



OPEN Multiscale diffusion-enhanced attention network for steel surface defect detection in Polysilicon Production

Yiwei Duan¹, Lizhen He¹, Zhisheng Wang¹, Jinhai Sa^{1✉}, Jiawen Yang², Xiaolong Chen³, Bingdong Shi⁴, Yangyang Zhang⁴ & Jiawen Sun^{5,6}

Surface defect detection on steel components is crucial for quality control in polysilicon production. However, this task remains challenging due to tiny defect sizes, irregular geometries, complex backgrounds, and low contrast. To address these issues, we propose MSEOD-DDFusionNet (Multi-Scale and Effective Object-Detection Diffusion Fusion Network), a novel multi-scale diffusion-enhanced attention network. The network integrates four specialized modules: MTECAAttention (Multi-Scale Texture Enhancement Channel-Aware Attention) for lossless multi-scale feature fusion, ODConv (Omni-Dimensional Dynamic Convolution) for dynamic adaptation to irregular geometries, LMDP (Local Multi-Scale Discriminative Perception) for selective noise suppression and micro-defect amplification, and DDFusion (Diffusion-Driven Feature Fusion) for scene-aware noise modeling. Pruning further reduces computational complexity while improving accuracy. Extensive experiments on the specialized DDTE dataset and public benchmarks demonstrate state-of-the-art performance. Our model achieves 82.6% mAP₅₀ and 61.6% mAP_{50–95} on DDTE, while maintaining a high inference speed of 193.5 FPS with only 8.46M parameters. It also shows excellent generalization across NEU-DET, GC10-DET, and cross-domain tasks, providing an efficient and accurate solution for industrial defect inspection.

Keywords Surface defect detection, Polysilicon production, Industrial defect inspection

The surface integrity of steel components is critical for the safety and longevity of polysilicon production equipment in photovoltaics. Surface defects, such as microcracks (Cr), silicon deposits (SD), pits (PT), and impurity spots (IS), originate from manufacturing or operational stress. These defects can severely compromise material strength and may lead to catastrophic failure (Fig. 1). Consequently, automated, precise detection of these defects is imperative for quality control.

Deep learning-based detectors are now the predominant solution, yet they struggle with the compounded challenges of real-world photovoltaic inspection. We identify three fundamental limitations: (1) Limited multi-scale discriminability, where minute defects are easily lost against complex backgrounds during feature fusion; (2) Insufficient geometric adaptability of static convolutional kernels to capture irregular shapes like Cr and SD; and (3) An inherent robustness-sensitivity trade-off, where enhancing noise robustness often diminishes sensitivity to low-contrast defects.

To address these limitations, we propose MSEOD-DDFusionNet, an integrated defect detection framework. Our principal contributions are threefold:

A lossless multi-scale fusion principle that preserves micro-defect signatures against complex backgrounds, establishing a high-fidelity feature foundation.

A multi-dimensional dynamic adaptation mechanism that enables convolutional kernels to precisely capture irregular defect geometries based on the fused features.

¹School of Software Engineering, Xinjiang University, Urumqi 830046, China. ²School of Public Administration, Public Administration Major, Xi'an University of Architecture and Technology, Xi'an 710055, China. ³Technical Department, Xinjiang Binghua Technology Co., Ltd., Urumqi 830000, China. ⁴Smart Manufacturing Division, Xinte Energy Co. Ltd., Urumqi 830011, China. ⁵School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China. ⁶State Grid Xinjiang Electric Power Company Economic Research Institute, Urumqi 830002, China. ✉email: super_sjh@163.com

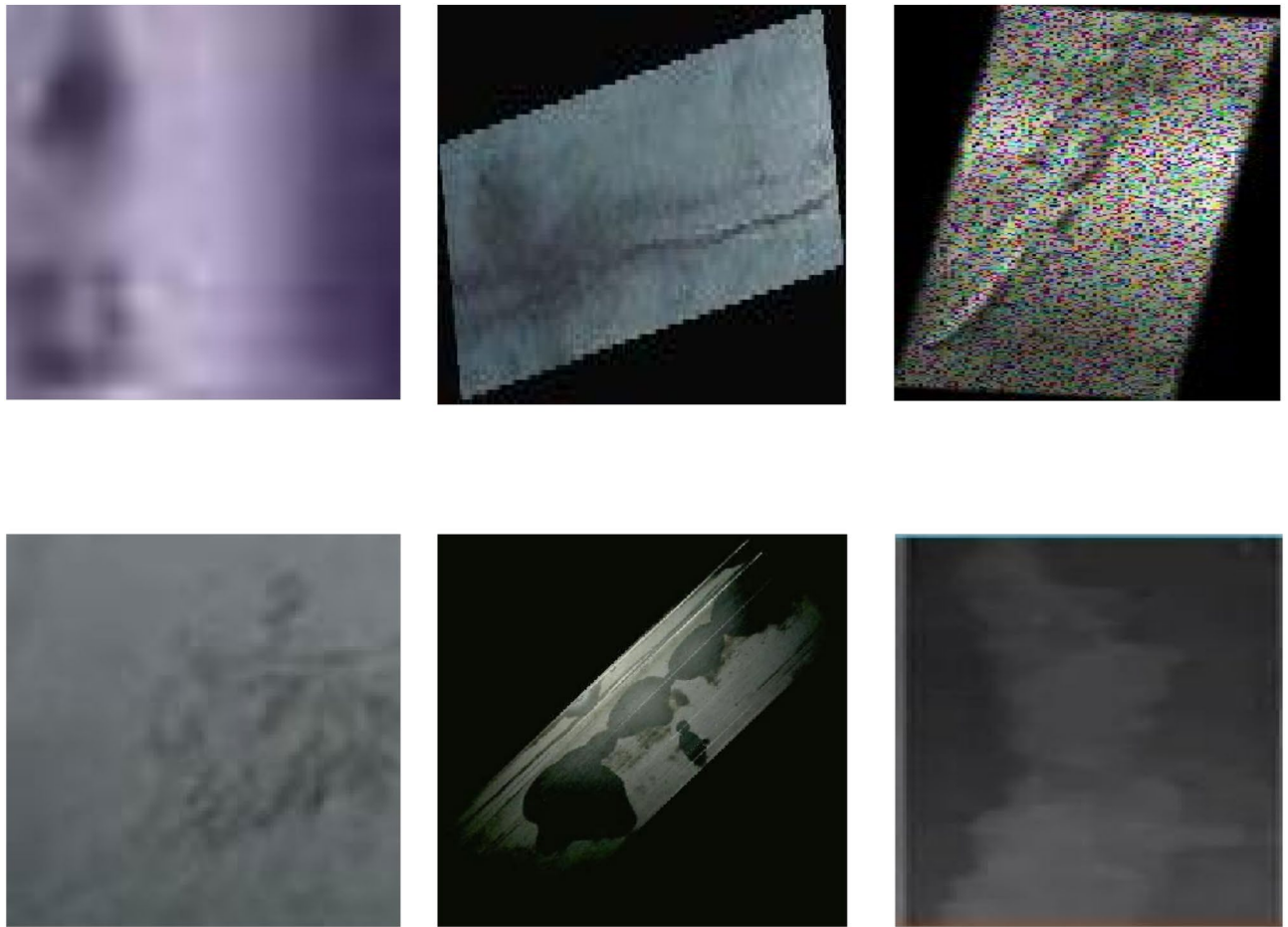


Fig. 1. Representative steel surface defects highlight the key challenges of detection. These challenges arise from the high diversity in morphology, scale, and background texture among defect categories such as microcracks (Cr), pits (PT), and silicon deposits (SD).

A decoupled noise robustness strategy that refines features through selective real-time suppression and training-time learned invariance, thereby resolving the sensitivity-robustness trade-off.

Additionally, we contribute the DDTE dataset, a specialized industrial benchmark providing high-resolution annotated images to address the scarcity of domain-specific data. Extensive experiments demonstrate that our framework achieves state-of-the-art accuracy and efficiency while exhibiting superior generalization across multiple domains.

The remainder of this paper is organized as follows: Section "[Related work](#)" reviews related work; Section "[Methodology and design](#)" details the network architecture; Section "[Experiments](#)" presents the experimental analysis; and Section "[Conclusion](#)" concludes the paper.

Related work

Multi-scale feature fusion

Constructing feature pyramids is fundamental for surface inspection, as it represents defects at different scales. Prior research, such as MSAF-YOLO¹ and UWSDNet², has balanced accuracy and efficiency through spatial transformations or re-parameterized designs. To further enhance feature quality, attention mechanisms have been widely integrated—from SENet³ focusing on channel relations, to variants incorporating spatial coordinates⁴, and to CBAM-based hybrids fusing both⁵. However, their core operation compresses spatial context into compact channel descriptors via global pooling. While this improves efficiency, the compression acts as a low-pass filter. It attenuates the weak, high-frequency activations that characterize micro-defects before fusion. Consequently, critical target features become submerged within complex background textures. This significantly reduces discriminability during multi-scale aggregation. Together, these effects reveal a fundamental limitation in the prevailing compression-based attention paradigm: it inevitably leads to channel-dimensional information loss. Therefore, a new multi-scale fusion principle is needed to avoid such loss.

Dynamic convolution

Dynamic convolution enhances geometric adaptability. Its development has evolved through several stages: from attention-guided kernel selection⁶ and constrained sampling optimization⁷ to cross-domain transfer⁸. Recent

advances, such as AKSNet⁹ and MDSCELAN¹⁰, further explore adaptive kernel width and multi-scale dynamics. However, these methods share a common shortcoming—incomplete dimensional adaptation. They typically optimize only one aspect at a time, such as spatial deformation, kernel weights, or channel importance. This fragmented approach is non-collaborative. Consequently, it fails to capture the complex coupling among spatial morphology, feature channels, and kernel parameters in irregular defects like silicon deposits and microcracks. Therefore, precise geometric feature extraction requires a new solution: a mechanism capable of unified, co-modulated adjustment of convolutional weights across spatial, channel, and kernel dimensions.

Noise robustness

Industrial noise demands robust feature learning. While adversarial training¹¹ enhances invariance, it can destabilize training. Alternative strategies, such as diffusion models¹² and integrated frameworks¹³, offer generative noise modeling, and CycleGAN-based methods augment scarce data¹⁴. However, these strategies primarily act as global regularizers. They improve overall robustness by adapting to a broad noise distribution, but they also introduce a key drawback: they uniformly raise feature activation thresholds. This reduces sensitivity to low-contrast micro-defect signals—the classic robustness-sensitivity trade-off. Standard diffusion models show potential but are limited by their fixed, scene-agnostic noise injection, which lacks targeted modeling of physical interferences like hot spots or motion blur. This limitation constrains feature enhancement precision. Therefore, advancing beyond this trade-off requires a dual approach. First, we need an immediate processing mechanism for selective noise suppression and conditional signal amplification. Second, an enhanced learning paradigm must simulate physical noise patterns and achieve feature-fidelity recovery to improve inherent robustness.

Our MSEOD-DDFusionNet framework bridges this gap by integrating a coherent sequence of specialized modules that directly address these foundational challenges.

Methodology and design

System overview

The overall architecture and data flow of MSEOD-DDFusionNet are depicted in Figs. 2 and 3. The network follows a cascading design from feature extraction to defect detection. The processing sequence begins as input images undergo basic feature extraction through the backbone network. The extracted features then enter a pipeline of four specialized modules. First, the MTECAAttention module performs multi-scale feature fusion to preserve fine details of micro-defects. Next, the ODConv module dynamically adjusts convolutional weights based on these fused features, adapting to the irregular geometries of defects. The LMDP module then refines the features using a dual-stream strategy. It selectively amplifies potential defect signals while suppressing background noise. Throughout this process, the DDFusion module continuously enhances feature robustness via scene-aware noise modeling and denoising learning. Finally, the processed features are fed into the detection head for classification and localization. This cascaded design ensures that each module's output directly informs

