CAM-DETR: An Efficient End-to-End Transformer for Real-Time Metal Surface Defect Detection

Zheyuan Lu, Chaoli Wang*, and Zhanquan Sun†
Institute of Optical-Electrical and Computer Engineering
University of Shanghai for Science and Technology
Shanghai, China
{13567661836@163.com, clwang@usst.edu.cn, sunzhq@usst.edu.cn}

***Abstract*—Real-time and precise metal surface defect detection is crucial for industrial quality control. Existing deep learning methods face challenges in practical applications: redundant backbones limit edge inference efficiency; attention mechanisms suffer from high computational overhead and low robustness; and insufficient cross-layer interaction hinders multi-scale defect handling. To address these, we propose CAM-DETR, a lightweight end-to-end detector based on RT-DETR. It features three core modules: (1) CE-RGELAN, a lightweight backbone that reduces complexity while preserving fine-grained features; (2) Additive Token Mixer (ATM), an encoder attention mechanism replacing quadratic self-attention with linear-complexity operations to cut costs and boost noise robustness; (3) MANet, a multi-scale fusion network that enhances localization via parallel branches and depthwise convolutions. Experiments show CAM-DETR reduces parameters by 10.97% and computation by 15.26%. It achieves 79.17% mAP@0.5 on NEU-DET (+2.29%, +57.0% speed) and 74.52% on GC10-DET (+5.59%, +89.3% speed), demonstrating an excellent accuracy-latency balance for industrial deployment.**

***Index Terms*—Object detection, real-time detection, end-to-end architecture, attention mechanism, multi-scale feature fusion**

I. Introduction

This section presents the industrial background and technical challenges of metal surface defect detection, establishing the motivation for developing efficient real-time detection algorithms. We analyze limitations of existing CNN and Transformer-based methods in handling complex industrial scenarios with high texture interference and multi-scale defects.

In the industrial production of metallic materials such as cold-rolled steel strips and aluminum foils, surface defect detection is a critical process for ensuring product integrity and production efficiency [1], [15]. Owing to harsh manufacturing environments and equipment precision limitations, various surface defects—including scratches, pits, inclusions, pockmarks, and oxide scale indentations—are prone to occur during rolling, coating, and stamping processes [2]. These defects not only degrade the mechanical properties and visual quality of products, but may also lead to failures in subsequent assembly, resulting in severe economic losses [19], [20]. With the rapid development of intelligent manufacturing, production lines are operating at increasingly higher speeds while product specifications are becoming more diverse. Consequently, achieving high-speed, stable, and accurate automated surface defect detection has become a core research direction for enhancing quality control in modern manufacturing systems [1], [2], [18].

In real industrial scenarios, metallic surface defects typically exhibit significant scale variations, complex background textures, low contrast, and high inter-class similarity [2], [11]. To address these challenges, deep learning–based object detection algorithms have been increasingly introduced into metallic surface defect detection and have demonstrated remarkable advantages in industrial applications. CNN-based detectors such as R-CNN [3], Fast R-CNN [4], and Faster R-CNN [5] achieve high detection accuracy but incur heavy computational overhead. One-stage detectors including SSD [10] and the YOLO series [8], [9], [12] provide faster inference through dense prediction. However, the inherent local receptive field of CNNs limits their ability to capture global geometric characteristics of large-scale defects, often resulting in high false alarm rates under periodic texture interference [1], [20].

Transformer-based end-to-end detection frameworks have been introduced to compensate for CNN limitations. DETR [12] first proposed direct interactions among queries via global self-attention and eliminated traditional non-maximum suppression through bipartite matching loss. RT-DETR [6] introduced an efficient Hybrid Encoder that combines intra-scale attention interaction with CNN-based cross-scale fusion, significantly improving real-time performance. Although these approaches exhibit notable advantages in global contextual modeling, existing Transformer-based methods generally face a trade-off between computational complexity and detection accuracy. Maintaining high-resolution feature maps causes the computational cost of self-attention to grow quadratically, making it difficult to meet real-time requirements of high-speed production lines [18].

Although aforementioned methods have made continuous progress, they still face several critical bottlenecks: (1) Backbone networks are often complex with redundant parameters, hindering real-time deployment on resource-constrained edge devices; (2) Multi-head self-attention mechanisms con-

This work was supported by the National Natural Science Foundation of China under Grant 62173232 and Grant 62003214. *†Corresponding authors: Chaoli Wang (clwang@usst.edu.cn) and Zhanquan Sun (sunzhq@usst.edu.cn).

sume immense computational resources when processing high-resolution features and exhibit redundancy when dealing with simple repetitive surfaces; (3) Cross-layer feature fusion mechanisms are insufficient, leading to inadequate interaction between shallow texture details and deep semantic information. To this end, based on the RT-DETR architecture, this paper proposes CAM-DETR, an efficient end-to-end detection model designed for high-speed production lines. The main contributions are summarized as follows:

- A lightweight backbone network, CE-RGELAN, is designed. By integrating controllable channel expansion and a reparameterized Ghost convolution mechanism, this network constructs low-redundancy feature evolution paths while significantly reducing parameters and computational complexity.
- The Additive Token Mixer (ATM) attention mechanism is introduced into the encoder. By utilizing additive operations of linear complexity and spatial-channel gating fusion units to replace traditional multi-head self-attention, the model filters high-frequency noise through a simplified interaction topology.
- A multi-scale feature aggregation network, MANet, is constructed. Through parallel branch processing and depthwise separable convolutions, the interaction and alignment between shallow texture information and deep semantic features are strengthened, effectively improving localization accuracy for defects with large scale spans.

