

YOLOv11-AHFE: Adaptive High-Frequency Signal Enhancement Mechanism for Surface detection

Nathan Du

Artificial Intelligence center, shanghai

Yanfang Meng

1000006368@ujs.edu.cn

Jiangsu University

Article

Keywords: YOLOv11, Adaptive High-Frequency Signal Enhancement, Metal surface defect detection

Posted Date: August 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10410763/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 3rd, 2026.
See the published version at <https://doi.org/10.1038/s41598-026-73404-y>.

YOLOv11-AHFE: Adaptive High-Frequency Signal Enhancement Mechanism for Surface detection

Nathan Du², Yanfang Meng¹

¹ School of Mechanical Engineering, Jiangsu University, 301# Xuefu Road, Jingkou District, Zhenjiang China, 200127 (Corresponding author) email: 1000006368@ujs.edu.cn;

² Artificial Intelligence center, Shanghai, China; 200127 Email: dunan131@163.com;

ABSTRACT:

Metal surface defect detection faces significant hurdles in modern intelligent manufacturing. Traditional image processing algorithms often struggle to identify complex and irregular defects, while relying solely on artificial intelligence algorithms face difficulties in achieving effective image preprocessing. This paper aims to explore an algorithm that integrates wavelet transform into deep vision to enhance detail features. Through the frequency domain decomposition of wavelet transform and attention mechanisms, high-frequency details are explicitly preserved and enhanced, making this approach particularly suitable for tasks requiring fine boundaries, such as object detection and image segmentation. The Haar wavelet achieves low computational complexity by using only integer addition and subtraction and can effectively improve model accuracy without sacrificing efficiency. The “dual-stream wavelet” is an important extension of the aforementioned Haar wavelet module, primarily reflected in two aspects: network architecture design and frequency-spatial feature fusion. While traditional wavelet modules only use Discrete Wavelet Transform Module (DWT) as a fixed preprocessing or down-sampling tool, the dual-stream wavelet significantly enhances feature representation capability through parallel architecture and learnable fusion. The Wavelet-Conv path effectively preserves details such as edges and textures by explicitly decomposing the image into LL, LH, HL, and HH sub-bands and using learnable weights to boost high-frequency components. Moreover, the standard down-sampling path learns abstract semantic features and fuses them with the former via concatenation, enabling complementary information exchange. In the dual-stream framework, the transformation from fixed wavelet bases and simple learnable weights to end-to-end adaptive fusion of two streams results in a frequency-aware, multi-receptive-field down-sampling module.

25 Particularly, the High-Freq-Enhance module enhances high-frequency details in feature maps through residual
26 learning. The residual connection ensures the fidelity of the original information, while the attention mechanism
27 determines, based on global information, which channels require high-frequency enhancement. Placed in the
28 deeper layers of the backbone, it compensates for high-frequency details attenuated by repeated down-sampling.

29 The core insight of this architecture is that down-sampling in traditional CNNs is "discard-based", whereas
30 wavelet-enabled down-sampling is "encoding-based" and "recoverable". Compared to standard YOLO, this
31 approach proposes three key improvements: (1) High-frequency information retention: Conventional down-
32 sampling discards high-frequency details, whereas the dual-stream wavelet transform encodes and preserves
33 them.(2) Multi-path feature fusion: Standard CNNs rely solely on single-path spatial features, while the
34 dual-stream design achieves effective multi-path information integration.(3) Small-target enhancement: After
35 repeated down-sampling, small-target signals are prone to being overwhelmed by noise; the proposed
36 high-frequency enhancement module selectively restores detailed responses to mitigate this issue.

37 **AUTHOR KEYWORDS:** YOLOv11; Adaptive High-Frequency Signal Enhancement; Metal surface defect detection

38 **INTRODUCTION:**

39 Metal defect detection is a routine procedure on industrial production lines, involving full or sampled
40 inspection of semi-finished goods to ensure process quality[1]. Specifically, classification and analysis the defects
41 are indispensable processes to formulate improvement implementation plans, thereby achieving a closed-loop
42 strategy for optimizing and iterating the entire automated production line.

43 Traditional vision algorithms rely on customized processing methods such as threshold segmentation, filtering,
44 and edge extraction. Despite being tailored to specific defect morphologies, these methods exhibit limited
45 generalization and struggle to cope with the coexistence of multiple defect types.. The parameter tuning
46 workload in later stages is heavy, and the algorithm robustness is poor. Deep learning algorithms can effectively
47 overcome the inherent limitations of traditional vision.

48 The core advantage of the YOLO (You Only Look Once) series is that it serves as an end-to-end deep learning
49 framework, systematically addressing these issues. Its superiority for industrial defect detection manifest in four
50 dimensions: global view, real-time efficiency, multi-scale fusion, and continuous iteration capability.

51 1.1 Core Advantage: Global View and "End-to-End" Detection

52 Traditional algorithms approaches typically rely on "sliding windows" or "handcrafted feature extraction"
53 methods, which are not only computationally inefficient but also prone to losing global information. In contrast,
54 YOLO's core innovation lies in treating object detection as a regression problem.

55 The working mechanism lies in treating object detection as a regression problem. Specifically, it divides the input
56 image into several grids. A single forward propagation simultaneously predicts bounding boxes, class
57 probabilities, and confidence scores. This "you only look once" mechanism endows the model with a large
58 receptive field. When detecting large-area scratches, deformations, or sparsely distributed imperfections on
59 metal surfaces, YOLO can better understand the contextual relationship between defects and the background,
60 significantly reducing false positives.

61 1.2 Real-time Advantage: Built for Industrial Production Lines

62 To meet stringent requirements for detection speed (FPS - Frames Per Second) in the process of industrial
63 production, the YOLO series have long been renowned for its balance of speed and accuracy [2-3].

