ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 380-401
DOI 10.5755/j01.itc.55.2.43872

An Improved YOLOv12 Algorithm with Multi-Stage Transfer Learning for Real-Time Wood Defect Detection

Received 2025/12/18 | Accepted after revision 2026/02/24

HOW TO CITE: He, S. (2026). An Improved YOLOv12 Algorithm with Multi-Stage Transfer Learning for Real-Time Wood Defect Detection. *Information Technology and Control*, 55(2), 380-401. https://doi.org/10.5755/j01.itc.55.2.43872

An Improved YOLOv12 Algorithm with Multi-Stage Transfer Learning for Real-Time Wood Defect Detection

Shidu He
College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, China

Corresponding author: hsd@hrbeu.edu.cn

Exploring efficient, precise, and cross-domain adaptable wood defect detection algorithms is essential for advancing automation and ensuring reliability in industrial quality inspection. To overcome the limitations of existing approaches, including difficulties in modeling complex defect morphologies, handling large-scale variations, and addressing data scarcity, this paper presents an enhanced solution that integrates the YOLOv12-SMD algorithm with a multi-stage transfer learning paradigm. Specifically, considering the shortcomings of YOLOv12 in aspect ratio modeling, feature representation, and information retention under fixed interpolation sampling, three improvements are introduced. Firstly, the Shape-Intersection over Union (Shape-IoU) loss function enhances sensitivity to variations in width, height, and center offsets. Secondly, the Mixed Local-Channel Attention (MLCA) module strikes a balance between global dependency modeling and local saliency extraction. Thirdly, dynamic upsampling (DySample) alleviates the loss of tiny details. To overcome the limitations of small sample sizes and distributional discrepancies across domain settings, a multi-stage transfer learning strategy is employed to achieve feature alignment and enhance generalisation. Experimental results demonstrate that on preprocessed source domain datasets, YOLOv12-SMD increases the baseline mean Average Precision at IoU of 0.5 (mAP@50) from 77.9% to 80.8%, outperforming other sophisticated detectors. On the two target domain datasets, the incorporation of multi-stage transfer learning paradigm improves mAP@50 by 4.8% and 9.4%, respectively, while reducing training time by 16.7% and 34.2%. Multiple runs confirm the stability of these improvements. In conclusion, the proposed paradigm strikes an optimal balance of accuracy, efficiency, and robustness across various domains, providing an effective and intelligent solution for wood defect detection in industrial quality inspection.

KEYWORDS: Wood defect detection, Transfer learning, YOLOv12, Shape-IoU, Mixed Local Channel Attention

1. Introduction

Wood, as a vital renewable structural and decorative material, plays a critical role in guaranteeing safety, economic effectiveness, and standardization in construction, furniture manufacturing, and prefabricated production [30]. Common defects include blue stain, cracks, dead knots, missing knots, and resin pockets [8], which arise from growth environments, pest infestations, and processing techniques. These defects not only reduce load-bearing capacity and cause machining failures but also lead to inaccurate grading and resource waste. Traditionally, wood defect inspection has relied on manual visual examination. However, the complex wood grain structures, characterized by strong anisotropy, illumination variability, and moisture-induced texture changes, significantly increase inspection difficulty. Consequently, manual inspection depends heavily on individual experience and subjective judgment, resulting in low efficiency, inconsistent accuracy, and poor reproducibility. These inherent limitations render traditional manual inspection inadequate for large-scale and high-throughput industrial quality control, thereby necessitating the development of accurate, robust, and automated wood defect detection methods[29].

With advances in deep learning, wood defect detection has attracted growing research interest, and data-driven object detection algorithms have been widely adopted to address challenges related to complex textures, scale variation, and irregular defect geometries [31]. From a methodological perspective, these algorithms are generally categorized into one-stage [17] and two-stage [16] detection frameworks. Two-stage methods follow a decoupled detection paradigm, in which region proposal generation is first performed, followed by fine-grained classification and bounding box regression. Owing to their strong representation capacity and precise localization, such methods are well known for achieving high detection accuracy. Representative approaches include the Faster Region-based Convolutional Neural Network (Faster R-CNN) [18] and Mask R-CNN [9], both of which have been extensively applied to wood defect detection tasks. Zou et al. [35] enhanced Faster R-CNN by integrating an improved ResNet-50 backbone, Focal Loss, and Soft Non-Maximum Suppression (Soft-NMS), resulting in a 6.76% accuracy improvement on a large-scale wood defect dataset. Shi et al. [19] combined neural architecture search with a multi-channel Mask R-CNN framework, enabling high-precision defect detection through preselection, feature fusion, and instance segmentation, and achieved 98.7% classification accuracy and 95.31% mean Average Precision (mAP) on an industrial veneer dataset while improving detection speed. Li et al. [13] proposed a generative data augmentation framework based on CycleGAN coupled with a hierarchical deformable Mask R-CNN, which alleviated class imbalance and enhanced the detection and segmentation of irregular defects such as cracks and insect holes, achieving 84.4% detection accuracy and 82.8% segmentation accuracy at an Intersection over Union (IoU) threshold of 0.5. Despite their superior accuracy and detailed instance-level representation, two-stage detection methods are often constrained by high computational complexity, limited real-time performance, and insufficient recall for small-scale or elongated defects, which restricts their applicability in high-throughput industrial inspection scenarios.

In contrast to two-stage detectors, one-stage detection algorithms adopt an end-to-end regression paradigm that directly predicts object categories and bounding boxes, thereby significantly improving inference speed and deployment efficiency. Among them, the YOLO (You Only Look Once) series has become one of the most representative frameworks and has been extensively applied in wood defect detection tasks. Benefiting from unified feature extraction and prediction, YOLO-based methods are particularly suitable for high-throughput industrial inspection scenarios that require real-time performance and lightweight deployment. Lopes et al. [15] applied YOLOv3 to the Southern Pine knot dataset and achieved an mAP at IoU 0.5 (mAP@0.5) of 0.80, demonstrating the feasibility of end-to-end detection frameworks for wood defect recognition. Fang et al. [7] employed YOLOv5 for automated knot detection in sawn timber and, by incorporating Mosaic data augmentation and the GIoU loss function, achieved F1 scores of 91.7% and 97.7% on two public datasets, outperforming both YOLOv3-

SPP and Faster R-CNN. An et al. [2] proposed the CWB-YOLOv8 algorithm and reported a 3.5% improvement in mAP@50 over the original YOLOv8 on a multi-species, multi-defect dataset, although the enhanced feature extraction capability was accompanied by increased inference latency. Kang et al. [12] introduced CFIS-YOLO, which strengthened multi-scale feature fusion and tiny-object localization, achieving an inference speed of 135 frames per second on edge devices together with a 4% accuracy gain. Jiang et al. [11] presented YOLOv11n-CBP for wood crack detection, improving mAP@50 from 70.6% to 86.4% while reducing the model size to 4.2 MB and reaching 416 FPS, thereby achieving a favorable balance between efficiency and accuracy. Despite these advances, existing YOLO-based methods still encounter notable challenges in wood defect detection scenarios. On the one hand, complex defect morphologies, extreme aspect ratios, and strong texture interference pose difficulties for accurate localization and robust feature representation, particularly for small-scale or elongated defects. On the other hand, industrial inspection environments are characterized by continuously evolving defect patterns and data distributions, which limits the efficiency and practicality of repeatedly training models from scratch. These challenges highlight the need for more effective strategies to enhance adaptability and robustness when applying YOLO-based detectors to real-world wood defect inspection tasks.

To address the challenges of limited labeled data and continuously evolving defect distributions in industrial inspection, the integration of transfer learning with YOLO-based object detectors has been extensively investigated. By leveraging knowledge learned from large-scale source domains, transfer learning aims to accelerate convergence and improve detection performance under small-sample conditions [34]. As a result, numerous studies have explored the combination of transfer learning and YOLO frameworks for defect detection tasks. Situ et al. [20] incorporated transfer learning into YOLOv2 and systematically compared eleven pretrained feature extractors, demonstrating improved accuracy and IoU performance, although complex defects such as cracks and root intrusions remained challenging. Situ et al. [21] applied transfer learning to YOLOv5s pretrained on COCO and optimized it through pruning for sewer defect detection, significantly reducing computational cost while preserving accuracy, though robustness for small targets was still limited. Zhu et al. [33] applied transfer learning with optimized YOLO to subsurface defect detection in radar data, confirming effectiveness under sparse labeling, though high interference still caused false detections. Dubey et al. [6] analyzed the impact of input resolution and network depth in YOLOv5 transfer learning using COCO and GC10-DET as source domains, showing that deeper models like YOLOv5x enhanced detection stability but at prohibitive computational cost for lightweight deployment. However, most current YOLO-based transfer learning approaches adopt pretraining schemes based on generic visual datasets, including ImageNet and COCO. Owing to the lack of wood-specific semantic structures and textural characteristics in these datasets, the transferred representations may not adequately capture the distinctive properties of wood defects, thereby limiting cross-domain generalization in industrial inspection tasks. This issue suggests that more domain-aware transfer strategies are required to improve practical applicability.

Motivated by the above limitations, this study concentrates on two closely related aspects of wood defect detection. First, existing YOLO detectors still exhibit notable deficiencies when applied to wood defect inspection, particularly in effectively handling complex defect morphologies, extreme aspect ratios, and fine-grained texture variations. These limitations constrain detection accuracy and robustness in real industrial environments. To overcome these challenges, an enhanced detection algorithm, termed YOLOv12-SMD, is developed to strengthen feature representation, improve localization precision, and better adapt to the complex surface characteristics of wood materials. Second, although transfer learning has been widely adopted to alleviate data scarcity, conventional transfer strategies often fail to effectively accommodate the progressive evolution and domain-specific characteristics of wood defect data. Models initialized from generic source domains still face difficulties in achieving stable generalization and efficient adaptation under limited target-domain samples. To overcome this issue, a multi-stage transfer learning paradigm is introduced to progressively align feature representations across

domains and improve training efficiency and robustness in practical deployment scenarios. By integrating YOLOv12-SMD with the proposed multi-stage transfer learning paradigm, the resulting framework simultaneously addresses architectural limitations and data scarcity challenges, enabling accurate, robust, and efficient wood defect detection in dynamic industrial environments.