We are pleased to make the code, datasets, and implementation details available upon acceptance, in compliance with applicable institutional and data-sharing agreements, to support reproducibility and further research.

II. Related Work

This section reviews metal surface defect detection from traditional methods to CNNs and Transformer-based models, highlighting RT-DETR as a strong industrial baseline and its limitations that motivate our improvements.

A. Traditional Defect Inspection Methods

Metal surface defect detection has evolved from conventional physical and vision-based methods—including magnetic particle testing, eddy current testing, ultrasonic testing, and optical inspection with handcrafted features [3], [11]—toward data-driven deep learning approaches. Traditional techniques suffer from limited adaptability to diverse defects, noise sensitivity, or high deployment costs in industrial settings [6], [7], [18].

B. CNN-Based Defect Detection

With the advent of deep learning, CNN-based detectors became dominant: two-stage frameworks (e.g., R-CNN series [3]–[5]) achieve high localization accuracy but incur substantial computational overhead, whereas one-stage detectors (e.g., SSD [10], YOLO family [8], [9], [14]) enable faster inference via dense prediction. To better capture metallic defect characteristics, EC-YOLO [18] introduced collaborative attention and channel shuffling [19] to enhance feature interaction in lightweight networks. Nevertheless, CNNs remain constrained by local receptive fields, limiting their capacity to model long-range dependencies between defects and structured backgrounds [1], [20].

C. Transformer-Based Detection Frameworks

Transformer-based architectures offer an alternative paradigm. DETR [12] reformulated detection as a set prediction task using global self-attention and bipartite matching loss, enabling end-to-end training without non-maximum suppression (NMS) [20]. RT-DETR [6] further improved efficiency through a Hybrid Encoder: its Attention-based Intra-scale Feature Interaction (AIFI) module performs global modeling on high-level semantics, while the CNN-based Cross-scale Feature Fusion (CCFF) module aggregates multi-scale features via 1×1 convolutions and reparameterized convolutions, effectively combining global perception with local spatial integration. The decoder employs an IoU-aware query selection mechanism followed by iterative refinement through multi-layer Transformers, directly predicting bounding boxes and categories without post-processing [12], [13]. However, on high-texture industrial datasets such as NEU-DET [16] and GC10-DET, the quadratic complexity of standard self-attention remains a computational bottleneck. Recent efforts have thus focused on lightweight attention mechanisms (e.g., additive attention, TSSA [21]) and dataset-specific optimizations (e.g., dual-label assignment in DATE [13], edge enhancement in MESC-DETR [2]), collectively advancing efficiency and defect-aware representation in Transformer detectors.

Given its favorable trade-off between end-to-end global modeling and real-time performance, RT-DETR has emerged as a promising baseline for industrial inspection. This work adopts the RT-DETR-R18 variant with ResNet-18 backbone [15], which maintains inference efficiency and cross-platform compatibility while constructing a multi-scale feature pyramid for complex detection tasks.

III. CAM-DETR Model

This section details the architectural innovations of CAM-DETR, focusing on three core components that address key bottlenecks in industrial defect detection. We present mathematical formulations and implementation strategies for each module to ensure both theoretical rigor and practical applicability.

Although RT-DETR has demonstrated favorable performance in general real-time object detection tasks, it still encounters difficulties in fully addressing challenges such as large defect scale spans, complex morphologies, and prominent local details in metal surface defect detection. To this end, while maintaining the overall end-to-end detection framework of RT-DETR, this paper proposes an improved model named CAM-DETR, whose architecture is shown in Fig. 1. The model undergoes targeted optimizations around three key

