I²GF-Net: Inter-layer Information Guidance Feedback Networks for Wood Surface Defect Detection in Complex Texture Backgrounds

Qiwu Luo, *Senior Member*, *IEEE*, Wenlong Xu, Jiaojiao Su, Chunhua Yang, *Fellow*, *IEEE*, Weihua Gui, *Member*, *IEEE*, Olli Silvén, *Senior Member*, *IEEE*, and Li Liu, *Senior Member*, *IEEE*

***Abstract*—Visual surface defect detection is crucial for product quality control in the large-scale wood manufacturing industry. This study focuses on how to assist deep learning model in surviving in the challenges brought by complex texture backgrounds. A novel visual defect detection model, inter-layer information guidance feedback networks (I²GF-Net), is proposed in this paper. To be specific, a top-down feedback encoder (TDFE) is proposed to guide the attention of the low-level feature map, enabling it to focus on the defect regions by incorporating enhanced high-level semantic information. This significantly reduces false positives triggered by intense textures. Meanwhile, a semantic feature texture enhancement (SFTE) method is designed to compensate for high-level semantic features with fine-grained local information, thereby avoiding frequently missed detections resulted from multiple down-sampling in deep models. Furthermore, we provide an option of dual-round feature refinement (DRFR) to pursue a higher mAP in scenarios where sacrificing a certain amount of time is acceptable. Experimental results demonstrate the I²GF-Net outperforms *13* state-of-the-arts on two benchmark datasets (VSB-DET and NEU-DET), as well as our newly opened wood dataset (OULU-DET), which will be publicly available at http://www.ilove-cv.com/oulu-wood/.**

***Index Terms*—Visual surface defect detection, wood surface defect, complex texture backgrounds, inter-layer information guidance.**

I. INTRODUCTION

WOOD material, with its natural green merits, plays vital role in various aspects of daily life and infrastructures. Any surface defects suffering on wood materials reduce its aesthetics and mechanical properties, which directly affect the quality of the end products [1]. Automated visual inspection (AVI) instrument targeting on surfaces emerges as a fundamental assurance for wood manufacturing industry to promote production quality and outputs [2][3]. Wood, as a natural material, possesses its unique features [4]. Notably, textures such as intensive tree rings and various knots normally exist on the wood products like sawn timbers, planks, etc. These textures may be what consumers expect in terms of wood defect detection precisely. The challenging textures are clearly evident in Fig. 1, and they pose two adverse impacts on the AVI system:1) *Frequent false positives*. The intense textures, referring to the green dash box in Fig. 1(a), may be incorrectly identified as defects. While texture feature-based methods aim to detect regions that disrupt homomorphic properties, they are not specialized in distinguishing between defects and textures. 2) *Inevitable missed detections*. The subtle defects, referring to the orange dash box in Fig. 1(b), are often concealed within texture backgrounds, making them difficult to detect or even overlooked by the AVI system. While the multiple down-sampling scheme of deep learning methods tends to lost the fine-grained image details, so then tiny objects is easy to be ignored. Furthermore, as shown in red dash box in the Fig. 1(a), some defects suffer with blurring boundaries, which might result in many coarse bounding boxes at the stage of defect localization.

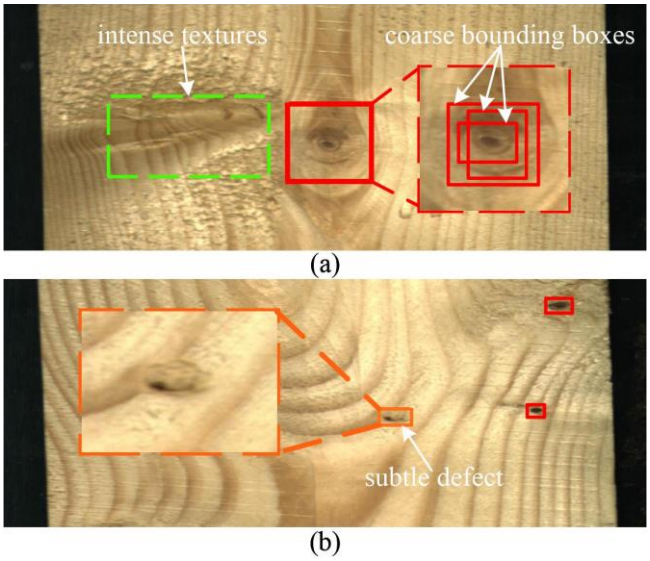

Fig. 1. Defect samples in complex wood texture backgrounds.

This work was supported in part by the National Natural Science Foundation of China under Grant 62322317 and Grant 6202781, and by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC3019. (Corresponding Author: Chunhua Yang: ychh@csu.edu.cn)

Qiwu Luo, Wenlong Xu, Jiaojiao Su, Chunhua Yang, and Weihua Gui are with the School of Automation, Central South University, Changsha 410083, China.

Olli Silvén is with the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, 90014 Oulu, Finland.

Li Liu is with the College of System Engineering, National University of Defense Technology, Changsha 410073, China.

In response to the aforementioned challenges, scholars are actively exploring ways to enhance the performance of detection models by effectively integrating low-level textural and high-level semantic information. The goal is to achieve a more balanced performance that minimizes both false positives and missed detections, while improving the accuracy of correct detections. Notably, Shi *et al.* [5] proposed a novel analysis-by-synthesis vision transformer (AbSViT) mainly based on top-down vision attention, the significant improvement on visual segmentation and classification is attributed to the strong priors mined from high-level semantic information. On the contrary, Zhu *et al.* [6] designed a statistical texture learning network, a novel quantization and counting operator (QCO) is proposed to describe the texture information, so as to help capturing the invisible small targets.

Building upon the aforementioned references, our investigation focused on the fundamental fact that deep learning models have the capability to extract complex and abstract feature representations, specifically semantic information, from wood surface images. This inherent capability strengthens detection networks, enabling them to effectively differentiate defects from intense texture backgrounds. Additionally, texture analysis can capture detailed features that may be overlooked by purely relying on semantic information. Following this roadmap, this paper proposes an innovative approach called Inter-Layer Information Guidance Feedback Networks (I²GF-Net) to address the challenges posed by complex wood texture backgrounds. The key idea is to leverage both semantic and texture information in an inter-layer manner, allowing for more comprehensive and robust defect detection. The contributions are as follows:

First, ***for lower false positives***, a top-down feedback encoder (TDFE) is proposed to guide the attention of the low-level features to focus on defect regions by enhancing defect semantics from high-level feature maps, which significantly reduces false positives triggered by intense textures.

Second, ***for lower missed detections***, a scheme of semantic feature texture enhancement (SFTE) is designed to compensate high-level semantic features with fine-grained information. The missed detection of subtle defects can be suppressed to a large extent.

In addition, a framework of dual-round feature refinement (DRFR) is developed to refines defect features by reusing the backbone networks, so as to further enhance the defect localization precision. Under the DRFR framework, two versions of I²GF-Net have been provided: I²GF-Net-d with DRFR for higher detection accuracy (Mean Average Precision, mAP), and I²GF-Net-s without DRFR for higher detection speed (Frames Per Second, FPS).

The subsequent sections are organized as follows: Section II reviews related research efforts. Section III delineates the framework of the proposed I²GF-Net. Section IV introduces extensive experiments and discussions. Finally, Section V presents the conclusions of this paper.

II. Related Works

A. Wood Defect Detection

Machine vision has become increasingly popular for wood defect detection in recent years due to its cost-effectiveness, high speed, accuracy, and user-friendly nature. Initial works in this field focused on extracting features from wood images including color, texture, and gray-level co-occurrence matrix. Then, classifiers such as clustering, support vector machines, genetic algorithms, and neural networks have been used to differentiate between defects and normal regions. However, these methods are vulnerable to noise interference [7], and defect attribute extraction heavily relies on human expertise.

Recently, CNN-based defect detection methods for wood images have been proposed. Hu *et al.* [8] used ResNet18 for wood image classification using deep learning. Ren *et al.* [9] introduced a classifier based on image patch features, followed by pixel prediction using the trained classifier. Jung *et al.* [10] utilized three different CNN architectures to detect defects in wood with surfaces featuring random textures. However, these methods exhibit low accuracy due to difficulties adapting to multi-scale variations in defects and distinguishing defects from normal background texture.

Despite extensive CNN-based defect detection studies, the complex wood surface texture and various defect types create a challenging, high-dimensional state space for extracting effective features.

B. Top-down Attention Mechanism

Top-down attention exhibits unique neural features in brain regions that process sensory signals [11] and is critical for selecting information relevant to behavioral goals. Inspired by human visual processing, several studies have applied top-down attention to object detection [5]. Zhang *et al.* [12] proposed an adaptive asymmetric fusion module to exchange multi-scale contexts between lower and higher levels, enriching the decoding of semantic information and spatial details. Liu *et al.* [13] devised a U-shaped architecture that progressively refines higher-level features through bottom-up and top-down pooling modules, enhancing the role of pooling in CNN models for Salient Object Detection.