The primary contributions of this work are the following:

1 This paper presents the YOLOv12-SMD algorithm, which is intended to overcome significant challenges in wood defect detection, including complex defect morphologies, large-scale variations, and vulnerability to texture interference. To this end, the method introduces the Shape-Intersection over Union (Shape-IoU) loss function, which enhances boundary delineation for irregular defects, incorporates the Mixed Local-Channel Attention (MLCA) attention mechanism to improve feature discrimination under complex textural conditions, and integrates the dynamic upsampling (DySample) dynamic upsampling strategy to optimize cross-scale fusion of features while preserving fine-grained information.

2 To overcome the challenges of limited annotated samples and significant cross-domain distribution differences in wood defect detection, a multi-stage transfer learning paradigm is developed. Instead of relying on generic public datasets for pretraining, this paradigm learns generalizable representations from wood-related source domains, followed by gradual fine-tuning in the target domain to accommodate distribution discrepancies. Combined with an optimization strategy, this staged transfer process enables faster convergence and improved performance, effectively enhancing the model's generalization capability and robustness in wood defect inspection tasks.

3 This paper validates the proposed YOLOv12-SMD algorithm within a multi-stage transfer learning paradigm using three wood defect datasets. Results demonstrate that this approach maintains high detection performance and stability in complex scenarios, significantly advancing the automation and intelligence of wood quality inspection. It also provides a generalizable technical pathway for surface defect detection.

The remainder of this paper is organized as follows. Section 2 details the proposed YOLOv12-SMD algorithm. Section 3 describes the multi-stage transfer learning paradigm. Section 4 provides detailed experimental settings and systematic result analysis. Finally, Section 5 concludes this research.

2. YOLOv12-SMD Algorithm

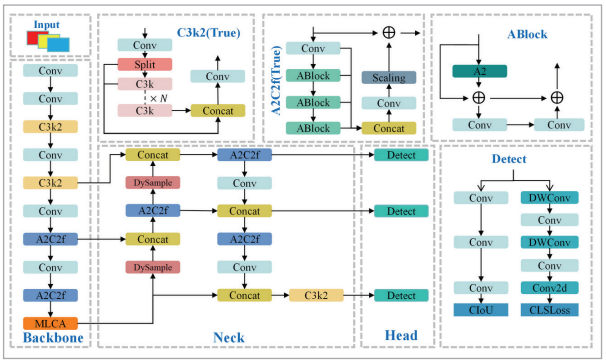
2.1. Overall Architecture

As the latest instalment in the YOLO series, YOLOv12 [22] emphasises the integration of attention mechanisms. It leverages the advantages of attention to enhance detection performance while maintaining inference speeds comparable to those of conventional CNN-based models. Although the typical YOLOv12 algorithm achieves strong performance in general object detection tasks, it has significant limitations in wood defect detection. These include the insufficient adaptability of localization regression to extreme aspect ratios and small objects, loss of fine-grained texture features at higher semantic levels, and low spatial alignment accuracy during cross-scale feature fusion. To overcome these challenges, this study presents an enhanced version of YOLOv12, referred to as YOLOv12-SMD. The overall architecture of the improved model is shown in Figure 1. First, to address the limitations of the CIoU loss function, namely its weak gradient sensitivity to short-edge offsets and limited effectiveness in regressing small objects, which often results in poor alignment between predicted and ground-truth boundaries, this paper introduces the Shape-IoU loss function [28]. By incorporating aspect ratio and scale weights of the ground-truth box into the regression process, this method enhances the model's ability to capture extreme shapes and accurately delineate small-target boundaries. Second, given the strong texture directionality and low contrast inherent in wood images, the backbone network often fails to preserve fine-grained defect cues and to establish effective position-aware channel dependencies at deeper semantic stages. To address this, the MLCA module [23] is embedded with high-level feature representations, thereby strengthening both channel interactions and local texture saliency modeling. It significantly improves feature discriminability

while mitigating background interference. Finally, in the neck network, fixed interpolation-based upsampling can cause boundary blurring and spatial misalignment during multi-scale fusion, thereby reducing detection performance for small objects and elongated cracks. To resolve this issue, the DySample dynamic upsampling method [14] is adopted. Building on bilinear initialization, DySample introduces learnable offsets to achieve more precise cross-scale feature alignment and finer structural fidelity. Through these three improvements, YOLOv12-SMD enhances localization accuracy, improves recall for small objects, and strengthens adaptability to complex and irregular defect morphologies, while maintaining computational efficiency.

Figure 1

Overall architecture of the YOLOv12-SMD model.



2.2. Shape-IoU Loss

In wood defect detection, targets such as cracks, filamentous resin, and edge defects often possess minute dimensions, extreme aspect ratios, and complex boundary geometries. These characteristics impose stringent requirements on bounding box regression accuracy. Conventional CIoU-based loss functions are insufficient in such scenarios, as they fail to effectively capture the geometric properties of irregularly shaped targets and struggle to characterize variations in boundary form and scale. This mismatch reduces the accuracy of alignment between predicted and ground-truth bounding boxes. To address this limitation, this study presents the Shape-IoU loss function [28]. By explicitly modeling the aspect ratio and scale attributes of ground-truth boxes, Shape-IoU applies anisotropic and adaptive penalties across different orientations and scales. This design enhances localization accuracy, improves regression stability, and increases recall for small objects, while boosting overall detection performance and robustness, all without altering the underlying detection framework.

In this study, the proposed Shape-IoU loss replaces the original CIoU loss function. Beyond the traditional IoU, Shape-IoU introduces a shape-scale-aware weight derived from the width-to-height ratio and scale information of objects, thereby imposing anisotropic constraints in both horizontal and vertical directions. Since objects with different scales exhibit significantly different sensitivities to center offset errors, it is necessary to normalize the offset penalty to achieve consistent optimization across scales. To this end, horizontal and vertical scale weighting coefficients, denoted as and $\omega\omega$, are hh introduced to ensure comparability of center offset constraints across different object scales, as defined in Equation (1):

$$\omega\omega = \frac{2\times\left(\omega^{gt}\right)^{scale}}{\left(\omega^{gt}\right)^{scale}+\left(h^{gt}\right)^{scale}}, hh = \frac{2\times\left(h^{gt}\right)^{scale}}{\left(\omega^{gt}\right)^{scale}+\left(h^{gt}\right)^{scale}}, \quad (1)$$

where, ω^{gt} and h^{gt} denote the width and height of the ground-truth bounding box, respectively. The parameter *scale* is a predefined hyperparameter that controls the sensitivity of the weighting coefficients to object size variations.

By incorporating these scale-aware coefficients, the shape-aware center offset term is expressed in Equation (2):

$$distance^{shape} = hh\times\left(x_c - x_c^{gt}\right)^2/c^2 + \omega\omega\times\left(y_c - y_c^{gt}\right)^2/c^2, \quad (2)$$

where (x_c, y_c) and (x_c^{gt}, y_c^{gt}) denote the center coordinates of the predicted and ground-truth boxes, respectively, and c represents the diagonal length of their minimum enclosing rectangle, used for normalization. This formulation imposes direction-sensitive penalties on center offsets, enabling anisotropic constraints along the horizontal and vertical directions.

To further capture discrepancies in width and height, Shape-IoU integrates a shape consistency term, expressed in Equation (3):

$$\Omega^{shape} = \sum_{t=\omega,h} \left(1 - e^{-\omega_t}\right)^{\theta}, \theta = 4 , \quad (3)$$

where ω_ω and ω_h represent weighted relative errors in the width and height directions, respectively, as defined in Equation (4):

$$\begin{cases} \omega_\omega = hh \times \dfrac{\left|\omega - \omega^{gt}\right|}{\max\left(\omega, \omega^{gt}\right)} \\ \omega_h = \omega\omega \times \dfrac{\left|h - h^{gt}\right|}{\max\left(h, h^{gt}\right)} \end{cases} . \quad (4)$$

This shape consistency term aggregates width-height discrepancies through an exponential saturation function, which prevents gradient over-amplification under large deviations while maintaining sensitivity for small deviations. The hyperparameter θ controls the degree of nonlinear scaling. Finally, the complete Shape-IoU loss function is defined in Equation (5):

$$L_{Shape-IoU} = 1 - IoU + distance^{shape} + 0.5 \times \Omega^{shape} , \quad (5)$$

where the coefficient 0.5 is a predefined weighting factor used to balance the contribution of the shape consistency term during optimization.

Overall, the Shape-IoU loss consists of three complementary components: the IoU term ensures spatial overlap between predicted and ground-truth boxes, the shape-aware center offset term enforces direction-sensitive localization constraints and the shape consistency term reinforces geometric coherence in width–height relationships. Together, these components yield improved regression stability and localization accuracy, particularly for small-scale and elongated defects, while remaining computationally efficient and fully compatible with YOLOv12 regression heads.

2.3. MLCA Module

Fine-grained wood defect detection often involves complex textures, heterogeneous backgrounds, and targets with small spatial extents or elongated structures. Defects such as cracks, resin filaments, and edge irregularities typically activate strong responses only within limited local regions, rather than across entire feature maps. This characteristic imposes stringent requirements on feature recalibration mechanisms, which must preserve localized saliency while maintaining robust semantic discrimination.