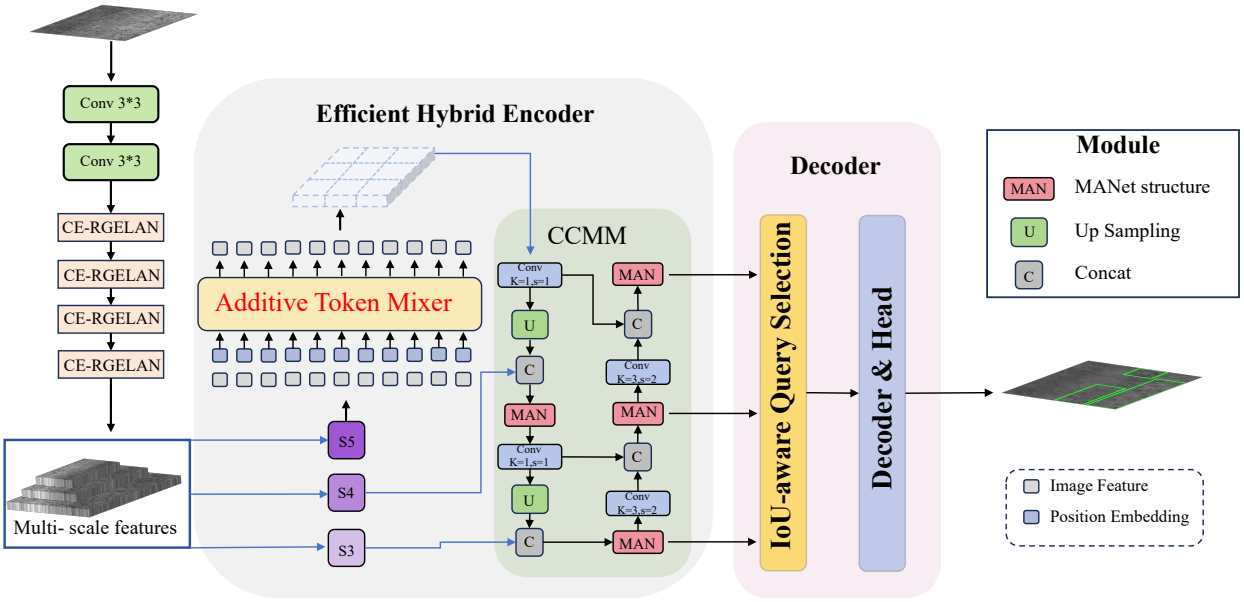

Fig. 1: Overall architecture of CAM-DETR:CE-RGELAN backbone extracts multi-scale features from metal surface images and feeds them into the Additive Token Mixer encoder to model global context with low computational cost. The MANet module further aggregates shallow texture details and deep semantic information, after which the RT-DETR decoder performs IoU-aware query selection and defect detection.

stages: feature extraction, feature interaction, and multi-scale fusion.

A. CE-RGELAN Lightweight Backbone Network

To balance feature extraction efficiency and representation depth under constrained computational budgets, we propose CE-RGELAN (Controllable Expansion Re-parameterized Ghost-ELAN), a lightweight backbone that constructs low-redundancy feature evolution paths by integrating residual re-parameterization with Ghost convolution. As illustrated in Fig. 2, given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, CE-RGELAN applies multi-branch parallel topology to yield output Y:

$$Y = \mathcal{F}_{\text{main}}(X) + \mathcal{F}_{\text{res}}(X), \tag{1}$$

where $\mathcal{F}_{\text{main}}(X)$ denotes the main branch stacking GhostConv and standard convolutions to generate high-dimensional features via intrinsic features and linear transformations, while $\mathcal{F}_{\text{res}}(X)$ represents the residual branch that preserves cross-stage feature interactions through concatenation and channel alignment. During inference, RepConv equivalently transforms the multi-branch structure into a single-path convolution, eliminating structural redundancy without sacrificing representational capacity.

To dynamically balance feature expressiveness and computational cost, CE-RGELAN introduces a controllable channel expansion mechanism with coefficient e:

$$C_{\text{mid}} = e \cdot C, \tag{2}$$

where C_{mid} is the intermediate channel dimension. Larger e values enhance semantic modeling capacity at the cost of latency, while smaller values prioritize inference speed. Based on ablation studies, we set $e = 0.5$ as the default configuration, achieving optimal trade-offs between discriminative power (79.17% mAP@0.5) and real-time performance (189.7 FPS) on NEU-DET.

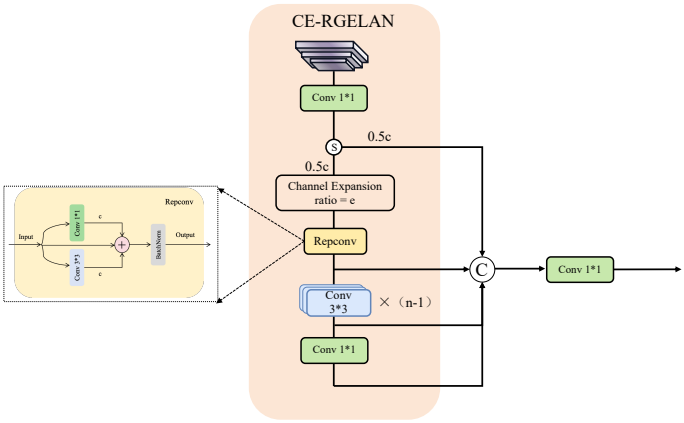

Fig. 2: Network architecture of CE-RGELAN module with controllable channel expansion mechanism.

B. Additive Token Mixer Attention Module

In the standard RT-DETR model, the Multi-Head Self-Attention (MHSA) mechanism in the AIFI module achieves cross-layer interaction via pairwise token correlations. However, its computational complexity of $\mathcal{O}(n^2 \cdot d)$ (where n is sequence length and d is feature dimension) causes significant latency on high-resolution feature maps, hindering real-time deployment.

To overcome this bottleneck, we introduce the Additive Token Mixer (ATM) mechanism [35] that achieves linear-complexity feature modeling. As illustrated in Fig. 3, ATM independently computes contextual representations through learned weight matrices W_q and W_k:

$$A = \text{Softmax}\big(W_v \cdot \sigma(W_q x_i + W_k x_j)\big), \tag{3}$$

where x_i and x_j denote feature vectors at query and key positions, respectively. All operations scale linearly with sequence length n, reducing time complexity from $\mathcal{O}(n^2 \cdot d)$ to $\mathcal{O}(n \cdot d)$ and eliminating the $n \times n$ attention matrix storage requirement.