64 Data Support: Lightweight variants (e.g., improved models based on YOLOv7) can achieve inference speeds
65 exceeding 90 FPS in metal surface detection tasks; optimized models for steel surfaces, including the
66 ASFRW-YOLO architecture can reach up to approximately 125 FPS on a GPU.

67 1.3 Technical Architecture Advantage: Multi-scale Fusion and Complex Texture Handling

68 Addressing challenges such as "sensitivity to illumination variations" and "significant background interference",
69 YOLO has effectively mitigated these challenges through architectural evolution:

70 1.3.1 Multi-scale Feature Fusion: Modern YOLO versions including YOLOv5, v8, v9, and beyond commonly
71 incorporate Feature Pyramid Networks (FPN) or BiFPN structures. This design solves the problem of large scale

72 variations in metal defects – enabling the detection of both large-scale anomalies (e.g."fold") defects and
73 extremely small detection (e.g"micro-cracks" or "pitting").

74 1.3.2 Attention Mechanism: By integrating attention modules like Convolutional Block Attention Module
75 (CBAM), ECA, or Transformer-style mechanisms, YOLO models can focus on defective regions, effectively
76 suppressing interference from complex backgrounds including metal brushed textures, oxide scale, or oil stains,
77 thereby enhancing detection robustness.

78 1.3.3 Anchor-Free Mechanism: The latest YOLO versions (e.g. v8, v10) adopt anchor-free detection heads. This
79 eliminates reliance on pre-set anchor box dimensions dimensions, allowing for better handling of defects with
80 extreme aspect ratios (e.g., slender scratches), which often required complex parameter tuning in traditional
81 algorithms.

82 1.4 Proposed Improvement: YOLOv11-AHFE

83 This paper proposes YOLOv11-AHFE, an improved detection framework that combines YOLOv11 with adaptive
84 wavelet transform. This strategically-designed approach aims to retain the generalization advantages of the YOLO
85 algorithm while improving the precision of edge detail feature extraction.

86 1.4.1 Wavelet transform

87 Wavelet analysis is a time-frequency domain analysis of radiofrequency ultrasound signals using a
88 mathematical model to assess local changes in the geometrical profile of time-series signals thereby enabling the
89 identification of plaque components [4].Wavelets offer distinct advantages of multi-resolution properties and
90 flexibility in the choice of basis functions.. More than three decades ago, Morlet introduced a novel framework
91 for wavelet analysis that achieves an adaptive trade-off between time and frequency resolution automatically[5],
92 and this proposal concept became the generation of ideas promoted by Haar in 1910 and Gabor in 1946[6], and
93 later in 1976, Zweig further developed the field based on Morlet's principles [7].

94 The wavelet basis representation generates a set of wavelet coefficients organized at multiple resolution levels.
95 Each parameter is associated with a specific level of accuracy and a corresponding point in the time domain. At

the lower resolution level, the coefficients describe coarser frequency features of data spanning wide time periods, while at the higher resolution level, the coefficients are associated with highly localized high-frequency features [8].

In the improved architecture, Wavelet transform is employed for feature enhancement by first discriminating high-frequency and low-frequency components without sacrificing spatial resolution, these components are then fed into different convolutional branches, and subsequently adaptively fuse them, which significantly improves sensitivity to weak edges, scratches, and cross-scale defects. Replacing the original convolutional modules with wavelet down-sampling reduces information aliasing and improves detection accuracy.

I. DWT (Discrete Wavelet Transform Module)

The discrete Wavelet transform (DWT) is considered one of the most powerful tools for image processing in both time and frequency domains which gives a better analysis of image localization features. The term wavelet was first presented in the early 1980s by Grossman and Morlett. The researchers used the French word "ondelette", as a representation of a small wave [9]. In DWT-based decomposition, one sub-band (LL) is for image zoom, and the other three sub-bands (LH, HL and HH) dedicated to highlighting the image edges [10].

The goals of high-frequency feature enhancement can be summarized in three aspects: (1) Amplitude enhancement: Increase the energy of high-frequency coefficients to make details more prominent; (2) Signal-noise separation: Distinguish "effective details" from "noise," enhancing the former while suppressing the latter; (3) Structure fidelity: Maintain edge integrity and naturalness during enhancement, avoiding artifacts.

II. Wavelet-Conv (Wavelet Convolution Module)

Combines wavelet transform and convolution to achieve high-frequency enhanced down-sampling.

III. WaveletC2f (Dual-Stream Wavelet C2f Module)

This module parallelly fuses the wavelet path and the standard down-sampling path to achieve multi-scale feature complementarity. Specifically, custom modules are embedded into the YOLO11 backbone network to achieve high-frequency information enhancement before the C3K2 module.

IV. High-Freq-Enhance (High-Frequency Feature Enhancement Module)

A custom module is embedded within the network to enhance high-frequency details in feature maps through residual learning while simultaneously introducing an attention mechanism to compensate for high-frequency signal loss caused by down-sampling.

1.4.2 CBAM (Convolutional Block Attention Module)

As a plug-and-play attention module, CBAM improves the feature representation capability and model performance. By enabling the neural network to focus on both "which feature channels are more important" and "which spatial locations in the image are more important." The CBAM module consists of two sequentially connected sub-modules:

I. Channel Attention Module

This module applies both global average pooling and global max pooling to the input feature map. The two pooled results are then fed into results into the same shared fully connected network (implemented with 1×1 convolutions). The outputs are element-wise added and passed through a Sigmoid function to obtain the weight for each feature channel.

Role: Helps the network learn "which types of features (channels) are more useful," for example, suppressing channels related to irrelevant backgrounds and enhancing those of target objects.