In our study, we consider defects as anomalous regions in an image. In order to distinguish defects and intense textures, we explore a method to enhance defect semantics from the highest-level feature maps to guide low-level feature maps in focusing on defect regions.

C. Texture Enhancement

Complex texture backgrounds on the surface of wood may obscure the features of subtle defects. Texture features efficiently capture the grayscale distribution and spatial organization of the image surface, proving beneficial for extracting features of subtle defects. Some studies have enriched defect representation by enhancing defect texture features [6]. Xu *et al.* [14] introduced sparse binary convolution filters for finer encoding of local textures. Fan *et al.* [15] referenced the texture structure of receptive fields in the human visual system and proposed a texture enhancement module. Liang *et al.* [16] employed wavelet

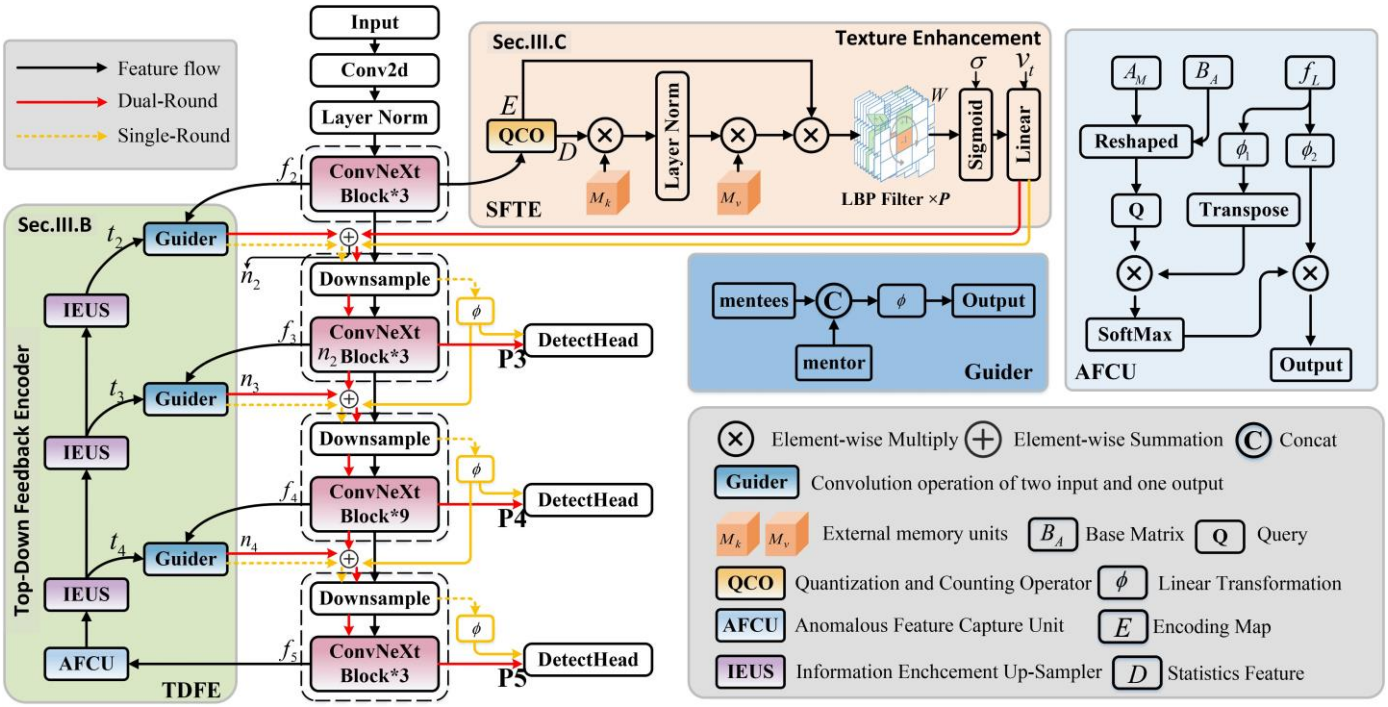

Fig. 2. Overall architecture of the proposed I²GF-Net.

transformation to weaken texture information and balance grayscale distribution.

However, as texture feature extraction is unsuitable for obtaining high-level image content, relying solely on texture can only partially reflect the intrinsic properties of objects. Therefore, we position texture features as a complement to high-level features. Inspired by [6] and Local Binary Patterns (LBP), we designed a texture enhancement method to compensate for high-level features by extracting fine-grained local information from shallow feature maps.

D. Feature Refinement

Various Defects distributed in varied texture backgrounds may exhibit similarity with the background texture, leading to a blurred boundary of the defects. Advanced fine-feature extraction methods have been widely employed to enhance the model's capability in recognizing defect boundaries. Liu *et al.* [17] designed a composite backbone network, iteratively refining edge details of target features by connecting adjacent backbone networks and passing feature output from the previous network to the next sequentially. Wang *et al.* [18] enhanced the feature content by compressing it and weighted encoding. Zhu *et al.* [19] employed feature-guided decoders to progressively refine multi-scale texture information, effectively focusing on blurry object boundaries. Wen *et al.* [20] introduced a feature-based domain unraveling and randomization framework, successfully segmenting cracks with ambiguous boundaries using multi-scale contextual features and cross-attention mechanisms.

Given that our priority is accurate defect recognition rather than precise defect localization, we present this potentially computationally demanding feature refinement scheme as an optional solution, primarily applied in scenarios requiring higher precision.

III. Proposed Method

A. Architecture of I²GF-Net

The structure of our proposed I²GF-Net is shown in Fig. 2. The design motivations are as follows:

Firstly, for a backbone network comprising L layers, we designate the input image as $X \in \mathbb{R}^{3\times H\times W}$ and the output of each layer as $f_i \in \mathbb{R}^{C_i\times H_i\times W_i}$, where C_i stands for the number of channels in the output of the *i*-th layer, H_i and W_i denote the spatial dimensions of the feature map, with $i = 1,\cdots,L$. The feature extraction process of the backbone network operates layer by layer, with f_{i+1} being equal to $M_i(f_i)$ for each layer output, where $M_i(\cdot)$ denotes the feature extraction module for the *i*-th layer. Furthermore, with an increase in the number of layers, C_i also increases, usually following the relation $C_{i+1} = 2\times C_i$. Due to the down-sampling process, the spatial dimensions of each layer output also change, typically resulting in $H_{i+1} = H_i/2$ and $W_{i+1} = W_i/2$. Considering the diverse morphologies of wood defects, we balance accuracy and computational cost by selecting ConvNeXt [21] as the backbone network for initial feature extraction.

Secondly, we introduce the abnormal feature capture unit (AFCU) to capture the anomalous regions, which focuses on analyzing and perceiving the feature distribution and differences among dataset samples. Additionally, to account for the sensitivity of feature maps at different levels to targets of various scales, we introduce the information enhancement up-sampler (IEUS) to extract and transmit guidance information to different scale levels. AFCU and IEUS together constitute the top-down

feedback encoder (TDFE).

Next, we introduce texture features to preserve the data's inherent structure and capture small yet crucial features. A texture enhancement branch is brought out at the shallow level, and a semantic feature texture enhancement (SFTE) method is designed to enhance defect texture features.

Then, combine high-level semantic information with low-level texture information. The feature maps undergo refinement through the backbone network in the dual-round solution, while the single-round solution does not. Both solutions generate three feature maps at shallow, medium, and deep levels.

Finally, the three feature maps are through 1×1 convolution to generate multiple anchor boxes. Subsequently, the CIOU loss between each anchor box and its corresponding ground truth is calculated. During the training process, the SGD function minimizes the loss function. In the detection stage, a non-maximum suppression method filters the bounding boxes.

B. Top-Down Feedback Encoder

1) Abnormal Feature Capture Unit: At the feature extraction stage of the backbone network, the input image undergoes multilevel feature extraction to generate a high-level feature map $f_L \in R^{C_L \times H_L \times W_L}$ containing rich semantic information. To accurately capture critical features, we introduce a learnable affinity graph $A_M \in \mathbb{R}^{C_A \times H_L W_L}$ to acquire statistical features of the image [22]. Typically, regions of the image that deviate from statistical features are considered potential anomaly features [23]. Therefore, we employ an attention mechanism to handle the affinity graph, enabling it to interact with every position in the wood image, thus precisely locating anomalous features. To address potential feature redundancy and mitigate overfitting issues related to large-weight tensors [24], we employ the basis matrix $B_A \in \mathbb{R}^{C_L \times C_A}$ to create a reshaped version of the affinity graph. This reshaped version is then incorporated into the query for the attention computation to obtain top-level guidance features.

$$\hat{f}_L = Softmax\left(\left(B_A \cdot A_M\right) \cdot \phi_1(f_L)^T\right) \cdot \phi_2(f_L) \tag{1}$$

Where ϕ_1 and ϕ_2 are two different linear transformation matrices and $Softmax(\cdot)$ is a normalization function.

