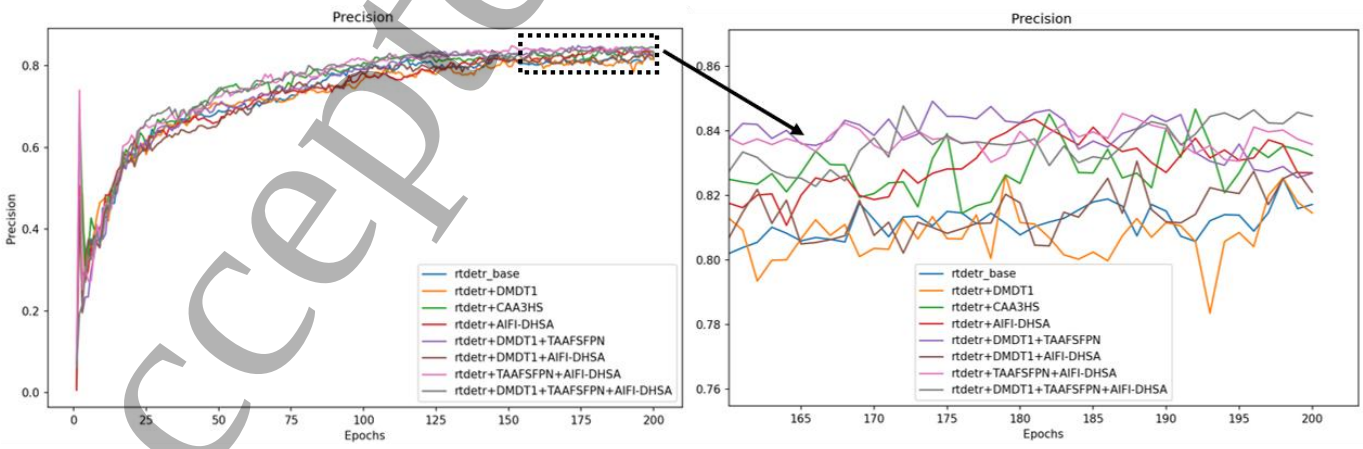

Fig. 13 Precision comparison curves

4.4.4 Validity of loss function improvements

In the loss function comparison experiments, multiple Inner-series loss functions are first systematically evaluated to investigate the correlation between the DMCA-Net algorithm's detection performance and InnerIoU computation methodologies.

Subsequently, the effectiveness and stability of the NWD-InnerDIoU combination are rigorously investigated to validate its contributions to small object defects detection accuracy enhancement and loss convergence improvement. Comparative experiments in Table 6 and Fig. 11 demonstrate that NWD-InnerDIoU (ratio=0.5) achieves the lowest Train Box Loss and Val Box Loss values through scaling factor ratio optimization. This observation, combined with the values of mAP50 and mAP50:95 metrics in Table 6, indicates a significant improvement in small object defects localization performance.

NWD-InnerDIoU fundamentally addresses the inherent challenge of 'small targets being submerged by large objects' in WT defect detection through probability distribution alignment and dynamic scale adaptation. A reusable loss function design paradigm is established by the ratio modulation mechanism for industrial multi-scale object detection.

4.4.5 Impact of Dataset Scale on Model Performance

The WT defect training set consists of 1,968 raw images, and is expanded to 2,392 through data augmentation thereafter. Although smaller than generic object detection benchmarks, this scale can accurately reflect real-world industrial inspection scenarios[7]. However, there are some inherent data acquisition constraints in these applications. For example, the limitation of high-altitude access can reduce image collection frequency, occurrence rates defect are low (e.g., the annual failure rate of wind turbine blades is only 0.3% [45]), and the specialised annotation is needed to support fine-grained defect categorisation.

To evaluate the impact of data volume on model performance, we construct multiple training sets based on the original 2,392 images. As illustrated in Fig. 14, we generate an additional seven training subsets through random sampling and augmentation techniques. These subsets vary in size and composition, which allows for a more comprehensive analysis of the impact of dataset size on the model[18]. The experimental setup and corresponding results are shown in Fig. 14 When trained on 50% of the base set (1,196 images), the model shows an 8.2% drop in mAP50 due to limited feature diversity. Performance saturates at 2,392 images, which explains that adequate learning signals were obtained. Detection performance is slightly improved by less than 1.0% when using images larger than those in the base set, consistent with the principle of diminishing returns. This indicates that our model can achieve better detection performance without relying on excessively large datasets.