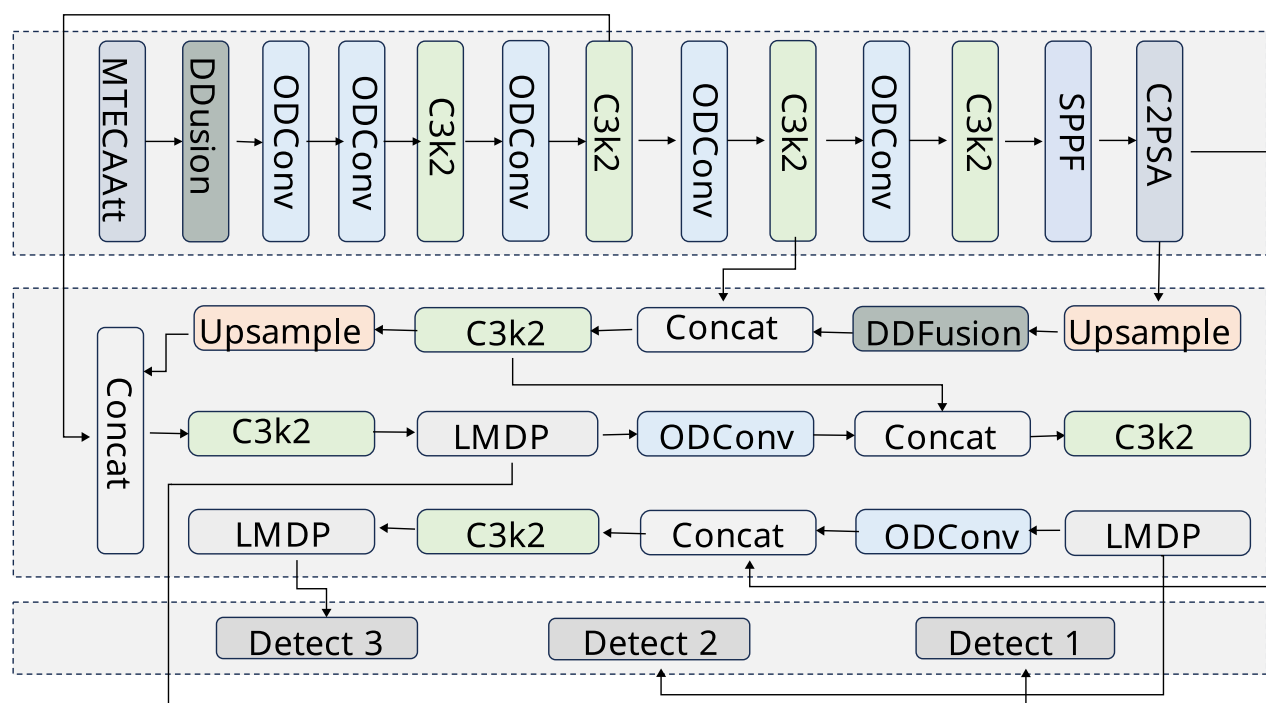


Fig. 2. The overall architecture of MSEOD-DDFusionNet. It features an MTECAAtt-equipped backbone for multi-scale feature extraction, a neck network that integrates ODConv and LMDP within a feature pyramid for enhancement and adaptation, a multi-scale detection head, and a pervasive DDFusion module employing a diffusion-driven strategy to boost robustness in low-contrast scenarios.

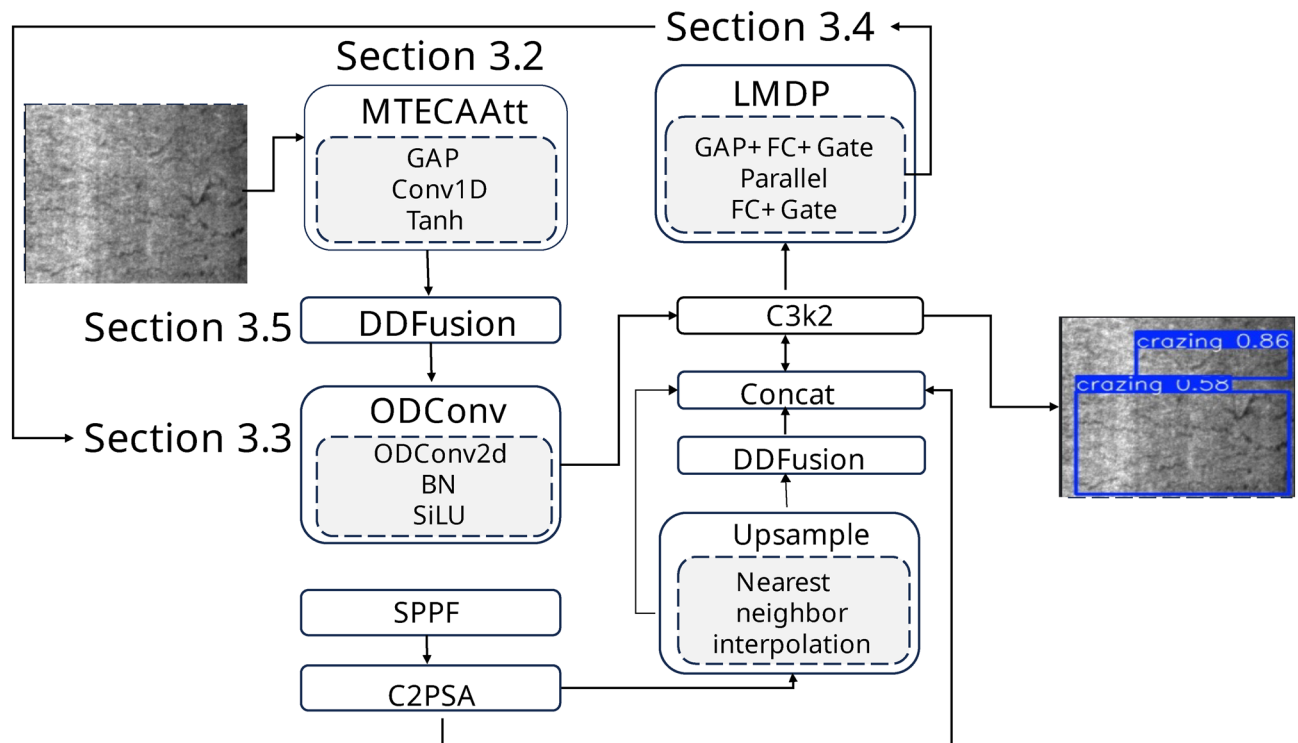


Fig. 3. The overall flowchart of MSEOD-DDFusionNet. The diagram depicts the complete pipeline from image input to defect detection output. The process begins with multi-scale feature extraction via the MTECAAttention module, followed by shape adaptation using ODConv. Subsequent stages involve feature fusion and optimization through SPPF, C2PSA, and LMDP modules. Simultaneously, the DDFusion module enhances robustness via noise injection and denoising. The pipeline concludes with defect identification and localization in the detection head.

the next, forming a coherent pipeline from feature preservation to robust detection. The specific designs and innovations of each module are detailed in the following subsections.

MTECAAttention module

The MTECAAttention module performs lossless multi-scale fusion via a three-stage workflow, as depicted in Fig. 4. Its design preserves all channel information during fusion, avoiding the information loss typical of compression-based methods. This produces a high-fidelity feature map that serves as the foundation for subsequent modules. The three stages are as follows:

First, multi-branch parallel feature extraction. This stage uses depthwise separable convolutions with different kernel sizes (3×3 , 5×5 , 7×7). These convolutions capture diverse information simultaneously: defect edge details, mid-scale morphology, and context with complex backgrounds. A key aspect is that all original channel information is preserved. This establishes the essential basis for the lossless fusion process.

Second, cross-scale fusion. A 1D convolution aggregates the channel responses from the multi-scale features. Critically, this operation does not reduce dimensionality. Instead, it integrates responses from the same original channel across different scales. The output is a fusion weight vector that retains the original channel count.

Finally, interpretable channel calibration. The Tanh function is applied to constrain the fusion weights to the interpretable range $[-1, +1]$. These weights then perform channel-wise weighting on the original input features: channels with weights near $+1$ are enhanced, while those near -1 are suppressed.

By fundamentally eliminating information loss at this initial stage, the MTECAAttention module resolves a bottleneck that would otherwise propagate and limit the effectiveness of the entire cascade.

ODConv module

The ODConv module performs omni-dimensional dynamic convolution via a three-stage process, as depicted in Fig. 5. Its core innovation is a four-dimensional dynamic weight co-modulation mechanism, enabling holistic adaptation to irregular geometries. This provides the adaptive convolutional foundation required for subsequent feature refinement. The three stages are as follows:

First, dual-path feature statistics extraction. This stage employs concurrent global average pooling and global max pooling to extract feature statistics. This dual-path approach captures both global contextual information and locally salient features. The combined statistics provide richer cues for the subsequent generation of adaptive weights.

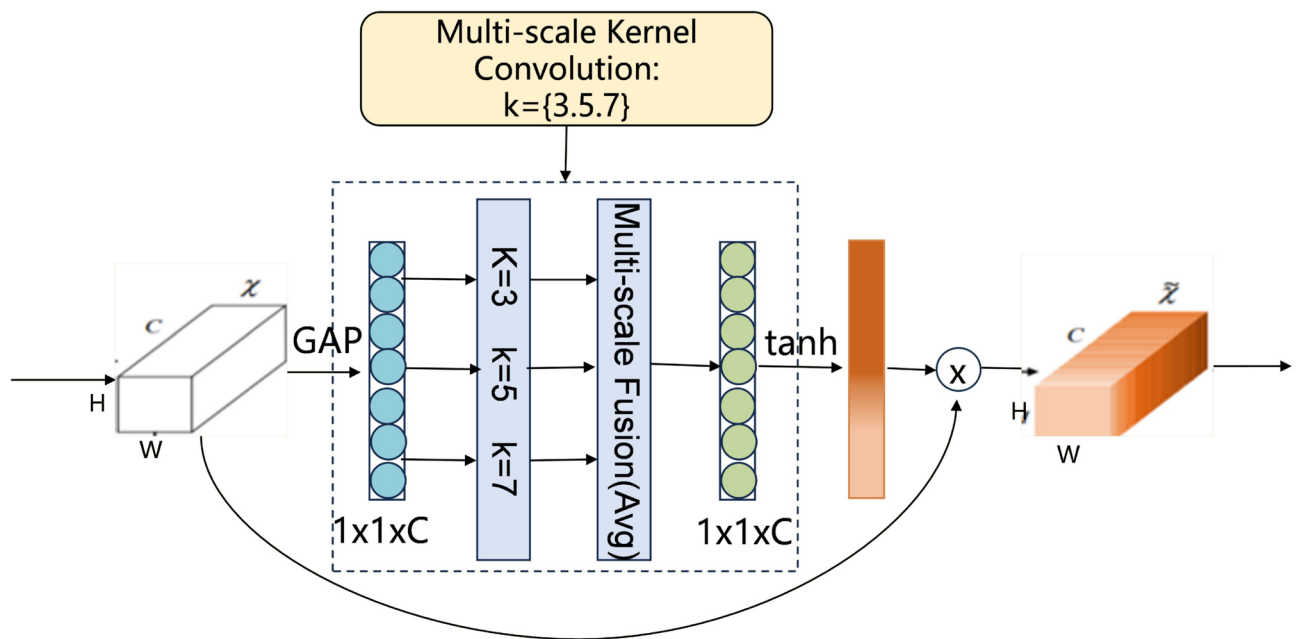


Fig. 4. The MTECAAttention module. The module performs multi-scale feature extraction via depthwise separable convolutions ($k=3,5,7$), followed by feature aggregation, channel weight generation (Conv1D + Tanh), and input feature recalibration — all while preserving original dimensions ($C \times H \times W$).

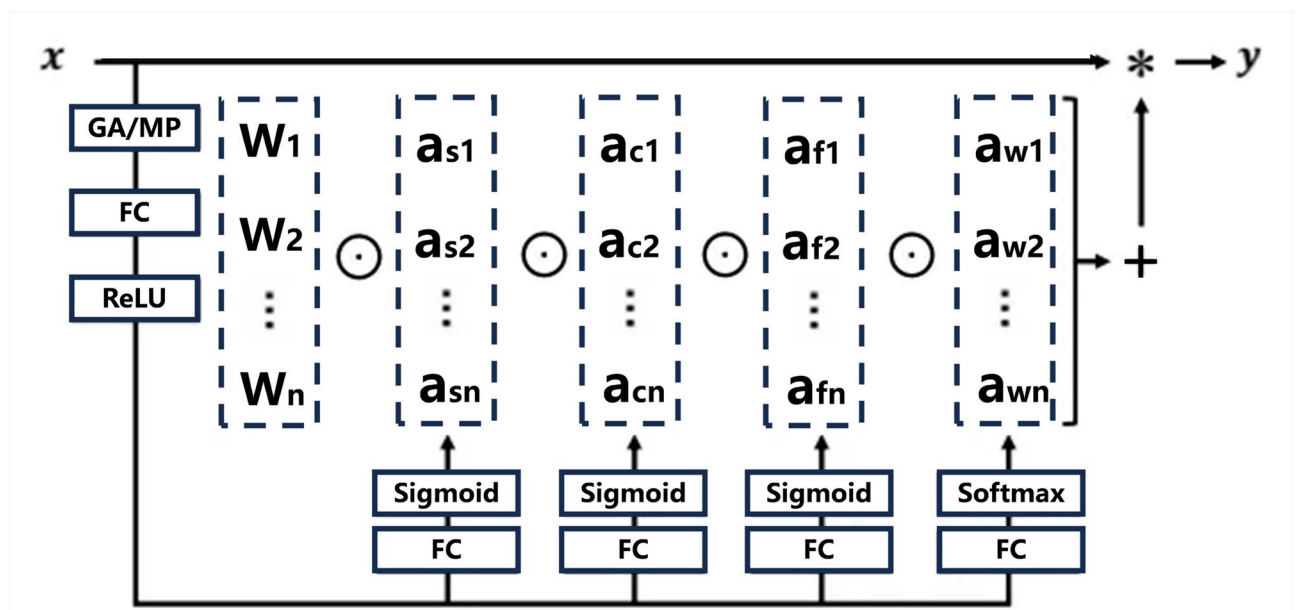


Fig. 5. The ODConv module. The ODConv module employs a multidimensional attention mechanism to simultaneously compute four types of attention across all four dimensions of the convolutional kernel space: a_{si} , a_{ci} , a_{fi} , and a_{wi} .

