ITC 4/54
Information Technology and Control
Vol. 54 / No. 4/ 2025
pp. 1227-1247
DOI 10.5755/j01.itc.54.4.42676

ACT-YOLO: An Efficient Multi-Module Fusion Object Detection Algorithm for Steel Surface Defect Detection

Received 2025/08/30 | Accepted after revision 2025/10/10

HOW TO CITE: Tang, P., Shi, Y., Chen, P., Ting, T., T., Zhao, J., Liang, Z., Hu, C., Xia, H. (2025). ACT-YOLO: An Efficient Multi-Module Fusion Object Detection Algorithm for Steel Surface Defect Detection. *Information Technology and Control*, 54(4), 1227-1247. https://doi.org/10.5755/j01.itc.54.4.42676

ACT-YOLO: An Efficient Multi-Module Fusion Object Detection Algorithm for Steel Surface Defect Detection

Peng Tang, Yunyin Shi

Guangxi Key Laboratory of Automatic Detection Technology and Instrument, School of Electrical Engineering and Automation, Guilin University of Electronic Technology, Guilin, 541001, China

School of Big Data and Computing, Hechi University, Hechi, 546300, China

Peng Chen

School of Big Data and Computing, Hechi University, Hechi, 546300, China

Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, 71800 Nilai, Malaysia

Tin Tin Ting

Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, 71800 Nilai, Malaysia

School of Information Technology, UNITAR International University, Selangor, Malaysia

Jiaqi Zhao, Zhixun Liang

School of Big Data and Computing, Hechi University, Hechi, 546300, China

Cong Hu*

Guangxi Key Laboratory of Automatic Detection Technology and Instrument, School of Electrical Engineering and Automation, Guilin University of Electronic Technology, Guilin, 541001, China

Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips, Guangxi Normal University, Guilin, Guangxi, 541004, China

Haiying Xia

Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips, Guangxi Normal University, Guilin, Guangxi, 541004, China

Corresponding author: tx1367426905@163.com

With the development of intelligent manufacturing, higher requirements for real-time performance and accuracy have been placed on the inspection of surface quality in industrial products. As a core basic material in manufacturing, steel has various complex surface defects, such as cracks, scratches, and scale, which are characterised by small size, diverse shapes, and strong background interference, posing significant challenges for automated inspection. To address the issues of insufficient detection accuracy, limited feature fusion capabilities, and coupling of classification and localisation tasks in existing YOLO models when processing fine-grained defects on steel surfaces, this paper proposes a high-performance object detection algorithm based on an improved YOLOv8m: ACT-YOLO (Adaptive Content-guided Task-aligned YOLO). This algorithm integrates three key modules: the AFMA module to enhance multi-scale perception capabilities for small objects; the CGAF module to achieve content-guided multi-attention feature fusion; and the TADD module to optimise the dynamic alignment between classification and regression tasks in the detection head. Evaluations on the NEU-DET steel surface defect benchmark dataset demonstrate that ACT-YOLO achieves an mAP@0.5 of 86.4% and a detection speed of 115 FPS. Compared to non-YOLO methods such as SSD (mAP@0.5: 61.0%, FPS: 41), RetinaNet (mAP@0.5: 69.5%, FPS: 15), and RT-DETR-r101 (mAP@0.5: 78.8%, FPS: 108), as well as other YOLO series models, ACT-YOLO exhibits significant advantages in both detection accuracy and real-time performance. Generalisation experiments on the GC10-DET dataset also validate its cross-scenario adaptability. ACT-YOLO balances detection accuracy, speed, and model lightweighting, making it suitable for the demand for efficient, real-time defect detection systems in actual industrial environments, with broad engineering application prospects and research value.

KEYWORDS: Steel Surface Defect Detection, Deep Learning, YOLOv8 Improvements, Attention Mechanism, Dynamic Task Alignment.

1. Introduction

Driven by digitalisation, the manufacturing industry is accelerating its transition towards smart production [32]. However, industries with complex processes, such as steel production, have seen relatively slow upgrades. This often results in structural defects on the surface of the product [27]. Steel is widely used in critical fields such as infrastructure, aerospace, and automotive [8], and its surface quality directly determines the structural safety, service life, and market competitiveness of the end product. As demand for high-performance steel continues to rise in sectors like aerospace, automotive manufacturing, and marine engineering, surface defects such as cracks, inclusions, and spots have become core hazards that can trigger structural failure and shorten service life. Therefore, achieving real-time, precise detection of surface defects in steel is not only an urgent requirement for ensuring the safe operation of major engineering projects but also a key technological pillar for driving the steel industry's transformation towards high-quality, intelligent development [37]. However, traditional manual inspection methods are inefficient, inaccurate, and lack real-time performance [25], making them unable to meet the stringent requirements of modern high-speed production lines. Early machine vision methods based on manually designed features and traditional image processing are limited by lighting changes, rolling scale patterns, and material texture interference, resulting in poor adaptability [5]. Even when combined with traditional machine learning schemes using HOG, LBP, or Gabor filter features and SVM classifiers, their recognition performance remains inadequate when dealing with micron-level scratches, morphologically variable cracks, and defects with significant grey-scale fluctuations in complex rolling texture backgrounds [19], making them unsuitable for stable and reliable industrial-level applications.

CNNs have advanced significantly in recent years in domains including object detection and picture recognition. A large number of research results have shown [40] that CNNs can achieve automatic learning and accurate identification of defect features through end-to-end feature extraction and classification, thereby improving detection accuracy and generalisation capabilities. The emergence of this technology marks the entry of defect detection into a

new stage centred on data-driven, feature self-learning. Single-stage detectors like YOLO [23] and RetinaNet [13] have improved inference speed and accuracy, gradually replacing the more computationally expensive two-stage models like Mask R-CNN [7] and Faster R-CNN [24]. The YOLO series, known for its single-stage design and effective trade-off between computational cost and accuracy, has become the preferred option for real-time detection [23].

Despite its advantages, YOLO still encounters challenges in detecting steel defects in industrial environments. First, there is the issue of degradation of small defect features. Industrial surface defects are typically small targets that account for a very low percentage of pixels. The downsampling in the YOLO backbone network results in the loss of detailed information. The neck FPN-PAN structure [11] further exacerbates the attenuation of spatial information during cross-scale feature transmission, reducing its ability to identify fine defects. Second, the issue arises from interference caused by variations within a class and similarities between classes. Defects within the same class exhibit diverse morphologies, while different defect classes such as rolled oxides (Rs) and pits share structural similarities. Additionally, due to variations in lighting and material properties, the grey-scale value composition of images of the same defect class fluctuates, increasing detection difficulty [26]. The backbone and neck networks of the YOLO model are centred around convolutional neural networks (CNNs), which excel in object detection but have limitations in extracting global features. Related studies have demonstrated that combining deep detectors with customised preprocessing or attention mechanisms can improve detection performance in complex manufacturing environments. While improved methods such as the CSPNet backbone [9], BiFPN [14], and self-attention layers [12] have partially addressed the aforementioned issues, the lack of explicit inter-layer interactions still limits the model's contextual reasoning capabilities; As a result, effective features may be weakened by noise or conflicting signals within the detection head, reducing the reliability of predictions in industrial applications.

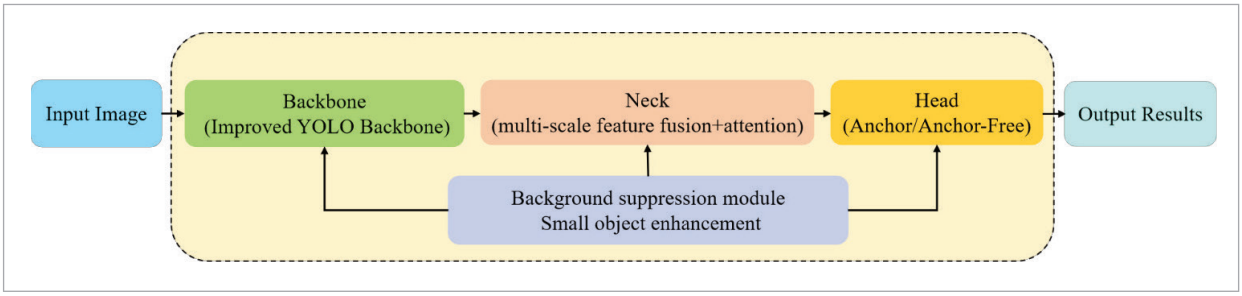
To tackle these challenges, this paper introduces ACT-YOLO, an enhanced YOLOv8m architecture incorporating three specialized modules: AFMA, CGAF, and TADD, optimised for small target accuracy improvement, cross-scale feature integrity preservation, and context-aware detection, respectively. Specifically, the Adaptive Feature Modulation and Aggregation (AFMA) module replaces the backbone network's C2f structure with a dynamic multi-branch structure and attention mechanism to retain and adaptively weight local texture details and global contextual information; the Content-Guided Attention Fusion (CGAF) module reconstructs the Neck structure by fusing spatial, channel, and pixel-level mixed attention, achieving scale adaptation and noise-resistant feature fusion; Finally, the Task-Aligned Dynamic Detection (TADD) detection head adopts a guided task interaction strategy to coordinate optimisation between the classification and localisation branches, minimising feature conflicts and enhancing detection robustness under severe grayscale and texture perturbations. Through these contributions, ACT-YOLO effectively bridges the gap between general object detection frameworks and the practical needs of industrial steel defect detection, providing theoretical references for multi-scale attention design and offering feasible technical solutions for high-speed, high-precision quality control.

2. Related Work

2.1. The Development History of the YOLO Series and Its Application in Defect Detection

Currently, the identification of flaws in steel surfaces using deep learning can be broadly categorised into two types: two-stage detection algorithms, represented by the RCNN series [21], and single-stage detection algorithms, represented by the SSD and YOLO series [28]. Two-stage methods first generate candidate regions, which are then subjected to fine-grained classification and bounding box regression. Although these methods achieve high detection accuracy, they are computationally intensive, resulting in slower detection speeds. In contrast, single-stage methods treat object detection as a single regression task, providing class probabilities and bounding box coordinates simultaneously, which improves detection efficiency.

Figure 1
Improved YOLO-based defect detection framework.