Conventional channel attention mechanisms, including Squeeze-and-Excitation (SE) and Efficient Channel Attention (ECA), compress each feature channel into a single scalar through global average pooling. Although this strategy effectively captures inter-channel dependencies, it inherently disregards spatial variability within channels. Consequently, localized but informative responses may be excessively smoothed, leading to the attenuation of critical defect cues. Alternatively, spatial attention mechanisms with finer granularity, such as Log-Sum-Exp pooling or coordinate-based attention variants, explicitly model spatial distributions to retain local details. However, these methods typically incur additional parameters and computational overhead, which compromises inference efficiency and limits their applicability in lightweight detection networks.

Motivated by these limitations, the MLCA [23] mechanism is introduced to balance representational capacity and computational efficiency through a hybrid attention formulation. MLCA jointly models global channel dependencies and local spatial saliency without performing channel dimensionality reduction or introducing significant computational burden. Global aggregation characterizes inter-channel semantic relevance, while local aggregation preserves intra-channel spatial heterogeneity by enhancing defect responses concentrated in localized regions. By fusing these complementary cues, MLCA generates a content-adaptive channel weight map that recalibrates backbone features in a lightweight and minimally intrusive manner. This design effectively reconciles semantic discrimination with spatial detail retention, enabling robust feature enhancement under complex textures and small-object detection conditions.

The structural design of the MLCA module is illustrated in Figure 2. For an input feature map $X \in {}^{\sim C \times k_s \times k_s}$, the MLCA mechanism constructs complementary channel-wise representations through a

dual aggregation strategy, aiming to jointly capture global semantic relevance and local spatial saliency. This design directly addresses the limitation of conventional channel attention mechanisms that rely solely on global pooling and consequently suppress localized but informative responses.

On one hand, MLCA employs Local Averaging Pooling (LAP) to compute channel-wise responses over partitioned spatial subregions. By aggregating features within local grids rather than across the entire spatial extent, this operation preserves intra-channel spatial heterogeneity and highlights defect cues that are confined to localized regions, such as fine cracks or filament-like structures. As a result, local saliency is explicitly retained during channel recalibration, mitigating the loss of fine-grained information caused by global aggregation.

On the other hand, Global Average Pooling (GAP) is applied to each channel over the full feature map to capture global contextual statistics. This global representation characterizes inter-channel semantic importance, enabling the model to emphasize channels that are globally informative for defect discrimination.

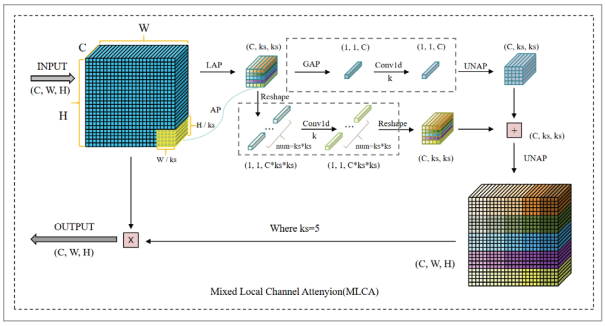
The local and global channel descriptors are subsequently processed through lightweight one-dimensional convolutions to model cross-channel dependencies at different spatial scales. The convolution kernel size k is proportional to the channel number C and is constrained to be an odd value, thereby ensuring sufficient cross-channel coupling while maintaining controlled parameter count and computational complexity. The corresponding calculation for k [25] is given in Equation (6):

$$k = \Phi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{6}$$

where y and b are hyperparameters with default values of 2. *odd* indicates that k is constrained to be odd. If an even value is encountered, it is adjusted to the nearest odd number by adding 1.Finally, the outputs of the local and global branches are fused to form a mixed channel attention representation. Since the local aggregation branch operates on partitioned spatial regions, its output is first restored to the original spatial resolution through an unpooling-based alignment operation, denoted as UNAP. UNAP redistributes locally aggregated responses back to their corresponding spatial locations in a parameter-free manner, thereby preserving spatial correspondence between localized saliency cues and the original feature map. In contrast, the global branch inherently maintains spatial consistency and can be directly broadcast across spatial dimensions without additional processing.

In summary, the MLCA module achieves an effective balance between representational capability and computational efficiency by jointly modeling global channel dependencies and localized spatial saliency. Through its hybrid attention formulation, MLCA enhances feature discrimination while preserving fine-grained defect cues that are critical under complex textural backgrounds and small-scale defect conditions.

Figure 2
Information flow schematic of the MLCA architecture.



2.4. DySample Module

In wood defect detection, multi-scale feature fusion in the neck network plays a critical role in preserving defect boundaries and spatial consistency. However, the fixed interpolation-based upsampling commonly adopted in lightweight detectors uniformly processes all feature regions, ignoring local geometric variations and texture heterogeneity. As a result, fine-grained structures such as thin cracks, filamentous resin, and edge defects are prone to boundary blurring, spatial misalignment, and detail degradation during upsampling, which directly undermines localization stability and recall for small-scale and elongated defects. To alleviate these limitations, this study integrates DySample [14], a content-adaptive dynamic upsampling strategy, into the neck network. Unlike deterministic interpolation schemes, DySample reformulates upsampling

as a point-sampling problem on a continuous feature surface, where sampling locations are adaptively adjusted according to feature content. By learning small, content-dependent offsets on top of a regular upsampling grid, DySample enables more accurate reconstruction of high-resolution features while preserving structural continuity across scales. This mechanism effectively improves feature alignment during multi-scale fusion and enhances the representation of fine defect boundaries without introducing heavy computational overhead.

From a methodological perspective, DySample decouples upsampling from kernel-based reconstruction. Instead of relying on large convolutional kernels or dynamically generated kernel weights, it learns a lightweight offset field that perturbs standard interpolation grids with micro-displacements. Conventional interpolation operators are then used for resampling, which preserves the numerical stability and geometric consistency of bilinear upsampling while selectively enhancing local textures and boundary transitions. This design achieves a favorable balance between adaptivity and efficiency, making DySample particularly suitable for real-time industrial inspection scenarios.

DySample can be categorized into four variants, defined by the range factor type (static or dynamic) and the offset generation method (linear+pixel shuffle (LP) or pixel shuffle+linear (PL)). In this study, the LP variant with a dynamic range factor is adopted to balance offset flexibility and optimization stability. Specifically, the dynamic range factor is generated in a pointwise manner from the input features and bounded by a sigmoid function with a fixed scaling coefficient. This bounded design constrains the magnitude of learned offsets, preventing excessive spatial displacement during training while preserving sufficient adaptivity for fine-grained structural reconstruction. As illustrated in Figure 3(a), the initial offset O_{raw} is computed as:

$$O_{raw} = 0.5 sigmoid\left(linear_1\left(X\right)\right) \cdot linear_2\left(X\right), O_{raw} \in R^{2gs^2 \times sH \times sW}. \quad (7)$$

To make the learned offsets compatible with the upsampling grid at the target resolution, DySample performs pixel shuffle to reorganize the offset representation into a spatially aligned offset O, enabling subsequent grid-based sampling at the desired scale.

Next, the offset O is combined with the original sampling grid g to generate the sampled set S, as expressed in Equation (8):

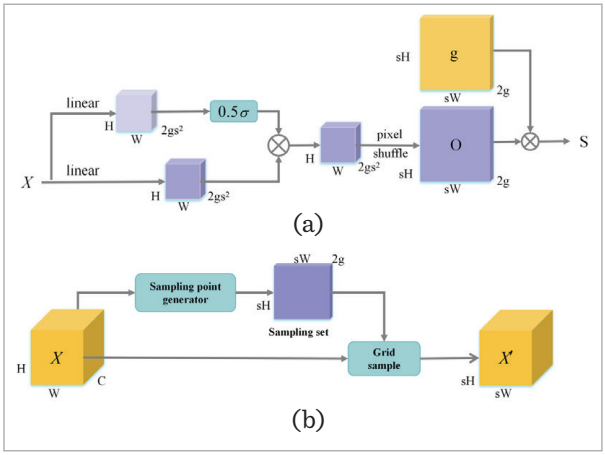
$$S = g + O, O \in R^{2g \times sH \times sW}. \quad (8)$$

Finally, DySample resamples the original feature map using the computed sampled set S via a standard grid-based interpolation operator, producing the upsampled feature map X' with improved spatial alignment and boundary fidelity, as shown in Figure 3(b). The process is formally expressed in Equation (9):

$$X' = grid_{sample}\left(X, S\right), X' \in R^{C \times sH \times sW}. \quad (9)$$

Figure 3

The architecture of the Dysample module, where (a) illustrates the upsampling component and (b) depicts the sample point generator equipped with a dynamic scope factor.