This unit provides critical statistical features for wood surface defect detection, enabling the model better to capture potential defect regions and attenuate intense texture regions.

2) Information Enhancement Up-Sampler: Considering the sensitivity of different levels of feature maps to targets of various scales, we introduce an up-sampler to handle these feature maps. As show in Fig. 3, the guidance feature $f_{i+1} \in \mathbb{R}^{C_{i+1} \times H_{i+1} \times W_{i+1}}$ is compressed to $f \in \mathbb{R}^{C' \times H_{i+1} \times W_{i+1}}$ by 1×1 convolution to reduce the computation, where $C' = C_{i+1} // r$, and r represents the channel compression factor. Then, the deformable content encoding [25] of the compressed feature is computed. To better adapt to the diverse shapes of defects on wood surfaces. Subsequently, the 2D size of the content encoding is expanded by the PixelShuffle method and normalized using Softmax to obtain the content reassembly kernel used to do up-sampling of the guidance feature.

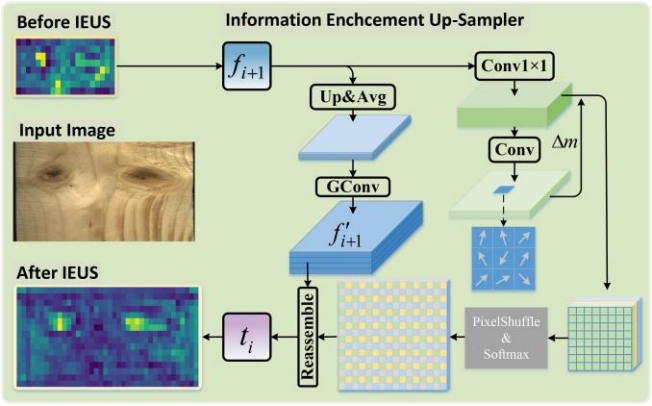

Fig. 3. Information Enhancement Up-Sampler.

$$\tilde{f}_{(d,i)} = \sum_{j \in \Omega(i)} W\left[p_i - p_j\right] f_{(d,j+\Delta j)} \cdot \Delta m_{(j+\Delta j)} \tag{2}$$

$$W_l = \mathrm{softmax}\left(S\left(\tilde{f}\right)\right) \tag{3}$$

Where $\tilde{f}_{(d,i)}$ represents the content encoding of the *i*-th pixel in the *d*-th channel. Δ_j denotes the learnable sampling offset component for the *j*-th pixel, and $\Delta m_{(j+\Delta j)}$ is the learnable feature modulation coefficient. $\Omega(i)$ represents a convolution window of size $k_c \times k_c$ centered around the *i*-th pixel. $W \in \mathbb{R}^{C' \times C' \times k_c \times k_c}$ is the content encoding convolution kernel, $W\left[p_i - p_j\right] \in \mathbb{R}^{C' \times C'}$ denotes the convolution kernel parameters corresponding to the positional offset between the *i*-th and *j*-th pixels. p_i represents the 2D pixel coordinates. $S(\cdot)$ denotes the PixelShuffle method, $W_l \in \mathbb{R}^{C'/4 \times 2H_{i+1} \times 2W_{i+1}}$.

After up-sampling the input feature map with nearest neighbor interpolation, the receptive field space features are extracted by 2D average pooling. Then, the feature map is unfolded by group convolution. Finally, the enhanced guidance information up-sampled features are obtained by content reassembly, and a linear transformation matrix changes the number of channels of the features to match the dimension of the low-level feature map.

$$f'_{i+1} = G\left(Avg\left(Up\left(f_{i+1}\right)\right)\right) \tag{4}$$

$$t_i = \sum_{n=-p}^{p} \sum_{m=-p}^{p} W_l(n,m) \cdot f'_{i+1}(n,m) \tag{5}$$

Where $Up(\cdot)$ denotes nearest neighbor interpolation, $Avg(\cdot)$ is 2D average pooling with a pooling window size of 5, $G(\cdot)$ denotes group convolution with a convolution kernel size of 1. $p = \lfloor k_c / 2 \rfloor$, $n_i \in \mathbb{R}^{C_i \times H_i \times W_i}$.

By cascading this up-sampler, the guidance information can be transmitted to different scale levels, thereby providing attention guidance to feature maps at various levels to accurately detect defects on wood surfaces and reduce the false positive rate.

C. Semantic Feature Texture Enhancement

Previous research has shown that histogram equalization [26] enhances image contrast by adjusting the grayscale distribution, thereby making defect regions more prominent. Additionally, Local Binary Pattern (LBP) [27] have been employed to capture minute textural variations in images, further improving the ability to identify defect regions. Motivated by these methods, we developed a Semantic Feature Texture Enhancement (SFTE) approach and introduced a branch at a shallow level to extract fine-grained local information.

We introduced a quantization and counting operator (QCO) [6] for multiple-level quantization to characterize the feature distribution statistically. The input feature map undergoes global average pooling, and feature vectors are computed using cosine similarity, generating encoding map $E \in R^{L\times HW}$ and statistics feature $D \in R^{L\times C}$. Where *L* denotes the quantization levels, and *C* represents the dimension of the statistical feature. These are expanded into a learnable graph to reconstruct each quantization level.

For the reconstruction process, we use two memory units, $M_k \in \mathbb{R}^{C\times d}$ and $M_v \in \mathbb{R}^{C\times d}$, implicitly considering the effects between all samples [28]. Attention maps of statistical features are inferred from dataset-level knowledge learned by M_k. The input features are updated based on similarities in the attention mapping to obtain the reconstructed quantization level, yielding the contrast-enhanced feature.

$$x = \left(Norm\left(DM_k^T \right) \cdot M_v \right) \cdot E \tag{6}$$

Where $Norm(\cdot)$ is the normalization function.

Traditional LBP compares the pixel values around the center pixel in clockwise order and then generates the final LBP pattern by linear combination. When applied to the entire image, a texture map is formed. The formula for LBP is:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s\left(g_p - g_c \right) 2^p \tag{7}$$

Where *P* represents the number of filters, which is usually set to 8 in traditional LBP, and $s(\cdot)$ represents the Heaviside step function.

We found that the computational process of traditional LBP has similarities with convolutional operations, so we attempt to provide an alternative to the traditional LBP formulation to further explore its potential in enhancing defect texture features.

$$y_{(d,i)} = \sum_{t=0}^{P-1} \sigma \left(\sum_{j\in\Omega(j)} W\left[p_i - p_j \right] x_{(d,j)} \right) \cdot v_t \tag{8}$$

Where $\sigma(\cdot)$ denotes a non-linear threshold operation, *W* denotes a fixed filter, *x* is the original input image, and v_t denotes a learnable linear weight parameter.

We define *P* fixed differential convolution filters distributed in clockwise order and *P* learnable weights. The contrast-enhanced feature passes through these fixed filters to produce *P* difference maps, activated by a nonlinear function. We replace the non-differentiable Heaviside step function [29] in LBP with a differentiable activation function (Sigmoid) for backpropagation. The *P* learnable weights linearly combine the disparity maps thus obtaining a texture feature map.

SFTE enhances the representation capability of texture features in the process of wood defect detection by introducing a shallow texture enhancement branch and compensating for high-level semantic features. This activation of minor defect characteristics reduces the omission rate of subtle defects on the wood surface.

D. Dual-Round Feature Refinement

We designed a framework of dual-round feature refinement (DRFR), which aims to fully leverage the texture and semantic information within images to achieve accurate detection and localization of surface defects in wood.

Under the framework, the backbone network initially generates a set of initial feature sets $F = \{ f_1, f_2, f_3, f_4, f_5 \}$. Subsequently, the SFTE processes shallow-layer features to extract texture features. Following this, the AFCU captures potential defect regions in the highest-level feature map f_5, yielding the top-level guidance features. These guidance features are then further optimized through three iterations of the IEUS to enhance their adaptability and expressive capability, thus obtaining more accurate guidance information sets $T = \{ t_2, t_3, t_4 \}$. Finally, a subset $F' = \{ f_2, f_3, f_4 \}$ of the feature set *F* is merged with the guidance information set *T* to form an image feature set containing guidance information.

$$n_i = CBR\left(Cat\left(f_i, t_i \right) \right) \tag{9}$$

Where $Cat(\cdot)$ denotes the concatenation operation, and $CBR(\cdot)$ represents the convolution, BatchNorm, and ReLU activation.