The identification of defects in steel surfaces typically involves semantic segmentation and object detection tasks. U-Net [15][17], which fully exploits contextual information, achieves strong performance in segmenting small-scale defects, particularly those densely distributed. When defects are sparsely distributed, as in the NEU-DET dataset, the YOLO family offers greater advantages due to its speed and global detection capability. In recent years, many YOLO-based variants have been introduced for surface defect detection, such as MSFT-YOLO [6], GDCP-YOLO [33], and RDD-YOLO [34]. As illustrated in Figure 1, most studies focus on optimizing feature extraction, feature fusion, and detection modules, aiming to enhance multi-scale representation and adaptability to complex backgrounds.

Improved YOLO variants have shown strong adaptability to fine-grained defects in steel surface detection. For instance, MSFT-YOLO [6] integrates a Transformer-based TRANS module into the backbone, fusing local and global information via multi-scale features to enhance robustness across object sizes. GDCP-YOLO [33] introduces DCNV2 into the backbone and incorporates channel attention with adaptive receptive fields in the C2f module, suppressing redundant information while strengthening feature representation. RDD-YOLO [34] incorporates a DFPN in the neck, improving cross-scale fusion and enriching contextual information. These approaches achieve promising trade-offs between accuracy and efficiency, yet challenges remain in preserving tiny defect features and mitigating complex background interference.

The YOLO family, known for its end-to-end design, low latency, and scalability, is the core framework for steel surface defect detection. To overcome its limitations in fine-grained defect recognition and background complexity, this study introduces improvements, including multi-scale fusion, feature refinement, and background suppression, aiming to enhance both accuracy and speed for efficient, real-time detection.

2.2. Research Progress on Multi-scale Feature Fusion Mechanisms

In the inspection of steel surface imperfections, defects like cracks, scratches, inclusions, and pits display distinct morphological variations and a broad range of scale distributions. These defects are influenced by factors like lighting variations, noise, complex backgrounds, and oxide layers, making feature extraction and detection challenging. Especially when detecting small-scale defects, the response of defects in high-level feature maps is typically weak. If the fusion strategy is inappropriate, shallow-level detail features may degrade or even be lost during network transmission. Therefore, establishing an effective multi-scale feature fusion mechanism has become a core technical approach to improving the recognition of minute objects in challenging environments.

Traditional multi-scale detection frameworks, like FPN and PAN, help bridge the gap between semantic information and feature resolution. Through top-down or bidirectional information integration, these frameworks transmit high-level semantic details to low-level spatial features. However, such structures still have limitations in practical applications: (1) insufficient cross-scale feature interaction, with information flow between different feature layers remaining constrained; (2) The absence of dynamic cross-layer semantic alignment limits

adaptability to complex backgrounds and large variations in object size; (3) during fusion, shallow detail features are often suppressed by stronger high-level semantics, reducing the effectiveness of small target detection.

To solve this problem, researchers have suggested several improvements: Liu et al. [16] introduced an enhanced FPN that focuses on reusing low-level features to better preserve detailed information. C. Baoyuan et al. [2] proposed FF-YOLO, which optimises feature extraction and information flow by integrating cross-layer features. Xie et al. [31] enhanced FPN with spatial position awareness (SPA) to improve defect localisation accuracy. Wang et al. [30] developed a cross-layer feature fusion (CFF) approach that integrates semantic information from different layers to improve the recognition performance for small defects. Additionally, the cross-layer attention mechanism of the Transformer was introduced into the multi-scale architecture, enabling adaptive weighting of features at different scales and global dependency modelling, thereby balancing the retention of fine-grained features with global context awareness.

Multi-scale feature fusion research is evolving from a unidirectional pyramid structure to a bidirectional interactive, cross-layer reweighting, and attention-driven fusion approach. This method tackles the challenge of identifying minute imperfections on steel surfaces within complex backgrounds by enhancing global perception while preserving and enhancing fine-grained details, leading to substantially improved performance.

2.3. Application of Attention Mechanisms in Feature Enhancement

The attention mechanism, as a feature-guided strategy, can dynamically adjust feature weights according to task objectives, highlight key information areas, and suppress redundant background noise. This method has found extensive use in industrial visual inspection. By assigning weights across feature channels, spatial dimensions, or at the pixel level, it strengthens the model's ability to perceive and represent regions of interest.

In industrial scenarios with complex backgrounds such as defects on steel surfaces, common attention mechanisms include: channel attention, which selectively emphasises the importance of different channels. spatial attention, which guides the model to focus on the spatial distribution of the target to reduce background interference. Pixel-level attention, which enhances the response of specific local areas and is suitable for defects with unclear contours. Content-guided attention, which combines contextual relationships to understand semantics and improve the accuracy of structural defect recognition. Compared to improving performance by increasing network parameters, attention mechanisms can optimise accuracy and efficiency with minimal or no increase in parameters. They also concentrate computational resources on task-relevant areas when processing large-scale features, thereby enhancing detection performance while ensuring real-time capability.

Research on steel surface defect detection has further validated the effectiveness of attention mechanisms. For example, G. Deepti Raj et al. [22] introduced a parameter-free SimAm attention module into the YOLOv7 framework and optimised the loss function, achieving a significant improvement in defect localisation accuracy. J. Zou et al. [39] enhanced the CBAM channel-spatial hybrid attention mechanism within the YOLOv9c baseline and incorporated deformable convolution modules within the core network, thereby improving feature adaptability and detection performance for small defect regions.

Therefore, the attention mechanism plays a key role not only in emphasizing crucial features and suppressing irrelevant backgrounds but also in achieving effective information concentration. Its synergistic application with technologies such as multi-scale feature fusion and deformable convolution provides an important direction for optimising the performance of industrial detection systems.

2.4. Detection Head Optimisation and Dynamic Alignment Technology

The detection head is the core module in a target detection model that performs classification and location regression. Its design plays a critical role in determining the model's capabilities regarding localization precision and class recognition. Traditional coupled detection heads often suffer from task interference, feature redundancy, or missing features when handling classification and regression

tasks simultaneously, making it challenging to balance localisation and classification performance in complex scenes. For defects with elongated shapes or blurred boundaries (such as cracks), the stability of boundary fitting is particularly inadequate, which is also one of the performance bottlenecks for small targets and fine-grained detection.

In high-precision industrial inspection tasks, inspection head optimisation has become the key path to performance breakthroughs. On the one hand, structural decoupling can reduce negative transfer between classification and regression; on the other hand, the use of dynamic alignment strategies can achieve adaptive matching of task requirements at the feature level, thereby improving feature utilisation and prediction stability. C. Zhao et al. [34] introduced a classification and regression decoupling structure in YOLOv5, improving the classification accuracy and boundary localization capabilities for small defects on steel surfaces; Zhou et al. [38] proposed the DPN detector based on the Dynamic Feature Pyramid module (DPN), embedding it into Faster R-CNN and adopting a dual detection head structure to separate and enhance the classification and regression paths, thereby improving overall classification accuracy; Kong et al. [10] designed the Task-aligned Detection Head (TDH), which enhances the consistency and robustness of multi-scale target processing by jointly optimising classification scores and localisation confidence. As such, detection head optimisation and dynamic alignment techniques not only alleviate conflicts across classification and bounding box regression but also markedly enhance identification precision and contour adaptation in complex defect environments, without substantially raising computational costs. This offers strong support for accurate and reliable identification of surface imperfections on steel.

Overall, industrial defect detection tasks place higher demands on detection models, particularly regarding multi-scale feature extraction, attention guidance, and task structure optimisation. The ACT-YOLO algorithm proposed in this paper builds upon previous work by integrating multiple mechanisms and performing in-depth structural optimisation of YOLOv8m, thereby providing a new algorithmic approach for high-performance defect detection in industrial settings.

3. Methodology

3.1. Overall Model Architecture: ACT-YOLO

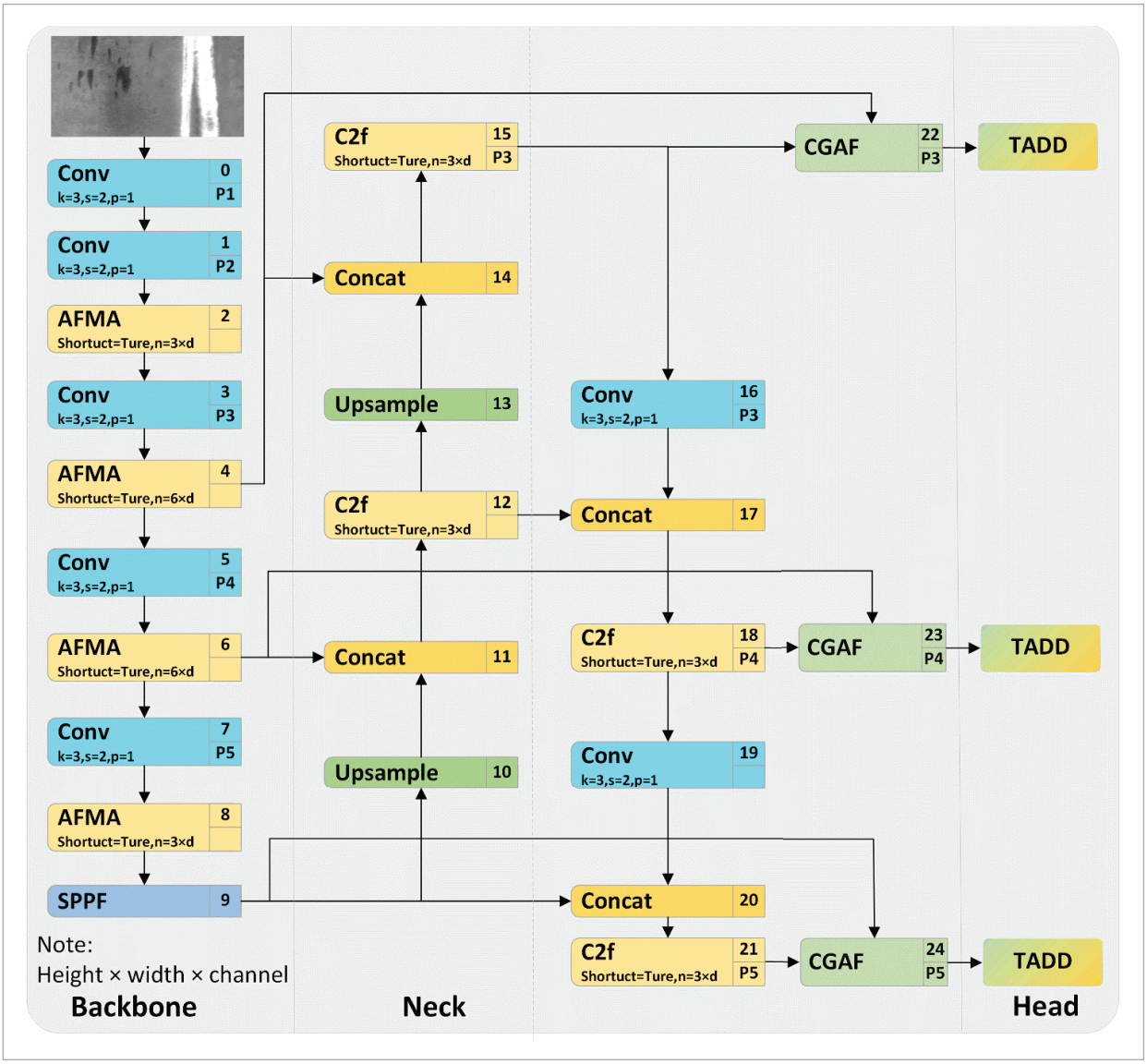
To improve the performance of YOLOv8m in steel surface defect detection, this paper proposes ACT-YOLO (Adaptive Content-guided Task-aligned YOLO). Addressing issues such as insufficient fine-grained defect perception, low feature fusion efficiency, and performance bottlenecks caused by the coupling of classification and regression tasks in the detection head, the paper designs three core improvement modules: the AFMA (Adaptive Feature Modulation Aggregation) module, the CGAF (Content-Guided Attention Fusion) module, and the TADD (Task-Aligned Dynamic Detection) module.