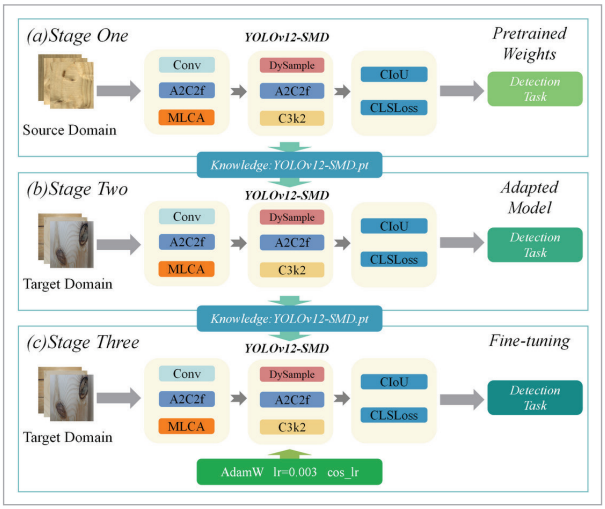
3. Multi-stage Transfer Learning Paradigm

To address the substantial domain discrepancies observed in wood defect detection, which arise from variations in acquisition conditions, background textures, and defect morphologies, this study proposes a multi-stage progressive transfer learning paradigm. Unlike conventional transfer learning strategies that rely on generic large-scale datasets (e.g., ImageNet or COCO) for pretraining, the pro-

posed paradigm is built upon a wood-specific source domain, enabling the model to acquire task-relevant structural and textural priors from the outset. Building upon this domain-consistent initialization, the paradigm progressively integrates knowledge across the source and target domains, preserving class-discriminative representations while effectively mitigating negative transfer caused by domain mismatch. The overall framework is illustrated in Figure 4.

Figure 4

Transfer learning training paradigm.



As shown in Figure 4(a), the first stage performs deep training of the model on a large-scale wood defect source domain dataset. The selected source domain comprises approximately 10,000 rigorously curated wood defect images, covering multiple categories of typical defects and diverse textural backgrounds. To balance the thoroughness of training with the stability of convergence, this phase is configured for 300 epochs, allowing the model to learn generalizable representations of wood defects from large-scale data effectively. At this stage, the improved YOLOv12-SMD architecture is employed to enhance feature extraction and parameter learning, providing a high-quality initialization tailored to wood defect characteristics rather than generic visual semantics.

As shown in Figure 4(b), the second stage leverages the pretrained YOLOv12-SMD weights obtained from the source domain as initialization for fine-tuning on the target domain. The target domain data are typically characterized by limited annotation coverage, strong texture variability, and pronounced distribution shifts relative to the source domain. Under such conditions, training a detector from scratch often leads to optimization instability, slow convergence, and a high risk of overfitting. By employing pretrained parameters as a "hot start," the model inherits texture and shape priors embedded in both the backbone and detection head. It narrows the parameter search space, improves sample utilization efficiency, and produces smoother loss descent trajectories in small-sample settings. This strategy accelerates convergence, stabilizes training, alleviates overfitting, and ultimately strengthens the model's generalization capacity and robustness when adapting to the target domain.

As shown in Figure 4(c), the third stage focuses on refining training strategies to ensure stable and efficient convergence under limited target domain samples. Specifically, Adam with Weight Decay (AdamW) [32] is employed as the optimizer with an initial learning rate of 0.003, in conjunction with a cosine learning rate scheduler (cos_lr) [27] to regulate learning rate decay dynamically. This design maintains a relatively high learning rate in the early training phase to encourage exploration, while gradually reducing it in later iterations to enable fine-grained parameter updates. Such a strategy enhances convergence stability and strengthens the model's generalization capability. During this phase, the model undergoes further optimisation, building on the results achieved in earlier transfer stages, and exhibits improved robustness and superior detection performance in the small-sample environment of the target domain.

Overall, the proposed multi-stage transfer learning paradigm differs fundamentally from conventional transfer learning approaches that rely on generic source domains, by anchoring knowledge transfer in a wood-specific source domain and progressively adapting feature representations to the target distribution. Through hierarchical optimization across source-domain pretraining, domain-aware transfer, and strategy-level refinement, the framework effectively alleviates training instability and enhances robustness and generalization. This design provides reliable technical support for wood defect detection in real industrial environments characterized by complex textures and limited annotated data.

4. Experiments and Analysis

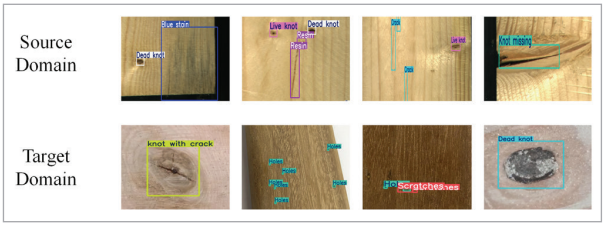
4.1. Description Source and Target Domain Datasets

The source domain dataset was selected from the Roboflow Universe platform and is titled Wood defect detection dataset [1]. It comprises 10,000 wood surface images, all meticulously annotated by hand, encompassing nine representative categories of wood defects. Specifically, blue stain results from fungal infection, causing bluish or gray discoloration of the wood surface. Cracks refer to structural fissures generated during drying or mechanical stress. Dead knots arise from residual dead branches and are prone to detachment during use. A knot missing denotes the complete loss of a knot, leaving a cavity in the wood. Knots with cracks contain visible internal fissures that compromise their integrity. Live knots remain firmly bonded to the wood matrix. Marrow manifests as exposed central trunk tissue. Quartzity appears as quartz-like streaks or bright inclusions within the grain. Resin pockets are sac-like structures formed by resin accumulation within the wood. For subsequent model training and objective evaluation, the 10,000 preprocessed images were randomly partitioned into training, validation, and test subsets at an 8:1:1 ratio, corresponding to 8,000, 1,000, and 1,000 images, respectively. This division strategy maintains consistent defect category distribution across subsets, ensures sufficient data for effective model learning, and preserves independent validation and testing sets to support reliable model tuning and performance assessment.

To evaluate the adaptability and robustness of the proposed method under realistic wood surface inspection scenarios, two target domain datasets were constructed to reflect practical industrial conditions, denoted as Target-A and Target-B. Target-A is formed by integrating two open-source datasets, Defect in Wood [4] and Wood Surface [26]. Specifically, 3,000 images are selected from the Defect in Wood dataset and 1,227 images are selected from the Wood Surface dataset, resulting in a total of 4,227 images after preprocessing. The integrated dataset covers five representative defect categories, including dead knot, knot with crack, marrow, holes, and scratches. By combining the complementary characteristics of these two datasets, Target-A exhibits increased diversity in defect morphology and surface appearance, making it more representative of practical wood defect inspection scenarios in real-world industrial environments. Target-B is built from the New-Defects in Wood Dataset [5], which contains 1,320 annotated images and exhibits more diverse defect appearances and distribution characteristics. The dataset includes five defect categories, namely crack, dead knot, holes, live knot, and knot with crack, representing newly emerging and more complex defect patterns commonly encountered in real industrial environments. For consistency and reproducibility, both Target-A and Target-B follow the same 8:1:1 train–validation–test split protocol as the source domain. A comparison of selected source and target domain samples is presented in Figure 5.

Figure 5

Partial image comparison between source and target domain datasets.



4.2. Simulation Environment

4.2.1. Experimental Setup

The experimental platform was configured on a Linux operating system. The hardware environment comprised a single NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM, paired with an Intel(R) Xeon(R) Gold 6430 processor (16 vCPUs), providing sufficient computational power and memory bandwidth to support large-scale parallel computation and tensor operations. The software stack employed Python 3.11 and PyTorch 2.2.2, with GPU acceleration enabled through CUDA 11.8. This configuration exhibited excellent stability and reproducibility throughout both the training and inference phases.

4.2.2. Experimental Parameter Settings

Model optimization was performed under a unified training configuration. The training process was conducted for 300 epochs, ensuring parameter stability through sufficient iterative convergence.

During training and validation, all input photos were scaled to 640 × 640 pixels. To balance throughput and memory usage, Automatic Mixed Precision (AMP) was enabled, combining 16-bit Floating Point (FP16) and FP32 arithmetic. This hybrid precision scheme accelerated both training and inference while maintaining numerical stability. Optimization was performed using stochastic gradient descent (SGD) with an initial learning rate of 0.01 and a weight decay of 5×10^{-4}, which provides effective regularization and helps prevent overfitting.

Regarding hyperparameter configuration, for the Shape-IoU loss, the parameter *scale* is set to 0, and the parameter θ is set to 4, with its selection and influence further analyzed in Section 4.3.1.

Overall, these settings achieve an effective equilibrium among reproducibility, computational efficiency, and stability, establishing a consistent foundation for subsequent experimental analysis.

4.2.3. Evaluation Metrics

To comprehensively evaluate the model's performance in terms of detection accuracy, localisation robustness, and computational efficiency, six metrics are employed: Precision (P), Recall (R), mean Average Precision (mAP) at 50% and 95%, Giga Floating-point Operations (GFLOPs), and Training Time. Evaluation follows a one-to-one matching protocol, with IoU as the criterion. Predictions are first sorted in descending order of confidence, after which matching and statistical calculations are performed. Detection quality is then quantified using both single-threshold and cross-threshold measures derived from precision-recall (PR) curves and their integral forms, specifically the mean average precision (mAP). Overall mAP is obtained by averaging across all categories. For computational cost, GFLOPs represents the theoretical inference complexity of a single forward pass at a fixed input resolution, while Training Time denotes the actual end-to-end optimization duration. All metrics are computed under a unified protocol with consistent dataset partitioning, input dimensions, hardware/software environment, threshold settings, and post-processing strategies to ensure fair comparisons and reproducibility. Precision and Recall at a fixed IoU threshold τ are defined as in Equations (10)-(11):

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

where TP denotes the number of correct detections matching ground-truth targets with IoU ≥ τ, FP denotes the number of false detections (unmatched predictions or IoU < τ), and FN represents the number of missed ground-truth targets not detected by any prediction with sufficient IoU. Average Precision (AP) and mAP@50 are derived from the precision-recall curve $p_\tau(r)$ computed at different confidence thresholds τ. The average AP_τ precision is defined as the area under its PR curve [0, 1], as shown in Equation (12):

$$AP_\tau = \int_0^1 p_\tau(r)\,dr. \tag{12}$$

The arithmetic mean of AP across all categories yields mAP@50 at threshold $\tau = 0.50$, as shown in Equation (13):

$$mAP@50 = \frac{1}{C}\sum_{c=1}^{C} AP_{\tau=0.50}^{(c)}. \tag{13}$$

where C denotes the number of categories, and $AP_\tau^{(c)}$ represents the AP of category c at threshold τ. The more stringent mAP@50:95 is obtained by averaging across multiple IoU thresholds, as expressed in Equation (14):

$$mAP@50:95 = \frac{1}{10C}\sum_{\tau\in\{0.50,0.55,\ldots,0.95\}}\sum_{c=1}^{C} AP_\tau^{(c)}. \tag{14}$$