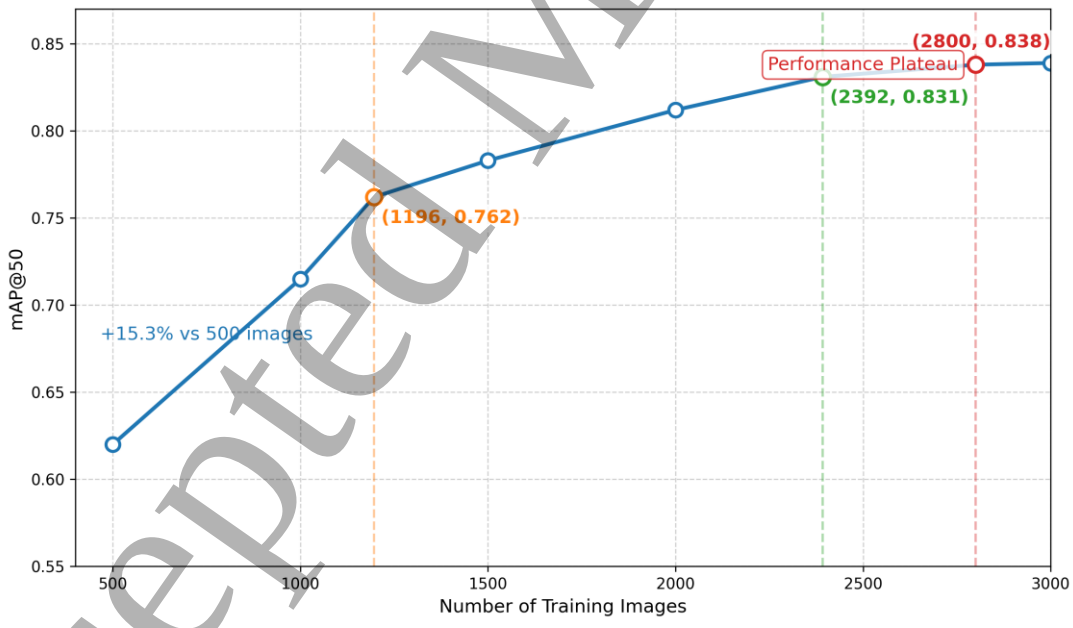

Fig. 14 Impact of Training Data Scale on Model Performance

4.4.6 Comparison experiments

To rigorously validate the superiority of our proposed DMCA-Net, comprehensive comparative experiments are conducted against both established object detection architectures (including the YOLOv5–YOLOv8 series) and the state-of-the-art YOLOv10. For a fair performance evaluation, models with similar computational complexity (GFLOPs) are systematically selected, as shown in Fig. 14 The hyperparameter settings for all of the above versions refer to open-source data published by the experimental developers.

Combination of Table 7 and Fig. 15, DMCA-Net demonstrates superior efficiency and real-time processing capability with 19.94M parameters and 12.3ms inference time (substantially which are lower than YOLOv8m, YOLOv9m[46], and YOLOv10m[47]).

Our proposed model achieves state-of-the-art detection performance, with a precision of 0.840, mAP50 of 0.831, and mAP50-95 of 0.449, surpassing all compared methods. Additionally, the recall rate remains highly competitive, with only a 0.3% gap compared to the top-performing YOLOv10m. These results confirm the model's strong accuracy and robustness in object detection tasks. Compared to the RT-DETR baseline model, it achieves 3.1% and 1.4% improvement in the mAP50 and mAP50:95 metrics, respectively.

The proposed detection system achieves an optimal precision-efficiency balance through novel architectural design. Specifically, it is designed for high-precision detection of small defects in computationally constrained environments. The proposed solution delivers outstanding performance in wind turbine surface inspection. It maintains a modest 7.8% increase in inference latency and significantly enhances real-world applicability in resource-limited settings.