Critically, the additive interaction pattern inherently suppresses high-frequency noise prevalent in metal surface textures. Unlike multiplicative attention that amplifies subtle background variations, ATM's linear topology naturally focuses on salient defect regions with higher contrast, enhancing robustness under complex illumination. To compensate for potential fine-grained detail loss, a Gated Fusion Unit applies dual spatial-channel recalibration: a spatial branch with 3×3 convolutions captures local context, while a channel branch with global average pooling extracts semantic information. Their gated fusion effectively suppresses texture interference while preserving defect-sensitive features.

Experimental results confirm that ATM integration significantly reduces memory consumption and boosts inference speed with minimal accuracy trade-off, providing discriminative features for subsequent multi-scale fusion.

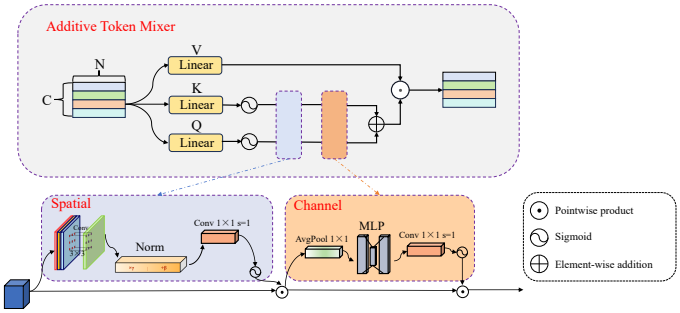

Fig. 3: Structure of the Additive Token Mixer (ATM) with additive interaction unit and gated fusion unit.

C. MANet Multi-Scale Feature Aggregation Network

Metal surface defects exhibit highly non-uniform scale distributions and low local contrast, challenging the RepC3 module's ability to balance receptive field coverage and feature precision under computational constraints. To address this, we utilize the Multi-Scale Feature Aggregation Network (MANet) [36] to reconstruct the feature fusion stage. MANet executes differentiated enhancement across three parallel branches: a top branch with cascaded ConvNeck strengthens high-level semantics; a middle branch with Depthwise Separable Convolutions (DSConv) captures local textures; and a bottom branch preserves spatial details through identity mapping.

For an input feature map of size $H \times W$ with C channels, standard convolution with $k \times k$ kernel incurs $\mathcal{O}(C \cdot C_{\text{out}} \cdot k^2 \cdot H \cdot W)$ complexity. MANet reduces this burden via channel-wise depthwise convolution followed by pointwise fusion:

$$Z_c = W_c^{\text{dw}} * X_c, \quad Y = W^{\text{pw}} * Z, \tag{4}$$

where X_c denotes the c-th input channel. This decoupling optimizes complexity to $\mathcal{O}(C \cdot k^2 \cdot H \cdot W + C \cdot C_{\text{out}} \cdot H \cdot W)$, significantly reducing parameter redundancy while preserving local texture modeling capability. Heterogeneous features from all branches are concatenated and recalibrated via 1×1 convolution, constructing a complete feature chain from microscopic textures to macroscopic semantics. This design effectively enhances contrast for subtle defects (e.g., pitting) while maintaining localization consistency for complex shapes, stably improving detection performance in multi-scale scenarios with controlled computational overhead.

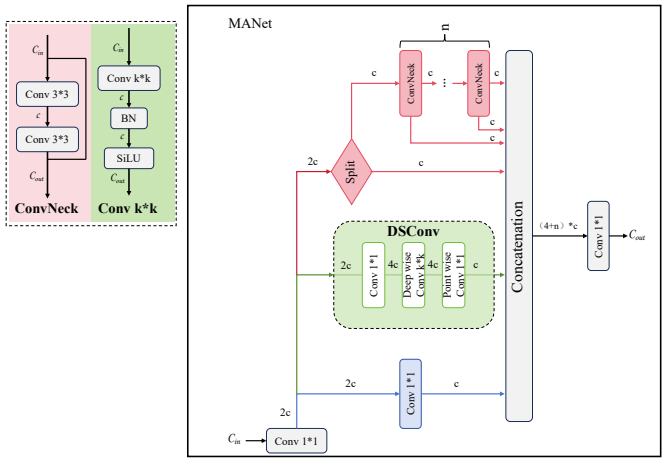

Fig. 4: Network architecture of MANet with parallel branch processing strategy for multi-scale feature fusion.

IV. Experimental Results and Analysis

This section presents comprehensive experimental validation of CAM-DETR on three industrial datasets, including ablation studies, comparative evaluations, and visualization analysis. We demonstrate the model's superior accuracy-speed balance and robustness across diverse defect types and backgrounds.

A. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090D GPU using PyTorch 1.13.1 and CUDA 11.6. For fair comparison, all models were trained with an input resolution of 640×640 using the AdamW optimizer for 400 epochs with a batch size of 8 and an initial learning rate of $1\text{e}{-4}$. A cosine annealing scheduler was applied, and weight decay was set to 5×10^{-4}.