II. Spatial Attention Module

After channel attention, feature map is processed along the channel dimension by computing both the average and the maximum values, generating two 2D feature maps. These are concatenated into a 2-channel feature map, then reduces it to 1 channel using a 7×7 convolution. A Sigmoid function is finally applied to produce spatial location weights.

Role: Directs the network to focus on "which spatial locations in the image are more important," for example, attending to the region containing the target object while ignoring the background.

143 The input feature first undergoes channel recalibration via channel attention, followed by spatial pixel location
144 recalibration via spatial attention. The final output is the element-wise multiplication of the original features with
145 the two attention weight maps. This mechanism, thereby enables the network to focus more on critical features,
146 thereby significantly improving performance on tasks such as classification, detection, and segmentation, with
147 minimal computational overhead.

148 In all, this work demonstrates these improvement.(1)High-frequency information retention: Conventional
149 down-sampling discards high-frequency details, whereas the dual-stream wavelet transform encodes and
150 preserves them. (2) Multi-path feature fusion: Standard CNNs rely solely on single-path spatial features, while the
151 dual-stream design achieves effective multi-path information integration.(3) Small-target enhancement: After
152 repeated down-sampling, small-target signals are prone to being overwhelmed by noise; the proposed high-
153 frequency enhancement module selectively restores detailed responses to mitigate this issue .

154

155 **2 Methodology**

156 Current mainstream deep learning models for visual representation, such as convolutional neural networks and
157 Vision Transformer architectures, operate exclusively operation within the spatial domain, lacking the capability
158 to directly assess the structural characteristics of images in the frequency domain. To overcome this inherent
159 shortcoming, research efforts have progressively integrated frequency-domain representations into deep learning
160 frameworks[11].

161 2.1 Overall Network Architecture of YOLO11

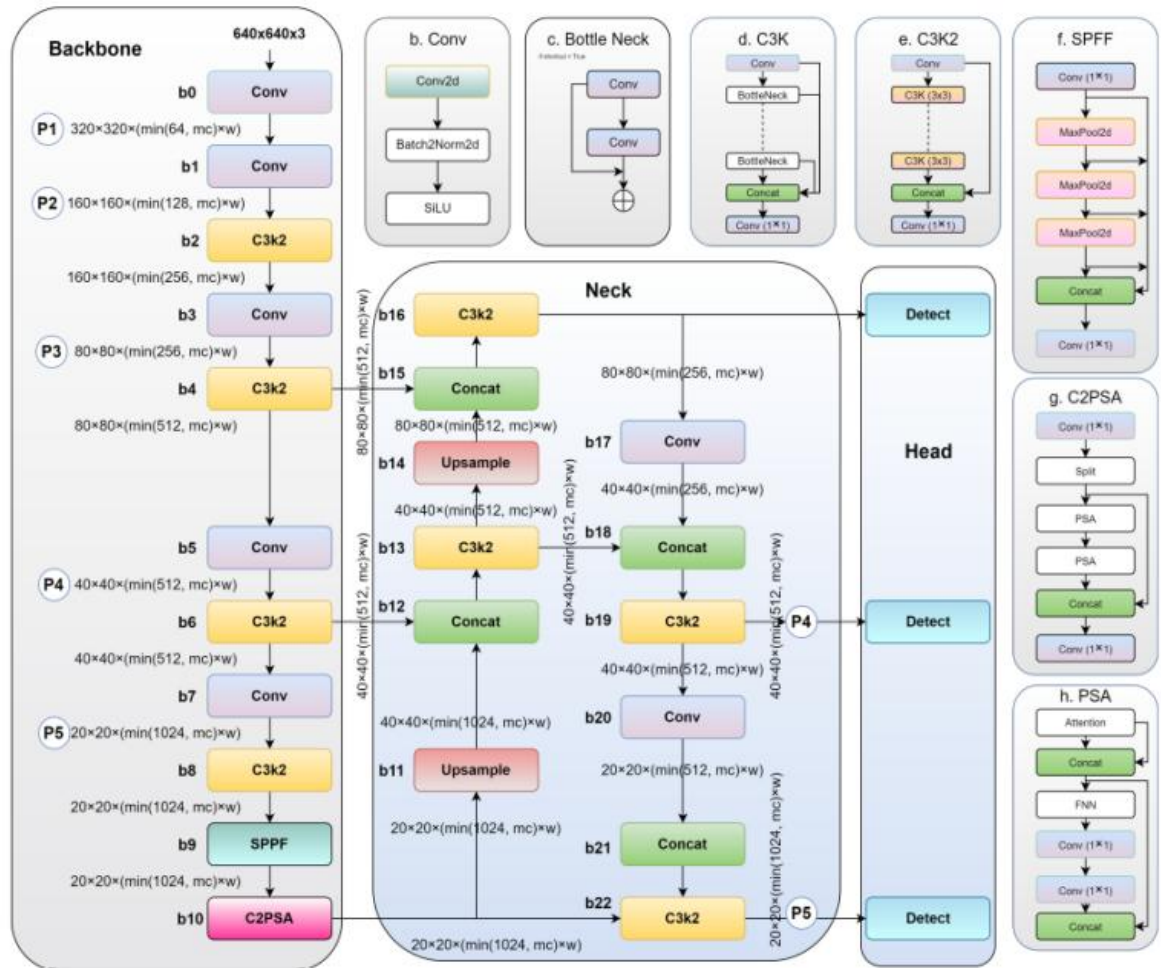


Figure 1. YOLOv11 architecture [12]

YOLO11 adopts the single-stage detection design characteristic of the YOLO series, and its network structure remains clearly three core components: Backbone, Neck, and Head (**Figure1**) [12].

2.1.1 Backbone: Feature Extractor

The backbone is a critical component of the YOLO architecture, responsible for extracting features from the input image at multiple scales. This process involves stacking convolutional layers and specialized blocks to generate feature maps at various resolutions [13].