As illustrated in Figure 2, ACT-YOLO retains the conventional three-stage layout of YOLOv8m, consisting of a Backbone, Neck, and Head. In the Backbone, the AFMA module substitutes the original C2f unit, integrating a multi-branch convolutional structure and a lightweight attention component to improve the perception and representation of subtle local defects. In the Neck, the embedded CGAF module employs triple attention (spatial, channel, and pixel) to direct the integration of multi-scale information derived from layers 4, 6, and 9, thereby harmonizing fine-grained details with high-level semantics and strengthening feature discriminability. In the Head, the TADD module is deployed to decouple the classification and regression branches via task alignment. It incorporates group normalisation and shared convolutions, minimising parameter count while enhancing prediction accuracy and consistency in boundary localisation. The synergistic effect of the three stages enables the network to simultaneously achieve high accuracy, low latency, and low computational cost when detecting surface defects in steel.

3.2. AFMA Module

The C2f module in YOLOv8 is primarily used for multi-scale feature extraction to enhance the network's feature representation capabilities. However, there is an imbalance in the processing of local features and global contextual information: the ability to capture local details is relatively limited, and such details are easily weakened in deeper features. This

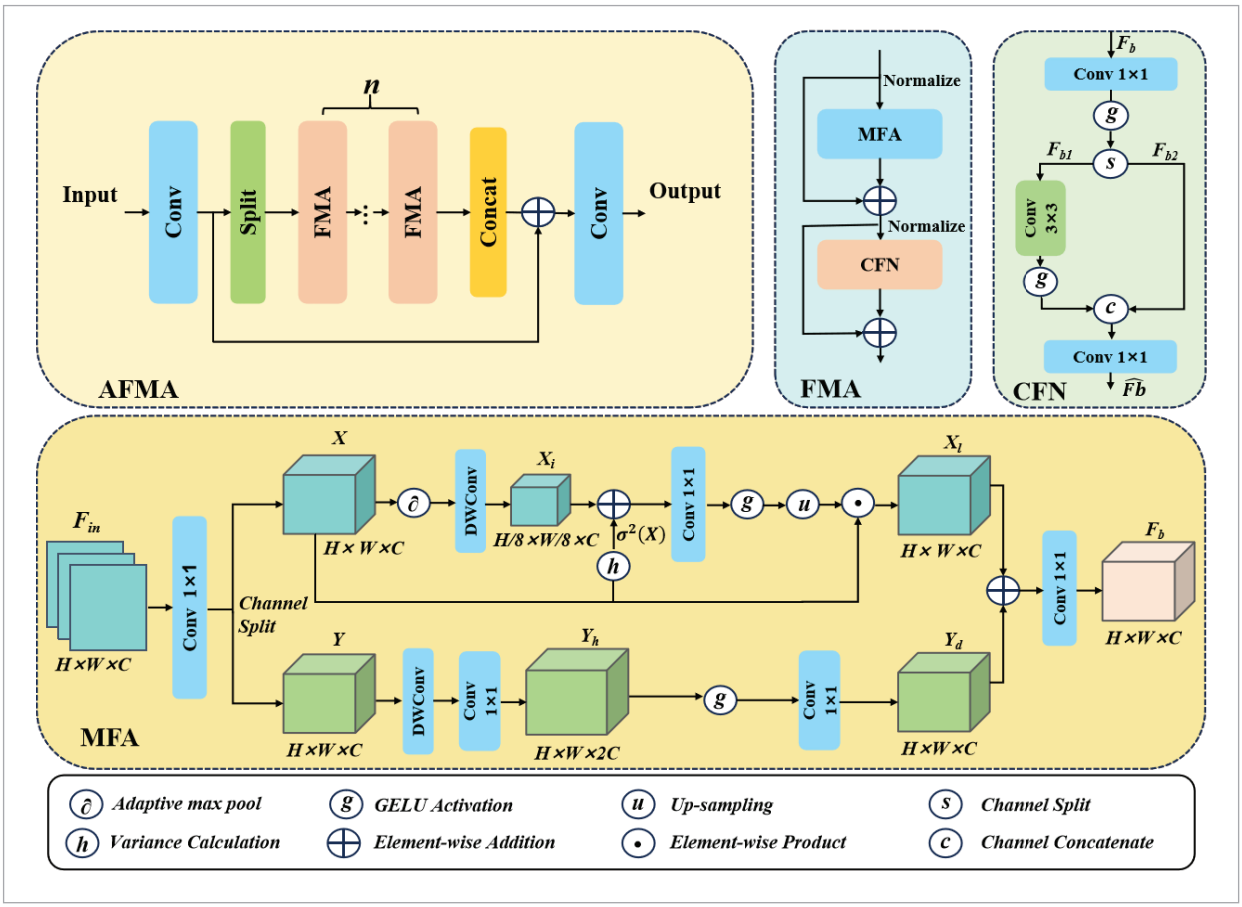
Figure 2
ACT-YOLO Overall Architecture Diagram.



imbalance directly impacts the model's accuracy in steel surface defect detection. Especially since steel surface defect images exhibit significant multi-scale features, where small-sized defects not only have low pixel occupancy but are also more susceptible to interference from complex background noise, making them difficult to accurately locate and identify. Therefore, the AFMA module was designed in the Backbone stage to replace the original C2f unit. AFMA achieves focused enhancement of small-sized defect regions by constructing a multi-path cross-fusion structure that organically combines global context with local detail features, thereby improving the perception capability and detection robustness of small targets.

Figure 3 shows the structural framework of AFMA. This framework inherits the residual fusion idea of C2f module in topology design and draws on the design concept of SMFANet [36]. Specifically, AFMA introduces a feature modulation unit (FMA) at the

Figure 3
AFMA module network architecture diagram.



Bottleneck position to replace the traditional dual 3x3 convolution structure. Each FMA module consists of Modulation Feature Aggregation (MFA) and Convolutional Feedforward Network (CFN), which can simultaneously embed spatial and channel context modeling mechanisms as well as lightweight feedforward expansion strategies into gradient propagation paths. The collaborative work of MFA and CFN not only optimizes the gradient transfer efficiency, but also generates more discriminative deep feature representations, thereby effectively improving the performance of downstream tasks.

This network takes steel images as input and uses 1×1 convolutional layers to extract shallow features F. Then, the extracted shallow features are divides them into n sub-feature groups F_i by channel, and enhances representation across both spatial and channel domains via multi-level feature integration and channel refinement pathways in each branch, leading to F_i', as follows:

$$F_i' = FMA(F_i) = MFA(F_i) \oplus CFN(F_i), \quad (1)$$

where the value of i ranges from 1 to n, F_i' is the branch feature of each FMA module, $\oplus$ is addition operation. all enhanced branch features $\{F_1', F_2', \ldots, F_n'\}$ are concatenated in the channel dimension to form F_{cat}':

$$F_{cat}' = Concat(F_1', F_2', \cdots, F_n'). \quad (2)$$

F_{cat}' is connected to the original features F via residual connections, thereby preserving the original low-level information and gradient smoothness.

This is then fused through a 1×1 convolution to yield the final output F_{out} of the AFMA module.

$$F_{out} = Conv\left(F'_{cat} \oplus F\right), \tag{3}$$

where the F_{out} is the output feature of the AFMA module. As shown in Figure 3, the defect feature map $F_{in} \in {}^{\sim C\times H\times W}$ of the steel material is provided within the MFA module. where $H\times W$ denotes the spatial dimension and C is the number of channels. The MFA module divides F_{in} into two branches, X and Y, for parallel processing. In the X branch, the input features first undergo deep separable convolution (DWConv) to extract primary spatial features, then variance enhancement and 1×1 convolution are introduced for feature compression, and nonlinear activation is used to increase expressive power. Next, upsampling is performed to restore spatial resolution under a larger receptive field, as follows:

$$X_l = u\left(g\left(Conv_{1\times1}\left(h\left(DWConv\left(X\right)\right)\right)\right)\right), \tag{4}$$

where u denotes the upsampling operation, g denotes the GELU activation function, h is variance modulation, X denotes the initial feature quantity, X_l is the output feature quantity of branch X. In branch Y, input features capture local spatial relationships through DWConv, and then 1×1 convolution and activation functions extract defect classification features, as follows:

$$Y_h = Conv_{1\times1}\left(g\left(DWConv\left(Y\right)\right)\right), \tag{5}$$

where Y_h is the output feature quantity for branch Y. this branch reinforces fine-grained edge and texture information through local detail modelling. Finally, the results of the two branches are merged through element-wise addition and feature compression and integration are performed using 1×1 convolution, yielding F_b.

$$F_b = Conv_{1\times1}\left(X_l \oplus Y_h\right), \tag{6}$$

where F_b is the feature output of the MFA module. this preserves the global context information while highlighting local structural features.

