Article

Not peer-reviewed version

An Integrated and Robust Vision System for Internal and External Thread Defect Detection with Adversarial Defense

Fu Liu , Leqi Li , Genpei Zhang , Zhihao Jiang *

Posted Date: 14 August 2025

doi: 10.20944/preprints202508.1070.v1

Keywords: thread defect detection; lightweight neural network; Alpha channel attack; YOLO-based optimization; image data augmentation

Article

An Integrated and Robust Vision System for Internal and External Thread Defect Detection with Adversarial Defense

Fu Liu [†], Leqi Li [†],Gengpei Zhang and Zhihao Jiang [*]

Yangtze University, Jingzhou, Hubei, 434100 China

[*] Correspondence: zhihaojiang800@gmail.com

[†] These authors contributed equally to this work.

Abstract

In industrial automation, threaded mechanical components present significant challenges for inspection due to their complex geometries and the high concealment of micro-defects. This paper proposes an integrated detection system for internal and external thread defects, combining image enhancement, data synthesis, lightweight detection, and adversarial defense. A unified image acquisition platform with fisheye lenses and high-definition industrial cameras enables synchronized imaging of both internal and external threads. By incorporating MLWNet-based dynamic deblurring and DarkIR-based low-light enhancement, image quality is significantly improved (PSNR 30.3 dB, SSIM 0.945). A Residual Diffusion Denoising Model (RDDM) is used to diversify samples, reducing the FID from 69.6 to 24.92. For detection, a lightweight enhanced architecture, SLF-YOLO, achieves a precision of 0.881 and mAP@0.5 of 0.813, outperforming multiple YOLO baselines. A dual defense mechanism—input perturbation suppression and output anomaly analysis—effectively mitigates over 95% of mAP loss under Alpha channel attacks. Experimental results demonstrate that the proposed system delivers robust, secure, and efficient performance, offering a practical pathway toward reliable, interpretable, and resilient industrial vision inspection.

Keywords: thread defect detection; lightweight neural network; Alpha channel attack; YOLO-based optimization; image data augmentation

1. Introduction

Threaded components are fundamental units widely used in modern industrial equipment for connection and power transmission, playing crucial roles across key sectors such as machinery, automotive, aerospace, energy, and rail transportation. In various assembly structures, threads are not only responsible for mechanical connection and sealing alignment but are also directly related to the overall structural reliability and safety. Due to their complex geometric structure and stringent machining precision requirements, even minute defects—such as broken threads, burrs, cracks, and corrosion—can lead to loosening, sealing failure, or even catastrophic mechanical breakdowns. Therefore, effective defect detection for industrial threaded parts is vital for ensuring product quality, improving assembly consistency, and reducing equipment failure rates.

Beyond the most primitive visual inspection, traditional defect detection methods for threads largely rely on contact-based measurements, such as thread gauges (e.g., plug and ring gauges), calipers, micrometers, and profilometers with mechanical probes [1,2]. These techniques involve manual or semi-automatic operations to identify issues like dimensional deviations or machining defects. However, they suffer from low efficiency, reliance on operator experience, and difficulty detecting micro or structural defects—especially in reflective surfaces, deep cavities, or mass production settings—making them inadequate for modern manufacturing demands of high precision, efficiency, and automation.

In recent years, non-contact inspection technologies for industrial components have seen substantial advancements, becoming mainstream due to their non-destructive, automated, and high-accuracy advantages. These include computer vision inspection [3–5], laser scanning [6–8], X-ray [9,10], and optical techniques [11–13]. Among them, computer vision-based defect detection stands out for its lightweight nature, easy deployment, high accuracy, and low cost, making it the most general and extensible solution within non-contact inspection methods.

Image processing is the foundational technology in computer vision and has been widely adopted across visual inspection tasks. Notably, dynamic image deblurring has made significant progress through deep learning. Jang et al. (2025) [14] proposed DSANet, a deep supervision attention network leveraging a ConvLSTM encoder-decoder and frequency-domain constraints for precise recovery of blurred regions. Ren et al. (2022) [15] introduced a spatially variant neural network for dynamic scene blur, combining CNN and RNN to better model complex blur patterns. Gao et al. (2019) [16] presented a parameter-sharing network with nested skip connections and released a benchmark dataset for dynamic deblurring. Zhang et al. (2023) [17] adopted a flow-guided multi-scale RNN to improve fine detail recovery, while Chen et al. (2023) [18] proposed a CNN–Transformer hybrid model using stripe attention and cross-layer feature fusion, achieving state-of-the-art performance on multiple benchmarks.

Illumination normalization, a critical step in image preprocessing, enhances robustness under suboptimal lighting for tasks such as enhancement, detection, and recognition. Vasluianu et al. (2024) [19] introduced Ambient Lighting Normalization (ALN) and a corresponding dataset Ambient6K, using frequency-domain fusion to restore shadowed regions. Dias Da Cruz et al. (2020) [20] proposed a learning framework with partially unachievable autoencoder objectives for better illumination-invariant representation. Huang et al. (2023) [21] proposed Transition-Constant Normalization (TCN) for stable enhancement under varying exposures. Rad et al. (2020) [22] developed Adaptive Local Contrast Normalization (ALCN), dynamically predicting normalization parameters to boost recognition in complex lighting. Goswami (2020) [23] designed a deployable deep method for correcting uneven lighting in RGB images, showing good generalization in real-world scenarios.