Core Module C3k2: YOLOv11 employs C3K2 blocks to handle feature extraction at different stages of the backbone. The adoption of smaller 3x3 kernels allow for more efficient computation while retaining the model's ability to capture essential features in the image[14].Efficient Pooling SPPF: Following YOLOv8, YOLO11 continues to use the Fast Spatial Pyramid Pooling module. Instead of using parallel large-kernel pooling with serial small-

174 kernel pooling, significantly improving running speed while maintaining an equivalent receptive field.Attention
175 Mechanism C2PSA: This module is newly introduced at the end of the backbone network. It employs a parallel
176 spatial attention mechanism, allowing the model to focus more on important regions within image, which is
177 particularly helpful for detecting small or occluded objects.

178 2.1.2 Neck Network

179 The neck network acts as a "bridge," responsible for efficient multi-scale feature fusion.

180 **Decoupled Structure:** Continuing the design of modern YOLO versions, classification and regression tasks use two parallel convolutional
181 branches, allowing the model to focus better on each task. **Depth-wise Separable Convolution:** Introduced in the classification
182 branch, Depth-wise Separable Convolution effectively reduces computational load while maintaining high accuracy, making it highly
183 deployment-friendly for deployment on edge devices. **Multi-task Capability:** Beyond basic object detection, YOLO11 has spawned variants
184 for tasks such as instance segmentation and pose estimation. For example, its segmentation version's detection head includes an
185 additional mask generation branch to handle pixel-level classification tasks. **Lightweight Design:** During feature fusion, YOLO11 also utilizes
186 modules like C3k2 to control computational overhead while ensuring fusion effectiveness.

187 2.1.3 Detection Head: Decoupled Design and Task Expansion

188 The detection head of YOLOv11 adopts a decoupled design, which separates the classification and regression tasks into different
189 branches to avoid feature conflicts and improve detection accuracy. Furthermore, it also employs an anchor-free strategy, directly
190 predicting key points instead of using preset anchor boxes, thereby simplifying model complexity.

191 **Anchor-Free Mechanism:** Unlike early YOLO versions that relied on preset anchor boxes, the anchor-free mechanism enables the
192 model to directly regress object locations, allowing for more flexible adaptation to objects of varying sizes and shapes.

193 **Multi-scale Prediction:** As a classic design in the YOLO series for handling variations in object sizes, YOLOv11 inherits this strategy by
194 deriving three detection heads from the P3, P4, and P5 layers of the feature pyramid network (FPN), which are responsible for predicting
195 small, medium, and large objects respectively, thereby ensuring model robustness across different scenarios.

196 **C3k2 Module within the Detection Head:** The C3k2 module is also embedded inside the detection head to refine and enhance the
197 feature maps received from the neck network. This design inherits the efficient feature processing capability of the backbone network,
198 further strengthening the accuracy of the final predictions.

199 Overall, the detection head of YOLOv11 is an efficient, decoupled, and multi-scale prediction structure that balances lightweight design
200 with high accuracy. Among its features, the separation of classification and regression tasks, along with the anchor-free design, are the key
201 improvements that enhance performance.

202 **2.1.4 Version selection**

203 YOLOv11 derives five versions (**Table 1**) with increasing parameter scales—Nano (n), Small (s), Medium (m), Large (l), and X-Large (x)—
204 by adjusting the network's depth (number of layers), width (number of channels), and input resolution. Given the limited GPU memory
205 available on the author's validation platform, the **YOLO11-n** model was selected for training, and this selection does not compromise the
206 comparative experimental results in any way.

207 Table 1 The five versions of YOLOv11

Model	Params	Model size	Use Cases
YOLO11n	2.6M	5.2M	Mobile devices, edge devices, ultra-low-latency real-time applications.
YOLO11s	20.0M	38.7M	Resource-constrained scenarios requiring a balance between speed and accuracy.
YOLO11m	25.3M	48.9M	The "sweet spot" for general-purpose use; offers excellent balance.
YOLO11l	25.3M	48.9M	Server-side deployment, professional applications with higher accuracy demands.
YOLO11x	56.8M	109.1M	Cloud or data center deployment, where maximum accuracy is prioritized over speed.

208

209 **2.2 Overall Network Architecture of YOLO11_AHFE**

210 Based on the YOLO11 network architecture, a dual-stream WaveletC2f module is inserted into the network backbone to enhance
211 detailed feature extraction. Additionally, three High-Freq-Enhance modules are added to the network neck to compensate for high-
212 frequency detail information lost due to multiple down-sampling operations, and to adaptively enhance fine structural features such as
213 edges and textures. Finally, a CBAM module is incorporated at the end of the network to focus on key semantic regions while suppressing
214 background or noise information (As shown in **Figure2**).