The CFN module is designed to achieve efficient information interaction and fusion at the channel dimension, avoiding the inadequacy of single convolution in modelling channel relationships. First, the channels are split into two branches: one branch enters the lightweight branch to preserve the original semantics, and the other branch enters the enhanced branch to extract more complex channel interaction relationships. The lightweight branch compresses the feature channels through 1×1 convolution, resulting in F_{b1}, enhance the branch by splitting the input features, then use 3×3 convolution and GELU non-linear activation to enhance and obtain F_{b2}, after concatenating the results of the two branches, perform further compression using a 1×1 convolution to obtain $\hat{F}_b$, as follows:

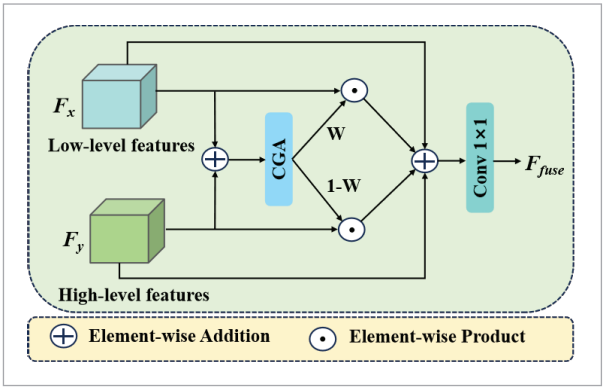
$$\begin{cases} F_{b1} = Conv_{1\times1}\left(F_b\right) \\ F_{b2} = g\left(Conv_{3\times3}\left(s\left(F_b\right)\right)\right) \\ \hat{F}_b = Conv_{1\times1}\left(c\left(\left[F_{b1}, F_{b2}\right]\right)\right) \end{cases}, \tag{7}$$

where c denotes feature channel concatenation, $\hat{F}_b$ is the output of the CFN module.

3.3. CGAF Modules

The CGAF module [3] adopts a bidirectional feature pyramid architecture design. The upsampling path transmits deep semantic features to shallow layers to guide detail enhancement, while the downsampling path transmits shallow detail

Figure 4
CGAF module network architecture diagram.



features to deep layers to supplement boundary and texture information. Each fusion node receives features from different layers and achieves cross-layer information interaction through feature splicing and C2f units.

As shown in Figure 4, F_x and F_y are concatenated as input to the attention module, and the CGA module generates a weight map W to weight F_x, while $1-W$ is used to weight F_y, achieving adaptive feature selection and fusion. The merged feature representations are then processed using a 1×1 convolutional layer to produce the final fused output, denoted as F_{fuse}, according to the following expression:

$$F_{\text{fuse}} = Conv\left(F_x \cdot W + F_y \cdot (1-W) + F_x + F_y\right) \quad (8)$$

where W is a weight map, F_x is the low-frequency feature, F_y is the high-frequency feature.

The key to fusion in the CGAF module lies in content-guided attention (CGA), which can dynamically generate weights based on feature content, thereby adaptively selecting information.

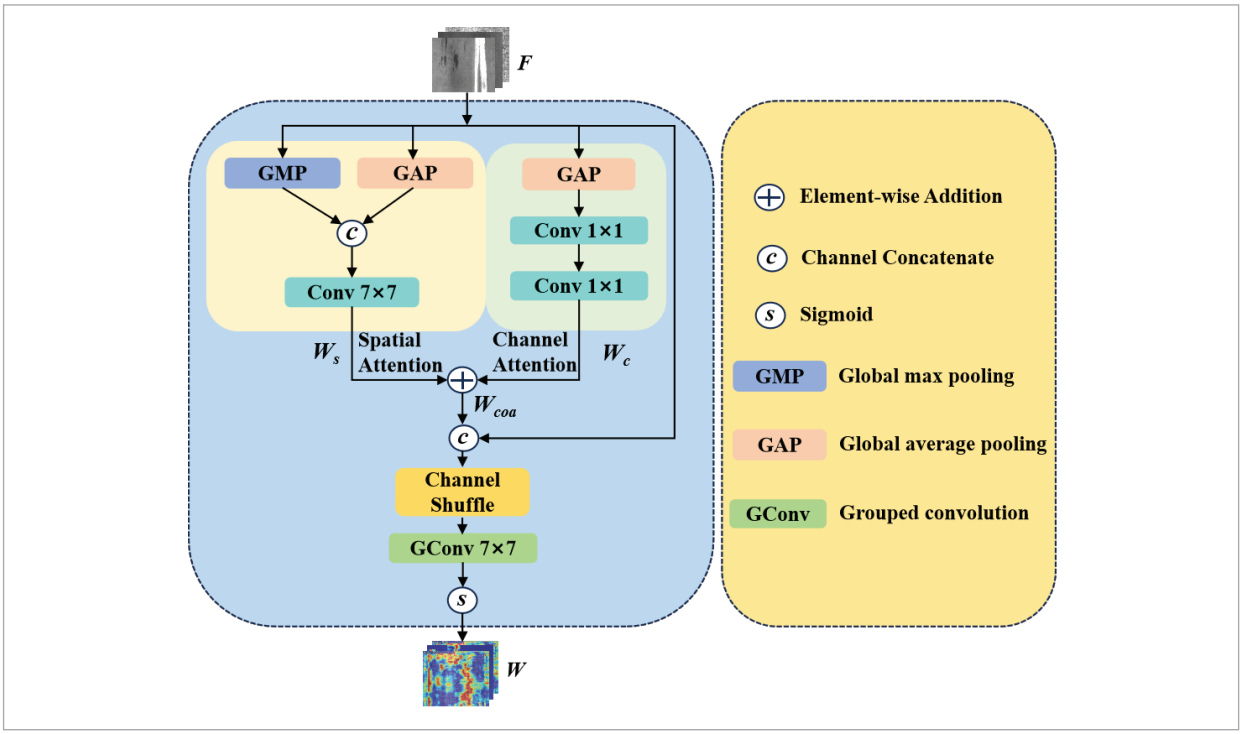
As depicted in Figure 5, the CGA module processes image features with two main goals: attention weighting and feature fusion of the input feature map F to generate a weight map W. The CGA module comprises multiple sub-modules, each responsible for extracting and fusing different aspects of the features. Ultimately, the sub-modules output normalised weights via the Sigmoid function.

1 **Global Pooling and Feature Extraction:** The input feature map F undergoes both global maximum pooling (GMP) and global average pooling (GAP). GAP captures overall statistical information, whilst GMP highlights the most prominent activation regions.

2 **Channel and Spatial Attention Mechanism:** Next, one of the GAP paths undergoes two 1x1 convolutions to generate channel attention weights W_c, which mainly determine the importance of each channel. At the same time, the GAP and GMP results are concatenated across channels and convolved with a 7×7 kernel to generate spatial attention weights W_s. This enables the model to focus on key regions in the local space.

Figure 5

CGA module network architecture diagram.



These two paths focus on channel attention and spatial attention, respectively, and are presented in parallel branches in Figure 5.

3 **Attention fusion:** The channel attention weights W_c and the spatial attention weights W_s are then fused via element-wise addition to obtain the joint attention weights W_{coa}. This fusion operation utilises both channel and spatial information to provide a more comprehensive feature weighting scheme.

4 **Feature mixing and grouped convolution:** The fused weight W_{coa} is concatenated with the previous result and then a channel shuffling operation is performed. This breaks the fixed intra-group order and allows information to be exchanged between different groups, thereby improving the richness of feature expression. Finally, a 7×7 grouped convolution (GConv) operation is performed to further extract and refine local features, thereby enhancing the model's expressive power while maintaining control over the number of parameters.

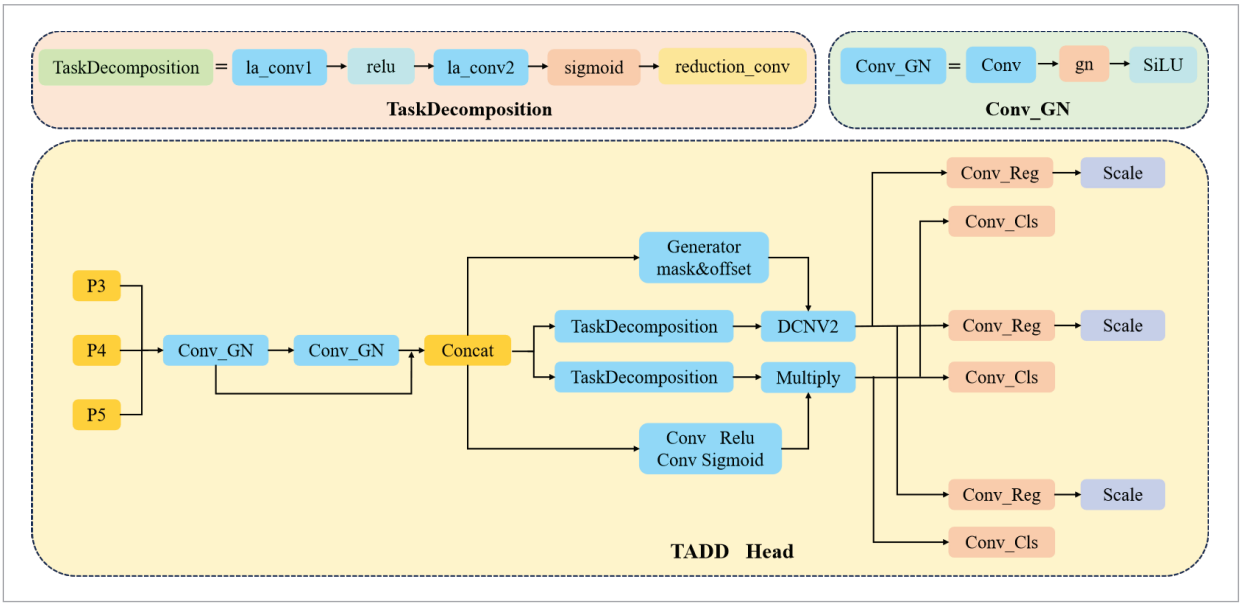
5 **Activation output:** After a series of feature extraction, fusion, and shuffling operations, the result undergoes a non-linear transformation using a sigmoid activation function to map the output weight W to the interval [0,1]. This result can be used as a weighted input for subsequent modules, or it can participate directly in object detection prediction.