Second, four-dimensional attention weight generation. A lightweight parallel network processes the extracted statistics to generate four sets of attention weights. Critically, one unique weight set is generated for each of the four convolution dimensions: spatial, input channel, output channel, and kernel. This design ensures holistic adaptation across all dimensions, directly addressing the partial adaptation limitation of prior methods.

Third, element-wise co-modulation. The four sets of attention weights are fused with the base convolution kernel through element-wise multiplication and broadcasting. This deep, co-modulated fusion allows the convolutional operation to dynamically reconfigure its focus during each forward pass. Consequently, it emphasizes spatial regions containing irregular defects while suppressing irrelevant background patterns.

By establishing this fine-grained, global control over the convolution process, the ODConv module achieves precise characterization of the complex and irregular geometries prevalent in photovoltaic steel surface defects. It thereby fulfills its role as the adaptive processing unit within the integrated cascade.

LMDP module

The LMDP module performs local multi-scale discriminative perception via a dual-stream architecture, as depicted in Fig. 6. Its core innovation is a decoupled processing strategy that separates noise suppression from signal amplification, enabling precise enhancement of subtle defects in high-noise environments. This provides the targeted feature refinement required for robust detection. The two parallel processing streams operate as follows:

Stream 1: Selective Noise Suppression. This stream first computes channel-wise fusion statistics and generates corresponding suppression weights. A key design is the inversion of these weights. This mechanism selectively attenuates only predefined noise-sensitive channels. Crucially, all other channels are preserved intact, including those carrying low-amplitude defect signatures. This prevents the accidental suppression of critical defect signals during noise removal.

Stream 2: Conditional Signal Amplification. This stream employs a separate 1×1 convolutional branch to produce feature scaling factors. Its key innovation is a threshold-based conditional mechanism. Amplification (e.g., by a factor of 1.2) is applied only to spatial regions where the feature response intensity falls below a predefined threshold. This ensures that only faint, potential defect features are strengthened, without amplifying already dominant background responses.

Through the synergistic operation of these two streams, the LMDP module effectively amplifies low-contrast micro-defect features while suppressing background noise. It thereby directly resolves the long-standing sensitivity-robustness trade-off in industrial visual inspection. The module fulfills its role as the dedicated feature refinement unit within the cascade, delivering noise-aware suppression and defect-aware enhancement.

DDFusion module

The DDFusion module is introduced to solve a fundamental limitation that preceding modules cannot address: learning inherent, generalized robustness against unseen and complex physical noise patterns. It enhances feature robustness through a diffusion-driven, scene-aware denoising process, as a core component of our framework. Its innovation lies in a dual-mechanism design that addresses two key challenges: injecting realistic physical noise and ensuring distortion-free feature recovery. This goes beyond simple filtering or augmentation; it embeds a physical-world noise resilience directly into the feature representation. This equips the entire pipeline with enhanced resilience against low-contrast, noisy industrial conditions. The process is built upon two coordinated mechanisms.

Mechanism 1: Scene-Aware Dynamic Noise Modeling. Unlike standard diffusion models that use fixed, data-agnostic noise, our approach injects noise that simulates realistic physical interference patterns. This key departure transforms the diffusion process from a generic regularizer into a targeted simulator of domain-specific degradations. For example, it applies stronger intensity fluctuations to simulate high-temperature zones and moderate blur for welding areas. This is achieved by dynamically mapping the noise coefficient β_t to specific equipment states. The forward process introduces this scene-specific noise ϵ to the original feature x_0 at step t :

$$\alpha_t = 1 - \beta_t \quad (1)$$

$$\bar{\alpha}_t = \prod_{i=1}^t 1 - \beta_i \quad (2)$$

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (3)$$

This teaches the model to preserve defect features under targeted, realistic corruption.

Mechanism 2: Time-Aligned Progressive Feature Decoupling. To recover clean features without distortion—a common failure point in naive denoising that smears defect signatures—this mechanism enforces strict temporal synchronization between the forward noise addition and the backward denoising paths. This ensures that the

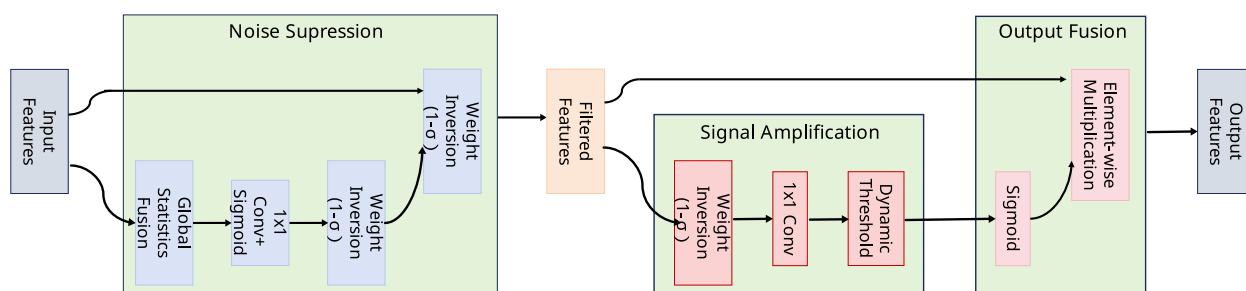


Fig. 6. LMDP overall structure. Schematic diagram showing the dual-stream architecture with parallel noise-suppression and signal-amplification paths, their interconnections, and output fusion.

recovery process is a precise inverse of the corruption, preserving critical high-frequency defect details. A U-Net, conditioned on the step t , predicts the noise component $\varepsilon_t = \text{UNet}(x_t, t)$. The denoised estimate x'_0 is then calculated by precisely reversing each forward step:

$$x'_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \varepsilon_t \right) \quad (4)$$

The alignment is maintained via $\bar{\alpha}_{t-1} = \frac{\bar{\alpha}_t}{\alpha_t}$, ensuring each denoising step accurately inverts its corresponding corruption. This enables the clean separation of noise from subtle defect signatures.

The DDFusion module introduces two core mechanisms to enhance robustness: scene-aware dynamic noise modeling and time-aligned progressive feature decoupling. During training, it injects feature-space noise that mimics real physical interference—such as thermal intensity fluctuations and welding motion blur—and learns the corresponding denoising process. This approach not only improves generalization to unseen noise but also strengthens localization accuracy and feature stability through feature-level restoration. Consequently, DDFusion acts as a generative noise-modeling and feature-recovery paradigm, providing subsequent detection heads with more discriminative and noise-invariant representations.

Experiments

Experimental environment

All experiments were conducted on a computational server equipped with an NVIDIA A30 GPU and an Intel Xeon Silver 4314 processor. Models were implemented in PyTorch 2.4.1 and accelerated with CUDA 12.4 and cuDNN 8.2.4.

Training was performed for 200 epochs using stochastic gradient descent. The optimizer was configured with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005.

To improve robustness against industrial variations, we employed a comprehensive data augmentation strategy. For the first 190 epochs, mosaic augmentation was applied by combining four randomly scaled images into a single 640×640 input to simulate multi-defect scenarios. This was supplemented with standard augmentations to mimic real-world conditions: random rotations ($\pm 45^\circ$), horizontal and vertical flips, brightness adjustments ($\pm 30\%$), contrast variations ($\pm 25\%$), and Gaussian blur ($\sigma = 0.5\text{--}1.5$). In the final 10 epochs, aggressive augmentations were disabled. Only minor brightness adjustments were retained to stabilize convergence and align the training process with inference conditions.

Data sets and assessment indicators

This study adopts a combined validation strategy, utilizing both a specialized industrial dataset and public benchmarks to assess generalization. The core evaluation relies on our proprietary Polysilicon Distillation Tower Equipment (DDTE) dataset, with additional validation from public datasets.

A. Specialized Industrial Defect Dataset: DDTE

To address the scarcity of defect data specific to polysilicon production equipment, we introduce the Polysilicon Distillation Tower Equipment (DDTE) dataset. It comprises 6,252 high-resolution images captured under extreme operational environments ($1000 - 1200^\circ\text{C}$ with corrosive vapors). These images cover six critical defect types: microcracks (Cr), silicon deposits (SD), pits (PT), scratches (SC), impurity spots (IS), and weld defects (WD). Data were collected using a DJI Matrice 350 RTK unmanned aerial vehicle (UAV) equipped with a Zenmuse H20T camera during maintenance periods. To mitigate intense specular reflections from high-temperature surfaces, a circular polarizing filter was employed. The UAV operated at a distance of 3–5 meters under varying natural lighting conditions, achieving a ground sampling distance (GSD) of 0.8–1.5 mm/pixel, a resolution sufficient for the detection of micro-defects. All images were annotated by three quality inspection experts following a standardized guideline, with tight bounding boxes drawn around each defect. To ensure consistency, a random subset of 500 images was independently annotated by all experts, resulting in high inter-annotator agreement ($\text{IoU} > 0.85$). Finally, the dataset was randomly split into training (5,001 images), validation (625 images), and test (626 images) sets at an 8:1:1 ratio, ensuring a balanced representation of all defect categories.

B. Public General-Purpose Verification Datasets

NEU-DET: This dataset contains 1,800 grayscale images of hot-rolled steel strips. It includes six common defect types: rolled-in scale (Rs), patches (Pa), cracks (Cr), pitting surfaces (Ps), inclusions (In), and scratches (Sc), with 300 samples per type.

GC10-DET: An industrial surface defect dataset comprising 3,570 grayscale images across 10 defect categories, such as punched holes, welds, and crescent gaps.

PASCAL VOC 2007: A widely adopted computer vision benchmark. It provides 20 object categories with detailed XML annotations, supporting tasks like object detection and image classification.

BCCD: A blood cell detection dataset annotated in the PASCAL VOC format, containing labels for red blood cells, white blood cells, and platelets. Its 640×480 -pixel JPEG images with XML annotations are used to evaluate the model's cross-domain performance.

All datasets were split into training, validation, and test sets at a consistent 8:1:1 ratio to ensure fair evaluation.

Evaluation Metrics

We employ a tiered evaluation strategy. The primary accuracy metrics are mean Average Precision at an IoU threshold of 0.5 (mAP_{50}) and over the range of 0.5 to 0.95 (mAP_{50-95}). Model efficiency is assessed through parameter count (params), computational complexity (Gflops), and inference speed (FPS). For domain-specific benchmarks (NEU-DET, GC10-DET), we perform diagnostic analysis using class-wise confusion

matrices and F1-confidence curves. These analyses primarily compare our model against a YOLOv11s baseline to highlight classification improvements. For the core DDTE dataset and cross-domain tests (PASCAL VOC, BCCD), evaluation focuses on practical detection capability and generalization. We report mAP₅₀ and mAP_{50–95}, supported by qualitative visualizations, as these metrics best reflect real-world localization accuracy and robustness. This structured approach ensures that the analysis of each dataset aligns with its specific validation purpose.

Analysis of ablation experiments

This section systematically assesses the contributions of each core component in MSEOD-DDFusionNet through module-level ablation experiments. By integrating quantitative metrics with qualitative visual evidence, we address two primary research questions: (1) whether the performance gains from individual modules are statistically significant, and (2) whether each module confers distinct functional advantages. The experiments further elucidate the synergistic interactions that emerge when these modules are integrated.

Table 1 summarizes the ablation results across three datasets: DDTE, NEU-DET, and GC10-DET. On the specialized DDTE dataset, the complete model achieves 82.6% mAP₅₀ and 61.6% mAP_{50–95}, representing a 2.7% relative improvement over the baseline configuration with only MTECAAttention (80.4% mAP₅₀). This improvement trend holds consistently across datasets: on NEU-DET, the complete model attains 80.5% mAP₅₀, a 7.3% relative gain over the baseline (75.0%); on the more challenging GC10-DET, performance rises from 60.4% to 67.4%, corresponding to an 11.6% relative increase. The cross-dataset consistency confirms that the functional advantages of each module are robust and generalize beyond the characteristics of any single dataset.

Table 2 provides a detailed, reproducibility-focused ablation analysis conducted on the DDTE dataset. To ensure statistical reliability and mitigate concerns that module-level gains might fall within the range of experimental noise, we report the mean and standard deviation derived from three independent experimental runs (using random seeds 0, 123, and 456). All performance improvements attributed to the modules exceed the observed experimental variance, thereby confirming their statistical significance.

Specifically, the MTECAAttention module yields a stable +0.25% relative gain in mAP₅₀ (80.4±0.3% → 80.6±0.2%). ODConv contributes an additional +0.5% relative gain (80.6±0.2% → 80.8±0.3%) while simultaneously reducing computational load by 25.4%. LMDP further enhances performance by +0.24% (80.8±0.3% → 81.0±1.0%). DDFusion delivers the most substantial improvement, elevating the localization-precision metric mAP_{50–95} from 58.9±0.8% to 61.6±0.9%, a 4.6% relative gain. DDTE was selected for this in-depth ablation due to its high industrial relevance: designed explicitly for polysilicon production equipment inspection, its extreme operating conditions and complex imaging backgrounds provide a realistic and demanding testbed for evaluating module utility in practical applications.