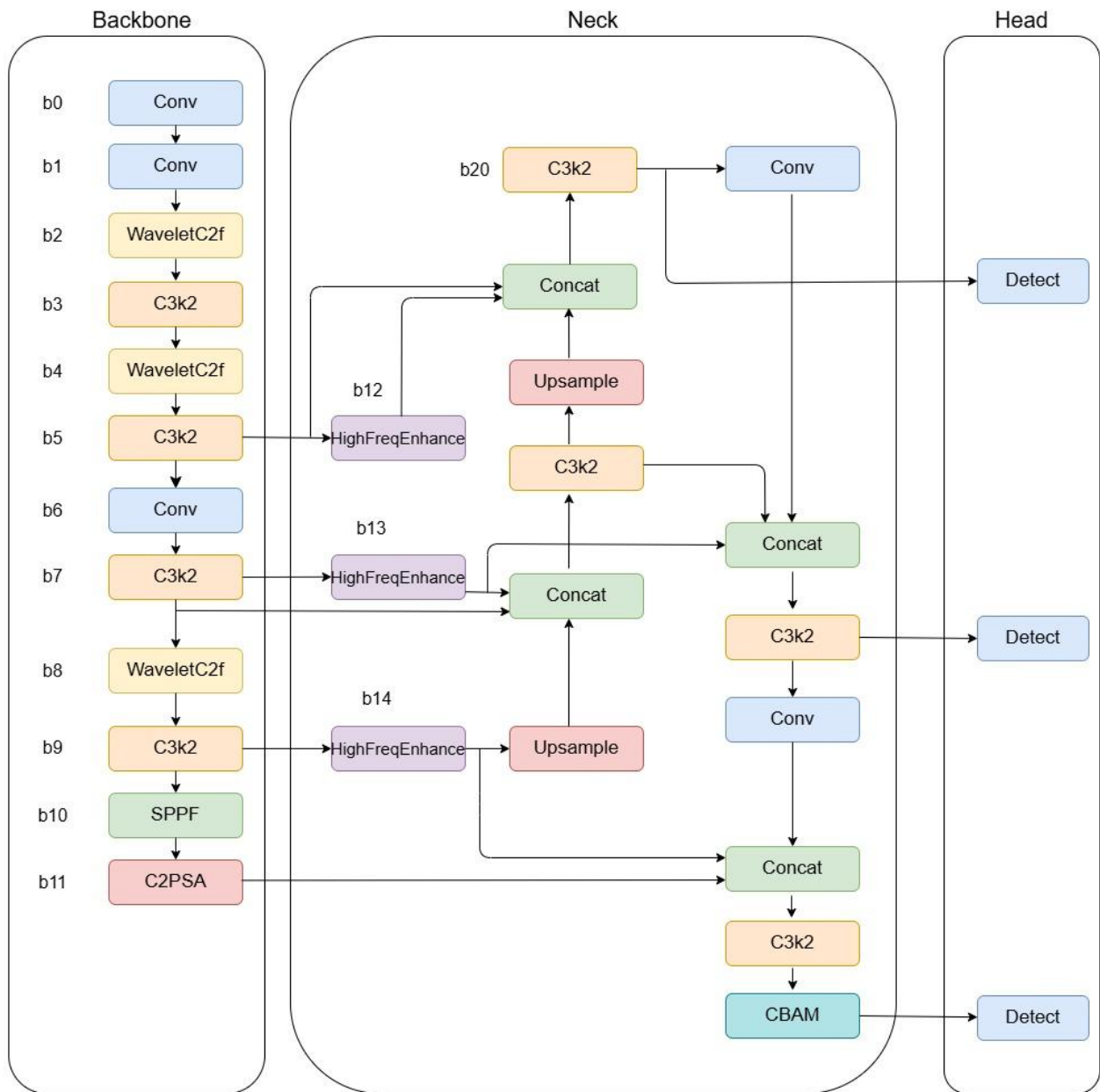


Figure 2. YOLOv11_AHFE architecture

2.2.1 DWT module

The two-dimensional Haar basis decomposes the image into four subspaces:

• LL (Approximation sub-band): Contains low frequencies within the level and vertical directions. It represents the coarser approximation of the original image and preserves the significant features and energy of the image.

• LH (Horizontal detail sub-band): Contains low frequencies within the even heading and high- frequencies within the vertical direction. It highlights the vertical edges and details of the original image.

• HL (Vertical detail sub-band): Contains high frequencies within the level direction and low frequencies within the vertical direction. It highlights the horizontal edges and details of the original image [15].

• HH (Diagonal detail sub-band): Contains high frequencies in both the horizontal and vertical directions. It highlights the diagonal edges and fine details of the image. All the sub-bands have Size of $M/2 \times N/2$. (For an image size $(M \times N)$ [10].

The discrete wavelet transform is then performed by projecting the discrete pixel values onto the Haar basis functions via the discrete inner product, as illustrated in **Table 2**:

Table 2 The functional relationship between sub-band and discrete computation

Sub-band	Discrete computation
LL (Approximation sub-band)	$(a + b + c + d) / 4$
LH (Horizontal detail sub-band)	$(a - b + c - d) / 4$
HL (Vertical detail sub-band)	$(a + b - c - d) / 4$
HH (Diagonal detail sub-band)	$(a - b - c + d) / 4$

2.2.2 Wavelet-Conv module encapsulates DWT into an end-to-end learnable neural network layer, adding sub-band weighting, channel transformation, and nonlinear processing, enables the network to learn frequency-specific features and dynamically emphasize or suppress the importance of each sub-bands.

2.2.3 WaveletC2f module

Multi-branch architecture is a common and effective approach in the design of deep learning networks [16-20].

WaveletC2f module adopts a dual-stream parallel architecture: one path preserves high-frequency details through Wavelet-Conv, while the other extracts semantic features via standard convolution. The outputs of the two paths are concatenated along the channel dimension and then adaptively fused through a 1×1 convolution, achieving complementary integration of frequency-domain and spatial-domain information during down-sampling (**Figure 3**).

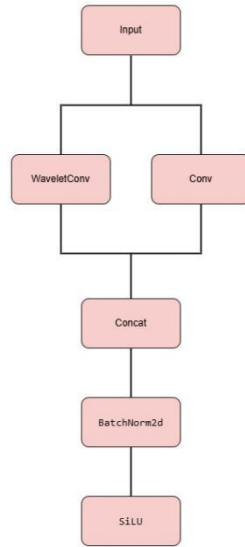


Figure 3. WaveletC2f architecture

2.2.4 High-Freq-Enhance module

The model enhance high-frequency details in the feature map through residual learning (Figure 4)..