Table 5 Results of the ablation experiments.

Group	DMDT1	TAAF-SFPN	HSAH	mAP50	mAP50:95	P	R	FPS	GFLOPs
1	-	-	-	0.800	0.435	0.817	0.782	87.70	56.9
2	√	-	-	0.813	0.435	0.820	0.786	78.13	59.4
3	-	√	-	0.817	0.449	0.835	0.801	88.50	54.3
4	-	-	√	0.817	0.42	0.833	0.777	91.74	57.3
5	√	√	-	0.821	0.439	0.831	0.786	80.65	56.8
6	√	-	√	0.820	0.429	0.821	0.769	81.96	59.7
7	-	√	√	0.823	0.444	0.840	0.798	90.10	54.3
8	√	√	√	0.828	0.445	0.841	0.802	81.30	57.1

Table 6 Loss function comparison experiment

R-Loss	mAP50	mAP50:95	P	R
Giou	0.828	0.445	0.841	0.802
InnerGiou+nwdloss	0.821	0.444	0.827	0.799
InnerEiou+nwdloss	0.825	0.440	0.832	0.793
InnerPiou+nwdloss	0.822	0.435	0.838	0.802
InnerSiou+nwdloss	0.819	0.435	0.843	0.797
NWD-InnerDIoU(ratio=0.5)	0.831	0.449	0.840	0.807
NWD-InnerDIoU(ratio=0.45)	0.817	0.444	0.828	0.809
NWD-InnerDIoU(ratio=0.6)	0.805	0.435	0.821	0.795
MPDiou	0.821	0.447	0.815	0.806
InnerMPDiou	0.827	0.432	0.83	0.806
Focaler_Giou	0.826	0.442	0.831	0.794

Table 7 Experimental comparison results of different algorithms

Models	P	R	mAP50	mAP50:95	Inference time	Parameters	GFLOPs
YOLOv5m	0.788	0.758	0.786	0.441	10.8	20.02	47.9
YOLOv6s	0.810	0.801	0.798	0.432	10.6	18.63	44.0
YOLOv8m	0.824	0.804	0.817	0.439	15.1	24.84	78.7
YOLOv9m	0.811	0.789	0.820	0.421	14.7	24.12	76.5
YOLOv10m	0.836	0.810	0.827	0.436	13.3	21.41	64.5
RT-DETR	0.817	0.782	0.800	0.435	11.4	19.87	56.9
Ours	0.840	0.807	0.831	0.449	12.3	19.94	57.1

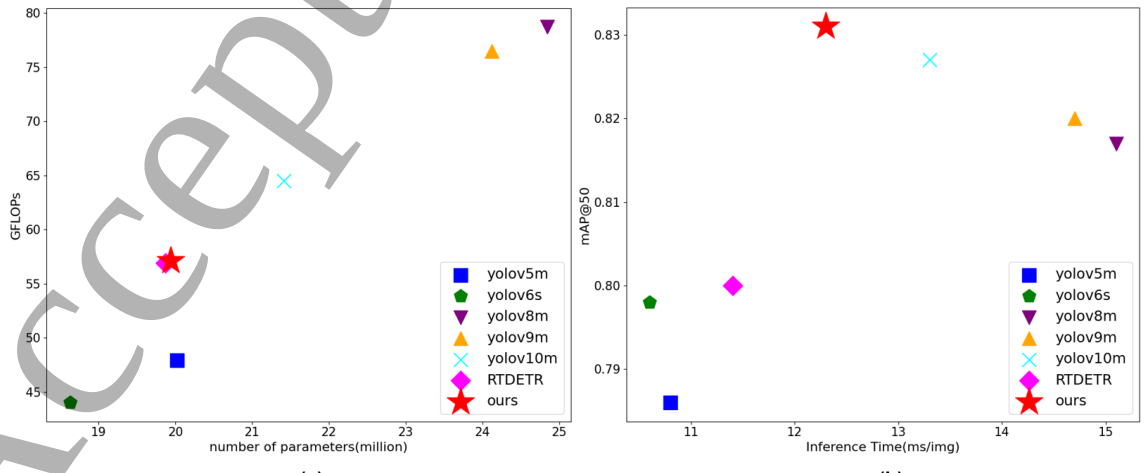

Fig. 15 A comparative evaluation of the latest models of the YOLO series and the proposed method based on four performance

metrics: (a) GFLOPs versus parametric quantities, (b) mAP50 versus inference speed.