6 In summary, the design of this module helps the model better capture the key features of small defects and provides accurate weight allocation for subsequent detection tasks.

3.4. TADD Modules

The TADD detection head module [29], illustrated in Figure 6, comprises a shared convolutional layer, a task decomposition module, a dynamic convolution alignment module, and a classification probability generation module, aiming to address the issues of insufficient task interaction and scarcity of reference samples when applying the YOLOv8 decoupled head to steel surface defect detection. The TADD module comprises two core components: shared feature extraction from the input, and subsequent feature fusion coupled with task decomposition. At this stage, multi-scale input features (P3, P4, P5) undergo processing via a dual-branch convolutional block with Group Normalization (Conv_GN). After fusion, the features from each scale are compressed into shared features F_s to reduce computational load.

Figure 6
TADD module network architecture diagram.



$$F_S = GN\left(Conv\left(P_i\right)\right),\ i \in \{3,4,5\}, \tag{9}$$

where GN and $Conv$ represent group normalisation and convolution operations, respectively. Subsequently, the task decomposition module maps F_s to classification features F_{cls} and regression features F_{reg} according to task requirements, sending them to the regression and classification paths, respectively. In the regression path, the lightweight generator (Generator mask & offset) predicts the sampling offset Δp and mask m of the deformable convolution and uses DCNV2 to achieve dynamic spatial alignment, as follows:

$$\begin{cases} F_{cls} = TD\_cls\left(F_s\right) \\ F_{reg} = TD\_reg\left(F_s\right) \\ \{\Delta p,\ m\} = Gen\left(F_{reg}\right) \\ F_{reg}{}' = DCN\left(F_{reg},\ \Delta p,\ m\right) \end{cases} \tag{10}$$

finally, the classification path is normalised by Sigmoid to obtain the classification probability P_{cls}, and the regression path is adjusted by Scale to output the prediction box parameters Breg B_{reg}.

$$\begin{cases} P_{cls} = Sigmoid\left(Conv\left(F_{cls}\right)\right) \\ B_{reg} = s \times Conv\left(F_{reg}{}'\right) \end{cases} \tag{11}$$

where s is a learnable scale coefficient. Through feature decomposition and cross-task information interaction, the TADD detector can not only improve the localisation and classification accuracy of complex defects such as cracks and star-shaped pits, but also alleviate the detection performance degradation problem when there are insufficient high-quality reference samples.

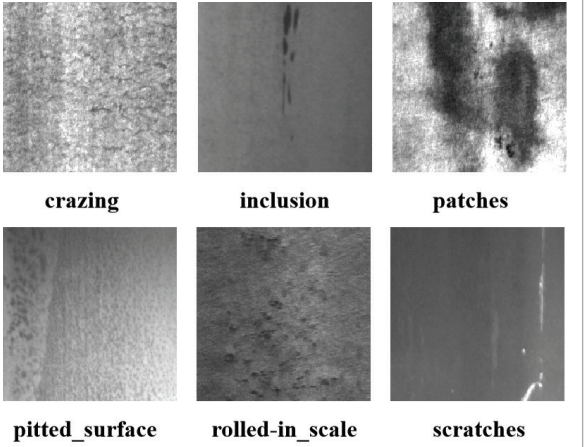
The ACT-YOLO model not only retains the efficiency of YOLOv8m in terms of structure, but also comprehensively enhances defect feature extraction, expression, and prediction by introducing the AFMA, CGAF, and TADD modules. The modules form a multi-dimensional synergistic relationship from micro-detail enhancement to semantic alignment, constructing a high-performance object detection framework for industrial defect detection.

4. Experiments and Results Analysis

4.1. Data Set and Experimental Setup

1 **Datasets:** To validate high practicality and transferability in complex industrial scenarios, the proposed model was rigorously evaluated on two datasets: NEU-DET for core performance benchmarks and GC10-DET for assessing generalization ability. NEU-DET (Northeast University Surface Defect Dataset) was constructed by Northeast University to support research and optimisation of metal surface quality detection algorithms in the field of intelligent manufacturing. It is a benchmark dataset for industrial surface defects, comprising 1,800 high-resolution grayscale images of uniform size (200×200 pixels), covering six typical defect categories: rolling scale, spots, cracks, pitted surfaces, inclusions, and scratches, with 300 images per category, ensuring balanced sampling. Some defect examples are shown in Figure 7. Its main characteristics include: small defect target sizes with sparse distribution, high similarity in features across different categories, complex background textures, and significant variations in lighting conditions.

Figure 7
Map of defect categories in the NEU-DET dataset.



The GC10-DET dataset, which was compiled by the Institute of Automation at the Chinese Academy of Sciences, contains 2,294 high-resolution images (2,048×1,000 pixels) that cover ten different defect categories. The dataset is designed to evaluate the model's

ability to generalise in scenarios involving unfamiliar defect patterns and increased image resolutions.

2 Experimental setup

All experiments in this study were conducted under the following software and hardware environments and training configurations, as shown in Table 1.

Table 1
Experimental setup and training configuration.

Configuration	Parameters
Language/Framework	Python 3.8, PyTorch 2.0
Hardware Configuration	Intel i5-12500H, NVIDIA RTX4090D (24GB)
Optimiser	SGD
Initial Learning Rate	0.001→0.0001 (Linear Decay)
Number of training laps	200 Epochs
Enter Dimensions	640×640

4.2. Ablation Experiment

An algorithm for steel defect identification, named ACT-YOLO, is introduced in this work, constructed within the YOLOv8m architecture. The model incorporates three novel modules: AFMA, CGAF, and TADD. To systematically quantify the contributions of these components, a series of eight ablation studies was conducted on the NEU-DET dataset under consistent training settings. As summarized in Table 2, × indicates that the corresponding module was not integrated, while √ denotes integration. All metrics, including Precision, Recall, and mAP@0.5, are presented with standard deviation ranges calculated from five experimental runs (each with a different random seed), thereby reflecting the model's robustness under stochastic conditions; similarly, the GFLOPs and FPS measurements include standard deviation ranges to demonstrate fluctuations in computational re-

Table 2
Ablation experiments on NEU-DET dataset.

Case	AFMA	CGAF	TADD	Precision/%	Recall/%	mAP@0.5/%	GFLOPs	FPS
1	×	×	×	75.2±0.023	71.7±0.015	75.7±0.002	78.7±1.0	151±2.0
2	√	×	×	84.4±0.022	75.9±0.013	84.5±0.003	75.6±1.2	126±1.5
3	×	√	×	86.5±0.012	78.9±0.006	85.3±0.003	80.3±1.1	138±2.3
4	×	×	√	86.7±0.015	79.0±0.012	85.4±0.004	102.2±1.5	136±2.0
5	√	√	×	87.2±0.017	78.2±0.015	85.7±0.005	101.5±1.0	115±2.2
6	√	×	√	87.6±0.016	77.1±0.012	85.9±0.004	103.7±1.3	120±2.1
7	×	√	√	87.6±0.015	77.3±0.010	85.8±0.005	102.5±1.0	110±1.8
8	√	√	√	88.2±0.018	78.9±0.016	86.4±0.001	88.2±0.8	115±2.0

Table 3
Ablation experiments for each defect category in the NEU-DET dataset.

Case	AP/(%)						Parameters
	Cr	In	Pa	Ps	Rs	Sc	
1	45.2	81.0	88.6	84.6	61.2	87.7	25.8M
2	60.1	87.3	97.3	84.6	80.6	97.3	25.3M
3	66.9	86.9	97.3	83.9	79.9	96.6	26.6M
4	56.3	86.1	95.3	79.8	75.2	97.9	27.8M
5	62.2	87.9	96.1	77.9	77.2	96.8	34.8M
6	58.0	89.3	96.6	82.4	75.3	96.2	27.0M
7	55.2	89.5	97.2	81.8	81.4	95.2	27.8M
8	69.1	91.0	97.4	85.3	82.3	97.5	32.2M

source consumption and real-time performance. Specifically, our experiments adopted an incremental enhancement strategy: Experiment 1 serves as the baseline model, achieving an mAP of 76.0%; Experiments 2–4 individually integrate the AFMA, CGAF, and TADD modules, respectively; Experiments 5–7 test pairwise combinations; finally, Experiment 8 incorporates all three modules to form the complete ACT-YOLO model. Table 3 presents the detection accuracies for different defect categories obtained from these ablation experiments.

1 **AFMA module**

In case 2, the AFMA module captured multi-scale contextual correlations by dynamically integrating features from different levels, resulting in an increased mean average precision (mAP) of 84.5%. This design particularly excels in detecting low-contrast defects or those that closely resemble the background textures (e.g., cracks), as evidenced by a recall improvement from 71.7% to 75.9%. By robustly fusing multi-resolution features, AFMA successfully mitigates the loss of information in small objects.

2 **CGAF module**

In case 3, the introduction of the CGAF module alone yielded an mAP of 85.3%. By enhancing the network's ability to focus on key regions through content-guided multi-attention fusion, CGAF optimizes feature representations while suppressing redundant background information.

3 **TADD module**

In case 4, the introduction of the TADD module alone yielded an mAP of 85.4%. TADD employs a dynamic feature alignment mechanism through deformable convolutions to spatially adjust features for improved capture of irregular or complex defects. This mechanism produced a significant 11.5% improvement over the baseline in complex defect detection, underscoring that dynamically compensating for misaligned features is essential for precise localisation and classification.

The experimental data in Tables 2 and 3 show that interactions between modules can yield non-linear improvements.

1 **AFMA and CGAF combination (case 5):** The combined application of AFMA and CGAF elevated the mAP to 85.7%, representing an improvement of merely 0.3% over the best-performing single module. This combination demonstrates that the multi-scale enhancement provided by AFMA complements CGAF's attention-based feature refinement. Together, they more suppress false detections in densely defective regions. Their complementary action indicates that AFMA enhances robust feature integration across different scales, while CGAF further refines these features to improve discriminative power.
2 **Combination of TADD and AFMA (case 6/7):** When TADD was combined with either AFMA (case 6) or CGAF (case 7), the mAP reached 85.9% and 85.8%, respectively. Notably, the TADD and AFMA combination (Experiment 6) improved cross-scale defect detection. TADD compensates for spatial misalignment, while AFMA aggregates multi-scale contextual information. Their combined effect fuses spatial adaptability with scale robustness.
3 **Full Tri-Module Integration (case 8):** The complete ACT-YOLO model, integrating AFMA, CGAF, and TADD, achieved an mAP of 86.4%, a 10.2 percentage point improvement over the baseline. This integrated approach demonstrates that dynamic spatial adaptability (TADD), multi-scale fusion (AFMA), and content-guided attention (CGAF) can work synergistically across defect categories (see Table 3), while adding only 6.4 million parameters. ACT-YOLO maintains real-time detection at 115 FPS, ensuring its practical application.