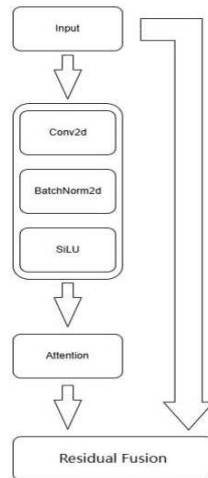


Figure 4. High-Freq-Enhance architecture

2.2.5 CBAM module

The significance of attention has been extensively investigated in the previous literature [21-26]. Attention Module (CBAM), a simple yet effective attention module for feed-forward convolutional neural networks. Given an intermediate feature map, our module sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps

are multiplied to the input feature map for adaptive feature refinement. Because CBAM is a lightweight and general module, it can be integrated into any CNN architectures seamlessly with negligible overheads and is end-to-end trainable together with base CNNs[27] (Figure 5)..

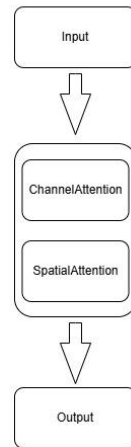


Figure 5.CBAM architecture

3 Experimental Evaluation

3.1 Implementation Environment

The proposed method was implemented using the Pycharm integrated development environment. Model training was conducted on windows system equipped with one NVIDIA GeForce RTX 3050 GPUs, with 6 GB of memory. The software environment included Python 3.12, PyTorch 2.5.1, CUDA 12.6, and cuDNN 9.0.

3.2 Datasets

This study evaluates the proposed method on the GC10-DET surface defect dataset [28].

GC10-DET Dataset: Collected from real industrial production environments, the GC10-DET dataset contains ten representative defect categories that reflect the complexity and diversity of practical inspection scenarios: punching holes (PH), welding lines (WL), crescent gaps (CG), water spots (WS), oil spots (OS), silk spots (SS), inclusions (In), roll pits (RP), creases (Cr),

and waist folding (WF). GC10-DET provides bounding-box annotations to enable objective and quantitative evaluation. Example images and annotations are presented in Figure7.

3.3 Evaluation Metrics

The comprehensive evaluation and quantitative comparisons have been implemented concerning detection accuracy, localization precision, and inference efficiency, with primary evaluation metrics Average Precision (mAP), mAP@0.50 (mAP50), mAP@0.75 (mAP75), and Frames Per Second (FPS), respectively (Figure 6).

Specifically, mAP is derived from the computation of averaging precision values over multiple IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05, thereby reflecting both detection accuracy and lo-calization quality. mAP50 evaluates detection performance at an IoU threshold of 0.50 and placing greater emphasis on recall, whereas mAP75 adopts a stricter IoU threshold of 0.75 and is more sensitive to bounding-box alignment accuracy. FPS evaluates inference efficiency by counting the number of images processed per second; higher FPS values indicate better suitability for real-time industrial deployment (Figure 7).

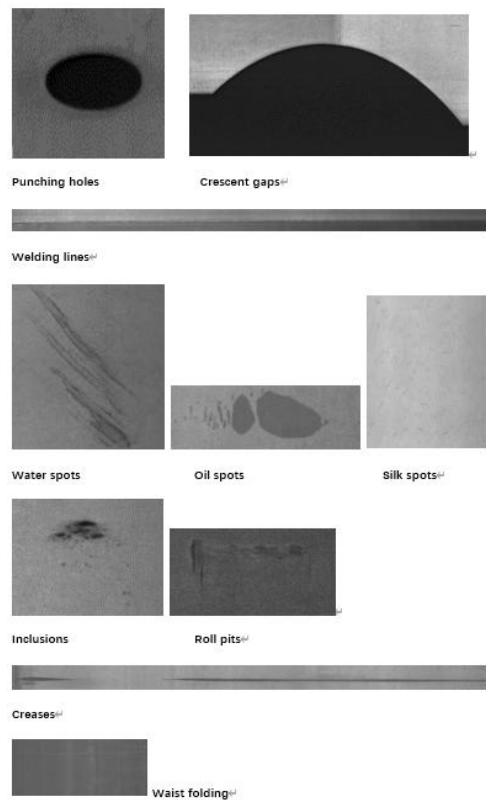


Figure6.Example images of ten typical defect categories in the GC10-DET dataset

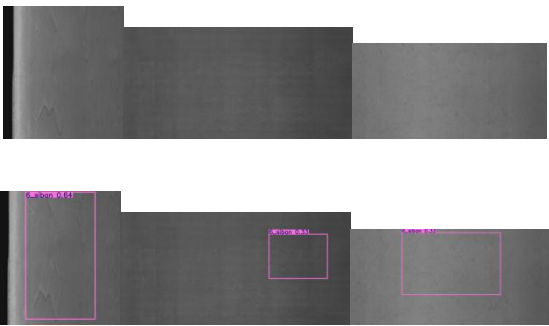


Figure7.Visual comparison of detection results between YOLOv11 and YOLOv11-AHFE on the GC10-DET dataset

This dataset suffers from significant class imbalance (e.g., silk spot category contains 884 samples, while rolled pit has only 85 samples), low contrast between defects and background, large intra-class variation, and complex morphological characteristics. For instance, inclusions can appear in various forms, including small spots, fish-scale patterns, strips, and blocks. These challenging factors place high demands on the generalization capability of detection models.

As shown in **Figure 8**, **Figure 9** and **Figure 10**: compared with YOLOv11, the proposed YOLOv11-AHFE exhibits fewer missed detections, and achieves significant improvements in detection accuracy for four defect categories in detection accuracy are achieved for the four defect categories: welding lines, inclusions, roll pits, and creases. These results demonstrate the effectiveness of enhancing high-frequency feature extraction in improving detection performance (**Table 3**).