It can be seen from Fig. 16 that larger box dispersion and lower robustness coefficients are caused in YOLOv5m and YOLOv6s with lightweight architecture due to their limited feature extraction capability. The proposed DMCA-Net is further improved through the integration of multi-scale feature enhancement and optimized feature fusion. This improvement yields superior performance over both YOLOv8m and YOLOv9m in WT defect detection tasks. However, both of them cannot deal with small object defects on complex backgrounds and WTs. The robustness is still insufficient. Therefore, a SPP-SAM hybrid module YOLOv10m is integrated to strengthen the defect edge response through spatial attention weighting. It significantly improves the mAP50 under complex backgrounds. The DMCA-Net model significantly outperforms other algorithms in terms of its stability and robustness distribution.

A comparison of the detection results of the seven models for the three types of defects on the test set is shown in Fig. 17. In the crack detection task, although YOLOv10m, RT-DETR and DMCA-Net show higher accuracies, ours can completely detect all three crack defects while the other models have leakages. For damage defects, YOLOv5m, YOLOv6s and YOLOv9m show significant underdetection and low localisation accuracy. It is demonstrated by our model's results that the highest bounding box matching accuracy is achieved on all five damage defect detections. In the detection task involving seven tiny erosion targets, YOLOv6s and YOLOv8m fail to detect four and three targets, respectively, while RT-DETR exhibits category misclassification. In contrast, our method achieves precise identification of all defects. Therefore, the experimental results show that our model achieves optimal detection accuracy in all three defect detection tasks.