The sequential integration of modules reveals clear synergistic effects. The combination of MTECAAttention and ODConv improves detection accuracy while lowering computational cost, demonstrating a strong

Datasets	MTECAAtt	ODConv	LMDP	DDFusion	mAP ₅₀	mAP _{50–95}	FPS	Gflops	params
DDTE					80.4	58.2	62.56	213.06	10.73
	✓				80.6	58.6	63.4	213.08	10.73
		✓			80.8	58.9	56.91	158.97	8.11
			✓		80.5	58.3	182.4	190.24	10.73
				✓	80.6	58.4	197.8	213.06	10.73
	✓	✓			80.9	58.7	58.5	158.99	8.11
	✓	✓	✓		81.0	58.9	183	159.0	8.46
	✓	✓	✓	✓	82.6	61.6	193.5	159.0	8.46
NEU-DET					75	43.1	62.02	213.06	10.73
	✓				78.5	45.3	58.7	213.08	10.73
		✓			78.6	45.7	52.19	158.97	8.11
			✓		77	45.4	105.5	190.24	10.73
				✓	77.4	44.6	118.9	213.06	10.73
	✓	✓			78.9	44.7	51.99	158.99	8.11
	✓	✓	✓		79.2	45.8	59.1	159.0	8.46
	✓	✓	✓	✓	80.5	46.9	156.1	159.0	8.46
GC10-DET					60.4	29.7	123.2	213.06	10.73
	✓				63.8	33.2	179.0	213.08	10.73
		✓			62.4	31.1	190.7	158.97	8.11
			✓		60.8	30.4	130.7	190.24	10.73
				✓	63.8	32	167.6	213.06	10.73
	✓	✓			65.2	33	52.0	158.99	8.11
	✓	✓	✓		65.1	33.9	159.8	159.0	8.46

Table 1. Ablation study results.

datasets	MTECAAtt	ODConv	LMDP	DDFusion	mAP ₅₀	mAP ₅₀₋₉₅	FPS
DDTE					80.4±0.3	58.2±0.5	62.56±1.2
	✓				80.6±0.2	58.6±0.4	63.4±1.0
		✓			80.8±0.3	58.9±0.6	56.9±0.8
			✓		80.5±0.4	58.3±0.5	182.4±1.1
				✓	80.6±0.3	58.4±0.6	197.8±0.6
	✓	✓			81.2±0.3	58.7±0.4	58.5±0.9
	✓	✓	✓		81.0±1.0	58.9±0.8	183±1.3
	✓	✓	✓	✓	82.6±0.7	61.6±0.9	193.5±1.5

Table 2. Statistical Ablation Analysis on DDTE (Mean ± Std, 3 runs).

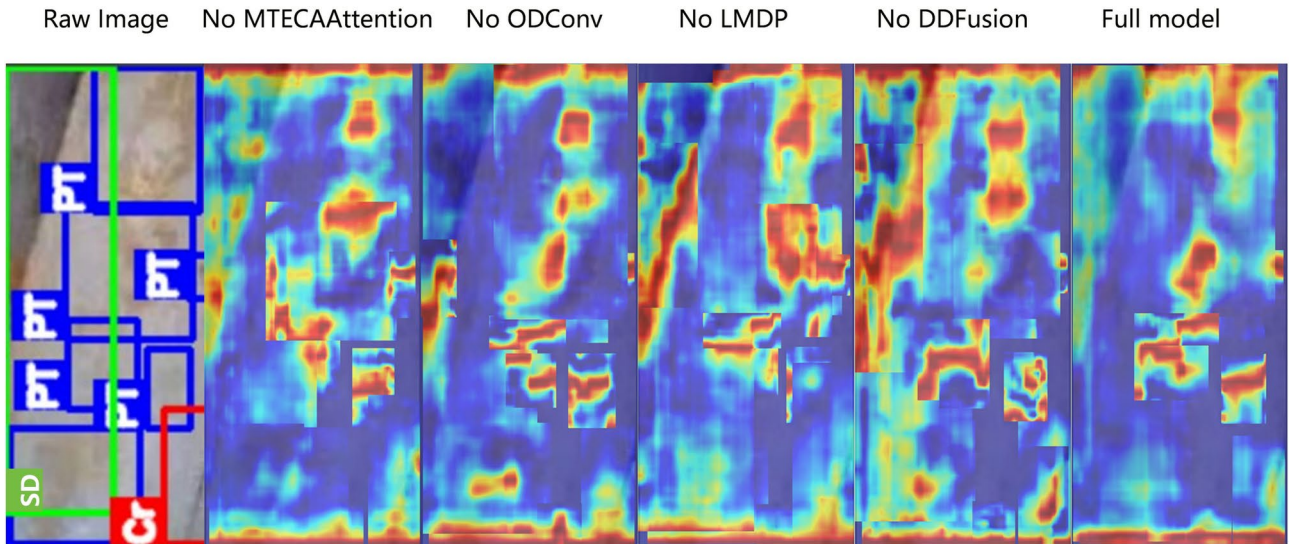


Fig. 7. Characteristic failure modes from module ablation. Visual comparisons on the DDTE dataset show the specific performance degradation from removing each core module: (a) Absence of MTECAAttention fragments micro-cracks; (b) Removing ODConv misaligns bounding boxes for irregular defects; (c) Without LMDP, low-contrast spots are missed under noise; (d) Omitting DDFusion reduces robustness under challenging illumination.

compatibility between lossless multi-scale feature fusion and dynamic geometric adaptation. Incorporating LMDP subsequently enhances the model’s sensitivity to low-contrast defects without compromising inference efficiency. Finally, the integration of DDFusion produces the most pronounced leap in localization precision (mAP₅₀₋₉₅), indicating that its scene-aware denoising mechanism effectively consolidates and refines the feature representations produced by the preceding modules. This stepwise performance enhancement validates the complementary roles of the modules across the core stages of the detection pipeline: multi-scale feature fusion, irregular shape adaptation, noise-aware refinement, and generalized robustness learning.

Collectively, the results presented in Tables 1 and 2 underscore the comprehensive advantages of the complete MSEOD-DDFusionNet architecture. On the core DDTE dataset, our model outperforms the widely adopted YOLOv11s baseline by 2.7% in mAP₅₀ and 5.8% in mAP₅₀₋₉₅, while utilizing 21.2% fewer parameters and delivering 209% higher inference speed. These concurrent gains across accuracy, efficiency, and model complexity demonstrate that our modular integration strategy effectively pushes the performance frontier of industrial visual inspection systems.

Figure 7 complements the quantitative analysis by providing visual evidence of the characteristic failure modes that emerge when individual modules are ablated. Removing MTECAAttention causes micro-cracks to fragment and disappear within cluttered backgrounds, underscoring the necessity of lossless multi-scale fusion for preserving fine defect structures. Ablating ODConv leads to bounding-box misalignment with the irregular contours of defects such as silicon deposits, confirming the module’s indispensable role in geometric adaptation. The absence of LMDP results in missed detections of low-contrast flaws under noisy conditions, validating the efficacy of its dual-path design for selective noise suppression and conditional signal amplification. Omitting DDFusion significantly degrades model robustness under challenging perturbations like intense glare or motion blur, highlighting the critical importance of its scene-aware denoising for reliable performance in variable industrial environments.

Datasets	Method	Cr	SD	PT	SC	IS	WD	mAP ₅₀
DDTE	YOLOv5s	79	84.8	76.6	67.9	91.9	81.9	80.4
	YOLOv8s	76.6	82.2	75.1	68.2	87.7	83.2	78.8
	YOLOv10s	79.4	80.3	72.1	66.7	83.5	81.4	77.5
	YOLOv11s	84.4	84.7	73.6	68.3	85.9	85.3	80.4
	FFDDNet ¹⁵	79.9	83.3	72.1	68.0	81.9	85.1	78.4
	Ours	84.2	86.4	78.9	74.6	85.0	87.9	82.6

Table 3. Per-class detection performance on DDTE (vs. YOLOv11s: +2.7% mAP₅₀; PT +7.2%, SC +9.2%, WD +3.0%).

Datasets	Methods	Params	Gflops	FPS	mAP ₅₀	mAP _{50–95}
DDTE	Faster R-CNN ¹⁶	41.364	90.913	-	77.5	59
	YOLOv5s	9.13	192.67	65.74	80.4	57.8
	YOLOv8s	11.14	229.5	65.15	78.8	56.9
	YOLOv10s	8.07	198.33	64.35	77.5	55.4
	YOLOv11s	10.73	213.06	62.56	80.4	58.2
	DAB-DETR ¹⁷	44	216	120.35	79.6	-
	Deformable-DETR ¹⁸	40	173	122.15	78.4	-
	DINO ¹⁹	218	-	113.76	79.7	-
	LFF-YOLO ²⁰	60.51	6.85	168.49	78.9	-
	FFDDNet ¹⁵	10.08	222.51	155.0	78.4	55.1
	LE-YOLOv5 ²¹	4.8	10.3	180.0	79.9	57.1
	Ours	8.46	159	193.5	82.6	61.6

Table 4. Overall performance comparison on DDTE (vs. YOLOv11s: +2.7% mAP₅₀, +209% FPS, −21.2% params).

In summary, the ablation study—supported by both rigorous quantitative metrics and illustrative qualitative visualizations—confirms that each module contributes meaningfully beyond the bounds of experimental variance. More importantly, their systematic integration yields a coherent and high-performance detection framework. This design not only effectively addresses the intertwined challenges of scale variation, geometric irregularity, and environmental noise inherent in industrial inspection but also establishes a scalable and efficient architectural foundation for real-world, precision defect detection.

Analysis of comparison experiments

MSEOD-DDFusionNet is evaluated through a three-tier strategy to verify its performance and generalization: on the specialized DDTE dataset (photovoltaic-specific), on public steel defect benchmarks (NEU-DET, GC10-DET), and on cross-domain datasets (PASCAL VOC 2007, BCCD). All comparisons are conducted with consistent metrics (mAP₅₀, mAP50 – 95, params, Gflops, FPS) to ensure a fair and thorough assessment.

Comparison of DDTE assay

We conducted a comprehensive evaluation of MSEOD-DDFusionNet on the DDTE dataset. As shown in Table 3 and Table 4, the model demonstrates significant advantages in detection accuracy, computational efficiency, and adaptability.

For per-class defect detection accuracy, MSEOD-DDFusionNet achieves optimal mAP₅₀ on four key defect types: silicon deposition (SD) at 86.4%, pits (PT) at 78.9%, scratches (SC) at 74.6%, and weld defects (WD) at 87.9%. These results validate the effectiveness of the modular architecture: the ODConv module effectively captures irregular geometric features, while the combination of LMDP and DDFusion significantly enhances detection capability for low-contrast defects. Compared with the specialized algorithm FFDDNet, our model shows clear improvements across multiple defect classes. Notably, the substantial 9.4% relative improvement in PT detection (78.9% vs. 72.1%) highlights the advantage of ODConv in adapting to complex pit geometries. The 3.3% relative gain in WD detection (87.9% vs. 85.1%) demonstrates the effectiveness of our noise-aware modules in handling challenging weld regions. For microcrack (Cr) detection, our mAP₅₀ of 84.2% is comparable to YOLOv11s (84.4%), while showing a 5.4% relative improvement over FFDDNet (79.9%). This indicates enhanced feature discrimination despite the slight performance gap with the strongest baseline. The 3.8% improvement in IS detection (85.0% vs. 81.9%) confirms the benefit of multi-scale fusion; however, the remaining gap compared to YOLOv5s (91.9%) suggests that detecting sparse, isolated defects requires further architectural refinement.

In overall performance, MSEOD-DDFusionNet achieves state-of-the-art results: 82.6% mAP₅₀ and 61.6% mAP_{50–95}. This represents a 2.7% relative improvement over the best baseline YOLOv11s (80.4% mAP₅₀) and a more substantial 5.4% improvement over the specialized FFDDNet (78.4% mAP₅₀). The consistent gains

across these benchmarks confirm the synergistic effect of integrating multiple specialized modules rather than relying on a single enhancement strategy. In computational efficiency, our model maintains an excellent balance between performance and practicality. With only 8.46M params and 159.0 Gflops, it achieves 193.5 FPS inference speed. This represents a 209% improvement over YOLOv11s (62.56 FPS) and a 24.8% improvement over FFDDNet (155.0 FPS). The speedup stems from the efficient design of our modules, particularly the lightweight LMDP and the structured pruning applied. Although LE-YOLOv5 is more parameter-efficient (4.8M params), our model delivers 3.4% higher mAP_{50} accuracy, offering a clearly superior accuracy-efficiency trade-off that is critical for real-time industrial inspection.

In conclusion, comprehensive experiments confirm that MSEOD-DDFusionNet achieves state-of-the-art performance in steel surface defect detection. It balances high accuracy across most defect categories with exceptional inference efficiency, making it highly suitable for industrial quality inspection in photovoltaic manufacturing.