This metric averages results across ten IoU thresholds, ranging from $\tau = 0.50$ to $\tau = 0.95$, thereby emphasizing precise localization performance under stricter conditions. GFLOPs, representing the floating-point operations for one forward pass, is defined as in Equation (15):

$$GFLOPs = \frac{FLOPs(f, s)}{10^9}, \tag{15}$$

where f denotes the model architecture s and indicates the input resolution. For CNN-based architectures dominated by 2D convolutions, the per-layer computational cost can be approximated by Equation (16):

$$FLOPs_{conv} \approx 2H_{out}W_{out}C_{out}C_{in}k_h k_w, \quad (16)$$

where H_{out} and W_{out} denote the height and width of the output feature map, C_{in} and C_{out} represent the counts of input and output channels, k_h and k_w specify kernel dimensions, and the coefficient 2 accounts for multiply-accumulate operations. Finally, Training Time measures the total wall-clock duration of model optimization. Let E be the total number of epochs and T_e the time consumed per epoch e. The total training time is then expressed as Equation (17):

$$T_{train} = \sum_{e=1}^{E} T_e. \quad (17)$$

4.3. Model Validation

4.3.1. Parameter Analysis of Shape-IoU

In the Shape-IoU loss, two key parameters are involved, namely the parameter *scale* and the parameter θ. In this study, the parameter *scale* is fixed to 0 to avoid introducing size-dependent bias into the weighting coefficients and to ensure stable optimization across defects with diverse scales. Under this setting, the loss emphasizes directional geometric consistency rather than absolute object size, which is more suitable for wood defect detection where defects exhibit large scale variation but similar elongated or irregular shapes. Based on this fixed scale configuration, the influence of the parameter θ, which controls the nonlinearity of the shape penalty, is further investigated within the YOLOv12-SMD framework. To determine the optimal parameter setting under the experimental conditions of this study, multiple values of θ are evaluated using an identical training and evaluation protocol, and the corresponding detection performance is reported in Table 1.

As shown in Table 1, setting θ = 4 achieves the highest mAP@50, outperforming both smaller and larger values. Lower values of θ provide insufficient sensitivity to width-height discrepancies, while higher values tend to overemphasize shape differences and adversely affect optimization stability. These results indicate that a moderate nonlinearity yields a better balance between geometric sensitivity and training robustness.

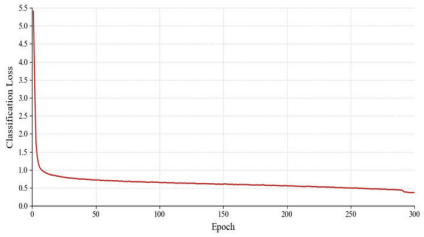
Table 1
Impact of parameter on detection performance.

θ	mAP@50(%)	θ	mAP@50(%)
2	78.7	6	79.7
4	80.8	8	79.6

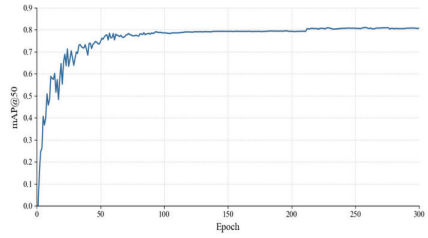
4.3.2. Training Process Analysis

To empirically assess the efficacy of the proposed improvements in wood defect detection, the YOLOv12-SMD model was systematically trained on the source domain dataset under unified input preprocessing and post-processing conditions, with sufficient training iterations to ensure convergence. During the training process, the evolution of classification loss and mAP@50 was continuously tracked, and their convergence behavior as well as performance improvement trends were visualized in curve form. The corresponding visualizations and statistical results are presented in Figure 6.

Figure 6
Convergence of the classification loss and evolution of mAP@50 during training.



(a) Classification loss function convergence curve



(b) mAP@50 evolution curve

As illustrated in Figure 6(a), the classification loss curve exhibits overall smooth convergence without marked oscillations, reflecting strong numerical stability throughout training. During the initial 10-20 epochs, the loss rapidly decreased from 5.4 to below 0.90, demonstrating that the YOLOv12-SMD model quickly established effective category discrimination boundaries and achieved preliminary feature alignment. Between epochs 20 and 150, the loss continued to decline more gradually, reaching approximately 0.60 around epoch 150, indicating a transition from coarse-grained parameter adjustments to fine-grained refinements. Between epochs 150 and 200, the curve stabilised within a narrow range, between 0.55 and 0.60, while trending downward, indicating steady-state convergence. Around epoch 245, the loss dropped below 0.50 for the first time, reflecting improved handling of boundary cases and long-tail samples. Collectively, the stable convergence observed in the final training phase demonstrates the model's effective feature learning capability and reliable optimization behavior. Correspondingly, as shown in Figure 6(b), the mAP@50 curve follows a complementary trend of rapid initial increase, gradual improvement, and eventual plateau. It rose sharply from 0 to 0.60 within the first 15 epochs, then steadily climbed to 0.75 by epoch 50. Between epochs 60 and 140, performance stabilized within a narrow range of 0.77 to 0.79. After epoch 200, the curve entered a high plateau between 0.80 and 0.81, where it gradually stabilized, indicating that the model's joint localization and classification capabilities had been effectively established and consolidated. In summary, the training dynamics demonstrate that YOLOv12-SMD achieves efficient feature learning, stable convergence, and reliable cross-scale fusion in the source domain. The decision boundaries progressively solidified, while residual errors were effectively suppressed, providing a strong foundation for subsequent transfer learning and target domain adaptation.

To comprehensively evaluate the category-level generalization and discrimination capability of YOLOv12-SMD in wood defect detection, a per-class assessment was conducted on the source domain dataset using four key metrics: Precision, Recall, mAP@50, and mAP@50:95. The results, summarized in Table 2, demonstrate consistent robustness and notable performance improvements, particularly for small-scale targets, elongated defects, and defects embedded within intensely textured backgrounds. These findings validate the model's effectiveness and stability in real-world timber inspection scenarios.

As reported in Table 2, the YOLOv12-SMD model achieves superior average performance, with overall metrics of Precision 79.2%, Recall 72.3%, mAP@50 80.8%, and mAP@50:95 48.1%. These results reflect both strong detection accuracy and robust recall capability in source domain tasks. Further analysis by

Table 2

Performance evaluation of the final model across different defect categories.

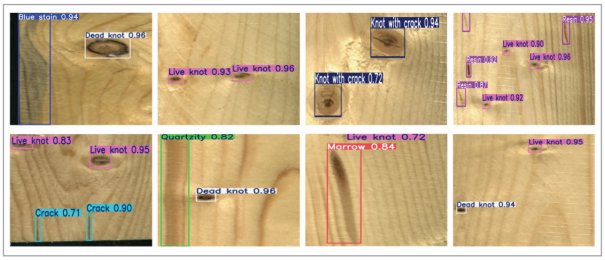
Class	Precision(%)	Recall(%)	mAP@50(%)	mAP@50:95(%)
Blue stain	86.4	100.0	99.5	73.2
Crack	74.8	74.3	74.3	38.0
Dead knot	75.3	77.5	83.0	44.2
Knot missing	100.0	59.5	78.7	42.3
Knot with crack	58.6	74.2	80.5	54.0
Live knot	83.0	80.8	87.6	40.5
Marrow	88.4	86.7	87.0	59.7
Quartzity	63.2	22.7	46.2	29.9
Resin	83.0	75.0	90.3	52.4
Total	79.2	72.3	80.8	48.2

class highlights particularly strong performance in typical structural defect categories. For Knot missing and Resin, the model achieves Precision values of 100% and 83.0%, Recall values of 59.5% and 75.0%, and mAP@50 scores of 78.7% and 90.3%, respectively. It demonstrates that in scenarios characterized by pronounced geometric anisotropy and sufficient boundary contrast, the model is capable of stable feature aggregation and precise boundary regression. In addition, categories such as Blue stain and Marrow also exhibit robust performance. Blue stain achieves a recall of 100% with an mAP@50 of 99.5%, while Marrow achieves a Precision of 88.4%, a Recall of 86.7%, and an mAP@50 of 87.0%. These results highlight the model's ability to achieve reliable feature alignment and scale adaptation in categories characterised by distinct textural segmentation or pronounced colour contrast. Overall, YOLOv12-SMD effectively leverages its detection advantages for elongated shapes and high-contrast targets under source domain conditions, thereby delivering reliable performance for wood defect identification. Representative detection results on typical wood defect categories are illustrated in Figure 7.

As shown in Figure 7, the YOLOv12-SMD model demonstrates outstanding performance in source domain wood defect detection. It achieves high detection accuracy for elongated objects with pronounced boundary contrast, while maintaining stable responses and elevated recall rates for small-scale defect identification. Overall, YOLOv12-SMD exhibits significant advantages in detection accuracy, feature representation capacity, and multi-scale adaptability, underscoring its robustness and applicability in complex wood surface environments. Furthermore, the model achieves rapid and stable convergence during source domain training while effectively extracting discriminative features. These capabilities not only provide dependable support for high-precision wood defect detection but also establish a strong performance foundation for subsequent cross-domain transfer studies.