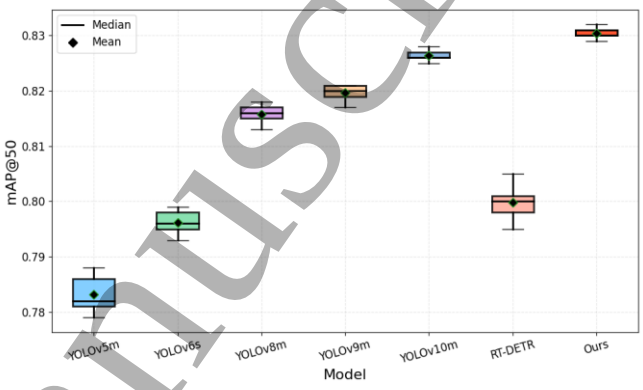

Fig. 16 Comparison of boxplots for multiple models

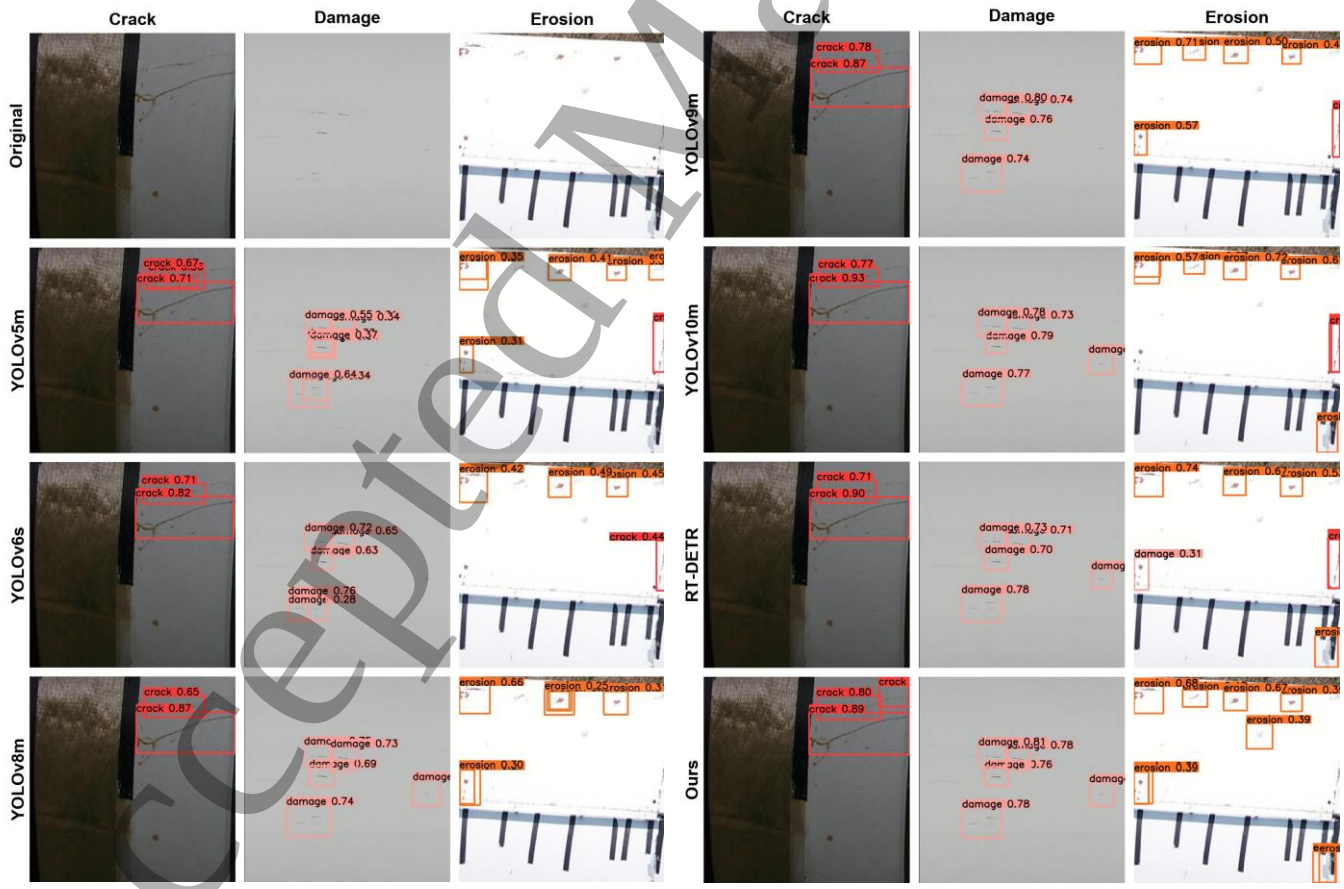

Fig. 17 Results of WTs surface defect detection.

5. Conclusion

The DMCA-Net detection model is proposed with the following core contributions:

First, the DMDT backbone network is constructed

to facilitate bidirectional complementarity between shallow edge details and deep semantic features. This is achieved through a triple mechanism comprising gradient-aware edge feature extraction, multi-scale edge pyramid generation, and cross-channel feature fusion.

Second, a novel HSAH encoder is introduced to dynamically capture target features in complex environments and perform intensity-based feature binning, thereby improving detection accuracy for small object defects.

Additionally, the TAAFSFPN feature fusion network is designed to enhance the performance of small defect detection via a dual-dimensional attention mechanism operating across both channel and spatial domains. This mechanism effectively balances global contextual modeling with fine-grained local detail perception.

Finally, a composite loss function named NWD-InnerDIoU is proposed to mitigate the limitations of traditional IoU metrics in small object defect detection scenarios.

While reducing detection loss, sensitivity has been improved and convergence speed has been accelerated. Experimental results show that DMCA-Net achieves a mAP50 of 0.831 on the WTs defect dataset. The proposed method outperforms the baseline RT-DETR, with a real-time detection rate of 81.3 FPS, and exceeds mainstream YOLO variants in inference speed.

This framework demonstrates a superior accuracy-efficiency balance compared to mainstream algorithms, e.g., the YOLO series. We also provide a computationally efficient and deployment-ready solution for industrial WTs inspection systems.

Author contributions: Yuxin Si: Conceptualization, Data curation, Methodology, Software, Visualization, Writing–original draft. Yunfei Ding: Methodology, Supervision, Software, Writing-review and editing. Fudi Ge: Methodology, Software. Xingtao Wu: Resources, Software. Jinglin Liu：Validation, Software. Dong Ding: Supervision. Hongwei Zhang: Methodology, Writing-review.