In dual-round solution, we reuse the backbone network for dual-round feature refinement.

$$P_i = M_i(n_{i-1} + P_{i-1}) \tag{10}$$

Where $M_i(\cdot)$ refers to the *i*-th feature extraction module and n_{i-1} represents the guidance feature from the preceding layer.

In each feature extraction iteration, the correlation between each module and the guidance features from the previous layer ensures a continuous focus of the network on defect regions, thereby enhancing detection accuracy.

During this iterative process, the initial input consists of texture-enhanced features extracted from shallow-layer feature maps. These features are combined element-wise with the corresponding guidance features to ensure that the resulting feature set encompasses both texture details and semantic information.

$$n_2 = SFTE\left(f_2 \right) + CBR\left(f_2, t_2 \right) \tag{11}$$

After the dual-round process, the final feature set $P = \{ P_3, P_4, P_5 \}$ is generated, characterized by rich semantic information and intricate texture features, providing a more comprehensive and accurate input for subsequent defect detection. Finally, the feature set is input to the subsequent detect head to obtain the prediction result.

Consider that the dual-round solution requires a more significant amount of computation. We provide a single-round solution with the output of each layer. This alternative approach processes the output of the backbone network through a linear transformation matrix to reduce computational costs and enhance operational speed.

$$P_i = \phi_i(n_{i-1} + P_{i-1}\downarrow) \tag{12}$$

Where ϕ_i is the linear transformation matrix and $\downarrow$ is the down-sample module in the backbone network.

IV. Experimental Results And Analysis

In this section, we provide a comprehensive description of the experimental setup, covering details about the experimental environment, the dataset used, and the evaluation metrics. Following this, we present and analyze the experimental results to validate the effectiveness of our proposed I²GF-Net.

A. Dataset

1) VSB-DET: We evaluated the performance of our study based on the publicly available wood surface image dataset [2] provided by the VSB-Technical University of Ostrava, referred to as VSB-DET. The dataset was developed through their hardware and software solution to acquire images on a conveyor belt in an industrial environment at a speed of 9.6 meters per second. The acquisition rate was 66 kHz, accompanied by constant and intense vibrations. The dataset comprises 20,275 wood surface images, each sized at 2800×1024 pixels, encompassing various surface defects. Considering the continuous nature of environmental variations on the production line, we randomly sampled 2800 images from the original dataset to expedite the experimental process. These images were divided into a ratio of 7:3 for the training and testing sets and were subsequently utilized for model training and evaluation. To ensure the breadth and adequacy of the dataset, we retained the types of prevalent defects and did not include defects that were too rare. More detailed information on the distribution of defects can be found in Table I, including Live Knot (LK), Dead Knot (DK), Marrow (Ma), Resin (Re), Knot with crack (KwC), Knot missing (KM), Crack (Cr).

2) OULU-DET: We further evaluated the results within the first publicly available wood defect database [30], referred to as OULU-DET. The dataset originates from the University of Oulu. It comprises 839 images of spruce wood captured using a 3200 K controlled halogen lamp. Each image is in RGB color, with 8 bits per channel, sized at 512×488 pixels, and the error due to lighting variations remains within 1%. The images in the dataset are annotated and segmented into rectangular regions, with each rectangle corresponding to an approximately 2.5 × 2.5 square centimeter area of the wood surface. Notably, the original dataset utilized images in ppm format, deviating from mainstream database standards, and the label information was presented in a slide-window manner with somewhat coarse defect coordinate positions, potentially impacting detection performance and limiting widespread usage. To address these issues, following confirmation from the original author, we converted the ppm format images to .jpg format and re-annotated the images in PASCAL VOC style, with each image accompanied by a .xml annotation file. The new labeling utilized a bounding box format better suited for defect detection. We also conducted data cleaning, rectified specific label errors, and divided the dataset into training. Testing sets with a ratio of 7:3. More detailed information on the distribution of defects can be found in Table II, including dry knot (DrK), sound knot (SK), encased knot (EK), knot hole (KH), horn knot (HK), leaf knot (LK), split (Sp), wane (Wa), core stripe (CS), decayed knot (DeK).

TABLE I
DEFECT STATISTICS IN VSB-DET

Defect	Occurrences	Defective Images	Frequency/%
LK	2943	1654	59.07
DK	1642	1144	40.85
Ma	161	144	5.14
Re	486	372	13.28
KC	317	260	9.28
KM	65	64	2.28
Cr	288	210	7.50

TABLE II
DEFECT STATISTICS IN OULU-DET

Defect	Occurrences	Defective Images	Frequency/%
DrK	304	236	28.13
SK	567	324	38.62
EK	167	138	16.45
KH	32	30	3.57
HK	40	37	4.41
LK	58	45	5.36
Sp	235	137	16.32
Wa	379	270	32.18
CS	67	55	6.55
DeK	24	20	14.89

3) NEU-DET: In addition to the wood dataset above, we included a steel surface defect dataset to evaluate our method's generalisation performance. NEU-DET [31] is a collection of surface defect images on steel strips, compiled and publicly shared by Northeastern University. This dataset comprises 1800 defect images at a resolution of 200×200 pixels. All defects are categorised into six major classes: Crazing (Cr), Inclusion (In), Patches (Pa), Pitted Surface (PS), Rolled-in Scale (RS) and Scratches (Sc), each containing 300 images per defect class. Within all defect images, defect positions and types are annotated in the VOC format. We randomly partitioned the 1800 defect images into training and validation sets using a 7:3 split ratio for our experiments.

B. Experimental Setup

1) Implementation Details: The hyperparameters for I²GF-Net are set as follows. The optimizer used is the Stochastic Gradient Descent (SGD) algorithm with a momentum of 0.9. The initial learning rate is set to 0.0005, optimized using the LambdaLR method. Based on the differences in image sizes across different datasets, we set the network input size to 1280×512 for VSB-DET, 640×640 for OULU-DET, and 224×224 for NEU-DET. For TDFE, we set C_A=32 and k_c =5. For STFE, we set L=128, d=128, and P=8. The hardware environment for experimentation

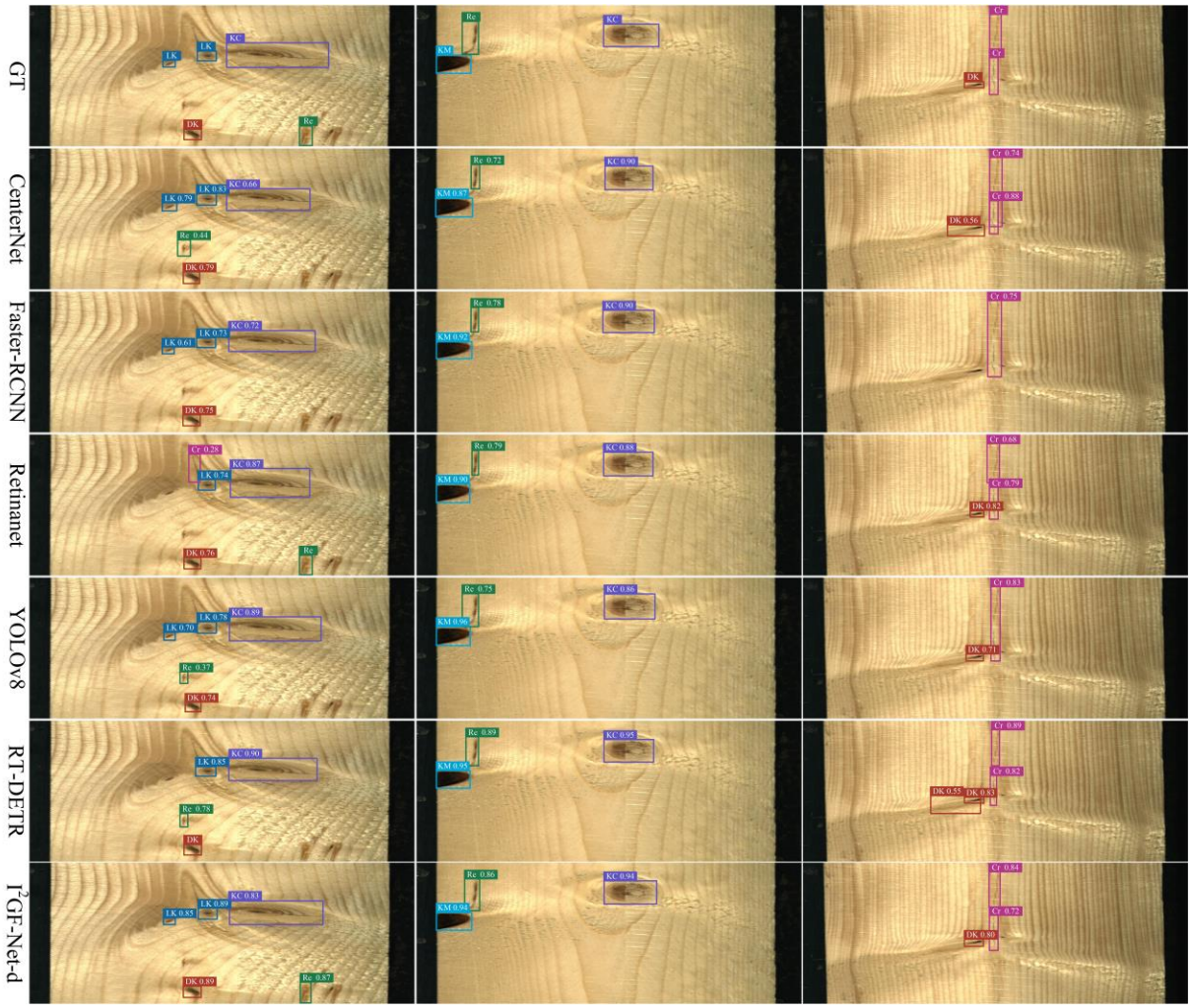

Fig. 4. Detection results of different algorithms on VSB-DET.

comprises 1 Intel(R) Xeon(R) W-2255 CPU @ 3.70GHz with 10 cores, 2 GeForce RTX 3090 graphics cards, 32 GB of memory, and a software environment of LINUX64 Ubuntu 20.04.6 with the PyTorch 1.12.1 framework.