Figure 7

Representative visualization of YOLOv12-SMD applied to wood defect detection.



4.3.3. Ablation Studies

To evaluate the effectiveness of progressively integrating the proposed improvement modules into YOLOv12, a series of systematic ablation studies was conducted on the source domain wood defect dataset under identical training, validation, and evaluation settings. The corresponding experimental results are summarized in Table 3.

As summarized in Table 3, the ablation results on the source-domain wood defect dataset clearly re-

Table 3

Ablation experiment results.

Method	Shape-IoU	MLCA	DySample	P(%)	R(%)	mAP@50 (%)	mAP@50:95 (%)	GFLOPs
YOLOv12	-	-	-	81.2	69.7	77.9	50.1	6.3
YOLOv12-S	√	-	-	74.8	73.4	78.5	47.2	6.3
YOLOv12-M	-	√	-	82.9	66.9	78.5	50.2	6.4
YOLOv12-D	-	-	√	71.8	76.8	78.9	49.2	6.3
YOLOv12-SM	√	√	-	84.9	67.2	79.6	46.9	6.4
YOLOv12-SD	√	-	√	84.8	69.4	79.3	48.5	6.3
YOLOv12-MD	-	√	√	72.8	74.4	78.6	50.1	6.4
YOLOv12-SMD	√	√	√	79.2	72.3	80.8	48.2	6.4

veal the functional role of each proposed module and the cooperative effects achieved through progressive integration.

When only Shape-IoU is introduced, the resulting YOLOv12-S variant primarily improves regression behavior. Compared with the baseline YOLOv12, mAP@50 increases from 77.9% to 78.5% and Recall rises from 69.7% to 73.4%, indicating enhanced localization stability and defect coverage, particularly for elongated and irregular targets. The slight decrease in Precision suggests that the stricter geometric constraints emphasize bounding box alignment over confidence calibration. When only MLCA is incorporated, the YOLOv12-M variant shows a clear improvement in semantic discrimination. Precision increases from 81.2% to 82.9% and mAP@50 reaches 78.5%, confirming that joint modeling of local spatial saliency and global channel dependencies effectively suppresses background interference. However, Recall decreases to 66.9%, implying that stronger feature selectivity may weaken responses to low-contrast or ambiguous defects.

When DySample is applied alone, the YOLOv12-D variant mainly enhances recall-oriented performance. Recall increases to 76.8% and mAP@50 reaches 78.9%, demonstrating that adaptive upsampling improves multi-scale feature alignment and small-defect recovery, albeit at the cost of reduced Precision due to increased sensitivity to fine-grained details.

Combining Shape-IoU and MLCA yields YOLOv12-SM, where Precision rises markedly to 84.9% and mAP@50 increases to 79.6%. This result indicates that geometry-aware regression and attention-based feature refinement jointly strengthen high-confidence predictions. Nevertheless, Recall drops to 67.2%, reflecting compounded selectivity from both constraints.

Combining Shape-IoU with DySample produces YOLOv12-SD, which achieves mAP@50 of 79.3% and Recall of 69.4%. This trend suggests that adaptive upsampling compensates for the stricter regression constraints imposed by Shape-IoU, resulting in a more balanced detection behavior. Similarly, integrating MLCA with DySample in YOLOv12-MD improves Recall to 74.4% while maintaining mAP@50 at 78.6%, indicating that DySample alleviates the recall suppression introduced by strong attention filtering.

When all three modules are jointly integrated, YOLOv12-SMD attains the best overall performance, achieving an mAP@50 of 80.8% and a Recall of 72.3%, while maintaining a competitive Precision of 79.2% and a computational cost of 6.4 GFLOPs. These results demonstrate that Shape-IoU stabilizes localization, MLCA enhances semantic robustness under complex textures, and DySample restores fine-grained spatial details. Their coordinated interaction effectively mitigates the individual trade-offs observed in isolated modules, leading to a balanced and robust detector for wood defect inspection.

Overall, the ablation study confirms that each module targets a distinct limitation of YOLOv12, and their progressive integration yields complementary gains in accuracy, recall, and robustness without introducing significant computational overhead.

4.3.4. Comparative Experiments

To thoroughly verify the performance advantages of the proposed YOLOv12-SMD model, six representative improved detection algorithms were selected for comparison: TRS-YOLO [24], BiFPN-CPCA-ShuffleNetV2 (BCS-YOLO) [10], GAM-HWD-BiFPN-PPA-NWD (GHBPN-YOLO) [3], InnerCIoU-TripletAttention-DBB (ITD-YOLO), SlideLoss-SENetV2-HWD (SSH-YOLO), and FocalerCIoU-SCSA-DynamicConv (FSD-YOLO). All models were trained and evaluated on the source domain wood defect dataset under identical settings to ensure fairness and comparability. The comparative experimental results are summarized in Table 4.

As shown in Table 4, the comparative results on the source domain dataset indicate that YOLOv12-SMD achieves consistently higher mAP@50 than all six benchmark models while maintaining competitive computational complexity. Compared with TRS-YOLO and FSD-YOLO, YOLOv12-SMD improves mAP@50 by 1.7%-1.8% with only marginal changes in Recall, suggesting that the proposed model enhances localization accuracy and confidence calibration rather than simply increasing detection aggressiveness. In comparison with BCS-YOLO and ITD-YOLO, YOLOv12-SMD achieves higher Precision and mAP@50 while preserving comparable Recall. This indicates improved discrimination between defect regions and complex wood textures, which can be attributed to strengthened feature representation

Table 4
Comparative experimental results.

Method	P(%)	R(%)	mAP@50(%)	mAP@50:95(%)	GFLOPs
TRS-YOLO	78.2	72.7	79.1	50.1	6.4
BCS-YOLO	75.2	74.2	78.2	48.8	6.5
GHBPN-YOLO	81.3	64.2	76.9	49.4	8.2
ITD-YOLO	74.7	70.1	78.1	46.7	6.9
SSH-YOLO	79.6	68.7	77.7	48.6	5.4
FSD-YOLO	76.6	73.2	79.0	48.9	6.3
YOLOv12-SMD	79.2	72.3	80.8	48.2	6.4

and more stable geometric regression. For GHB-PN-YOLO, although high Precision is achieved, its low Recall reveals limited coverage of diverse defect instances. YOLOv12-SMD significantly improves Recall (+8.1%) and mAP@50 (+3.9%) while reducing GFLOPs, demonstrating a more balanced trade-off between detection completeness and efficiency. Compared with SSH-YOLO, YOLOv12-SMD yields higher Recall and mAP@50 at a moderate computational increase, reflecting enhanced sensitivity to small-scale and elongated defects.

Overall, the consistent mAP@50 gains of 1.7%-3.9% across all baselines confirm that YOLOv12-SMD improves detection accuracy through more effective feature utilization and localization modeling, rather than relying on increased computational cost. These results validate its robustness and practical suitability for source domain wood defect detection.

4.4. Analysis of Transfer Learning Results

4.4.1. Validation of Transfer Paradigm

To validatethe effectiveness of the proposed multi-stage transfer learning paradigm and its adaptability across different practical inspection conditions, comparative experiments were conducted on two target domain datasets (Target-A and Target-B) before and after applying the transfer strategy. All experiments were conducted using the proposed YOLOv12-SMD framework under identical training and evaluation settings to ensure fair comparison. To reduce the influence of random initialization and training stochasticity, each experiment was repeated multiple times, and the mean performance values were reported. This evaluation protocol enhances statistical reliability and provides a more objective assessment of the model's detection performance in target domain scenarios.The corresponding experimental results are summarized in Table 5.

As shown in Table 5, the proposed multi-stage transfer learning paradigm consistently improves detection performance on both target domain datasets while substantially reducing training time. For Target-A, transfer learning leads to a notable increase in Recall and mAP@50, with mAP@50 improving from 52.1% to 61.5%, indicating more accurate and comprehensive defect detection. At the same time, the average training time is reduced from 0.856 h to 0.713 h, demonstrating more efficient convergence under limited target-domain data. For Target-B, the model achieves simultaneous improvements in Precision, Recall, and mAP@50, with mAP@50 increasing from 66.8% to 71.6%. Notably, the average training time is reduced by more than 30%, reflecting the strong optimization efficiency enabled by source-domain initialization and progressive adaptation.

Overall, these results indicate that the proposed transfer learning paradigm not only enhances detection accuracy across different target domain conditions but also significantly accelerates training, thereby improving both effectiveness and practicality for real-world wood defect detection.

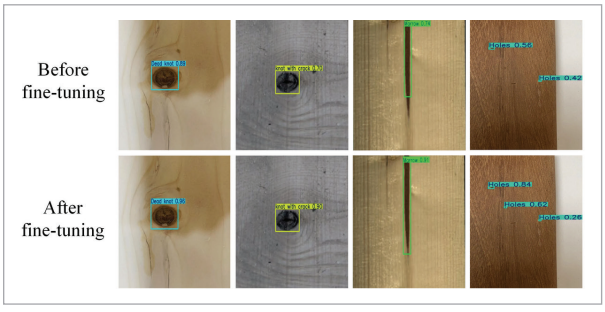
As illustrated in Figure 8, transfer fine-tuning significantly improves detection performance across multiple defect categories. For dead knots, the detection rate increased from 89% to 96%, indicating enhanced capability in distinguishing knot-type defects. For

Table 5
Comparison of detection results before and after applying the transfer learning paradigm.