B. Datasets

Experiments were conducted on three representative datasets: NEU-DET [20], GC10-DET, and PKU-PCB. NEU-DET contains six typical surface defect categories from hot-rolled steel strips, while GC10-DET includes ten defect types

with greater intra-class variation and background complexity. PKU-PCB serves as a cross-domain benchmark for PCB inspection, featuring small targets and cluttered textures. NEU-DET and GC10-DET were split into training, validation, and test sets with a ratio of 7:2:1, while PKU-PCB was divided using an 8:1:1 ratio.

C. Evaluation Metrics

This paper constructs a multi-dimensional evaluation system considering both detection accuracy and real-time requirements. Mean Average Precision (mAP) quantifies comprehensive detection accuracy, while Frames Per Second (FPS) measures inference speed. Parameters (Params) and GFLOPs evaluate model lightweighting degree.

Precision (P) measures the proportion of correctly predicted results among all detection results: $P = \text{TP}/(\text{TP} + \text{FP})$, while Recall ($R$) reflects the model's ability to detect actual positive samples: $R = \text{TP}/(\text{TP} + \text{FN})$, where TP (True Positive) denotes correctly predicted positives, FP (False Positive) denotes negatives incorrectly judged as positive, and FN (False Negative) denotes undetected positives. Average Precision (AP) is defined as the area under the Precision-Recall curve. Mean Average Precision (mAP) represents the average AP across all categories. This paper uses mAP@0.5 to denote detection accuracy at an IoU threshold of 0.5.

D. Ablation Experiments

1) Ablation Experiments of Improved Modules: To verify the effectiveness of proposed modules, ablation experiments were designed using RT-DETR-R18 as the baseline. The proposed CE-RGELAN, ATM, and MANet modules were progressively integrated, with independent validation on NEU-DET and GC10-DET datasets. Results are shown in Table I.

Quantitative analysis reveals that progressive integration of the three core modules demonstrates significant positive additive effects. First, replacing the baseline backbone with CE-RGELAN significantly reduces parameters and computational complexity while substantially increasing inference speed. On NEU-DET, GFLOPs dropped from 57.0 to 44.5 and FPS rose from 120.86 to 161.01. The introduction of ATM further enhances performance with minimal overhead—on GC10-DET, mAP@0.5 improves from 68.93% to 74.33% with FPS reaching 230.13. MANet provides the most significant precision gains when combined with other modules, demonstrating its crucial role in enhancing perception of small, polymorphous defects.

2) Ablation Experiments on Backbone Hyperparameters: To explore the influence of channel expansion coefficient e within CE-RGELAN, ablation studies were conducted on NEU-DET dataset. Results in Fig. 5 and Table II show that when e increases from 0.25 to 0.50, mAP@0.5 significantly improves from 77.64% to 79.17%, while FPS rises from 136.9 to 189.7. However, as e further increases to 0.75 and 1.00, mAP@0.5 drops to 77.28% and 77.21%, accompanied by decreased inference speed. This indicates that excessive channel expansion introduces redundant features, increasing computational burden and potentially weakening feature discriminability. Considering both accuracy and real-time metrics, $e = 0.5$ achieves optimal performance balance.

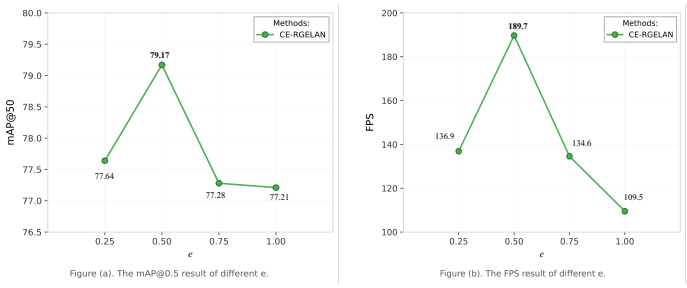

Fig. 5: Testing performance under different values of channel expansion coefficient e: (a) mAP@0.5 and (b) FPS trends.

E. Comparative Experiments

1) Comparison of Backbone Networks: Maintaining consistent training strategies, RT-DETR-R18 baseline was used to compare performance of emerging backbones on NEU-DET dataset. Results in Table III and Fig. 6 show significant differences among backbones regarding detection accuracy, computational complexity, and inference efficiency. While CSwin and LSNet achieve certain improvements in mAP@0.5, their significantly increased computational volume leads to reduced inference speeds. The proposed CE-RGELAN demonstrates best comprehensive performance with 79.17% mAP@0.5 and 189.74 FPS inference speed at lower parameter and computational scales.