2) Comparison Methods: The thirteen following methods are compared. 1) CenterNet [32], which transforms the object detection problem into a task of predicting object centers and sizes. 2) Faster-RCNN [33], which integrates a region proposal network to generate candidate regions and achieves remarkably high precision in detection. 3) RetinaNet [34], which introduces 'Focal Loss,' a specialized loss function addressing class imbalance issues. 4) YOLOv5 [35], which features a more efficient network structure and incorporates diverse data augmentation techniques. 5) YOLOv7 [36], which introduces model re-parameterization into the network architecture and proposes a training method for auxiliary heads. 6) YOLOv8 [37], which performs object detection through unique dual-path prediction and closely connected convolutional networks. 7) DETR [38], which approaches object detection as a set prediction problem and presents a highly concise object detection pipeline. 8) RT-DETR [39], which designs an efficient hybrid encoder and proposes an IoU-aware query selection mechanism. 9) PP-YOLOE [40], which introduces an ET head aimed at speed and accuracy. 10) RTMDet [41], which balances the computational load of various model components and employs dynamic soft labels to optimize training strategies. SFTE-Net [42], which proposes a multi-stage approach based on a pooling attention mechanism and cross-scale shallow feature reinforcement. CCG-YOLOv7 [43], which introduces a rapid supervised attention module to connect the backbone layers with the head layers, while simplifying the head layers. EAE-YOLOX [44], which introduces an efficient channel attention mechanism and adaptive spatial feature fusion mechanism while improving confidence loss and localization loss functions.

C. Evaluation Metric

We employ precision (P) and recall (R) as evaluation metrics to assess the model's accuracy in detecting wood defects. An excellent detection model should demonstrate both high recall and high precision. The F_1-score serves as a measure of this criterion. Additionally, the mean average precision (mAP) provides an indication of the average recognition accuracy across all wood defect categories.

TABLE III
DEFECT DETECTION RESULTS ON VSB-DET DATA

Model	mAP/%	FPS	AP/%							F_1-score	P/%	R/%
			LK	DK	Ma	Re	KC	KM	Cr			
CenterNet [32]	73.1	78.5	73.5	79.5	88.6	71.1	65.5	79.7	48.2	70.8	72.2	69.4
Faster-RCNN [33]	75.3	31.5	75.6	81.0	90.2	72.4	70.1	**86.1**	51.5	73.9	73.7	74.1
Retinanet [34]	73.0	48.9	73.6	79.7	86.5	72.2	68.5	83.7	46.9	70.2	68.4	72.1
YOLOv5 [35]	73.4	92.4	73.0	76.1	88.9	74.3	70.2	80.0	51.2	72.3	73.5	71.1
YOLOv7 [36]	74.1	103.2	74.5	79.1	90.2	73.4	66.9	77.0	57.5	73.2	73.8	72.6
YOLOv8 [37]	74.6	**114.5**	72.8	**81.7**	90.6	73.3	69.6	80.6	53.5	73.6	73.6	73.6
DETR [38]	72.4	33.5	71.7	73.7	87.5	73.1	68.4	79.3	53.2	71.4	71.7	71.1
RT-DETR [39]	74.5	38.6	73.2	78.6	88.1	72.3	70.8	81.6	56.9	72.3	72.7	71.9
PP-YOLOE [40]	72.4	108.7	74.1	79.0	86.0	69.6	67.5	81.0	49.5	71.5	71.0	72.0
RTMDet [41]	72.6	96.3	74.3	78.6	87.5	69.9	68.2	81.4	48.6	71.8	71.6	72.0
STFE-Net* [42]	76.2	80.4	76.1	79.4	91.3	74.9	72.5	81.8	57.4	74.6	75.9	73.4
CCG-YOLOV7* [43]	74.9	94.2	72.4	79.2	89.7	71.8	71.9	80.3	59.3	73.8	74.4	73.2
EAE-YOLOX* [44]	75.7	98.6	75.8	80.1	90.3	72.2	71.3	82.4	57.6	74.2	73.8	74.6
I²GF-Net-s	76.5	83.1	76.5	79.6	91.4	74.2	72.4	81.6	58.7	75.1	77.5	72.8
I²GF-Net-d	**78.4**	52.3	**77.6**	80.5	**92.3**	**75.7**	**74.6**	84.7	**63.7**	**77.3**	**78.8**	**75.8**

Results with '*' are re-implemented by authors in this paper, and others are implemented with the open source codes from the corresponding references.

D. Analysis of Results

1) VSB-DET: We compared our I²GF-Net with thirteen state-of-the-art methods on the VSB-DET dataset, and the results are shown in Table III. It can be observed from the results that I²GF-Net-d achieved the highest mAP of 78.4%. Additionally, the mAP is 3.1% higher than the best-performing Faster-RCNN among the general models. Some faster detection methods, such as YOLOv8, use a lightweight network structure design that reduces redundancy in the computation process, thus increasing the detection speed. Although I²GF-Net-d is less than half the speed of YOLOv8, it still meets real-time requirements. Additionally, our mAP outperforms YOLOv8 by 4.3%. We also evaluated our I²GF-Net-s, which is a faster single-round detection solution. While it sacrifices some accuracy for speed, I²GF-Net-s achieved a competitive mAP of 76.5% compared to CCG-YOLOV7 and EAE-YOLOX, which are designed for wood defect detection. Compared to STFE-Net, which enhances statistical texture features, our method exhibits slightly higher precision and speed. The detection speed of I²GF-Net-s is 83.1 FPS, representing a 30.8 FPS improvement over I²GF-Net-d. Our method focuses on enhancing defect information extraction through top-down information guidance and fine-grained local detail supplementation, achieving optimal AP values for the majority of defect categories. As shown in Fig. 4, our method is visually compared with several well-performing methods. It can be seen that in the case of wood defects, Re is often hidden in the background texture, and Cr often appears to be very thin and may be overlooked in the down-sampling process. Our method proved to be effective in detecting these defects. Although the performance of our method may not be the best in detecting DK and KM, it is close to the best. Additionally, our F_1-score is highest at 77.3, indicating very low false positives. This is attributed to our I²GF-Net guide the attention of the low-level feature map to focus on the defect regions through TDFE, significantly reducing false positives caused by intense textures, achieved the highest precision (P) of 78.8%. Subsequently, compensate high-level semantic features with fine-grained local information through SFTE, substantially reducing the risk of missing subtle defects, achieved the highest recall (R) of 75.8.

2) OULU -DET: As shown in Table IV, although our method performed less effectively in the OULU-DET dataset compared to its performance in the VSB-DET dataset, I²GF-Net still achieved the best mAP of 63.6% among the other thirteen advanced methods, with I²GF-Net-s following closely at 61.7%. The OULU-DET dataset presents challenges due to its limited sample size of 839 images, encompassing ten defect types, with a notably sparse quantity of Dek samples and very faint defect features in Sp. These factors contribute to the overall challenge all detection methods face in accurately identifying Dek and Sp. However, addressing these challenges, I²GF-Net employs a top-down information guidance mechanism to capture anomalous features and refine the edge details of wood defects through iterative feature extraction. This approach allows I²GF-Net to perform excellently in detecting the majority of defect types while, through texture enhancement techniques, elevating the detection performance of challenging Dek and Sp beyond other methods. Compared to defect detection algorithms CCG-YOLOV7 and EAE-YOLOX designed for wood surfaces, I²GF-Net-d exhibits a competitive improvement. Furthermore, both precision (P) and recall (R) surpass those of other detection algorithms, suggesting that our model exhibits lower false positives and missed detections compared to other methods.