Based on the visualization results and quantitative data in Fig. 8, our model demonstrates superior localization accuracy for irregular defects such as SD and PT compared to the baseline. This validates the effectiveness of the multi-scale feature fusion and dynamic convolution modules in handling complex geometries. For defects with weak features like Cr and IS, the model's performance remains competitive yet highlights an inherent challenge: distinguishing subtle targets from complex backgrounds. This observation directly points to a direction for future work: specifically, enhancing the model's sensitivity to faint defect signatures and its robustness against background interference, while fully preserving the existing architecture's strengths in multi-scale and geometric adaptation.

Comparison of NEU-DET, GC10-DET assays

Analysis based on Table 5 and Table 6 demonstrates the leading performance of MSEOD-DDFusionNet on both general-purpose industrial datasets. On NEU-DET, the model achieves an overall mAP_{50} of 80.5%, with notable gains on challenging irregular defects: cracks (Cr, 58.8%) and inclusions (In, 86.6%). The mAP_{50} for Cr shows a 43.4% relative improvement over the best baseline (FFDDNet), while In detection improves by 20.0% over YOLOv5s. On GC10-DET, the model maintains a leading overall mAP_{50} of 67.4%, excelling on complex-shaped defects such as welds (Wf, 92.2%) and oxide scale (Os, 81.1%). Specifically, Wf detection achieves a 22.6% relative improvement over YOLOv5s, and Os detection improves by 4.8%. These results validate the strong generalization of our multi-scale fusion and dynamic convolution modules for irregular defects, while slightly lower performance on regular or slender defect points to future refinement of the dynamic receptive field strategy.

The comparison in Table 7 (comprehensive performance metrics) shows that our model achieves an excellent balance between efficiency and accuracy. On NEU-DET, the model reaches 80.5% mAP_{50} and 46.9% mAP_{50-95} while operating at 156.1 FPS with a computational cost of 159 Gflops. This represents a 7.3% improvement in mAP_{50} over YOLOv11s (75.0%). On GC10-DET, the model maintains a leading mAP_{50} of 67.4% and mAP_{50-95} of 34.3% at 192.2 FPS under the same computational budget, corresponding to an 11.6%

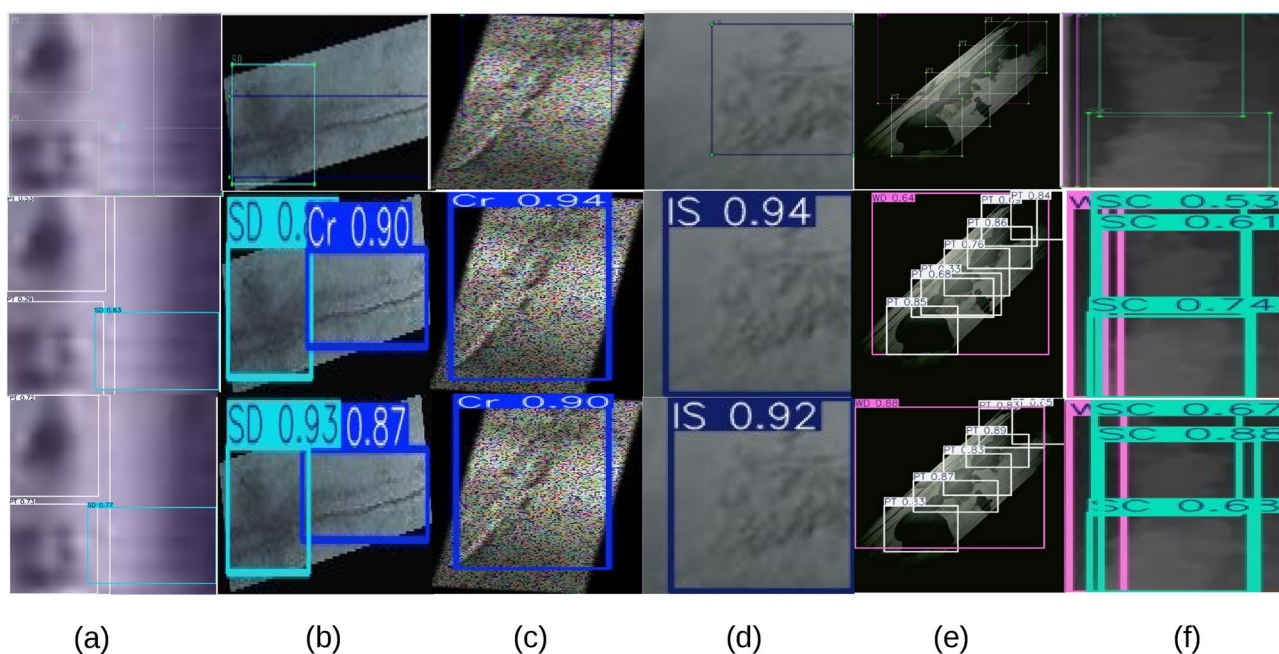


Fig. 8. Visual detection comparison. In vertically arranged examples, MSEOD-DDFusionNet (bottom) produces tighter and more accurate bounding boxes than YOLOv11s (top), particularly for irregular defects like SD and PT.

Datasets	Method	Cr	In	Pa	Ps	Rs	Sc	mAP ₅₀
NEU-DET	YOLOv5s	39.8	72.2	96.6	87.7	71.0	94	76.9
	YOLOv8s	30.3	82.9	94.3	85.4	61.0	90.8	75.2
	YOLOv9s	39	76.2	95.6	90.2	68.8	95.8	77.6
	YOLOv10s	32	70.9	94.4	77.6	68	81.4	70.7
	YOLOv11s	39.5	72.1	95	88	66.4	90.8	75
	FFDDNet ¹⁵	41	75.2	94.3	87.1	71.5	92.0	76.8
	Ours	58.8	86.6	93.6	93.8	68.5	92.7	80.5

Table 5. Per-class performance on NEU-DET (vs. best baseline: +3.7% mAP₅₀; Cr +43.4%, In +4.46%, Ps +3.99%).

Datasets	Method	Pu	Wl	Cg	Ws	Os	Ss	In	Rp	Cr	Wf	mAP ₅₀
GC10-DET	YOLOv5s	91.1	83.8	92.4	70.8	77.4	62.9	34.4	0	28.5	75.2	61.7
	YOLOv8s	93.8	79.2	93.5	73.2	74.1	56.4	27.8	4.5	20	74.9	61.3
	YOLOv9s	92.6	83.1	94.3	76.3	75.7	63.8	28	4.9	31.7	78	62.9
	YOLOv10s	92.7	75.1	85.1	62.1	70.8	53.6	30.2	0	18	66.2	55.4
	YOLOv11s	94.3	74.1	94.8	68.1	73.3	53.7	40.7	0	31.6	73.2	60.4
	FFDDNet ¹⁵	95.5	85.3	90.2	71.5	62.8	54.4	23.5	24	42.7	92.1	64.2
	Ours	95.3	93.8	92.8	74.1	81.1	53.7	32.3	25	41.4	92.2	67.4

Table 6. Per-class performance on GC10-DET (vs. best baseline: +5.0% mAP₅₀; Os +4.8%, Wl +10.0%, Wf +0.1%).

improvement in mAP₅₀ over YOLOv11s (60.4%). These results confirm that the lightweight design effectively supports the model’s potential for real-time industrial deployment.

Classification diagnostics using the confusion matrices in Fig. 9 (NEU-DET) and Fig. 10 (GC10-DET) reveal that our model exhibits higher diagonal purity and more concentrated distributions on both datasets. It significantly reduces misclassification rates for difficult defects such as Cr and Ps on NEU-DET, and achieves effective recognition of challenging categories, including Wf, Cg, and the rarely detected Rp on GC10-DET. This demonstrates the model’s superior feature discriminability in complex multi-class scenarios.

The F1-confidence curves in Fig. 11 (NEU-DET) and Fig. 12 (GC10-DET) further verify the prediction reliability of the model. On NEU-DET, the model attains a higher overall F1-score (0.75) and maintains high performance over a wider confidence range. On GC10-DET, it reaches a higher performance ceiling (F1-score 0.71 vs. 0.60 for YOLOv11s) and shows more stable results at high confidence thresholds. These findings indicate that the model’s output confidence aligns better with actual detection accuracy, which helps reduce false alarms in practical deployment.

Pixel-level and instance-level visual evidence is provided by the heatmaps in Fig. 13 and the detection visualizations in Fig. 14. On NEU-DET, the model produces more focused activation on Cr defects while suppressing background interference more effectively. On GC10-DET, it delivers more complete localization and higher confidence for defects such as Wf. The visual results are consistent with the quantitative analyses above, offering intuitive validation of the synergistic advantages of the model in feature enhancement, geometric adaptation, and small-target detection. For slender defects, the localization completeness remains to be improved, which aligns with the conclusions drawn from the tabular data.

In summary, MSEOD-DDFusionNet exhibits outstanding generalization performance, efficient inference capability, and reliable prediction quality on both NEU-DET and GC10-DET datasets. The experiments thoroughly validate the effectiveness and practicality of the core modules for general industrial defect detection, while also identifying clear directions for future improvement, particularly in the detection of slender targets.

Comparison of generalisation experiments

On PASCAL VOC 2007 (Table 8), the model achieves a leading overall mAP₅₀ of 75.5%, outperforming all compared methods. It excels in key categories such as aeroplane (86.9% mAP₅₀) and bicycle (90.8% mAP₅₀), showing 6.9% and 2.4% relative improvements over YOLOv11s, respectively. For the challenging low-contrast category “pottedplant”, the model attains 46.7% mAP₅₀, a 11.7% relative improvement over YOLOv11s. These gains highlight the model’s ability to integrate multi-scale features and adapt to diverse object geometries across domains.

In medical imaging validation on BCCD (Table 9), the model achieves 93.8% overall mAP₅₀, the highest among all compared methods. It detects platelets with 93.7% accuracy, a 6.0% relative improvement over YOLOv11s, and red blood cells with 88.1% mAP₅₀. The model effectively suppresses background interference and adapts to cellular shape variations, demonstrating strong applicability in medical image analysis.

Table 10 summarizes the efficiency and accuracy balance across domains. On PASCAL VOC 2007, the model operates at 265.26 FPS with 159.0 Gflops, 178% faster than YOLOv11s while maintaining a 0.3% higher mAP₅₀.

Datasets	Methods	Params	Gflops	FPS	mAP ₅₀	mAP _{50–95}
NEU-DET	Faster R-CNN ¹⁶	41.364	90.913	1.0	65.4	34.8
	YOLOv5s	9.13	192.67	13.42	76.9	43.4
	YOLOv7	43.5	130.2	3.797	70.5	40.3
	YOLOv8s	11.14	229.5	52	75.2	43
	YOLOv9s	7.29	219.17	16.26	77.6	44.5
	YOLOv10s	8.07	198.33	42.36	70.7	40.8
	YOLOv11s	10.73	213.06	62.02	75	43.1
	DAB-DETR ¹⁷	44	216	18	68.6	33.9
	Deformable-DETR ¹⁸	40	173	18	64.3	-
	DINO ¹⁹	218	-	14	55.2	43
	LFF-YOLO ²⁰	60.51	6.85	42	73.1	-
	FFDDNet ¹⁵	10.08	222.51	38	76.8	44.4
	LE-YOLOv5 ²¹	4.8	10.3	55.1	79.1	41.0
	Ours	8.46	159	156.1	80.5	46.9
GC10-DET	Faster R-CNN ¹⁶	41.364	90.913	27	55.4	/
	YOLOv5s	9.13	192.67	60.35	61.7	32
	YOLOv7	43.5	130.2	35	58.0	/
	YOLOv8s	11.14	229.5	48	61.3	31.4
	YOLOv9s	7.29	219.17	45	62.9	32.1
	YOLOv10s	8.07	198.33	88.54	55.4	28.8
	YOLOv11s	10.73	213.06	123.23	60.4	29.7
	DAB-DETR ¹⁷	44	216	18	52.6	-
	Deformable-DETR ¹⁸	40	173	16	54.3	-
	DINO ¹⁹	218	-	13	54.3	-
	LFF-YOLO ²⁰	60.51	6.85	40	41.3	-
	FFDDNet ¹⁵	10.08	222.51	35	64.2	32.8
	LE-YOLOv5 ²¹	4.8	10.3	-	63.8	28.9
	Ours	8.46	159	192.2	67.4	34.3

Table 7. Comprehensive efficiency-accuracy trade-off (vs. YOLOv11s: NEU-DET +7.3% mAP₅₀, +151% FPS; GC10-DET +11.6% mAP₅₀, +56% FPS).

On BCCD, it achieves 93.8% mAP₅₀ at 82.19 FPS, 40% faster than YOLOv11s with a 3.0% accuracy gain. This confirms the lightweight design's effectiveness in real-time cross-domain deployment.

Visual results in Figs. 15 and 16 further validate the model's performance. On PASCAL VOC 2007, it produces tighter bounding boxes and detects smaller, distant objects missed by baseline methods. On BCCD, it reduces false positives and improves separation between adjacent cells, especially in dense clusters. These visualizations confirm the advantages of the multi-module design for tasks beyond industrial inspection.