Acknowledgement: The authors wish to express their thanks for the support of the Chenguang Program of the Shanghai Education Development Foundation and Shanghai Municipal Education Commission with Grant No. 23CGA76.

Data availability: Data will be made available on request.

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Fei, Y. Gao, H. Gu, Y. Sun, Y. Tian, YOLOv5_CDB: A Global Wind Turbine Detection Framework Integrating CBAM and DBSCAN, 17 (2025) 1322.

[2] Y. Hu, Y. Zhang, Y. Li, Y. Wang, G. Li, X. Liu, Wind turbine blade recycling: A review of the recovery and high-value utilization of decommissioned wind turbine blades, Resources Conservation and Recycling, 210 (2024).

[3] Li T, Zhu C, Li J, et al. A real-time insulator condition detection model for UAV inspection based on FG-YOLO[J]. Measurement Science and Technology, 2025, 36(5): 056208.

[4] T. Matsui, K. Yamamoto, J. Ogata, Anomaly detection for wind turbine damaged due to lightning strike, Electric Power Systems Research, 209 (2022).

[5] R. Girshick, J. Donahue, T. Darrell, J. Malik, Ieee, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, 2014, pp. 580-587.

[6] Y. Xu, X. Luo, M. Yuan, B. Huang, J.M. Malof, Soft-masks guided faster region-based convolutional neural network for domain adaptation in wind turbine detection, Frontiers in Energy Research, 10 (2023).

[7] T. Zhang, W. Xu, B. Luo, G.J.N. Wang, Depth-wise convolutions in vision transformers for efficient training on small datasets, 617 (2025) 128998.

[8] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, Ieee, You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, 2016, pp. 779-788.

[9] I. Gohar, W.K. Yew, A. Halimi, J. See, Review of state-of-the-art surface defect detection on wind turbine blades through aerial imagery: Challenges and recommendations, Engineering Applications of Artificial Intelligence, 144 (2025).

[10] J.-Y. Kim, Comparison analysis of YOLOv10 and existing object detection model performance, Journal of The Korea Society of Computer and Information, 29 (2024) 85-92.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213-229.

[12] P. Diaz, P.J.S. Tittus, Image, V. Processing, Fast detection of wind turbine blade damage using Cascade Mask R-DSCNN-aided drone inspection analysis, 17 (2023) 2333-2341.

[13] M. Davis, E. Nazario Dejesus, M. Shekaramiz, J. Zander, M.J.A.S. Memari, Identification and localization of wind turbine blade faults using deep learning, 14 (2024) 6319.

[14] Gu Z, Peng Y, Sun C. WT-YOLO: A High-Accuracy Model for Wind Turbine Target Detection[C] International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2025: 170-181.

[15] Zhang Z, Dong C, Wei Z, et al. GCB‐YOLO: A Lightweight Algorithm for Wind Turbine Blade Defect Detection[J]. Wind Energy, 2025, 28(6): e70029.

[16] Yu X, Yan P, Zheng S, et al. YOLOv8-WTDD: multi-scale defect detection algorithm for wind turbines[J]. The Journal of Supercomputing, 2025, 81(1): 32.

[17] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Detrs beat yolos on real-time object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 16965-16974.

[18] Zhao B, Li X, Wang G, et al. End-to-end wind turbine damage detection model based on multi-branch feature sensing and contextual information reuse in harsh environments[J]. Renewable Energy, 2025: 123489.

[19] M. McGugan, L. Mishnaevsky, Damage Mechanism Based Approach to the Structural Health Monitoring of Wind Turbine Blades, Coatings, 10 (2020).

[20] Wang Z, Ma K, Qin B, et al. Hard sample mining-enabled supervised contrastive feature learning for wind turbine pitch system fault diagnosis[J]. Measurement Science and Technology, 2024, 35(11): 116203.

[21] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. Van Esesn, A.A.S. Awwal, V.K.J.a.p.a. Asari, The history began from alexnet: A comprehensive survey on deep learning approaches, (2018).

[22] L. Wang, S. Guo, W. Huang, Y.J.a.p.a. Qiao, Places205-vggnet models for scene recognition, (2015).