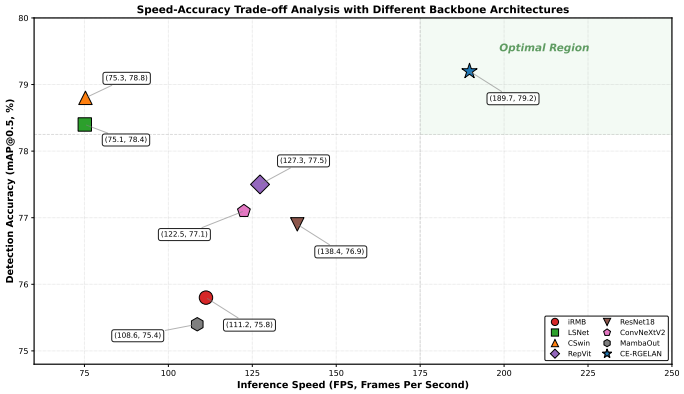

Fig. 6: Speed-accuracy trade-off diagram for different backbone networks on NEU-DET dataset.

2) Comparison with State-of-the-Art Algorithms: Systematic comparisons were conducted against mainstream object detection algorithms on NEU-DET, GC10-DET, and PKU-PCB datasets. The results are summarized in Table IV. For models not originally evaluated on a specific dataset, we report – (results not available in the referenced work).

On NEU-DET, our method achieves 79.17% mAP@0.5 (2.27% improvement over RT-DETR) with 189.74 FPS inference speed, outperforming other methods in both accuracy and

TABLE I: Ablation results of different module combinations on NEU-DET and GC10-DET datasets

CE-RG	ATM	MANet	NEU-DET			GC10-DET		
			GFLOPs ↓	mAP@0.5 (%) ↑	FPS ↑	GFLOPs ↓	mAP@0.5 (%) ↑	FPS ↑
–	–	–	57.0	76.88	120.86	57.0	68.93	105.52
✓	–	–	44.5	77.83	161.01	44.5	72.05	133.25
–	✓	–	57.1	77.14	190.82	57.1	74.33	230.13
–	–	✓	59.3	78.92	102.68	59.3	72.42	116.21
✓	✓	–	44.6	77.51	197.36	44.7	73.28	192.18
✓	–	✓	48.1	78.97	94.55	48.2	73.57	174.30
–	✓	✓	59.5	79.04	190.30	59.5	71.99	196.93
✓	✓	✓	48.3	**79.17**	**189.74**	48.3	**74.52**	**212.83**

TABLE II: Ablation results of different e values on NEU-DET dataset

e	GFLOPs ↓	mAP@0.5 (%) ↑	mAP@0.5:0.95 (%) ↑	FPS ↑
0.25	47.6	77.64	44.66	136.9
0.50	48.3	**79.17**	44.92	**189.7**
0.75	49.2	77.28	45.06	134.6
1.00	49.8	77.21	45.45	109.5

TABLE III: Performance comparison of different backbone networks on NEU-DET dataset

Model	FPS ↑	mAP@0.5 (%) ↑	GFLOPs ↓
ResNet18 (baseline) [28]	138.37	76.88	57.0
CSwin [29]	79.88	78.80	105.5
LSNet [30]	75.12	78.40	56.8
RepVit [31]	127.33	77.52	52.0
ConvNeXtV2 [32]	134.25	77.14	47.5
CE-RGELAN (ours)	**189.74**	**79.17**	48.3

real-time performance. On GC10-DET, CAM-DETR reaches 74.52% mAP@0.5 (7.02% improvement over RT-DETR) at 212.83 FPS. On the cross-domain PKU-PCB dataset, our approach attains 96.65% mAP@0.5 with 160.10 FPS, achieving the best performance among all compared models while maintaining competitive efficiency and the lowest parameter count. These results demonstrate the superior accuracy-latency balance and strong generalization capability of CAM-DETR across diverse industrial scenarios.

F. Visualization Analysis

To intuitively verify the effectiveness of CAM-DETR, we perform visual analysis using Grad-CAM++ on the NEU dataset. As shown in Fig. 7, the baseline RT-DETR exhibits dispersed attention on background textures, while CAM-DETR demonstrates a more stable and concentrated focus on actual defective regions, effectively suppressing background noise. Furthermore, qualitative comparisons in Fig. 8 show that CAM-DETR significantly reduces false positives and missed detections, producing more accurate bounding boxes even for challenging samples with complex backgrounds.

V. Conclusion

This paper proposes CAM-DETR, a lightweight end-to-end framework for real-time metal surface defect detection. By synergistically integrating the CE-RGELAN backbone, ATM attention mechanism, and MANet fusion module, our method achieves a superior balance between detection accuracy and inference speed. Extensive experiments on NEU-DET, GC10-DET, and PKU-PCB datasets demonstrate its state-of-the-art performance and robust cross-domain generalization. Future work will focus on validating its deployment on embedded devices and exploring stable training strategies for few-shot scenarios.

References

[1] Z. Hao, B. Liu, and B. Xu, "A lightweight RT-DETR model for metal surface defect detection using multi-scale network and additive attention mechanism," *J. Nondestruct. Eval.*, vol. 44, no. 3, pp. 107–119, 2025.

[2] S. Zhou, Y. Cai, Z. Zhang, *et al.*, "MESC-DETR: An improved RT-DETR algorithm for steel surface defect detection," *Electronics*, vol. 14, no. 11, p. 2232, 2025.

[3] R. Girshick, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 580–587.

[4] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440–1448.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[6] Y. Zhao, W. Lv, S. Xu, *et al.*, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16965–16974.