In conclusion, MSEOD-DDFusionNet proves effective not only for industrial defect detection but also for general object recognition and medical image analysis. It provides a reliable, high-speed solution for high-precision detection across diverse domains, demonstrating strong cross-domain generalization capability.

Comparison of pruning experiments

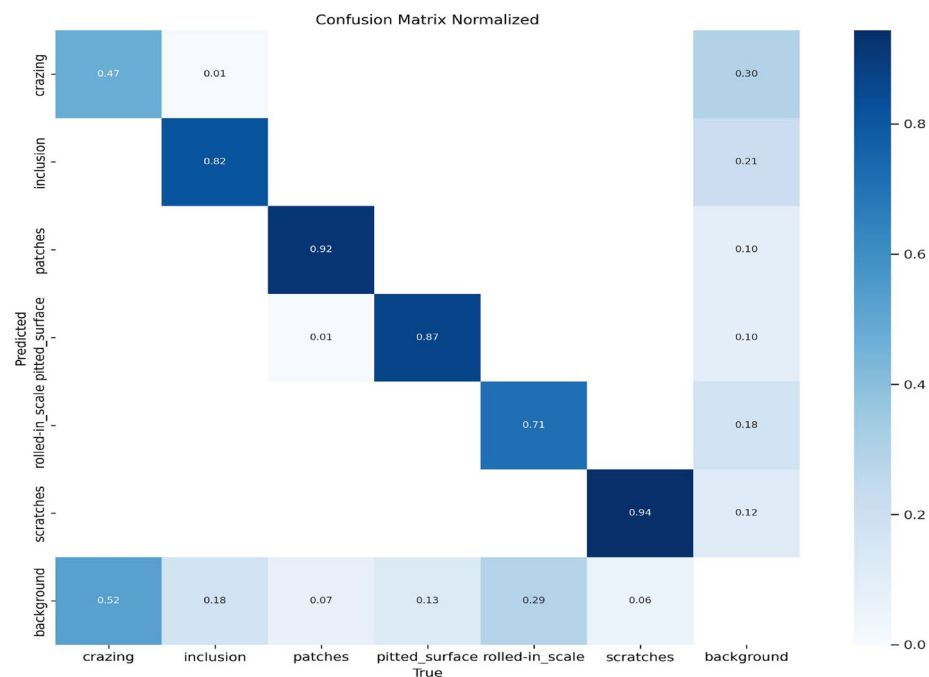
This study implements a structured pruning scheme to enhance model deployment efficiency. The method evaluates channel importance using the absolute values of BatchNorm layer weights, with a global pruning threshold set at the 75th percentile. The pruning process follows a sequential strategy, automatically adjusting connected layers to maintain dimensional consistency.

To preserve model capability, each convolutional layer retains at least 8 channels through an iterative threshold adjustment mechanism. If the channel count falls below this minimum, the threshold is automatically halved until the requirement is met.

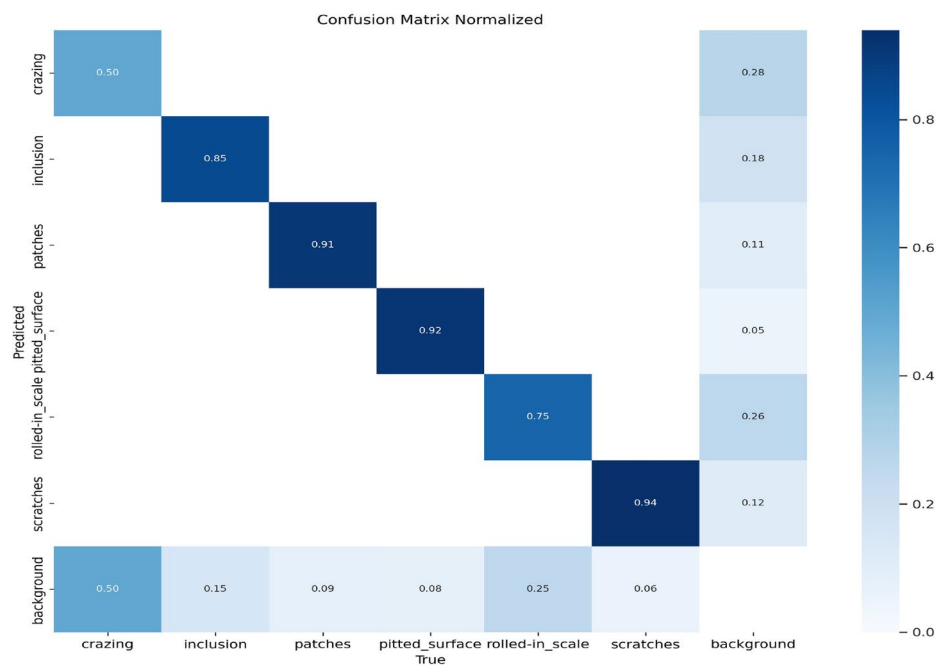
After pruning, the model is fully retrained with the original parameters: an initial learning rate of 0.01 decaying cosinely to 0.0001, a momentum of 0.937, and weight decay of 0.0005. Training runs for 100 epochs with a batch size of 16, including a 3-epoch warm-up. Data augmentation is progressively reduced and disabled in the final 10 epochs.

Table 11 demonstrates that pruning consistently improves both computational efficiency and detection accuracy. A key clarification is required: our pruned model retains 8.46M parameters—identical to the complete MSEOD-DDFusionNet. The −21.2% parameter reduction reported in the table is calculated relative to YOLOv11s (10.73M), underscoring the inherent parameter efficiency of our architecture.

On the DDTE dataset, the pruned model achieved a 1.9% improvement in mAP₅₀ and a 4.3% improvement in mAP_{50–95} compared to YOLOv11s, while reducing computational effort by 9.2%. This efficiency extends



(a) The YOLOv11s confusion matrix diagram

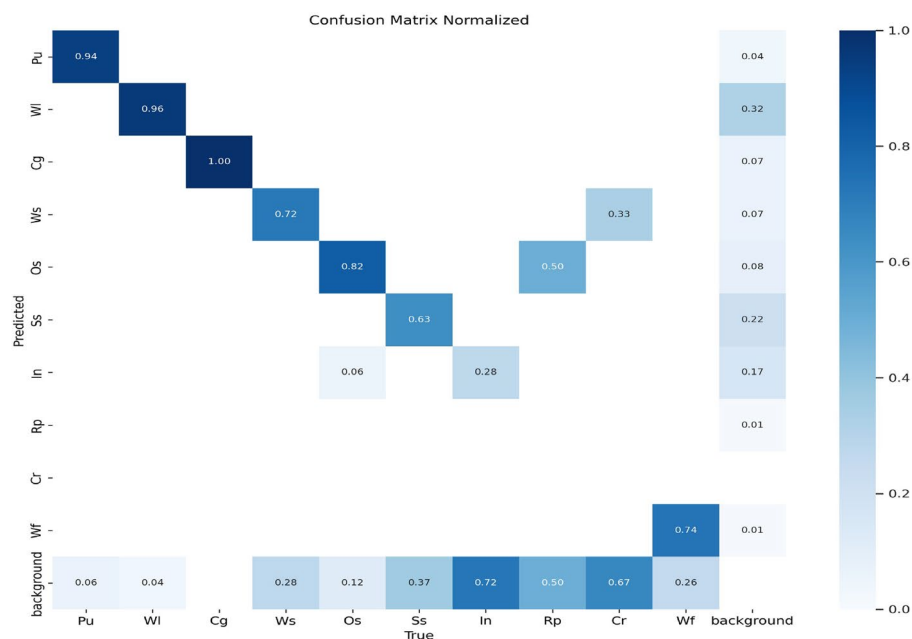


(b) The MSEOD-DDFusionNet confusion matrix diagram

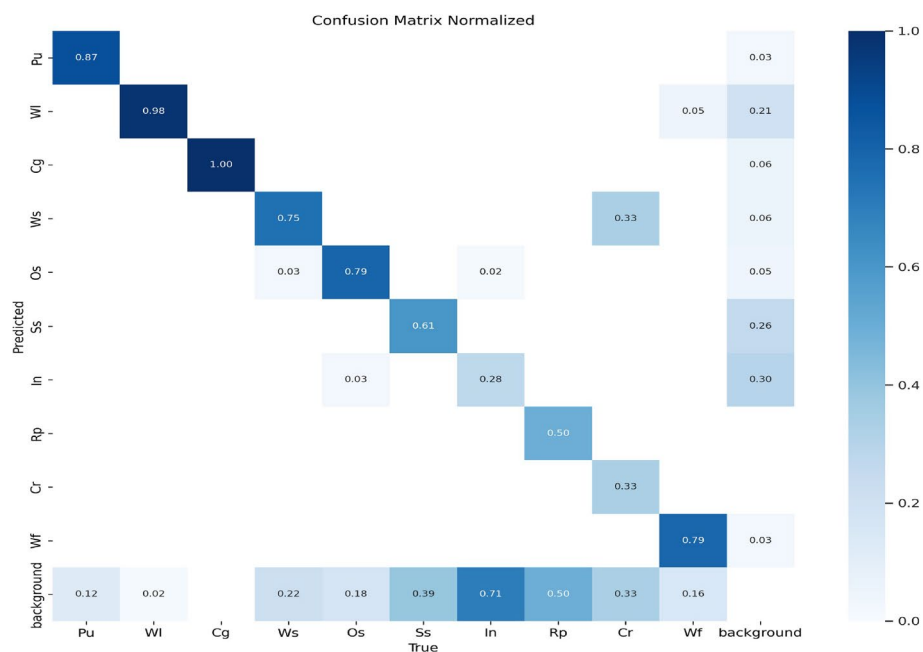
Fig. 9. Confusion matrix comparison on NEU-DET. MSEOD-DDFusionNet (b) demonstrates superior classification over YOLOv11s (a), with higher diagonal values and reduced misclassification for difficult categories such as Cr and Ps.

to other benchmarks: on the NEU-DET dataset, the model achieves a 3.5% improvement in mAP_{50} while reducing computational cost by 25.6%; on the GC10-DET dataset, mAP_{50} increases by 11.1% with a 25.4% reduction in computational cost. Cross-domain robustness is maintained, with performance comparable to the baseline on the PASCAL VOC 2007 dataset and a 2.8% improvement in mAP_{50} on the BCCD dataset.

The pruning scheme employs a clear criterion (75th percentile), a structured workflow, and a full retraining protocol. It preserves the original parameter count while substantially lowering computational cost, offering a



(a) The YOLOv11s confusion matrix diagram



(b) The MSEOD-DDFusionNet confusion matrix diagram

Fig. 10. Confusion matrices on GC10-DET. MSEOD-DDFusionNet (b) outperforms YOLOv11s (a) with higher diagonal purity, better recognition of Wf and Cg, and effective detection of the challenging defect Rp.

practical solution for deployment. The consistent gains across datasets suggest that pruning serves as an effective regularizer, sharpening the model's focus on salient features.

Conclusion

This study proposes the MSEOD-DDFusionNet framework to address key challenges in steel surface defect detection: scale variance, irregular geometry, and low-contrast noise. The core contributions are threefold: (1) a lossless multi-scale fusion principle to preserve micro-defect features, (2) a four-dimensional dynamic

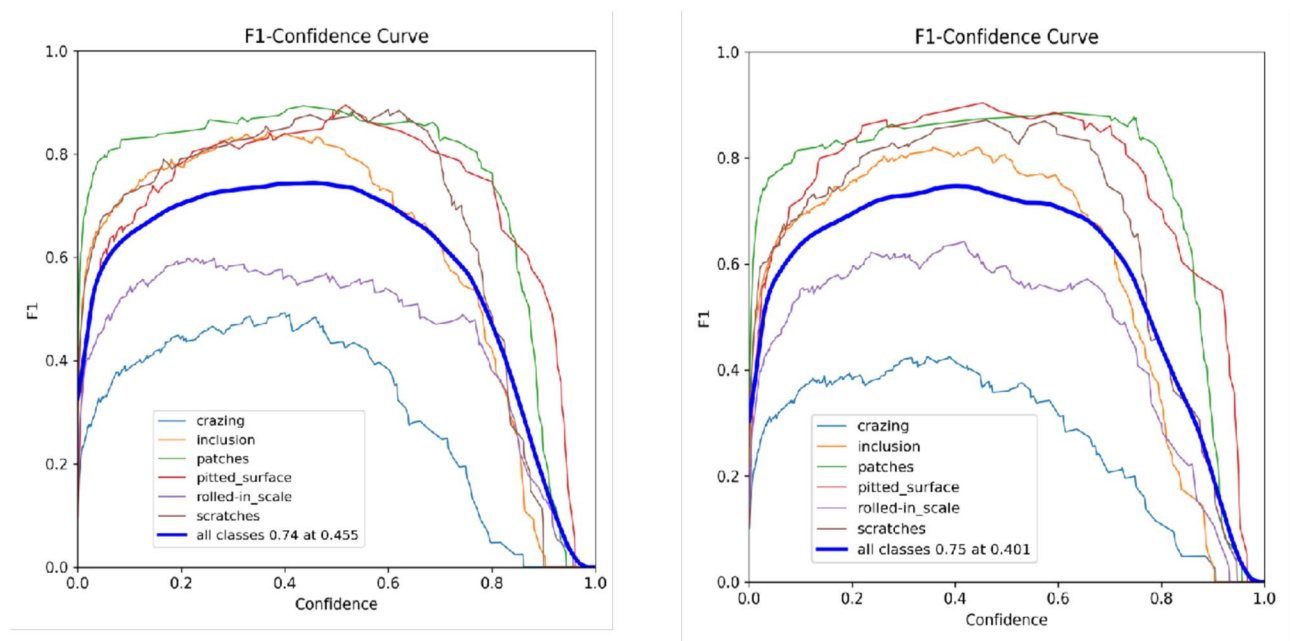


Fig. 11. F1-confidence curves on NEU-DET. MSEOD-DDFusionNet outperforms YOLOv11s with a higher F1-score (0.75 vs. 0.74) and more consistent performance across confidence levels.