Ablation experiments show that the ACT-YOLO model improves performance in steel defect detection by organically combining three modules: AFMA, CGAF, and TADD. The AFMA module first enhances multi-scale feature fusion to build a more comprehensive and hierarchically rich feature representation, which supports effective capture of defects of various sizes. Based on this output, the CGAF module uses an attention mechanism to filter and select key regions, enabling the model to focus on prominent defect areas while suppressing background interference. Further, the TADD module performs fine-grained localization optimization on the previous results, particularly enhancing the detection of small and tiny defects and reducing the risk of missing these targets.

Ablation experiments validate the independent contributions of each module in terms of multi-scale fusion, key region focusing, and fine-grained localization. Combined experiments show that the synergy among these modules yields a more significant performance boost. In summary, the complete ACT-YOLO architecture demonstrates stable performance in steel defect detection tasks. It maintains high detection accuracy while introducing only limited extra computational cost, meeting the strict requirements of real-time industrial detection.

4.3. Performance of Various Types of Detection

As presented in Table 4 and Figure 8, the per-category detection accuracy of ACT-YOLO on the NEU-DET dataset varies. Notably, the model achieves its highest AP values on pitted surface (Pa, 97.4%) and scratch (Sc, 97.5%) defects, demonstrating superior feature extraction and localization capabilities for flaw types with well-defined morphologies and clear contours. The detection accuracy for inclusions (In, 91.0%) and patches (Ps, 85.3%) also remains at a high level, indicating that the model's spatial perception and category discrimination capabilities are relatively stable for targets of moderate scale with relatively stable texture features. In contrast, the AP values for cracks (Cr, 69.1%) and rolling scale (Rs, 82.3%) are relatively low. This is primarily due to the fine, elongated, and irregular structural characteristics of cracks, which are easily affected by background texture interference; while scale and some other defect categories exhibit high similarity in grey-scale distribution and texture patterns, increasing the difficulty of category distinction. Overall, ACT-YOLO maintains high detection performance even for defect categories with unclear structural features (e.g., cracks) and high inter-class similarity (e.g., scale), demonstrating its robust perception and classification capabilities in complex industrial surface scenarios.

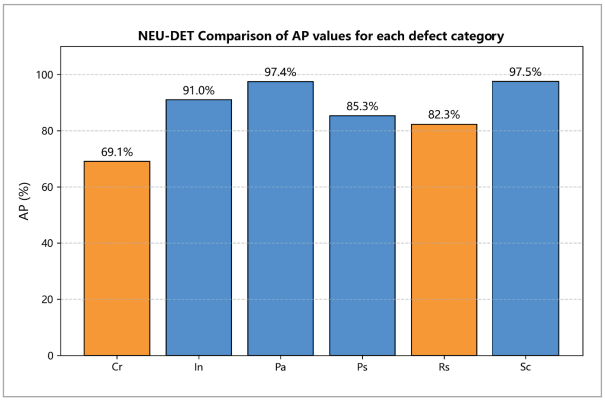
Table 4

Comparison of AP values for various categories of NEU-DET.

Method	AP/(%)					
	Cr	In	Pa	Ps	Rs	Sc
ACT-YOLO	69.1	91.0	97.4	85.3	82.3	97.5

Figure 8

NEU-DET Columnar comparison chart of AP values for various categories.



4.4. Comparative Experiments

A comparative analysis was conducted on the NEU-DET dataset to assess the industrial detection performance of ACT-YOLO against several detectors, including SSD, RetinaNet, RT-DETR and the YOLO family (from YOLOv3-Tiny to YOLOv10m). Following the same configuration outlined in Section 4.1, the optimal results in Table 5 are indicated in bold.

Based on the data in Table 5, the experimental results show that ACT-YOLO excels in multiple metrics. In terms of detection accuracy, ACT-YOLO achieves a precision of 88.2% and an mAP of 86.4%. This is the best performance among the compared models. Traditional detectors like SSD (with 67.5% precision and 61.0% mAP) and RetinaNet (with 70.3% precision and 69.5% mAP) perform worse. ACT-YOLO shows clear advantages in detecting complex industrial defects. Even compared to Transformer-based models such as the RT-DETR series and improved models like CDN-YOLOv7 and MSFT-YOLO, ACT-YOLO maintains higher detection accuracy. Its unique design architecture makes it particularly robust when detecting defects with complex backgrounds and varying shapes.

In terms of model parameters and speed, ACT-YOLO has 32.2M parameters. This number is much lower than that of RT-DETR-r50 (82.1M), RT-DETR-r101 (146.6M), and MSFT-YOLO (90.8M). A lower parameter count reduces both storage and computational resource requirements. Regarding inference speed, ACT-YOLO runs at 115 FPS. Although this is slight-

Table 5
Comparative experimental results.

Method	Precision/%	mAP@0.5/%	Params(M)	FPS
SSD	67.5	61.0	41.1	41
RetinaNet	70.3	69.5	18.3	15
YOLOv3-tiny	52.3	46.5	8.6	172
YOLOv5m	73.1	74.4	8.7	161
YOLOv6m	72.9	73.6	9.1	154
YOLOv7m	67.6	73.3	12.3	148
YOLOv8m	75.2	75.7	25.8	152
YOLOv9m	72.1	74.2	25.3	120
YOLOv10m	73.7	74.7	18.4	155
CDN-YOLOv7 [4]	-	80.3	73.4	60
DCN-YOLO [18]	-	76.5	16.3	13
Literature [1]	-	74.1	23.9	75
CBE-YOLOv5 [35]	-	75.5	**8.12**	96
MSFT-YOLO [6]	-	75.2	90.8	30
RT-DETR-r18	78.4	77.5	38.6	**275**
RT-DETR-r34	81.1	79.6	60.1	218
RT-DETR-r50	78.2	78.4	82.1	153
RT-DETR-r101	78.8	78.2	146.6	108
ACT-YOLO	**88.2**	**86.4**	32.2	115

ly lower than lightweight models like YOLOv3-tiny (172 FPS) and RT-DETR-r18 (275 FPS), 115 FPS is sufficient for real-time industrial detection. Overall, ACT-YOLO achieves an ideal balance between accuracy and speed compared to models that focus solely on accuracy at the expense of speed.

4.5. Visualisation of Comparative Experiments

The performance difference between the ACT-YOLO model and the YOLOv8m baseline on six different types of steel surface defects in the NEU-DET dataset is visually compared in this research utilizing PR curves (Figure 9) and a visual representation of the detection findings (Figure 10).

Figure 9 shows that compared to YOLOv8m, the PR curve of ACT-YOLO is closer to the upper right corner, highlighting its improved accuracy and recall. More specifically, the crack AP value recorded by ACT-YOLO is 0.661, the inclusion is 0.910, the plaque is 0.974, the rough surface is 0.809, the rolling mark is 0.811, and the scratch is 0.960, thus obtaining the overall mAP@0.50.864 In contrast, the AP values of YOLOv8m in these categories are 0.452, 0.810, 0.886, 0.846, 0.612, and 0.878, respectively, totaling mAP@0.50.757 These findings indicate that ACT-YOLO effectively enhances the detection of defects such as cracks, inclusions, and plaques. Its feature modulation and multi-scale fusion techniques help to capture local detail features more effectively.

A comparison of the detection outcomes for ACT-YOLO and YOLOv8m is displayed in Figure 10. It has been demonstrated that YOLOv8m has a comparatively low degree of confidence in identify-

Figure 9
YOLOv8m and ACT-YOLO model PR curves.

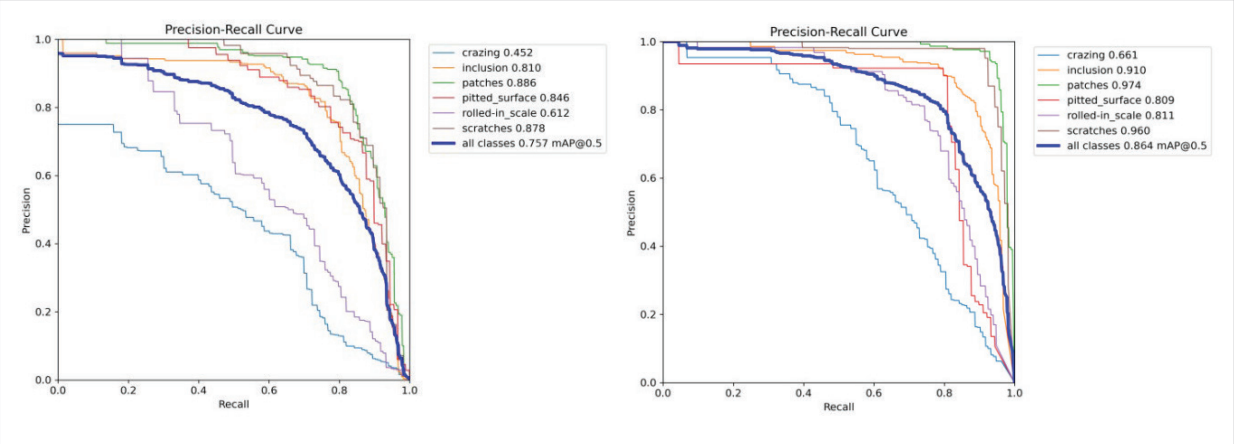
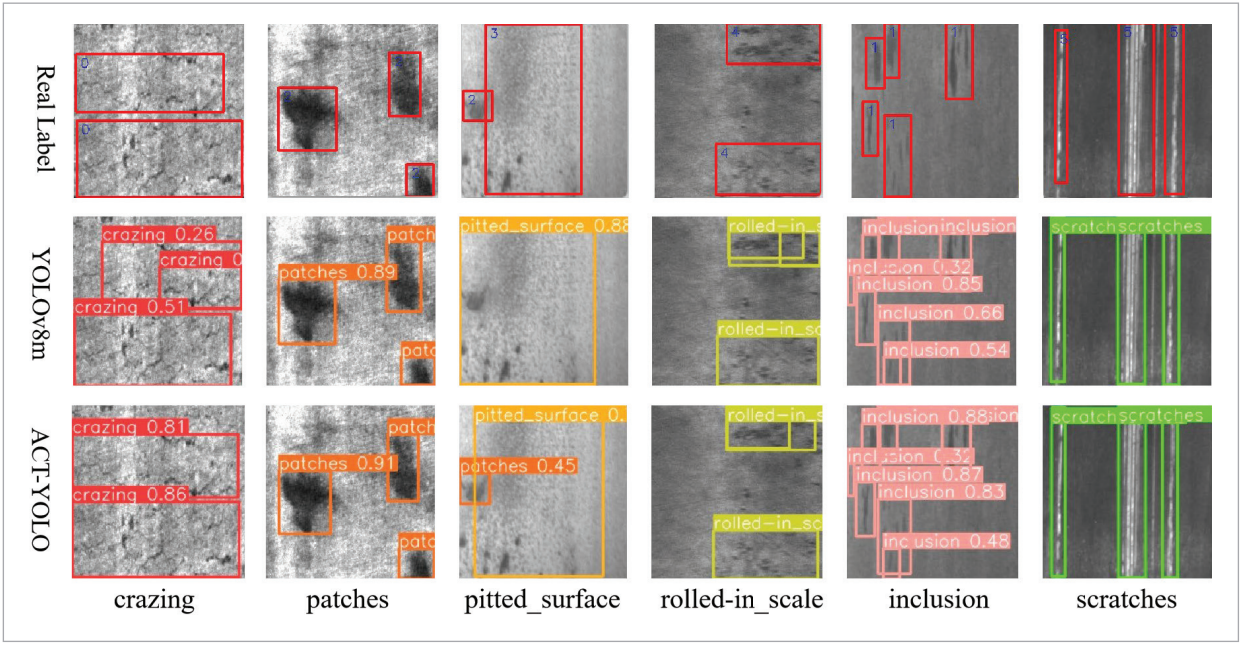