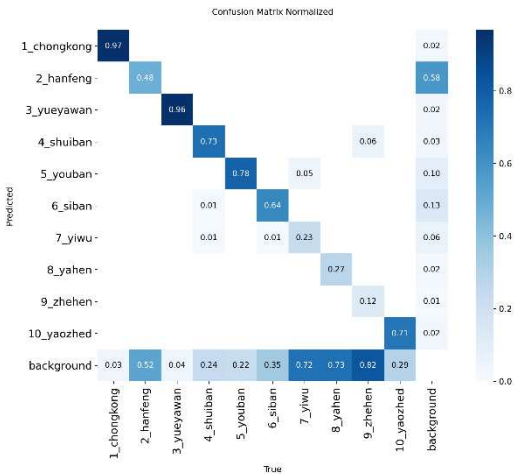


Figure 8. Confusion Matrix of YOLO11

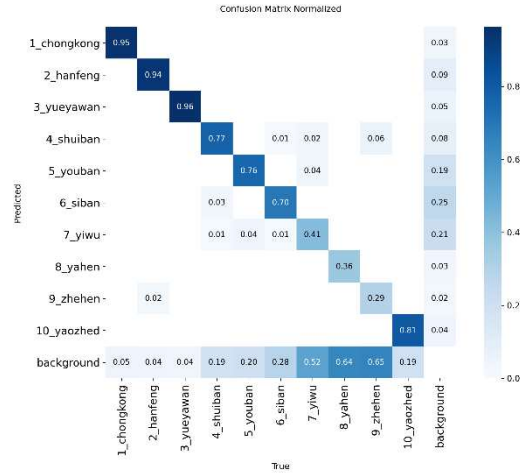


Figure 9. Confusion Matrix of YOLO11-AHFE

Table3 Quantitative comparisons of different detection methods on the GC10-DET dataset.

Model	mAP50-95	mAP50
YOLO11	0.309	0.607
YOLO11_AHFE	0.352	0.672

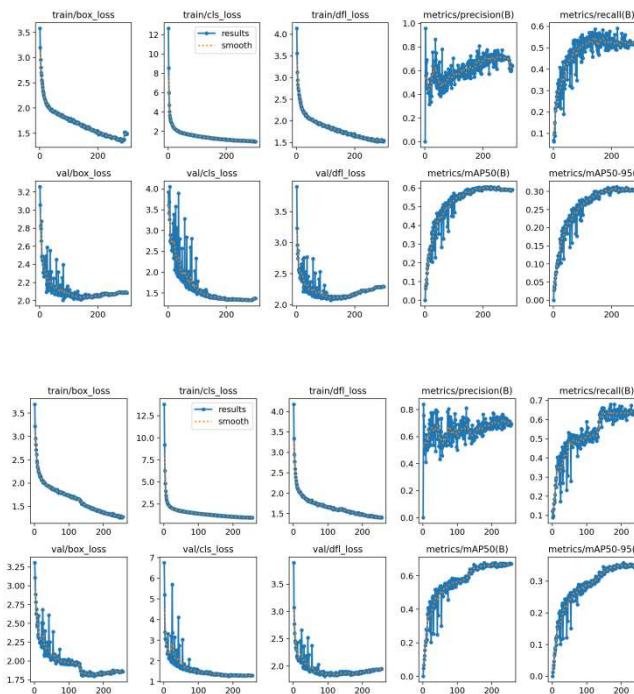


Figure 10 YOLO11 and YOLO11-AHFE training results

298

299 **CONCLUSIONS:**

300 This paper proposes YOLOv11-AHFE, a detection framework strategically designed to address the
301 challenges existing in small-scale, low-contrast, and morphologically irregular surface defects in metals.
302 The proposed method integrates the WaveletC2f module into the network backbone to capture high-
303 frequency detail information, and introduces the High-Freq-Enhance module in the neck to selectively
304 transmit cross-scale feature signals. In addition, the CBAM module is incorporated before the detection
305 head to enhance feature representation capability and overall model performance.

306 Experimental results on the GC10-DET dataset demonstrate that the proposed method achieves
307 significant performance improvements over the baseline model YOLO11 in terms of mAP@0.5 and
308 mAP@0.5:0.95, while maintaining competitive inference speed. These results validate the effectiveness
309 and deployability of YOLOv11-AHFE in real-world industrial defect inspection scenarios.

310

311 **DATA AVAILABILITY STATEMENT:**

312 All data, models, or code that support the findings of this study are available from the corresponding
313 author upon reasonable request.

314 **FUNDING:**

315 No funding was received for this manuscript.

316

317 **REFERENCES:**

318 1 Guo, Y., Guo, S., & Fu, J. Research on a novel heat treatment process for boron steel used for soil-
319 engaging components of tillage machinery. *Agriculture-Basel*, 14(9), 1555, (2024).

320 2 Ma, Z., Yang, S., & Qi, J. Research on SLAM localization algorithm for orchard dynamic vision based on
321 YOLOD-SLAM2. *Agriculture-Basel*, 14(1), 9, (2024).

322 3 Deng, L., Miao, Z., Zhao, X., Yang, S., Gao, Y., Zhai, C., & Zhao, C. HAD-YOLO: An accurate and effective
323 weed detection model based on improved YOLOV5 network. *Agronomy*, 15(1), 57, (2025).

324 4 Grant, A. D., Upton, T. J., Terry, J. R., Smarr, B. L., & Zavala, E. Analysis of wearable time. (n.d.).
325 [Manuscript in preparation].

326 5 Li, B., & Chen, X. Wavelet-based numerical analysis: A review and classification. *Finite Elements in*
327 *Analysis and Design*, 81, 14–31, (2014).

328 6 Dahmen, W. Wavelet methods for PDEs—Some recent developments. *Journal of Computational and*
329 *Applied Mathematics*, 128(1-2), 133–185, (2001).