[7] X. Zheng, *et al.*, "Automatic metal surface defect detection method based on machine vision," *IEEE Access*, vol. 11, pp. 4567–4578, 2023.

[8] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023.

[9] A. Wang, H. Chen, L. Liu, *et al.*, "YOLOv10: Real-time end-to-end object detection," 2024.

[10] W. Liu, *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[11] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for steel surface defect classification," *Appl. Surf. Sci.*, vol. 285, pp. 858–864, 2013.

[12] N. Carion, F. Massa, *et al.*, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[13] Y. Chen, Q. Chen, Q. Hu, *et al.*, "DATE: Dual assignment of labels for end-to-end fully convolutional object detection," *Pattern Recognit.*, vol. 169, p. 111877, 2026.

[14] X. Zhu, W. Su, *et al.*, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[15] S. Chen, S. Jiang, *et al.*, "An efficient detector for detecting surface defects on cold-rolled steel strips," *Eng. Appl. Artif. Intell.*, vol. 138, p. 109325, 2024.

[16] K. Song, *et al.*, "NEU-DET: A benchmark dataset for metal surface defect detection," *Sensors*, vol. 20, pp. 7412–7428, 2020.

[17] B. Tang, L. Chen, W. Sun, *et al.*, "Review of surface defect detection of steel products based on machine vision," *IET Image Process.*, vol. 17, pp. 303–322, 2023.

TABLE IV: Performance comparison with state-of-the-art methods on NEU-DET, GC10-DET, and PKU-PCB datasets. "–" indicates the model was not evaluated on that dataset in the original work (or not reported).

Model	**NEU-DET**		**GC10-DET**		**PKU-PCB**		**Params (M)** ↓
	mAP@0.5 (%) ↑	FPS ↑	mAP@0.5 (%) ↑	FPS ↑	mAP@0.5 (%) ↑	FPS ↑	
RT-DETR [6]	76.90	120.86	67.50	112.35	95.10	115.30	19.88
RT-DETR-L [6]	71.10	111.00	–	–	–	–	–
Faster R-CNN [5]	72.30	18.10	–	–	–	–	–
YOLOv7 [22]	74.20	138.88	–	–	–	–	70.99
YOLOv8l [8]	75.40	70.92	–	–	–	–	83.32
YOLOv10m [9]	74.30	121.95	–	–	–	–	31.45
MSRFE-RTDETR [1]	76.40	158.73	72.20	140.84	–	–	16.94
GFYOLOv7 [23]	–	–	72.10	61.00	–	–	–
EC-YOLO [18]	–	–	71.00	64.10	–	–	–
YOLOv5s [26]	–	–	–	–	89.30	156.00	7.20
YOLOv5m [26]	–	–	–	–	94.60	97.40	21.20
YOLOv8s [8]	–	–	–	–	92.50	128.70	11.20
Ours (CAM-DETR)	**79.17**	**189.74**	**74.52**	**212.83**	**96.65**	**160.10**	**17.70**

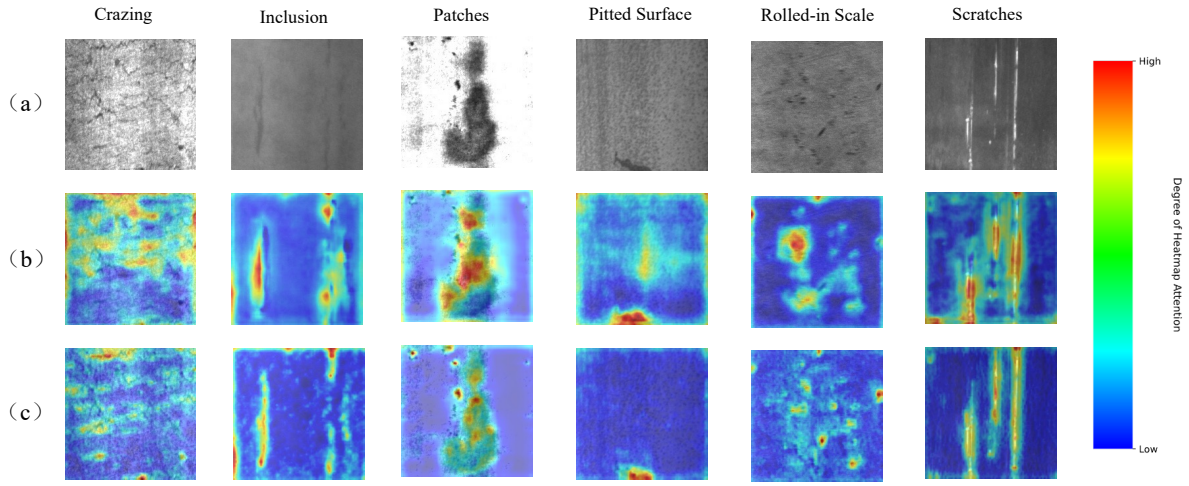

Fig. 7: Representative heatmap visualizations of six defect categories on NEU dataset: (a) input images, (b) RT-DETR results, (c) CAM-DETR results.