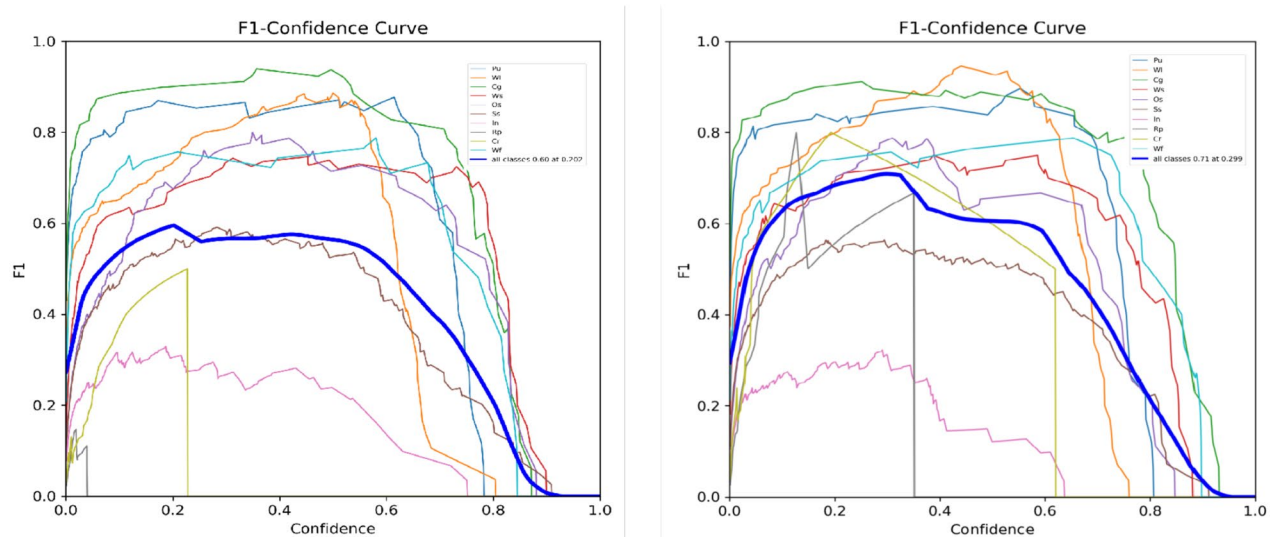


Fig. 12. F1-confidence curves on GC10-DET. MSEOD-DDFusionNet achieves a higher F1-score than YOLOv11s (0.71 vs. 0.60) and maintains better performance across confidence thresholds.

convolution mechanism to adapt to irregular shapes, and (3) a decoupled noise robustness strategy to balance sensitivity and robustness. Experiments demonstrate state-of-the-art performance on the specialized DDTE dataset (82.6% mAP_{50} , 61.6% mAP_{50-95}) and excellent generalization across public benchmarks and cross-domain tasks. While this modular design is a necessary and efficient solution to these compound challenges, we acknowledge the inherent architectural complexity. The current structure, justified by its performance, points to a future research direction: exploring more unified or automatically searched architectures that maintain high accuracy with reduced complexity. Furthermore, the model's limitations in detecting sparse defects (e.g., Cr) suggest another direction: enhancing sensitivity to subtle features and robustness against sparse interference. Together, these efforts aim to advance the efficacy and practicality of industrial visual inspection systems.

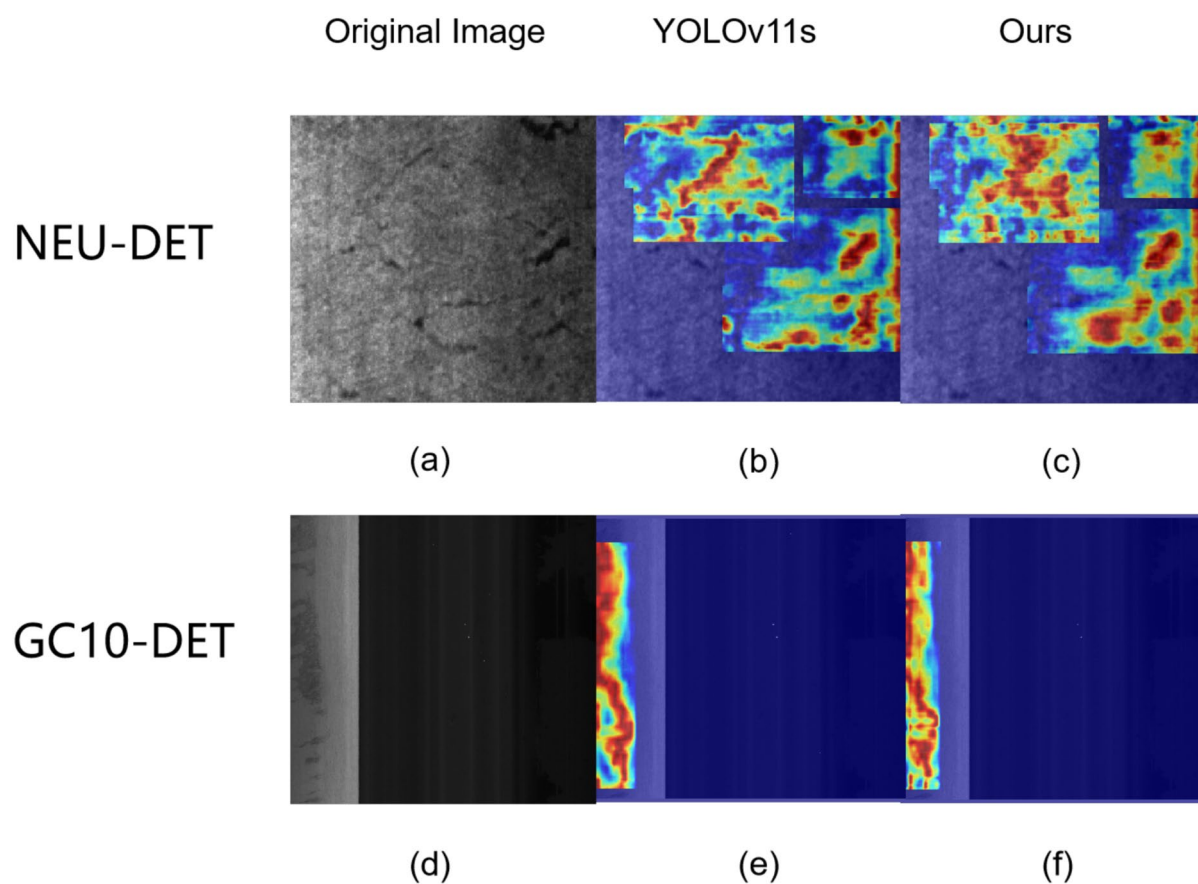


Fig. 13. Heatmap comparison on NEU-DET and GC10-DET. MSEOD-DDFusionNet shows more focused activation on Cr defects (NEU-DET) and more complete coverage on Wf defects (GC10-DET) compared to YOLOv11s.

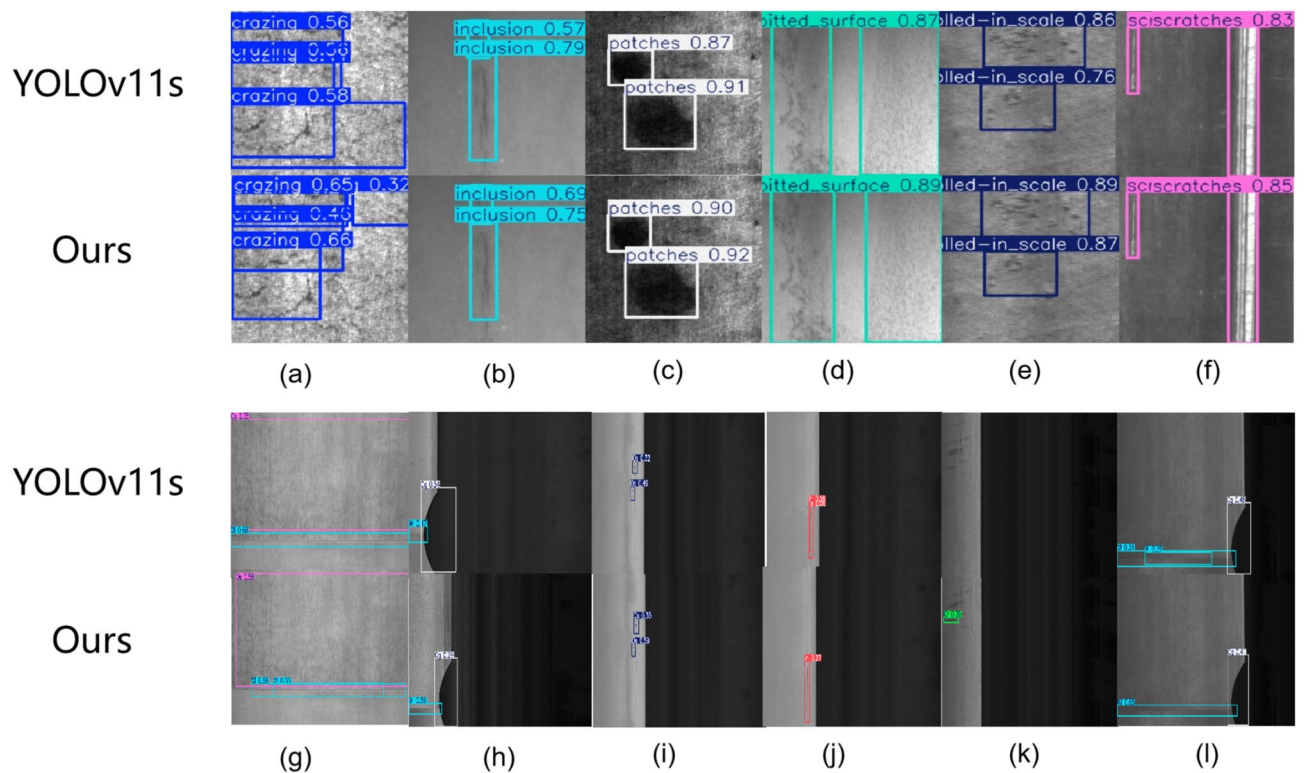


Fig. 14. Qualitative detection comparison. Visual results illustrate our model's improved localization for irregular defects (Cr, Wf) and highlight performance variations on regular/slender defects (Pa, Cg, In).

	Classes	YOLOv5s	YOLOv7t	YOLOv8s	YOLOv10s	YOLOv11s	FFDDNet ¹⁵	Ours
PASCAL VOC 2007	aeroplane	82.1	70.2	79.5	80	81.3	83.8	86.9
	bicycle	89	79.4	80.9	86.3	88.7	89.8	90.8
	bird	62.3	52.0	47.5	57.3	58.7	63.6	65.2
	boat	76.2	49.9	58.3	74.3	79	64.6	75.2
	bottle	63	40.3	50.4	63.3	66.3	70.8	60.5
	bus	78.9	70.8	73.7	81.1	78.1	80.2	77.1
	car	89.4	84.8	85.2	88.3	89	88.7	90.1
	cat	83.6	65.1	74.7	86	82.2	79.1	84.8
	chair	58.5	48.4	44.8	57	56.9	57.4	54
	cow	73.9	66.1	57	78.7	78.2	75.3	78.4
	diningtable	74.5	48.2	60.3	68.4	75.9	71.9	69.7
	dog	74.3	61.9	59.9	73.7	76.4	74.8	79.9
	horse	84.9	71.5	79	87.7	87.5	84.8	85
	motorbike	91.6	76.8	87.5	89.9	93.4	89.6	93
	person	87.1	79.9	82.6	85.7	88.1	86.7	87.6
	pottedplant	37.6	41.7	27.5	43.2	41.8	46.0	46.7
	sheep	57.6	63.2	45.3	51.5	59.7	60.7	56.1
	sofa	64.9	60.0	47.6	65.6	68.8	69.4	64.6
	train	86.1	77.6	79.5	85.3	87.1	88.1	89
	tvmonitor	72.7	62.1	57.3	67.3	66.7	65.6	75.0
	mAP ₅₀	74.4	63.6	63.9	73.5	75.2	74.5	75.5

Table 8. Performance on PASCAL VOC 2007 (vs. YOLOv11s: +0.4% mAP₅₀, pottedplant +11.7%, tvmonitor +12.4%).