330 7 Canuto, C., Tabacco, A., & Urban, K. The wavelet element method part II: Realization and additional
331 features in 2D and 3D. *Applied and Computational Harmonic Analysis*, 8(2), 123–165, (2000).

332 8 Chui, C. K. An introduction to wavelets (2nd ed.). Academic Press, (1992).

333 9 Grossmann, A., & Morlet, J. Decomposition of Hardy functions into square integrable wavelets of
334 constant shape. *SIAM Journal on Mathematical Analysis*, 15(4), 723–736, (1984).

335 10 Alickovic, E., Kevric, J., & Subasi, A. Performance evaluation of empirical mode decomposition,
336 discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure
337 detection and prediction. *Biomedical Signal Processing and Control*, 39, 94–102, (2018).

- 338 11 Aslan, M. F., Aslan, B., & Sabanci, K. Frequency-domain vision transformers: Architectures,
339 applications, and open challenges. *Applied Sciences*, 16(4), 2024, (2026).
- 340 12 Rasheed, A. F., & Zarkoosh, M. YOLOv11 optimization for efficient resource utilization. *The Journal of*
341 *Supercomputing*, 81, 1085, (2025).
- 342 13 Khanam, R., & Hussain, M. YOLOv11: An overview of the key architectural enhancements. (n.d.).
343 [Manuscript submitted for publication].
- 344 14 Sulake, N. R. YOLOv11 demystified: A practical guide to high-performance object detection.
345 [Technical report].
- 346 15 Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., & Zebari, D. A. Machine learning and region growing
347 for breast cancer segmentation. In *Proceedings of the International Conference on Advanced Science*
348 *and Engineering (ICOASE)*, Duhok, Iraq, 88, (2019).
- 349 16 Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. CSPNet: A new backbone
350 that can enhance learning capability of CNN. [Conference paper].
- 351 17 Gu, H., Fu, G., Li, J., & Zhu, J. Auto-ReID plus: Searching for a multi-branch ConvNet for person re-
352 identification. *Neurocomputing*, 435, 53–66, (2021).
- 353 18 Ma, H. M., & Wang, Y. Parallel feature extraction system. *Chinese Journal of Electronics*, 12(4), 598–
354 601, (2003).
- 355 19 Li, G., Zhang, M., Li, J., Lv, F., & Tong, G. Efficient densely connected convolutional neural networks.
356 *Pattern Recognition*, 109, 107610, (2021).

- 20 Zhu, H. B., & Shibata, T. A real-time motion-feature-extraction VLSI employing digital-pixel-sensor-based parallel architecture. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(10), 1787–1799, (2014).
- 21 Mnih, V., Heess, N., Graves, A., et al. Recurrent models of visual attention. In *Proceedings of the Neural Information Processing Systems (NIPS)*. (n.d.).
- 22 Lee, S., Kim, K., Kim, J.-Y., Kim, M., & Yoo, H.-J. Familiarity based unified visual attention model for fast and robust object recognition. *Pattern Recognition*, 43(3), 1116–1128, (2010).
- 23 Bahdanau, D., Cho, K., & Bengio, Y. Neural machine translation by jointly learning to align and translate. (n.d.). [Conference paper].
- 24 Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. (n.d.). [Conference paper].
- 25 Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. DRAW: A recurrent neural network for image generation. [Conference paper].
- 26 Zhang, C., Chen, X.-P., Han, G.-Q., & Liu, X.-J. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6), e13244, (2023).
- 27 Wu, Y., Liang, W., Fang, J., Zhou, C., & Sun, X. VTC-Net: A semantic segmentation network for ore particles integrating transformer and convolutional block attention module (CBAM). *Sensors*, 26(4), 787, (2026).
- 28 Lv, X., Duan, F., Jiang, J.-j., Fu, X., & Gan, L. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6), 1562, (2020).
- TABLES:**

Table 1 The five versions of YOLOv11

Model	Params	Model size	Use Cases
YOLO11n	2.6M	5.2M	Mobile devices, edge devices, ultra-low-latency real-time applications.
YOLO11s	20.0M	38.7M	Resource-constrained scenarios requiring a balance between speed and
YOLO11m	25.3M	48.9M	The "sweet spot" for general-purpose use; offers excellent balance.
YOLO11l	25.3M	48.9M	Server-side deployment, professional applications with higher accuracy
YOLO11x	56.8M	109.1M	Cloud or data center deployment, where maximum accuracy is prioritized

378

379

Table 2 The functional relationship between sub-band and discrete computation

Sub-band	Discrete computation
LL (Approximation sub-band)	$(a + b + c + d) / 4$
LH (Horizontal detail sub-band)	$(a - b + c - d) / 4$
HL (Vertical detail sub-band)	$(a + b - c - d) / 4$
HH (Diagonal detail sub-band)	$(a - b - c + d) / 4$

380

381

Table3 Quantitative comparisons of different detection methods on the GC10-DET dataset.

Model	mAP50-95	mAP50
YOLO11	0.309	0.607
YOLO11_AHFE	0.352	0.672

382

FIGUREURE CAPTION LIST:

Figure 1. YOLOv11 architecture

Figure 2. YOLOv11_AHFE architecture

Figure 3. WaveletC2f architecture

387 Figure 4. High-Freq-Enhance architecture

388 Figure 5.CBAM architecture

389 Figure 6.Example images of ten typical defect categories in the GC10-DET dataset

390 Figure 7.Visual comparison of detection results between YOLOv11 and YOLOv11-AHFE on the GC10-
391 DET dataset

392 Figure 8. Confusion Matrix of YOLO11

393 Figure 9. Confusion Matrix of YOLO11-AHFE

394 Figure10. YOLO11 and YOLO11-AHFE training results

395