Figure 10
Visualisation of model detection results.



ing surface flaws in steel and is prone to missed or inaccurate detections. In contrast, ACT-YOLO improves the confidence level of defect identification and performs particularly well with regard to typical defects, such as cracks and pitting.

Visual heatmaps for six defect types from the NEU-DET dataset, generated by the ACT-YOLO algorithm and the baseline model, are compared in Figure 11 to analyze their respective focal regions. The generated heatmaps align well with the corresponding detection results. Furthermore, the ACT-YOLO model exhibits a strong thermal response across most defect areas, showing notable proficiency in detecting and localizing small-target defects. This feature reduces the probability of false positives and false negatives, thereby enhancing overall detection performance.

Figure 11
Visualisation of model detection heat map.

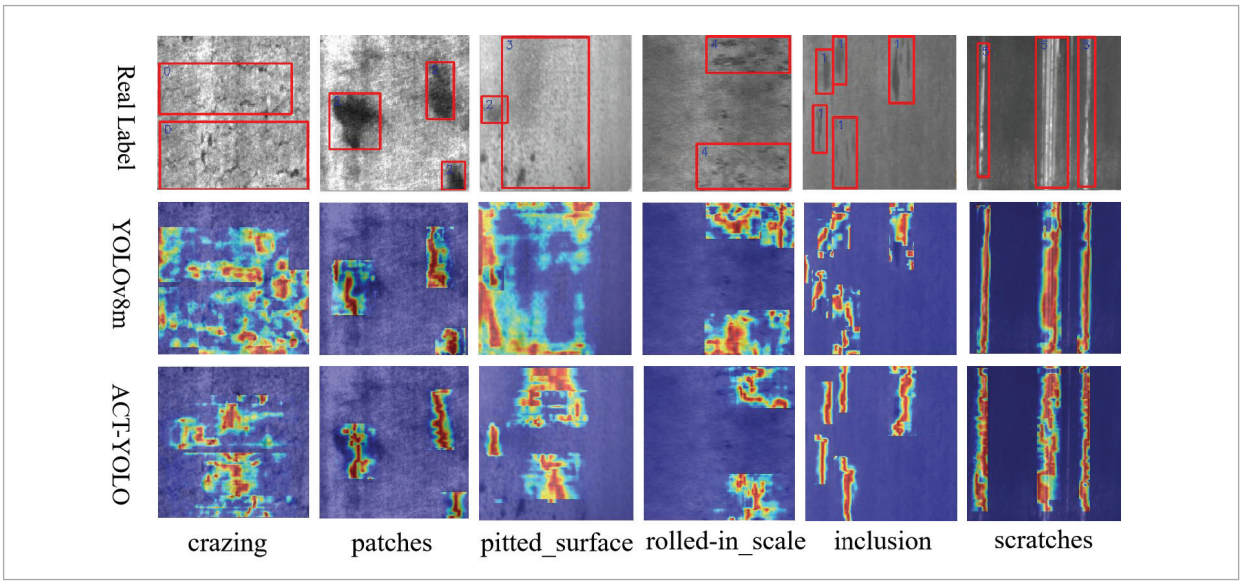
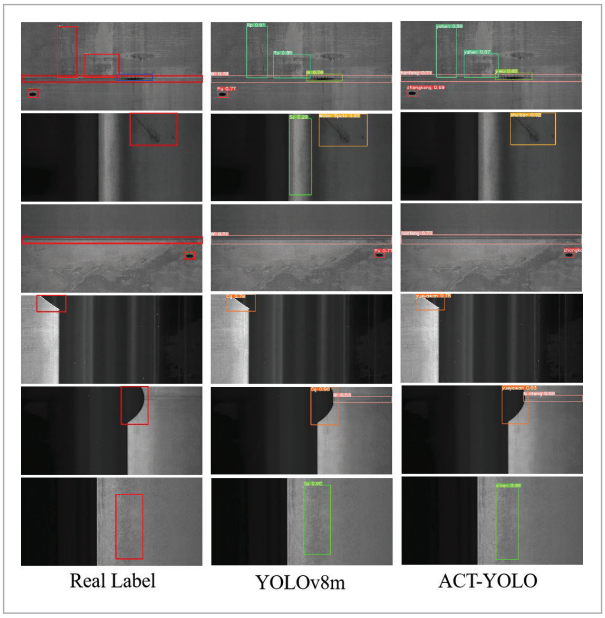


Figure 12
Visualisation of the generalised experimental model detection results.



4.6. Generalisability Validation Experiments

This experiment used the publicly available steel surface defect dataset GC10-DET [20] released by the Institute of Automation, Chinese Academy of Sciences (IAAS). The division of the dataset, the setting of experimental parameters, and the evaluation indicators were consistent with those of the NEU-DET dataset.

As shown in the generalisation experiment results in Figure 12 and Table 6, the ACT-YOLO algorithm achieves higher detection accuracy than the YOLOv8 model for all types of steel surface defects across completely different datasets. In particular, YOLOv8m is prone to misclassifying adjacent normal areas as water spots (Ws), an error that the proposed algorithm suppresses. ACT-YOLO demonstrates higher detection accuracy and robustness than YOLOv8m, it can be used to detect steel surface flaws in real time in a variety of industrial settings.

Table 6
Generalisation experiments.

Method	Precision/%	mAP@0.5/%	Params(M)	FPS
YOLOv8m	66.8	69.0	25.8	232
ACT-YOLO	75.8	70.4	32.2	211

5. Conclusions

This paper addresses three key bottlenecks in current YOLO models used for detecting surface defects on steel: insufficient capability for detecting small defects, inadequate fusion of multi-scale information, and overly tight coupling between the detection task and the detection head. To overcome these issues, a modular improved model based on YOLOv8m called ACT-YOLO is proposed. By integrating multi-dimensional strategies such as feature modulation, attention guidance, and task alignment, the model achieves varying degrees of improvement in both detection accuracy and inference speed, making it a promising candidate for industrial real-time detection applications.

On the NEU-DET dataset, ACT-YOLO achieves an mAP@0.5 of 86.4%, a 10.2% improvement over the baseline YOLOv8m, with precision and recall rising to 88.2% and 78.9%, respectively, while maintaining an inference speed of 115 FPS that meets industrial real-time detection requirements. In cross-scenario generalization tests on the GC10-DET dataset, its mAP further surpasses YOLOv8m by 1.4%, demonstrating robust and stable performance across different scenarios. Moreover, ACT-YOLO exhibits excellent performance in differentiating confusing classes and suppressing false alarms, effectively countering interference in complex industrial visual environments.

Compared to traditional detection models, ACT-YOLO introduces improvements tailored to their limitations. Models like SSD and RetinaNet often employ simple multi-scale prediction strategies, making it difficult to accurately identify tiny defects in complex backgrounds; early YOLO series models, constrained by static convolutions and fixed fusion paths, struggle to adapt to defects with deformation or scale variations. To address these challenges, ACT-YOLO introduces three core modules. First, the AFMA module dynamically adjusts and aggregates multi-level features, achieving adaptive multi-scale fusion to enhance detection capability across defects of varying sizes. Second, the CGAF module utilizes a content-driven attention mechanism to suppress irrelevant background interference, thereby strengthening the feature representation of defect regions and reducing false positives. Finally, the TADD module employs deformable convolutions to achieve local dynamic alignment, improving the accuracy in capturing irregular or elongated defect shapes.

Compared with other enhanced YOLO variants and Transformer-based models like RT-DETR, ACT-YOLO separately optimizes feature fusion, attention guidance, and spatial alignment, thereby avoiding the overfitting and increased computational burden that may arise simply from adding more parameters. Although RT-DETR possesses advantages in global modeling, its global attention mechanism tends to weaken the extraction of local features that are crucial for detecting tiny defects. In contrast, ACT-YOLO places a greater emphasis on local details, achieving a better balance between accuracy and speed.

Looking ahead, there is still potential to optimize ACT-YOLO for ultra-small target detection and false alarm suppression in complex environments. Future research will focus on developing a lightweight edge-deployable version (Edge-ACT-YOLO) to suit low-power devices and on-site detection scenarios, as well as incorporating self-supervised pre-training techniques to improve sample efficiency, thereby further enhancing the model's generalization capability and stability.