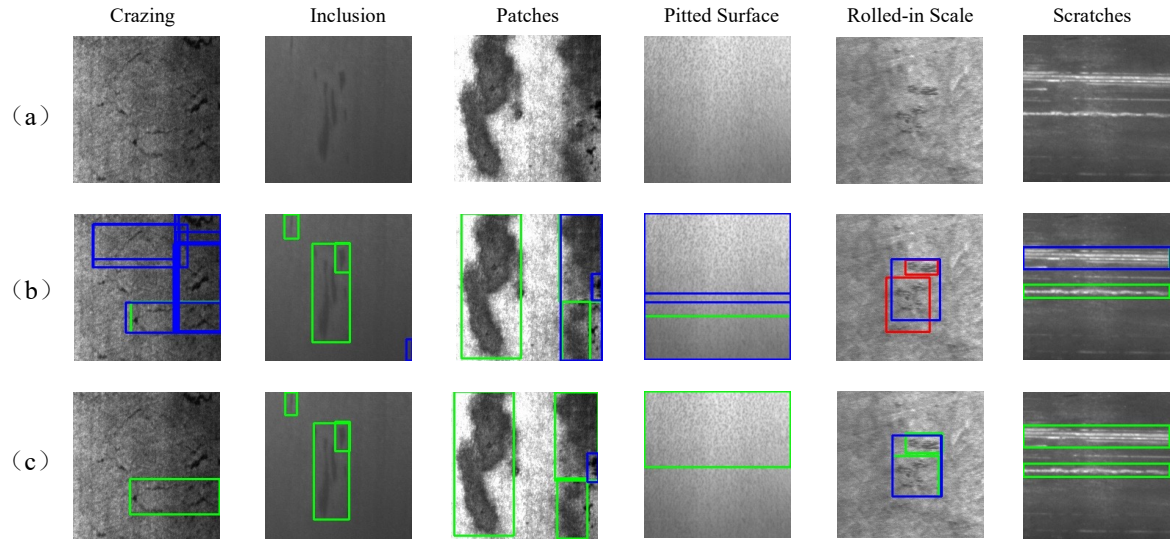

Fig. 8: Qualitative comparison of detection results: (a) RT-DETR predictions, (b) CAM-DETR predictions. Green, blue, and red bounding boxes indicate true positives, false positives, and false negatives, respectively.

[18] Z. Cheng, *et al.*, "EC-YOLO: Effectual detection model for steel strip surface defects based on YOLO-V5," *IEEE Access*, vol. 12, pp. 62765–62778, 2024.

[19] M. Yasir and H. Ahn, "Faster metallic surface defect detection using deep learning with channel shuffling," *Comput. Mater. Continua*, vol. 75, no. 1, pp. 1847–1861, 2023.

[20] X. Li, M. Cai, X. Tan, *et al.*, "An efficient transformer network for detecting multi-scale chicken in complex free-range farming environments via improved RT-DETR," *Comput. Electron. Agric.*, vol. 224, p. 109160, 2024.

[21] C. Wu, F. Wu, T. Qi, *et al.*, "Fastformer: Additive attention can be all you need," 2021.

[22] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7462–7471.

[23] X. Lv, F. Duan, J. J. Jiang, *et al.*, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors*, vol. 20, no. 6, p. 1562, 2020.

[24] S. Chan, S. Li, H. Zhang, *et al.*, "Feature optimization-guided high-precision and real-time metal surface defect detection network," *Sci. Rep.*, vol. 14, pp. 31941–31956, 2024.

[25] L. Liao, C. Song, S. Wu, *et al.*, "A novel YOLOv10-based algorithm for accurate steel surface defect detection," *Sensors*, vol. 25, no. 3, p. 769, 2025.

[26] G. Jocher, "Ultralytics YOLOv5," 2020.

[27] Z. Ge, S. Liu, F. Wang, *et al.*, "YOLOX: Exceeding YOLO series in 2021," 2021.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[29] X. Chu, Z. Tian, Y. Wang, *et al.*, "CSwin Transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12124–12134.

[30] W. Zhou, Y. Zhu, J. Lei, *et al.*, "LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.

[31] A. Wang, H. Chen, Z. Lin, *et al.*, "RepViT: Revisiting mobile CNN from ViT perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 15909–15920.

[32] S. Woo, S. Debnath, R. Hu, *et al.*, "ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16133–16142.

[33] J. Zhang, X. Li, J. Li, *et al.*, "Rethinking mobile block for efficient attention-based models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1389–1400.

[34] W. Yu and X. Wang, "MambaOut: Do we really need Mamba for vision?" 2024.

[35] T. Zhang, L. Li, Y. Zhou, *et al.*, "CAS-ViT: Convolutional additive self-attention vision transformers for efficient mobile applications," *IEEE Trans. Image Process.*, vol. 34, pp. 1254–1268, 2025.

[36] Y. Feng, J. Huang, S. Du, *et al.*, "Hyper-YOLO: When visual object detection meets hypergraph computation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2388–2401, 2025.

[37] Z. Lv, Z. Lu, K. Xia, *et al.*, "LAACNet: Lightweight adaptive activation convolution network-based defect detection on polished metal surfaces," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108482, 2024.

[38] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024.