	Methods	WBC	RBC	Platelets	mAP ₅₀
BCCD	YOLOv5s	99.5	82.6	83.7	88.6
	YOLOv7t	97.6	87.8	92.6	92.7
	YOLOv8s	98.3	83.6	93.4	91.8
	YOLOv10s	98	85.7	83.5	84.9
	YOLOv11s	99.4	84.6	88.4	90.8
	FFDDNet ¹⁵	99.4	83.9	87.7	90.3
	Ours	99.5	88.1	93.7	93.8

Table 9. Performance on BCCD medical dataset (vs. best baseline: +1.2% mAP₅₀, Platelets +1.2%).

Datasets	Methods	Params	Gflops	FPS	mAP ₅₀	mAP _{50–95}
PASCAL VOC 2007	Faster R-CNN ¹⁶	41.4	71.7	-	71	39.8
	YOLOv5s	9.13	192.67	98.21	74.4	51
	YOLOv7t	6.02	13.1	-	68.5	42.8
	YOLOv8s	11.14	229.5	71.79	63.9	42.2
	YOLOv10s	8.07	198.33	105.95	73.5	53.8
	YOLOv11s	10.73	213.06	95.23	75.2	55
	FFDDNet ¹⁵	10.08	222.51	88	70.1	49.0
	LE-YOLOv5 ²¹	4.8	10.3	-	73.8	54.1
	LF-YOLO ²²	7.25	16.2	-	53.9	27.8
	RT-DETR-R18 ²³	19.8	57.0	-	59.3	40.7
	Ours	8.46	159.0	265.26	75.5	54.9
BCCD	YOLOv8s	11.14	229.21	33.47	91.8	61.6
	YOLOv10s	8.07	198.33	99.13	84.9	58.3
	YOLOv11s	10.73	213	58.75	90.8	62
	FFDDNet ¹⁵	10.08	222.51	54.9	90.3	60.1
	Ours	8.46	159.0	82.19	93.8	64.6

Table 10. Cross-domain efficiency-accuracy balance vs. mainstream methods (vs. YOLOv11s: PASCAL VOC +0.4% mAP₅₀, +178% FPS; BCCD +3.3% mAP₅₀, +40% FPS).



Fig. 15. Detection examples on PASCAL VOC 2007. MSEOD-DDFusionNet produces tighter and more complete bounding boxes than YOLOv11s, with improved detection of smaller and more distant objects.

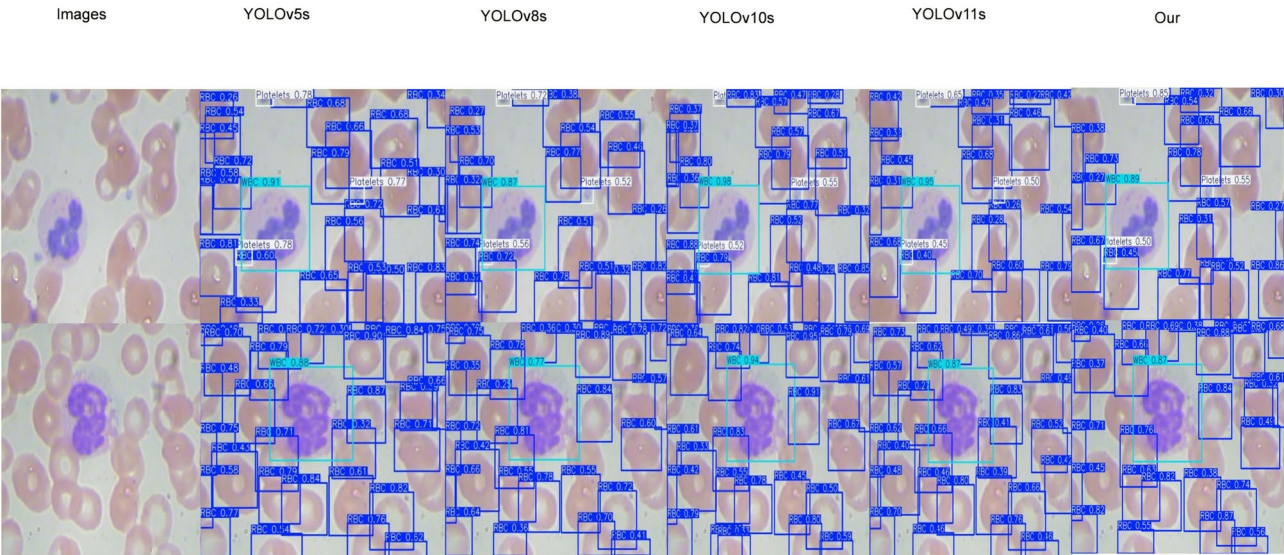


Fig. 16. Detection examples on BCCD. Paired visual examples show our model’s lower false-positive rate, improved cell separation, and more accurate platelet localization versus YOLOv11s.

Datasets	Methods	params	Gflops	mAP ₅₀	mAP _{50–95}
DDTE	YOLOv11s	10.73	213.06	80.4	58.2
	Prune(Ours)	8.46 (−21.2%)	193.51(−9.2%)	81.9 (+1.9%)	60.7 (+4.3%)
NEU-DET	YOLOv11s	10.73	213.06	75	43.1
	prune(Ours)	8.46(−21.2%)	158.5(−25.6%)	77.6(+3.5%)	45.0 (+4.4%)
GC10-DET	YOLOv11s	10.73	213.06	60.4	29.7
	prune(Ours)	8.46(−21.2%)	158.94(−25.4%)	67.1(+11.1%)	33.1(+11.4%)
PASCAL VOC 2007	YOLOv11s	10.73	213.06	75.2	55
	prune(Ours)	8.46(−21.2%)	198.46(−6.9%)	75.4(+0.3%)	54.7(−0.5%)
BCCD	YOLOv11s	10.73	213	90.8	62
	prune(Ours)	8.46(−21.2%)	158.94(−25.4%)	93.3(+2.8%)	63.6(+2.6%)

Table 11. Computational efficiency vs. accuracy: pruned model vs. YOLOv11s (DDTE: −9.2%, +1.9% mAP₅₀; GC10-DET: −25.4%, +11.1% mAP₅₀).

Data availability

The specific industrial dataset is subject to privacy restrictions and is not publicly available. To ensure reproducibility and enable further application, we provide the complete implementation, including code and pre-trained models, at: [https://github.com/jiunian158/DDTE_Anonymous_Subset/blob/main/DDTE_Anonymous_Subset%20\(3\).zip](https://github.com/jiunian158/DDTE_Anonymous_Subset/blob/main/DDTE_Anonymous_Subset%20(3).zip)

Received: 30 July 2025; Accepted: 8 January 2026
Published online: 16 January 2026

References

1. Wang, Z., Zhou, W. & Li, Y. MSAF-YOLO: An Efficient Multi-Scale Attention Fusion Network for high-precision steel surface defect detection. Measurement 118640, <https://doi.org/10.1016/j.measurement.2025.118640> (2025).
2. Liu, R. et al. A lightweight model based on multi-scale feature fusion for ultrasonic welding surface defect detection. Eng. Appl. Artif. Intell. 161, 112208. <https://doi.org/10.1016/j.engappai.2025.112208> (2025).
3. Wang, S. et al. Research on steel surface defect detection system based on YOLOv5s-SE-CA model and BEMD image enhancement. Nondestruct. Test. Eval. 1–20, <https://doi.org/10.1080/10589759.2024.2393205> (2024).
4. Pan, Y. & Zhang, L. Dual attention deep learning network for automatic steel surface defect segmentation.. Comput.-Aided Civ. Inf. 37, 1468–1487. <https://doi.org/10.1111/mice.12792> (2022).
5. Hosseini, S. M., Ebrahimi, A., Mosavi, M. R. & Shahhoseini, HSh. A novel hybrid CNN-CBAM-GRU method for intrusion detection in modern networks. Results Eng. 28, 107103. <https://doi.org/10.1016/j.rineng.2025.107103> (2025).
6. Yeung, C.-C. & Lam, K.-M. Efficient fused-attention model for steel surface defect detection. IEEE Trans. Instrum. Meas. 71, 1–11. <https://doi.org/10.1109/TIM.2022.3176239> (2022).
7. Ziadlou, G., Emami, S. & Asadi-Gangraj, E. Network configuration distributed production scheduling problem: A constraint programming approach. Comput. Ind. Eng. 188, 109916. <https://doi.org/10.1016/j.cie.2024.109916> (2024).

8. Wen, J., Zheng, Y., Zhang, Y. & Yu, W. Enhanced dual-channel feature fusion approach for rolling bearing fault diagnosis. *Nondestruct. Test. Eval.* **40**, 3309–3337. <https://doi.org/10.1080/10589759.2025.2507761> (2025).
9. Ye, Z. & Yu, J. AKSNet: A novel convolutional neural network with adaptive kernel width and sparse regularization for machinery fault diagnosis. *J. Manuf. Syst.* **59**, 467–480. <https://doi.org/10.1016/j.jmsy.2021.03.022> (2021).
10. Xu, X., Li, X., Ming, W. & Chen, M. A novel multi-scale CNN and attention mechanism method with multi-sensor signal for remaining useful life prediction. *Comput. Ind. Eng.* **169**, 108204. <https://doi.org/10.1016/j.cie.2022.108204> (2022).
11. Zhou, H. et al. Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction. *Neurocomputing* **445**, 298–308. <https://doi.org/10.1016/j.neucom.2021.03.024> (2021).
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 10684–10695. <https://doi.org/10.48550/arXiv.2112.10752> (2022).
13. Zhai, C., Wang, L. & Yuan, J. New Fusion Network with Dual-Branch Encoder and Triple-Branch Decoder for Remote Sensing Image Change Detection. *Appl. Sci.* **13**, 6167. <https://doi.org/10.3390/app13106167> (2023).
14. Fang, T. et al. Human-Guided Data Augmentation via Diffusion Model for Surface Defect Recognition Under Limited Data. *IEEE Trans. Instrum. Meas.* **74**, 1–16. <https://doi.org/10.1109/TIM.2025.3541684> (2025).
15. Lin, Z., Li, Z., Yu, J., Hu, M. & Wang, X. FFDDNet: Flexible focused defect detection network. *IEEE Trans. Instrum. Meas.* **74**, 5019812. <https://doi.org/10.1109/TIM.2025.3551459> (2025).
16. Wang, D., Shang, K., Wu, H. & Wang, C. Decoupled R-CNN: Sensitivity-specific detector for higher accurate localization. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 6324–6336. <https://doi.org/10.1109/TCSVT.2022.3167114> (2022).
17. Fang, R. et al. FeatAug-DETR: Enriching one-to-many matching for DETRs with feature augmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 6402–6415. <https://doi.org/10.1109/TPAMI.2024.3381961> (2024).
18. Yang, K. et al. DETA: A point-based tracker with deformable transformer and task-aligned learning. *IEEE Trans. Multimed.* **25**, 7545–7558. <https://doi.org/10.1109/TMM.2022.3223213> (2023).
19. Zhang, M. et al. Oriented-DINO: Angle decoupling prediction and consistency optimizing for oriented detection transformer. *IEEE Trans. Geosci. Remote Sens.* **62**, 5638315. <https://doi.org/10.1109/TGRS.2024.3450200> (2024).
20. Li, Q., Shao, Y., Li, L., Li, J. & Hao, H. Weak surface defect detection for production-line plastic bottles with multi-view imaging system and LFF YOLO. *Opt. Laser Eng.* **181**, 108369. <https://doi.org/10.1016/j.optlaseng.2024.108369> (2024).
21. Zhu, C., Sun, Y., Zhang, H., Yuan, S. & Zhang, H. LE-YOLOv5: A lightweight and efficient neural network for steel surface defect detection. *IEEE Access* **12**, 195242–195255. <https://doi.org/10.1109/ACCESS.2024.3519161> (2024).
22. Liu, M., Chen, Y., Xie, J., He, L. & Zhang, Y. LF-YOLO: A lighter and faster yolo for weld defect detection of X-ray image. *IEEE Sens. J.* **23**, 7430–7439. <https://doi.org/10.1109/JSEN.2023.3247006> (2023).
23. Sun, J., Shen, X., Huang, H., Wang, Q. & Zhao, H. RT-SLAM: A Real-Time Visual SLAM System Integrating Enhanced RT-DETR and Optical Flow Techniques. *IEEE Internet Things J.* **12**, 12803–12814. <https://doi.org/10.1109/JIOT.2024.3522490> (2024).

Author contributions

Yiwei Duan contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Others. The first draft of the manuscript was written by Yiwei Duan; all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This research is funded by the Autonomous Region Science and Technology Plan Project “Research and Application of Artificial Intelligence Technology for Silicon-Based New Material Manufacturing” (Project No. 2023B01033).

Declarations

Competing interests

Each named author has substantially contributed to conducting the underlying research and drafting this manuscript. Additionally, to the best of our knowledge, the named authors have no conflict of interest, financial or otherwise.

Additional information

Correspondence and requests for materials should be addressed to J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026