Target-A	Before				After			
Trail	P(%)	R(%)	mAP@50(%)	T_{train}(h)	P(%)	R(%)	mAP@50(%)	T_{train}(h)
1	81.8	45.0	52.2	0.894	68.4	61.4	61.4	0.711
2	76.3	48.1	51.1	0.886	72.0	58.9	61.7	0.721
3	78.3	50.6	53.1	0.804	68.3	60.2	61.8	0.798
4	80.1	47.8	52.7	0.886	77.1	57.2	61.5	0.758
5	55.9	48.4	51.4	0.808	70.1	58.1	61.0	0.577
Mean	74.5	48.0	52.1	0.856	71.1	59.2	61.5	0.713
Target-B	Before				After			
Trail	P(%)	R(%)	mAP@50(%)	T_{train}(h)	P(%)	R(%)	mAP@50(%)	T_{train}(h)
1	69.4	64.2	66.7	0.418	80.2	67.1	72.4	0.326
2	74.5	58.3	66.6	0.415	71.2	70.5	71.5.	0.198
3	78.0	57.9	66.7	0.412	75.0	67.6	71.2	0.254
4	70.6	66.6	67.7	0.419	75.4	68.4	71.2	0.265
5	65.9	63.8	66.5	0.412	67.0	73.5	71.6	0.320
Mean	71.7	62.2	66.8	0.415	73.8	69.4	71.6	0.273

knots with cracks, the detection rate rose from 70% to 90%, underscoring the positive effect of fine-tuning on feature representation for complex crack-type defects. In the case of marrow detection, performance improved from 74% to 91%, reflecting the strengthened ability to capture fine-grained features and achieve more accurate localization. Notably, in the hole detection task, the pre-fine-tuned model identified only two instances with detection rates of 56% and 42%, whereas the fine-tuned model successfully detected three instances with rates of 84%, 26%, and 62%. This outcome not only demonstrates improved coverage of defect instances but also highlights enhanced object perception in complex textured backgrounds, reflecting the dual advantage of fine-tuning in increasing detection counts while ensuring localization reliability. Overall, transfer fine-tuning yields consistent performance gains across diverse defect categories and scales, simultaneously improving detection rates and strengthening the model's adaptability and robustness in handling complex backgrounds and small-object scenarios.

Figure 8
Comparison of detection performance before and after fine-tuning.



4.4.2. Ablation Experiments on Transfer Learning

To further validate the effectiveness and necessity of the proposed multi-stage transfer learning paradigm for wood defect detection on the two target domain datasets (Target-A and Target-B), three ablation experiments were carried out using the upgraded YOLOv12-SMD model. All experiments were performed under identical experimental set-

Table 6
Results of ablation experiments on transfer learning.

Method	Target-A				Target-B			
	P(%)	R(%)	mAP@50(%)	Ttrain(h)	P(%)	R(%)	mAP@50(%)	Ttrain(h)
YOLOv12-SMD	74.5	48.0	52.1	0.856	71.7	62.2	66.8	0.415
YOLOv12-SMD +SMD.pt	71.3	60.1	60.6	0.699	77.5	66.1	70.9	0.242
YOLOv12-SMD +SMD.pt +Fine-tuning	71.1	59.2	61.5	0.713	73.8	69.4	71.6	0.273

tings to ensure fair comparison and reproducibility. In the first scheme, the baseline model was trained from scratch solely on the target domain dataset, without incorporating pretrained weights or transfer fine-tuning. The second scheme employed pretrained weights obtained from source domain training, which were used for model initialization, but no fine-tuning was performed on the target domain. In the third scheme, the model was not only initialized with pretrained weights but also underwent fine-tuning with the target domain dataset. This hierarchical comparative design systematically isolates and reveals the contributions of pretrained weights and transfer fine-tuning to performance improvement. The corresponding experimental results are summarized in Table 6.

As shown in Table 6, different stages of the proposed multi-stage transfer learning paradigm exert a clear and consistent impact on detection performance across both target domain datasets. For Target-A, which represents a small-sample target domain, training the YOLOv12-SMD model from scratch results in limited detection coverage, as reflected by a low Recall of 48.0% and a moderate mAP@50 of 52.1%, despite a relatively high Precision of 74.5%. When the model is initialized with source-domain pretrained weights, Recall increases substantially to 60.1% and mAP@50 rises to 60.6%, indicating that pretrained representations provide effective prior knowledge for expanding defect coverage in the target domain. At the same time, the average training time is reduced from 0.856 hours to 0.699 hours, corresponding to an approximate 18% reduction, demonstrating improved convergence efficiency. After applying target-domain fine-tuning, the model achieves its best overall performance on Target-A, with mAP@50 further improving to 61.5%, while Precision and Recall stabilize at 71.1% and 59.2%, respectively. Although the training time increases slightly to 0.713 hours compared with weight initialization alone, it remains about 17% shorter than training from scratch, confirming the efficiency of the complete transfer process. A similar progressive trend is observed on Target-B, where baseline training already yields relatively higher performance. Initializing the model with pretrained weights leads to clear gains in Precision, Recall, and mAP@50, while reducing the average training time from 0.415 hours to 0.242 hours, representing a reduction of more than 40%. Subsequent fine-tuning further improves detection accuracy and Recall, with the training time remaining significantly lower than that required for training from scratch. These results indicate that the proposed multi-stage transfer learning paradigm not only enhances detection accuracy but also substantially accelerates optimization under more complex target domain conditions.

Overall, the stage-wise ablation results demonstrate that source-domain pretraining primarily contributes to faster convergence and improved defect coverage, while target-domain fine-tuning further refines performance. The combined effect yields both higher detection accuracy and markedly reduced training time, validating the effectiveness and necessity of the proposed multi-stage transfer learning paradigm.

4.4.3. Transfer Learning Comparison Experiments

To verify that the proposed YOLOv12-SMD algorithm integrated with the multi-stage transfer learning paradigm is more effective than conventional approach-

es that rely on pretraining on large-scale generic datasets, particularly under new working conditions with limited target-domain samples, a comparative evaluation was conducted. YOLOv12 provides five pretrained variants (n, s, m, l, and x), enabling flexible selection according to computational resources and latency requirements. Based on this framework, six representative improved modules were selected for comparison: TRS-YOLO+m.pt, BCS-YOLO+n.pt, GHBPN-YOLO+l.pt, ITD-YOLO+n.pt, SSH-YOLO+s.pt, and FSD-YOLO+x.pt. All comparison models were initialized using weights pretrained on large-scale public datasets, following common practice in existing transfer learning methods. Comparative experiments were performed on two target domain datasets (Target-A and Target-B), and all models were trained and evaluated under identical experimental settings to ensure fair and reproducible comparison. The corresponding experimental results are summarized in Table 7.

As summarized in Table 7, models pretrained on large-scale generic datasets exhibit clear limitations when transferred to wood defect detection, particularly in terms of Recall and mAP@50. This performance degradation mainly stems from the semantic and structural mismatch between generic pretraining data and wood defect characteristics, such as elongated geometries, weak texture contrast, and fine-scale boundary patterns, which leads to insufficient defect coverage in the target domain. In contrast, the proposed multi-stage transfer learning paradigm, when applied with the YOLOv12-SMD model, consistently achieves superior performance across both target domain datasets. On Target-A, YOLOv12-SMD attains a Recall of 59.2% and an mAP@50 of 61.5%, exceeding all comparison models. On Target-B, it further achieves a Recall of 69.4% and an mAP@50 of 71.6%, maintaining the leading performance under complex target domain conditions. In addition to accuracy improvements, YOLOv12-SMD demonstrates clear advantages in training efficiency. Across both target domains, its training time is reduced by approximately 17% to over 40% compared with the comparison models. These results indicate that the proposed multi-stage transfer learning paradigm enables more effective feature adaptation and faster convergence than conventional transfer strategies based on generic pretraining, thereby offering superior accuracy and efficiency for practical wood defect inspection under data-scarce conditions.

5. Conclusions

This study addresses key challenges in wood defect detection, including the difficulty of identifying small-scale defects, strong interference from complex textures, and limited cross-domain adaptability, by proposing an enhanced YOLOv12-SMD framework integrated with a multi-stage transfer learning paradigm. From the perspective of model

Table 7
Performance comparison of various models.

Model	Target-A				Target-B			
+Pretrained Weights	P(%)	R(%)	mAP@50(%)	T_{train}(h)	P(%)	R(%)	mAP@50(%)	T_{train}(h)
TRS-YOLO+m.pt	71.4	47.6	55.0	1.701	72.1	68.9	70.4	0.583
BCS-YOLO+n.pt	78.4	52.2	53.8	1.183	70.1	65.1	67.7	0.436
GHBPN-YOLO+l.pt	74.1	46.1	47.5	2.659	60.5	62.1	61.4	0.781
ITD-YOLO+n.pt	73.9	51.3	53.2	0.855	72.9	65.0	69.0	0.538
SSH-YOLO+s.pt	77.2	53.8	54.8	1.679	73.6	63.8	67.6	0.501
FSD-YOLO+x.pt	75.2	54.4	54.2	1.019	0.769	0.627	69.3	0.323
YOLOv12-SMD+SMD.pt	71.1	59.2	61.5	0.713	73.8	69.4	71.6	0.273

design, YOLOv12-SMD introduces three targeted improvements tailored to wood defect characteristics. First, the Shape-IoU loss incorporates shape- and scale-aware regression constraints, effectively improving localization accuracy for elongated and irregular defects. Second, the MLCA attention mechanism jointly models local spatial saliency and global channel dependencies, enhancing feature discriminability and robustness under complex texture interference. Third, the DySample dynamic upsampling strategy improves the adaptability and precision of multi-scale feature fusion, thereby strengthening the detection of small-scale and fine-grained defects. From the perspective of training strategy, a multi-stage transfer learning paradigm is developed to address practical constraints such as limited data availability, high annotation cost, and pronounced cross-domain distribution discrepancies. Unlike conventional transfer learning approaches that rely on generic public datasets, the proposed paradigm leverages a wood-specific source domain for pretraining, followed by progressive fine-tuning and optimization in the target domain. This hierarchical strategy effectively mitigates negative transfer under small-sample conditions while improving convergence stability and generalization capability.