3) NEU-DET: As shown in Table V, I²GF-Net also exhibits outstanding detection performance in the NEU-DET dataset, demonstrating its potential for generalization. Specifically, compared to thirteen other state-of-the-art methods, our I²GF-Net-d achieved the highest mAP of 83.9%, followed by I²GF-Net-s with a mAP of 81.6%. Compared to the most accurate detection algorithm YOLOv8 within the general model, I²GF-Net-d exhibited a notable improvement of 6.3% in mAP, while I²GF-Net-s showed an increase of 4%. This enhancement primarily resulted from the significant improvement in detecting weak defects (Cr and Rs) by I²GF-Net. These weak defects share similarities with Ma and Re in wood defect datasets, displaying subtle features that

TABLE IV
DEFECT DETECTION RESULTS ON OULU-DET DATA

Model	mAP/%	AP/%										F_1-score	P/%	R/%
		DrK	SK	EK	KH	HK	LK	Sp	Wa	Cs	DeK			
CenterNet [32]	54.4	78.3	70.0	51.1	75.0	60.5	46.3	24.9	44.0	58.2	35.7	57.3	56.8	57.8
Faster-RCNN [33]	59.1	80.4	66.6	56.9	82.6	79.5	57.9	24.2	45.5	62.3	34.9	63.5	60.8	66.4
Retinanet [34]	56.7	79.1	64.3	49.8	74.6	56.7	58.8	28.9	50.6	68.0	35.9	60.1	58.5	61.8
YOLOv5 [35]	57.4	81.4	63.4	54.4	81.8	73.2	48.5	26.3	48.0	62.4	34.7	58.2	55.5	61.1
YOLOv7 [36]	59.2	79.9	66.0	56.3	82.5	**80.2**	58.0	24.6	46.0	63.9	34.9	63.6	64.2	63.0
YOLOv8 [37]	59.9	79.7	69.8	64.4	82.0	64.0	**59.5**	29.0	48.3	67.5	35.0	64.7	64.0	65.4
DETR [38]	57.7	78.2	65.1	58.9	81.4	59.6	51.5	33.0	47.5	65.3	36.2	61.3	60.1	62.6
RT-DETR [39]	58.4	78.4	64.0	58.5	**84.4**	63.7	55.1	31.6	45.7	66.9	35.8	62.4	63.4	61.4
PP-YOLOE [40]	55.0	77.1	64.5	60.2	68.6	59.0	47.1	28.1	47.6	61.7	36.5	60.8	61.0	60.6
RTMDet [41]	56.8	81.0	66.3	49.0	68.2	73.3	49.3	30.5	48.9	65.4	36.0	60.7	60.7	60.7
STFE-Net* [42]	61.4	80.1	68.7	64.4	81.7	74.3	57.5	32.2	52.6	67.3	35.4	63.4	63.3	63.5
CCG-YOLOV7* [43]	59.7	79.8	67.3	58.2	81.4	78.3	57.6	26.7	48.6	64.2	35.3	61.4	61.8	61.0
EAE-YOLOX* [44]	60.2	80.0	69.4	63.6	81.3	69.8	57.2	32.3	48.3	64.8	35.8	62.8	62.5	63.1
I²GF-Net-s	61.7	80.2	69.5	62.7	81.3	76.1	57.4	35.8	53.1	66.5	34.6	63.7	63.7	63.7
I²GF-Net-d	**63.6**	**81.9**	**70.3**	**66.2**	82.1	75.8	59.3	**38.7**	**56.3**	**68.3**	**36.8**	**66.9**	**66.0**	**67.8**

Results with '*' are re-implemented by authors in this paper, and others are implemented with the open source codes from the corresponding references.

TABLE V
DEFECT DETECTION RESULTS ON NEU-DET DATA

Model	mAP/%	AP/%						F_1-score	P/%	R/%
		Cr	In	Pa	PS	RS	Sc			
CenterNet [32]	76.7	55.4	75.0	93.5	88.9	62.9	84.4	73.9	72.9	74.9
Faster-RCNN [33]	79.6	47.3	84.2	94.9	85.3	70.4	**95.3**	75.5	76.0	75.0
Retinanet [34]	67.6	48.8	76.1	95.3	83.7	71.6	30.2	51.3	53.9	49.0
YOLOv5 [35]	77.1	42.5	85.2	95.3	84.4	61.7	93.2	74.3	73.3	75.3
YOLOv7 [36]	77.9	46.5	85.1	**96.3**	82.5	63.2	93.7	74.8	76.3	73.3
YOLOv8 [37]	77.6	45.3	83.6	94.2	84.2	68.0	90.1	74.7	73.6	75.9
DETR [38]	71.3	30.7	80.6	92.4	74.1	56.7	93.4	64.7	64.3	65.1
RT-DETR [39]	73.0	37.6	78.1	91.7	79.4	60.7	90.3	71.3	70.4	72.2
PP-YOLOE [40]	74.9	41.7	80.2	92.5	83.7	59.4	91.7	72.8	75.6	70.2
RTMDet [41]	73.4	39.8	83.4	89.4	84.7	52.3	90.5	71.6	72.4	70.8
STFE-Net* [42]	82.2	58.5	85.7	95.4	88.6	72.3	92.8	79.3	81.1	77.6
CCG-YOLOV7* [43]	78.9	55.2	83.6	92.4	83.9	68.1	90.4	76.4	76.1	76.7
EAE-YOLOX* [44]	80.1	56.3	84.1	93.7	84.8	69.5	92.2	78.5	79.5	77.5
I²GF-Net-s	81.6	56.9	84.4	96.1	85.7	73.7	93	79.2	80.3	78.1
I²GF-Net-d	**83.9**	**60.7**	**86.7**	95.3	**89.3**	**77.8**	93.5	**80.7**	**81.3**	**80.1**

Results with '*' are re-implemented by authors in this paper, and others are implemented with the open source codes from the corresponding references

TABLE VI
ABLATION STUDY OF DIFFERENT PARTS

ConvNeXt	TDFE	SFTE	DRFR	Params/M	GFLOPs	mAP/%
√				35.7	73.5	70.2
√	√			37.6	83.4	73.7
√	√	√		38.5	86.2	76.5
√	√	√	√	38.4	145.8	78.4

are challenging for conventional detection methods to discern. I²GF-Net effectively amplifies the representation of weak defects by utilizing SFTE to extract fine-grained local information from the shallow branch, thereby enhancing the accuracy of detecting subtle defects. Compared to the defect detection algorithm STFE-Net designed specifically for metal surfaces, I²GF-Net-s demonstrated similar effectiveness, while I²GF-Net-d showed an increase of 2.2% in mAP, with precision (P) increasing by 0.2% and recall (R) by 2.5%. This is attributed to I²GF-Net-d refining defect features through DRFR, thereby enhancing defect detection accuracy.

E. Ablation Study

In this section, we take the ConvNext backbone network as the baseline and conduct ablation experiments on the VSB-DET dataset to assess the effectiveness of each module in the proposed I²GF-Net model. The experimental results are presented in Table VI.

1) Baseline Network: The Top-Down Feedback Encoder we designed shares a similar concept with commonly used feature fusion methods in current state-of-the-art approaches, such as FPN and PAFPN. To systematically analyze the impact of our proposed method on the detection model, we removed the feature fusion module from our baseline network, detecting wood defects solely through the ConvNext backbone network's output at three layers $\{f_3, f_4, f_5\}$. After removing all the modules, the detection

performance significantly lagged behind ten advanced methods, with a mAP of only 70.2%.

2) Effect of Top-Down Feedback Encoder: As shown in Table VI, the introduction of TDFE has led to an improvement in the performance of the baseline model, with the mAP increasing from 70.2% to 73.7%, a gain of 3.5%. To visually demonstrate the guiding role of TDFE on low-level feature maps, Fig. 5 illustrates the input image feature map extracted by the backbone network, the guidance information generated by TDFE, and the combined effect of guidance information and feature map. Specifically, highlighted regions in the feature map represent extracted features, with brighter colors (yellow) indicating more prominent features. As shown in Fig. 5(b), the features extracted by the backbone network do not effectively distinguish between defects and intense textures. In Fig. 5(c), the guidance information obtained through TDFE effectively highlights the defect areas and almost does not activate the intense texture region on the left. This indicates that TDFE, leveraging high-level semantic information, can robustly differentiate between defects and intense textures. Fig. 5(d) shows that the fused guidance information has a specific inhibitory effect on complex texture backgrounds and activates the defect areas. These results demonstrate that TDFE can activate defect regions in the feature map, thereby avoiding false positives caused by intense textures.