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012-10022.

[24] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie, Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 16133-16142.

[25] J. Redmon, A.J.a.p.a. Farhadi, Yolov3: An incremental improvement, (2018).

[26] D. Thuan, Evolution of Yolo algorithm and Yolov5: The State-of-the-Art object detention algorithm, (2021).

[27] R. Varghese, S. M, YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness, in: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024, pp. 1-6.

[28] Zhao C, Wang Q, Tan J, et al. HSC-YOLO: steel surface defect detection model based on improved YOLOv10n[J]. Measurement Science and Technology, 2025.

[29] Bai X, Wang R, Pi Y, et al. DMFR-YOLO: an infrared small hotspot detection algorithm based on double multi-scale feature fusion[J]. Measurement Science and Technology, 2024, 36(1): 015422.

[30] L. Zhang, Y. Li, H. Chen, W. Wu, K. Chen, S.J.E.s.w.a. Wang, Anchor-free YOLOv3 for mass detection in mammogram, 191 (2022) 116273.

[31] C. Du, P.X. Liu, X. Song, M. Zheng, C.J.I.T.o.I. Wang, Measurement, A two-pipeline instance segmentation network via boundary enhancement for scene understanding, (2024).

[32] Y.-F. Lu, Q. Yu, J.-W. Gao, Y. Li, J.-C. Zou, H.J.N. Qiao, Cross stage partial connections based weighted bi-directional feature pyramid and enhanced spatial transformation network for robust object detection, 513 (2022) 70-82.

[33] Y. Chen, C. Zhang, B. Chen, Y. Huang, Y. Sun, C. Wang, X. Fu, Y. Dai, F. Qin, Y.J.C.i.b. Peng, medicine, Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases, 170 (2024) 107917.

[34] Chi Z, Wang H, Wei Y, et al. Sonar-YOLO: Acoustic Target Detection in Shallow Water Based on the Fusion of Improved Lightweight Networks and Attention Mechanism[J]. Measurement Science and Technology, 2025.

[35] Y. Lei, D. Peng, P. Zhang, Q. Ke, H.J.I.T.o. I.P. Li, Hierarchical paired channel fusion network for street scene change detection, 30 (2020) 55-67.

[36] Y. Chen, C. Zhang, B. Chen, Y. Huang, Y. Sun, C. Wang, X. Fu, Y. Dai, F. Qin, Y.J.C.i.b. Peng, medicine, Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases, 170 (2024) 107917.

[37] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, Y. Yao, Poly kernel inception network for remote sensing detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27706-27716.

[38] Zhang T, Xu W, Luo B, et al. Depth-Wise Convolutions in Vision Transformers for efficient training on small datasets[J]. Neurocomputing, 2025, 617: 128998.

[39] J. Wang, C. Xu, W. Yang, L.J.a.p.a. Yu, A normalized Gaussian Wasserstein distance for tiny object detection, (2021).

[40] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI conference on artificial intelligence, 2020, pp. 12993-13000.

[41] H. Zhang, C. Xu, S.J.a.p.a. Zhang, Inner-iou: more effective intersection over union loss with auxiliary bounding box, (2023).

[42] A. Shihavuddin, X. Chen, V. Fedorov, A. Nymark Christensen, N. Andre Bro-gaard Riis, K. Branner, A. Bjorholm Dahl, R. Reinhold Paulsen, Wind turbine surface damage detection by deep learning aided drone inspection analysis, Energies 12 (4) (2019) 676.

[43] Liu Y, Zheng Y, Wei T, et al. Lightweight algorithm based on you only look once version 5 for multiple class defect detection on wind turbine blade surfaces[J]. Engineering Applications of Artificial Intelligence, 2024, 138: 109422.

[44] Hu W, Fang J, Zhang Y, et al. Digital twin of wind turbine surface damage detection based on deep learning-aided drone inspection[J]. Renewable Energy, 2025, 241: 122332.

[45] Zhao S, He J, Zhu Y, et al. Theoretical Modeling and Identification Method for Surface Crack Damage in Wind Turbine Blades[J]. Measurement Science and Technology, 2025.

[46] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.

[47] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.