Extensive experiments validate the effectiveness of the proposed approach. On a wood-specific source domain dataset, YOLOv12-SMD achieves strong detection performance, reaching an mAP@50 of 80.8% and outperforming representative baseline detectors by 1.7% to 3.9% in mAP@50. Further transfer learning experiments conducted on two target domain datasets demonstrate that the proposed multi-stage transfer learning paradigm maintains high detection accuracy under limited-sample conditions while reducing training time by approximately 16.7%, confirming its efficiency and robustness in practical deployment scenarios.

Despite these advantages, several limitations remain. The proposed method still relies on the availability of a reasonably related wood-specific source domain; its performance may degrade when the source and target domains exhibit extremely large semantic gaps or when entirely novel defect types are introduced. In addition, while the current framework effectively balances accuracy and efficiency, further improvements are required to enhance adaptability under more extreme imaging conditions, such as severe illumination variations or highly ambiguous defect boundaries. Future work will focus on extending the transfer learning paradigm to more diverse wood defect scenarios and exploring adaptive mechanisms to further improve robustness in highly dynamic industrial environments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Agrawal B. Wood Defect Detection Dataset. Roboflow Universe, 2023. Available at, Roboflow Universe.
2. An H., Liang Z., Qin M., Huang Y., Xiong F., Zeng G. Wood Defect Detection Based on the CWB-YOLOv8 Algorithm. Journal of Wood Science, 2024, 70(1), 26. https://doi.org/10.1186/s10086-024-02145-1
3. Chen Z., Liu B. A High-Accuracy PCB Defect Detection Algorithm Based on Improved YOLOv12. Symmetry, 2025, 17(7), 978. https://doi.org/10.3390/sym17070978
4. Detect D. Defect in Wood Dataset. Roboflow Universe, 2024. Available at, Roboflow Universe.
5. Detect Detect. New-Defects in Wood Dataset. Roboflow Universe, 2024. Available at, Roboflow Universe.
6. Dubey P., Miller S., Günay E. E., Jackman J., Kremer G. E., Kremer P. A. You Only Look Once v5 and Multi-Template Matching for Small-Crack Defect Detection on Metal Surfaces. Automation, 2025, 6(2), 16. https://doi.org/10.3390/automation6020016

7. Fang Y., Guo X., Chen K., Zhou Z. Accurate and Automated Detection of Surface Knots on Sawn Timbers Using YOLO-V5 Model. BioResources, 2021, 16(3), 5390. https://doi.org/10.15376/biores.16.3.5390-5406
8. Han S., Jiang X., Wu Z. An Improved YOLOv5 Algorithm for Wood Defect Detection Based on Attention. IEEE Access, 2023, 11, 71800-71810. https://doi.org/10.1109/ACCESS.2023.3293864
9. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN. In, Proceedings of the IEEE International Conference on Computer Vision, 2017, 2961-2969. https://doi.org/10.1109/ICCV.2017.322
10. Ji Y., Ma T., Shen H., Feng H., Zhang Z., Li D., He Y. Transmission Line Defect Detection Algorithm Based on Improved YOLOv12. Electronics, 2025, 14(12), 2432. https://doi.org/10.3390/electronics14122432
11. Jiang Q., Zhu X., Wang H., Chong X., Li L., Li W. YOLOv11n-CBP: A Lightweight Model for Crack Detection in Wooden Components Under Complex Backgrounds. Structures, 2025, 80, 110077. https://doi.org/10.1016/j.istruc.2025.110077
12. Kang J., Cen Y., Cen Y., Wang K., Liu Y. CFIS-YOLO: A Lightweight Multi-Scale Fusion Network for Edge-Deployable Wood Defect Detection. Wood Material Science & Engineering, 2025, 1-14. https://doi.org/10.1080/17480272.2025.2556340
13. Li D., Xie W., Wang B., Zhong W., Wang H. Data Augmentation and Layered Deformable Mask R-CNN-Based Detection of Wood Defects. IEEE Access, 2021, 9, 108162-108174. https://doi.org/10.1109/ACCESS.2021.3101638
14. Liu W., Lu H., Fu H., Cao Z. Learning to Upsample by Learning to Sample. In, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 6027-6037. https://doi.org/10.1109/ICCV51070.2023.00557
15. Lopes Jr D. V., Bobadilha G. D. S., Grebner K. M. A Fast and Robust Artificial Intelligence Technique for Wood Knot Detection. BioResources, 2020, 15(4), 9351. https://doi.org/10.15376/biores.15.4.9351-9361
16. Mohammed S. Y. Architecture Review: Two-Stage and One-Stage Object Detection. Franklin Open, 2025, 100322. https://doi.org/10.1016/j.fraope.2025.100322
17. Mustapha A. A., Yoosuf M. S. Exploring the Efficacy and Comparative Analysis of One-Stage Object Detectors for Computer Vision: A Review. Multimedia Tools and Applications, 2024, 83(20), 59143-59168. https://doi.org/10.1007/s11042-023-17751-2
18. 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, 2016, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
19. Shi J., Li Z., Zhu T., Wang D., Ni C. Defect Detection of Industry Wood Veneer Based on NAS and Multi-Channel Mask R-CNN. Sensors, 2020, 20(16), 4398. https://doi.org/10.3390/s20164398
20. Situ Z., Teng S., Feng W., Zhong Q., Chen G., Su J., Zhou Q. A Transfer Learning-Based YOLO Network for Sewer Defect Detection in Comparison to Classic Object Detection Methods. Developments in the Built Environment, 2023, 15, 100191. https://doi.org/10.1016/j.dibe.2023.100191
21. Situ Z., Teng S., Liao X., Chen G., Zhou Q. Real-Time Sewer Defect Detection Based on YOLO Network, Transfer Learning, and Channel Pruning Algorithm. Journal of Civil Structural Health Monitoring, 2024, 14(1), 41-57. https://doi.org/10.1007/s13349-023-00681-w
22. Tian Y., Ye Q., Doermann D. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv Preprint arXiv:2502.12524, 2025. DOI: 10.48550/arXiv.2502.12524.
23. Wan D., Lu R., Shen S., Xu T., Lang X., Ren Z. Mixed Local Channel Attention for Object Detection. Engineering Applications of Artificial Intelligence, 2023, 123, 106442. https://doi.org/10.1016/j.engappai.2023.106442
24. Wang J., Li B. TRS-YOLO: A Lightweight Insulator Defect Detection Method Based on Enhanced YOLOv12. Journal of Real-Time Image Processing, 2025, 22(5), 1-18. https://doi.org/10.1007/s11554-025-01754-3
25. Wang Q., Wu B., Zhu P., Li P., Zuo W., Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 11534-11542. https://doi.org/10.1109/CVPR42600.2020.01155
26. WOOD. Wood Surface Dataset. Roboflow Universe, 2025. Available at, Roboflow Universe.
27. Zhang C., Shao Y., Sun H., Xing L., Zhao Q., Zhang L. The WuC-Adam Algorithm Based on Joint Improvement of Warmup and Cosine Annealing Algorithms. Mathematical Biosciences and Engineering, 2024, 21(1). https://doi.org/10.3934/mbe.2024009
28. Zhang H., Zhang S. Shape-IoU: More Accurate Metric Considering Bounding Box Shape and Scale. arXiv Preprint arXiv:2312.17663, 2023. DOI: 10.48550/arXiv.2312.17663.

29. Zhang M., Song G., Ma N. A Mechanism for Upgrading the Global Value Chain of China's Wood Industries Based on Sustainable Green Growth. Journal of Cleaner Production, 2024, 449, 141717. https://doi.org/10.1016/j.jclepro.2024.141717

30. Zhang Q., Liu L., Yang Z., Yin J., Jing Z. WLSD-YOLO: A Model for Detecting Surface Defects in Wood Lumber. IEEE Access, 2024, 12, 65088-65098. https://doi.org/10.1109/ACCESS.2024.3395623

31. Zheng Y., Wang M., Zhang B., Shi X., Chang Q. GB-CD-YOLO: A High-Precision and Real-Time Lightweight Model for Wood Defect Detection. IEEE Access, 2024, 12, 12853-12868. https://doi.org/10.1109/ACCESS.2024.3356048

32. Zhou P., Xie X., Lin Z., Yan S. Towards Understanding Convergence and Generalization of AdamW. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9), 6486-6493. https://doi.org/10.1109/TPAMI.2024.3382294

33. Zhu J., Zhao D., Luo X. Evaluating the Optimised YOLO-Based Defect Detection Method for Subsurface Diagnosis with Ground Penetrating Radar. Road Materials and Pavement Design, 2024, 25(1), 186-203. https://doi.org/10.1080/14680629.2023.2199880

34. Zielinski K. M., Scabini L., Ribas L. C., da Silva N. R., Beeckman H., Verwaeren J., Bruno O. M., De Baets B. Advanced Wood Species Identification Based on Multiple Anatomical Sections and Using Deep Feature Transfer and Fusion. Computers and Electronics in Agriculture, 2025, 231, 109867. https://doi.org/10.1016/j.compag.2024.109867

35. Zou X., Wu C., Liu H., Yu Z., Kuang X. An Accurate Object Detection of Wood Defects Using an Improved Faster R-CNN Model. Wood Material Science & Engineering, 2025, 20(2), 413-419. https://doi.org/10.1080/17480272.2024.2352605