Acknowledgements

This work was funded through the following grants: the Guangxi Natural Science Foundation (2025GXNSFAA069383), the Ministry of Education Key Laboratory of Equipment Data Security and Protection Technology (GDZB2024060100), the Guangxi Key Laboratory of Automatic Detection Technology and Instruments (YQ23102, YQ23210), the Guangxi Key Laboratory of Brain-Inspired Computing and Intelligent Chips (BCIC-23-K7) and the Guangxi Higher Education Institutions' Mid-Career Faculty Basic Research Capacity Enhancement Project (2025KY0710).

References

1. Cao, Y. Q., Wu, M. L., Xu, L. Steel Surface Defect Detection Based on Improved YOLOv5 Algorithm. Journal of Graphics, 2023, 44(2), 335-345.
2. Chen, B., Liu, Y., Sun, K. Research on Object Detection Method Based on FF-YOLO for Complex Scenes. IEEE Access, 2021, 9, 127950-126060. https://doi.org/10.1109/ACCESS.2021.3108398
3. Chen, Z., He, Z., Lu, Z. DEA-Net: Single Image Dehazing Based on Detail-Enhanced Convolution and Content-Guided Attention. IEEE Transactions on Image Processing, 2024, 33, 1002-1015. https://doi.org/10.1109/TIP.2024.3354108
4. Gao, C. Y., Qin, S., Li, M. H., Zhang, Y., Zhao, W. Research on Steel Surface Defect Detection with Improved YOLOv7 Algorithm. Computer Engineering and Applications, 2024, 60(7), 282-291.
5. Ghorai, S., Mukherjee, A., Gangadaran, M., Bhattacharya, A. Automatic Defect Detection on Hot-Rolled Flat Steel Products. IEEE Transactions on Instrumentation and Measurement, 2013, 62(3), 612-621. https://doi.org/10.1109/TIM.2012.2218677
6. Guo, Z., Wang, C., Yang, G., Chen, J., Sun, X. MSFT-YOLO: Improved YOLOv5 Based on Transformer for Detecting Defects of Steel Surface. Sensors, 2022, 22(9), 15. https://doi.org/10.3390/s22093467
7. He, K., Gkioxari, G., Dollár, P., Girshick, R. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2), 386-397. https://doi.org/10.1109/TPAMI.2018.2844175
8. He, L., Zheng, L., Xiong, J. FMV-YOLO: A Steel Surface Defect Detection Algorithm for Real-World Scenarios. Electronics, 2025, 14(6), 20. https://doi.org/10.3390/electronics14061143
9. Kim, J., Kim, N., Won, C. Global-Local Feature Learning for Fine-Grained Food Classification Based on Swin Transformer. Engineering Applications of Artificial Intelligence, 2024, 133, 7. https://doi.org/10.1016/j.engappai.2024.108248
10. Kong, S., Kong, Y., Chi, X., Zhou, L., Huang, W., Wang, F. A TBD-YOLO-Based Surface Defect Detection Method for Hot Rolled Steel Strips. Russian Journal of Nondestructive Testing, 2025, 61(1), 137-149. https://doi.org/10.1134/S1061830924603192
11. Kou, X., Liu, S., Cheng, K., Xu, P., Zhang, T. Development of a YOLOv3-Based Model for Detecting Defects on Steel Strip Surface. Measurement, 2021, 182, 9. https://doi.org/10.1016/j.measurement.2021.109454
12. Li, T., Zhang, Z., Zhu, M., Wang, L., Zhao, J., Huang, P., Chen, X. Combining Transformer Global and Local Feature Extraction for Object Detection. Complex & Intelligent Systems, 2024, 10(4), 4897-4920. https://doi.org/10.1007/s40747-024-01409-z
13. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P. Focal Loss for Dense Object Detection. Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22-29, 2017. IEEE, 2017. https://doi.org/10.1109/ICCV.2017.324
14. Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21-26, 2017. IEEE, 2017. https://doi.org/10.1109/CVPR.2017.106
15. Liu, B., Yang, B., Zhao, Y., Zhang, F., Li, C. Low-Pass U-Net: A Segmentation Method to Improve Strip Steel Defect Detection. Measurement Science and Technology, 2022, 34(3), 9. https://doi.org/10.1088/1361-6501/aca34a
16. Liu, C., Wu, Y., Liu, J., Zhang, H., Wang, F. Insulator Faults Detection in Aerial Images from High-Voltage Transmission Lines Based on Deep Learning Model. Applied Sciences (Basel), 2021, 11(10), 20. https://doi.org/10.3390/app11104647
17. Liu, Z., Zeng, Z., Li, J., Wang, Y., Zhou, D., Chen, X. Automatic Detection and Quantification of Hot-Rolled Steel Surface Defects Using Deep Learning. Arabian Journal for Science and Engineering, 2023, 48(8), 10213-10225. https://doi.org/10.1007/s13369-022-07567-x
18. Lu, J. Z., Zhang, C. Y., Liu, S. P., Chen, L., Xu, D., Wang, Y. Lightweight DCN-YOLO for Strip Surface Defect Detection in Complex Environments. Computer Engineering and Applications, 2023, 59(15), 318-328.
19. Luo, Q., Fang, X., Liu, L., Li, H., Wang, Y. Automated Visual Defect Detection for Flat Steel Surface: A Survey. IEEE Transactions on Instrumentation and Measurement, 2020, 69(3), 626-644. https://doi.org/10.1109/TIM.2019.2963555
20. Lv, X., Duan, F., Jiang, J., Zhang, Y., Zhao, L. Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network. Sensors, 2020, 20(6), 15. https://doi.org/10.3390/s20061562

21. Qu, Y., Wan, B., Wang, C., Zhang, R., Liu, X. Optimization Algorithm for Steel Surface Defect Detection Based on PP-YOLOE. Electronics, 2023, 12(19), 18. https://doi.org/10.3390/electronics12194161

22. Raj, G., Prabadevi, B. MoL-YOLOv7: Streamlining Industrial Defect Detection with an Optimized YOLOv7 Approach. IEEE Access, 2024, 12, 117090-117101. https://doi.org/10.1109/ACCESS.2024.3447035

23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, June 27-30, 2016. IEEE, New York, 2016. https://doi.org/10.1109/CVPR.2016.91

24. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

25. Song, F., Bao, K., Deng, M., Zhang, L., Li, J. Steel Surface Defect Detection Based on PSO-Gabor and an Improved Faster R-CNN. Nondestructive Testing and Evaluation, 2025, 24. https://doi.org/10.1080/10589759.2025.2512553

26. Song, K., Yan, Y. A Noise Robust Method Based on Completed Local Binary Patterns for Hot-Rolled Steel Strip Surface Defects. Applied Surface Science, 2013, 285, 858-864. https://doi.org/10.1016/j.apsusc.2013.09.002

27. Song, X., Cao, S., Zhang, J., Li, M., Huang, F. Steel Surface Defect Detection Algorithm Based on YOLOv8. Electronics, 2024, 13(5), 18. https://doi.org/10.3390/electronics13050988

28. Tang, B., Song, Z. K., Sun, W., Liu, P., Chen, D. An End-to-End Steel Surface Defect Detection Approach via Swin Transformer. IET Image Processing, 2023, 17(5), 1334-1345. https://doi.org/10.1049/ipr2.12715

29. Tang, P., Ding, Z., Jiang, M., Zhao, W., Xu, L. LBT-YOLO: A Lightweight Road Targeting Algorithm Based on Task-Aligned Dynamic Detection Heads. IEEE Access, 2024, 12, 180422-180435. https://doi.org/10.1109/ACCESS.2024.3509694

30. Wang, P., Li, L., Sha, B., Zhou, Q., Zhang, D. A Lightweight Image-Level Segmentation Method for Steel Surface Defects Based on Cross-Layer Feature Fusion. Insight - Non-Destructive Testing and Condition Monitoring, 2024, 66(3), 167-173. https://doi.org/10.1784/insi.2024.66.3.167

31. Xie, Y., Hu, W., Xie, S., Zhang, L., Chen, Q. Surface Defect Detection Algorithm Based on Feature-Enhanced YOLO. Cognitive Computation, 2023, 15(2), 565-579. https://doi.org/10.1007/s12559-022-10061-z

32. Xu, D., Wu, Y. Improved YOLO-V3 with DenseNet for Multi-Scale Remote Sensing Target Detection. Sensors, 2020, 20(15). https://doi.org/10.3390/s20154276

33. Yuan, Z., Ning, H., Tang, X., Li, J., Zhao, K. GDCP-YOLO: Enhancing Steel Surface Defect Detection Using Lightweight Machine Learning Approach. Electronics, 2024, 13(7), 19. https://doi.org/10.3390/electronics13071388

34. Zhao, C., Shu, X., Yan, X., Wang, J., Liu, H. RDD-YOLO: A Modified YOLO for Detection of Steel Surface Defects. Measurement, 2023, 214, 11. https://doi.org/10.1016/j.measurement.2023.112776

35. Zhao, L. R., Zhen, G. Y., Zu, C. Q., Li, S. P., Wang, T. H. Detection Method of Steel Surface Defects Based on CBE-YOLOv5. Electronic Measurement Technology, 2023, 46(15), 73-80.

36. Zheng, M., Sun, L., Dong, J., Wang, H., Liu, Y. SMFANet: A Lightweight Self-Modulation Feature Aggregation Network for Efficient Image Super-Resolution. In European Conference on Computer Vision (ECCV), Cham: Springer Nature Switzerland, 2024, 359-375. https://doi.org/10.1007/978-3-031-72973-7_21

37. Zheng, X., Liu, W., Huang, Y., Zhang, D., Li, R. LWMS-Net: A Novel Defect Detection Network Based on Multi-Wavelet Multi-Scale for Steel Surface Defects. Measurement, 2025, 252. https://doi.org/10.1016/j.measurement.2025.117393

38. Zhou, X., Wei, M., Li, Q., Fang, Y., Chen, D. Surface Defect Detection of Steel Strip with Double Pyramid Network. Applied Sciences (Basel), 2023, 13(2), 17. https://doi.org/10.3390/app13021054

39. Zou, J., Wang, H. Steel Surface Defect Detection Method Based on Improved YOLOv9 Network. IEEE Access, 2024, 12, 124160-124170. https://doi.org/10.1109/ACCESS.2024.3453931

40. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J. Object Detection in 20 Years: A Survey. Proceedings of the IEEE, 2023, 111(3), 257-276. https://doi.org/10.1109/JPROC.2023.3238524