3) Effect of Semantic Feature Texture Enhancement: Analyzing the experimental results, we observed that the multiple down-sampling processes led to the loss of local detailed information in high-level semantic features, making it challenging for TDFE to activate some subtle defects effectively. As shown in Fig. 6(b), this specific defect has relatively low contrast compared to other defects. It is concealed in the texture backgrounds and cannot be activated by the guidance information (Fig. 6(e)), resulting in a missed detection. To address this issue, we introduced SFTE. This method extracts fine-grained local information from a shallow feature map (Fig. 6(d)). It integrates it with the activated feature map, thereby compensating for high-level semantic information. Consequently, defects that were initially missed due to activation loss are successfully detected (Fig. 6(f)). With the inclusion of SFTE, the detection performance further improved, with the mAP increasing from 73.7% to 76.5%, representing a 2.8% enhancement.

4) Effect of Dual-Round Feature Refinement: With the introduction of DRFR, we achieved more precise localization of defect boundaries. This dual-round feature refinement solution integrates high-level semantic information with low-level texture information, gradually enhancing the expressive power of defect features and aiding in the accurate localization of defect boundaries. As shown in Fig. 7(d), TDFE activates a relatively large area of defect regions, resulting in a broader defect localization range. DRFR refines the defect boundaries, bringing the detection results closer to the edges of the defects (Fig. 7(e)). This refinement improves the accuracy of defect localization, enhancing the detection performance of I²GF-Net, with the mAP increasing from 76.5% to 78.4%, representing a 1.9% improvement. It is worth noting that while DRFR avoids a significant increase in overall parameters, it must be acknowledged that DRFR has an impact on

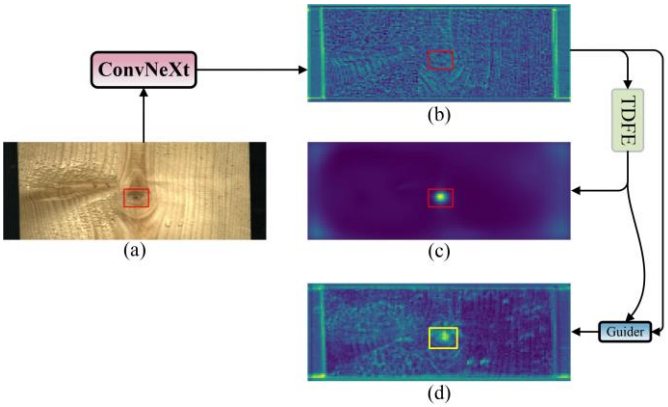

Fig. 5. The feature map illustrates the effect of the TDFE. (a) Input image. (b) Features of the baseline model. (c) Guidance information. (d) Detection results.

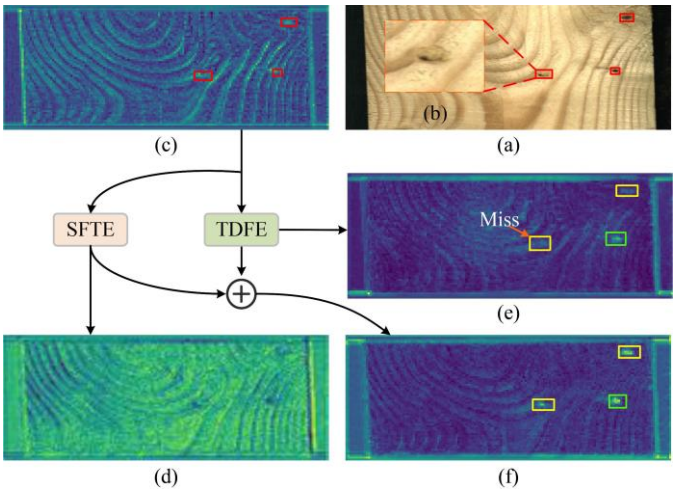

Fig. 6. The feature map illustrates the effect of the TDFE and SFTE. (a) Input image. (b) Features of the baseline model. (c) Texture-enhanced results. (d) Missed detection in TFED. (e) Successful detection of subtle defects through SFTE.

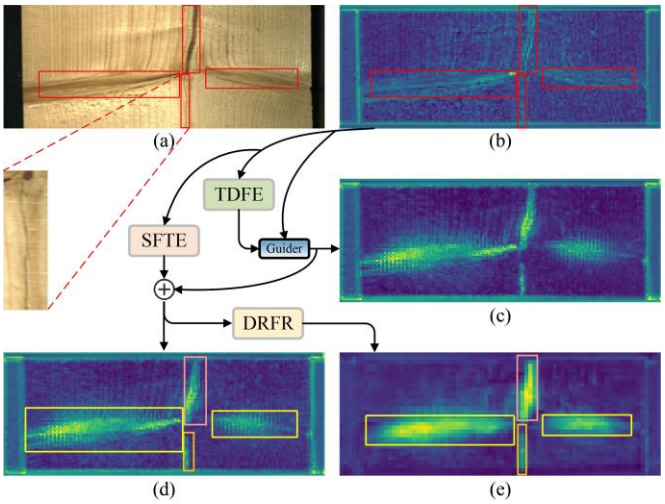

Fig. 7. The feature map illustrates the effect of the TDFE, SFTE, and DRFR. (a) Input image. (b) Features of the baseline model. (c) TDFE activates defect areas. (d) Detection results after texture enhancement with SFTE. (e) Detection results after refining.

computational resources, reflected in the model's GFLOPs increasing from 86.2 to 145.8.

F. Analysis of Illumination Effects

In real-world industrial environments, fluctuations in brightness are commonplace. Factors such as glare can arise at specific detection points, potentially hindering the detection process. To investigate the robustness of our method under varying lighting conditions, we present several wood surface images affected by different illumination settings, along with the corresponding feature maps generated by our detection model.

As shown in Fig. 8(a)-(f), our results demonstrate that even in challenging lighting scenarios, such as the presence of bright spots near defects (Fig. 8(a)-(b)), light spots resembling defect shapes (Fig. 8(c)), direct light shining onto the defect. (Fig. 8(d)), and large bright areas (Fig. 8(e)-(f)), our method reliably detects surface defects without significant sensitivity to illumination conditions. Through analysis of the feature maps produced by inputting these images into the I²GF-Net (Fig. 8(g)-(l)), we observe that while lighting affects the visual appearance of the original images, the guidance provided by the top-down feedback encoder (TDFE) on the low-level feature maps enables the model to focus attention on defect regions, thus mitigating interference from visually prominent illuminated areas on the detection process.

Consequently, our model demonstrates a certain degree of robustness to illumination conditions. We attribute this resilience to the correlation between high brightness and intense textures, which both represent visually prominent areas but do not necessarily indicate actual defects.

G. Future Works

Although our network performs well, it still needs to be immune to performance degradation due to small data samples. In VSB-DET, the mAP of I²GF-Net is 78.4%, but in the more diverse and sample-limited OULU-DET, it drops to 63.6%. Besides enhancing the feature extraction capabilities of the model, the introduction of Few-Shot Learning could be a valuable approach. Annotate the most informative samples to maximize model performance under limited labeling conditions.

Additionally, while DRFR does not add a substantial parameter overhead, it involves many floating-point operations. The data in TABLE VI indicates a sharp increase in GFLOPs from 86.2 to 145.8 with the inclusion of DRFR. While this improves detection performance, the computational cost remains considerable. Therefore, exploring a dynamic detection architecture seems a promising avenue. By analyzing the difficulty of defect detection in current images, adaptive triggering of whether to engage in second-level feature extraction could significantly enhance the average detection speed while ensuring detection effectiveness, which will be reflected in our further work.

V. Conclusion

This paper addresses the tricky challenges faced by Automated Visual Inspection (AVI) systems in the wood manufacturing industry, specifically in detecting defects in complex wood texture backgrounds. We propose a solution called Inter-Layer Information Guidance Feedback Networks

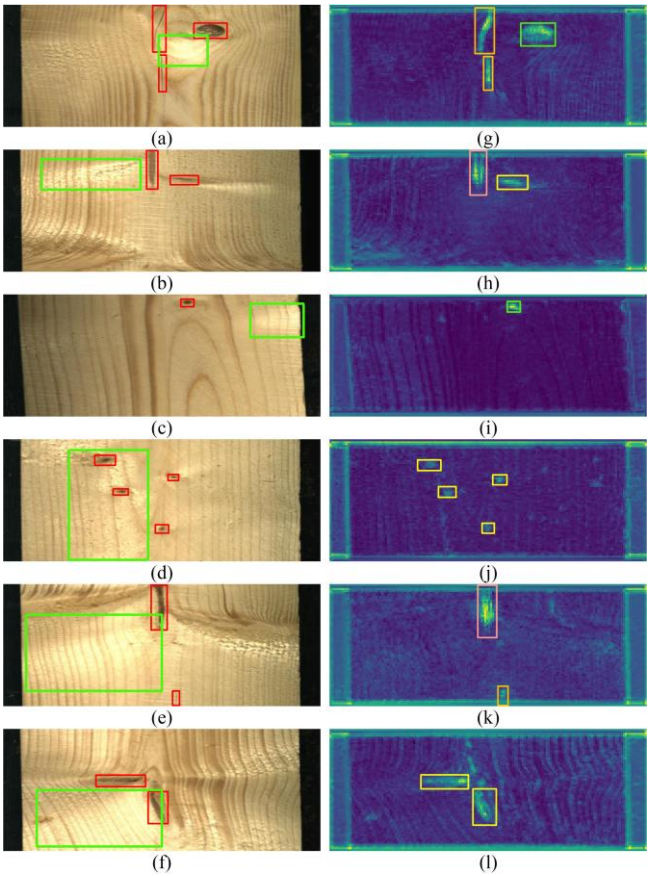

Fig. 8. Illustration of detection results in varied lighting conditions. (a) and (b) Bright spots near defects. (c) Light spots resembling defect shapes. (d) Direct light shining onto the defect. (e) and (f) Large bright areas in the images. (g)-(l) Feature maps.

(I²GF-Net) that utilizes both semantic and texture information in an inter-layer manner. The I²GF-Net introduces a top-down feedback encoder (TDFE) that effectively reduces false positives by guiding low-level features to focus on defect regions through enhanced semantics from high-level feature maps. Additionally, we employ a semantic feature texture enhancement (SFTE) scheme that compensates high-level semantic features with fine-grained information, significantly reducing missed detections of subtle defects. To further refine defect features and enhance localization precision, we propose the dual-round feature refinement (DRFR) framework. Experimental results demonstrate that the dual-round solution (I²GF-Net-d) with DRFR outperforms thirteen state-of-the-art methods, achieving fewer false positives, fewer missed detections, and more precise defect localization. On the other hand, the single-round solution without DRFR (I²GF-Net-s) is faster and demonstrates acceptable detection performance. Specifically, I²GF-Net-d achieved competitive mAP values of 78.4%, 63.6%, and 83.9% on VSB-DET, OULU-DET, and NEU-DET, respectively. I²GF-Net-s achieved mAP values of 76.3%, 61.7%, and 81.6% on the three datasets. The detection speeds were measured at 83.1 FPS for the single-round and 52.3 FPS for the dual-round.

REFERENCES

[1] Y. Tu , Z. Ling, *et al.*, "An Accurate and Real-Time Surface Defects Detection Method for Sawn Lumber," *IEEE transactions on instrumentation and measurement*, vol. 70, no. 2501911, 2021.

[2] P. Kodytek and A. Bodzas, *et al.*, "A large-scale image dataset of wood surface defects for automated vision-based quality control processes," *F1000Research*, vol. 10, pp. 581, 2021.

[3] Ding F, Zhuang Z, Liu Y, *et al.*: Detecting Defects on Solid Wood Panels Based on an Improved SSD Algorithm. *Sensors*. 2020; 20(18): 5315.

[4] J. Jiang and Y. Shi, *et al.*, "Utilizing adsorption of wood and its derivatives as an emerging strategy for the treatment of heavy metal-contaminated wastewater," *Environmental Pollution*, vol. 340, pp. 122830, 2024.

[5] B. Shi and T. Darrell, *et al.*, "Top-Down Visual Attention from Analysis by Synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2102-2112.

[6] L. Zhu and D. Ji, *et al.*, "Learning statistical texture for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12537-12546.

[7] J. Zhang and H. Wen, *et al.*, "Improved smoothing frequency shifting and filtering algorithm for harmonic analysis with systematic error compensation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9500-9509, 2019.

[8] J. Hu and W. Song, *et al.*, "Deep learning for use in lumber classification tasks," *Wood Science and Technology*, vol. 53, pp. 505-517, 2019.

[9] R. Ren and T. Hung, *et al.*, "A generic deep-learning-based approach for automated surface inspection," *IEEE transactions on cybernetics*, vol. 48, no. 3, pp. 929-940, 2017.

[10] S. Y. Jung and Y. H. Tsai, *et al.*, "Defect detection on randomly textured surfaces by convolutional neural networks," in *2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2018, pp. 1456-1461.

[11] F. Baluch and L. Itti, "Mechanisms of top-down attention," *Trends in neurosciences*, vol. 34, no. 4, pp. 210-224, 2011.

[12] F. Zhang and S. Lin, *et al.*, "Global attention network with multiscale feature fusion for infrared small target detection," *Optics & Laser Technology*, vol. 168, pp. 110012, 2024.

[13] J. J. Liu and Q. Hou, *et al.*, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3917-3926.

[14] F. Juefei-Xu and V. Naresh Boddeti, *et al.*, "Local binary convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 19-28.

[15] D. P. Fan and G. P. Ji, *et al.*, "Concealed object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 6024-6042, 2021.

[16] Y. Liang and K. Xu, *et al.*, "Automatic defect detection of texture surface with an efficient texture removal network," *Advanced Engineering Informatics*, vol. 53, pp. 101672, 2022.

[17] Y. Liu and Y. Wang, *et al.*, "Cbnet: A novel composite backbone network architecture for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, pp. 11653-11660.

[18] J. Wang and K. Chen, *et al.*, "Carafe: Content-aware reassembly of features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3007-3016.

[19] J. Zhu and X. Zhang, *et al.*, "Inferring camouflaged objects by texture-aware interactive guidance network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 3599-3607.

[20] X. Wen and S. Li, *et al.*, "Multi-scale context feature and cross-attention network-enabled system and software-based for pavement crack detection," *Engineering Applications of Artificial Intelligence*, vol. 127, pp. 107328, 2024.

[21] Z. Liu and H. Mao, *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976-11986.

[22] X. Zhu and C. Tang, et al., "Saliency detection via affinity graph learning and weighted manifold ranking," *Neurocomputing*, vol. 312, pp. 239-250, 2018.

[23] W. Jia and R. M. Shukla, et al.,"Anomaly detection using supervised learning and multiple statistical methods," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, 2019, pp. 1291-1297.

[24] H. Zhang and F. Cricri, et al.,"Learn to overfit better: finding the important parameters for learned image compression," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, 2021, pp. 1-5.

[25] X. Zhu and H. Hu, et al.,"Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9308-9316.

[26] X.Gao and L. Gao, et al., "A multilevel information fusion-based deep learning method for vision-based defect recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 3980-3991, 2019.

[27] X. YongHua and W. Jin-Cong, "Study on the identification of the wood surface defects based on texture features," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 19, pp. 2231-2235, 2015.

[28] H. Zhou and J. Yu, et al.,"Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 3769-3777.

[29] A. Iliev and N. Kyurkchiev, et al., "On the approximation of the step function by some sigmoid functions," *Mathematics and Computers in Simulation*, vol. 133, pp. 223-234, 2017.

[30] O. Silvén, M. Niskanen and H. Kauppinen, "Wood inspection with non-supervised clustering," *Machine Vision and Applications*, vol. 13, pp. 275-285, 2003.

[31] Y. He and K. Song, *et al.*, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE transactions on instrumentation and measurement*, vol. 69, no. 4, pp. 1493-1504, 2019.

[32] K. Duan and S. Bai, *et al.*, "Centernet: Keypoint Triplets for Object Detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569-6578.

[33] S. Ren and K. He, *et al.*, "Faster R-CNN: Towards realtime object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[34] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV).*, pp. 2980–2988, 2017.

[35] Ultralytics. (Jul. 2023). YOLOv5. Github. [Online]. Available: https://github.com/ultralytics/yolov5

[36] C. Y. Wang and A. Bochkovskiy, *et al.*, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464-7475.

[37] Ultralytics. (Jul. 2023). YOLOv8. Github. [Online]. Available: https://github.com/ultralytics/ultralytics

[38] N. Carion and F. Massa, *et al.*, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020, pp. 213-229.

[39] W. Lv and S. Xu, *et al.*, "DETRs Beat YOLOs on Real-time Object Detection," *arXiv:2304.08069*, 2023.

[40] S. Xu and X. Wang, *et al.*, "PP-YOLOE: An evolved version of YOLO," *arXiv:2203.16250*, 2022.

[41] C. Lyu and W. Zhang, *et al.*, "RTMDet: An Empirical Study of Designing Real-Time Object Detectors," *arXiv:2212.07784*, 2022.

[42] H. Zhong and D. Fu, *et al.*, "STFE-Net: A multi-stage approach to enhance statistical texture feature for defect detection on metal surfaces," *Advanced Engineering Informatics*, vol. 61, pp. 102437, 2024.

[43] W. Cui and Z. Li, *et al.*, "CCG-YOLOv7: A Wood Defect Detection Model for Small Targets Using Improved YOLOv7," *IEEE Access*, 2024.

[44] D. Li and Z. Zhang, *et al.*, "Detection method of timber defects based on target detection algorithm," *Measurement*, vol. 203, pp. 111937, 2022.