With the advancement of deep learning, many vision-based defect detection systems have integrated deep neural networks to improve recognition accuracy and robustness in complex settings. Jiang et al. (2024) [24] combined GAN and YOLO for generating and detecting internal thread defects, achieving 94.27% and 93.92% accuracy for internal and external threads, respectively. Dou et al. (2024) [25] developed a multi-camera inspection system incorporating lighting optimization and cylindrical image stitching, enabling efficient and visual thread defect localization. Xu et al. (2023) [26] proposed an enhanced YOLOv5 for bearing defect detection, using C2f, SPD modules, and CARAFE upsampling, achieving 97.3% mAP and 100 FPS. Wu et al. (2025) [27] introduced RBS-YOLO, a lightweight version of YOLOv5 for casting defects, balancing accuracy and complexity. Patil (2024) [28] compared YOLOv5 and YOLOv8 for nut thread presence detection, highlighting YOLOv8's superior speed and accuracy. Lang et al. (2022) [29] proposed MR-YOLO by integrating MobileNetV3, SE attention, and Mosaic augmentation, improving efficiency and accuracy. Tabernik et al. (2019) [30] designed a DNN-based segmentation model with few-shot learning support. Wang et al. (2022) [31] introduced Defect Transformer (DefT), a hybrid CNN-Transformer model that captures both local details and global semantics, enhancing detection robustness.

Despite the accuracy gains from deep learning, recent studies highlight the critical vulnerability of such systems to adversarial perturbations. Attackers can embed imperceptible but structured noise into input images, misleading models and causing misclassification or missed detections. Notably, Alpha channel attacks inject disruptions into the image transparency layer, altering model behavior without any perceptual changes. Since most industrial vision models accept RGBA inputs without explicitly removing the Alpha channel, such attacks are easily deployable yet difficult to detect, making them one of the most severe threats to industrial vision safety. In high-security applications involving automated thread inspection and sorting, undetected adversarial inputs may trigger

doi:10.20944/preprints202508.1070.v1

assembly faults or quality control failures. Therefore, integrating adversarial defense into defect detection frameworks is not only a robustness requirement but also a vital security foundation in industrial AI deployment.

This study extends prior work in small-object detection, multi-scale adaptability, and edge-device deployment while introducing a security-aware perspective to strengthen real-world robustness. The main contributions are as follows:

1. A dual-mode industrial image acquisition setup is constructed for internal and external threads, integrating fisheye lenses and HD cameras to solve structural complexity and switching inefficiencies in traditional systems.
2. MLWNet and DarkIR are employed for dynamic deblurring and illumination normalization, ensuring high-quality inputs. A residual diffusion denoising model (RDDM) is introduced for generating and augmenting thread defect samples.
3. A novel detection model, SLF-YOLO, is developed by integrating SC_C2f, Light-SSF_Neck, and FIMetal-IoU loss, outperforming YOLOv5s to YOLOv10s while remaining suitable for real-time edge deployment.
4. A defense mechanism is proposed against Alpha channel attacks, using histogram overlap and MSE-based detection to effectively identify and mitigate input-level adversarial perturbations.

2. Principle and System Overview

Figure 1 illustrates the overall processing workflow of the proposed internal and external thread defect detection system. From the initial data acquisition stage, the system integrates multi-level image enhancement and security defense mechanisms to improve model robustness and detection stability under complex industrial interference conditions.

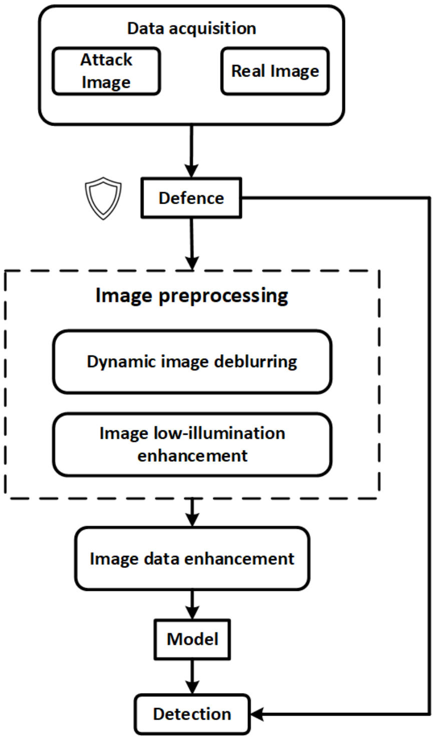


Figure 1. Overall system architecture.

The system begins by acquiring multi-source input data from industrial environments, which includes both normal images and potential adversarial samples. To counteract threats such as alpha-channel attacks, transparent padding interference, and adversarial patches, the input data is first processed by a lightweight security defense module. This module performs perturbation suppression, anomaly filtering, and input resizing to preliminarily eliminate explicit attack characteristics and prevent malicious samples from entering the core model pipeline.

Next, the data flows into the Image Preprocessing submodule, which consists of two processing paths:

(1) Dynamic deblurring, designed to mitigate motion blur caused by device vibration or camera instability, thereby enhancing the visibility of thread edges and defect boundaries;
(2) Low-light enhancement, targeted at restoring image quality in dark cavities such as internal threads, utilizing brightness normalization and edge detail enhancement to improve model perception.

The preprocessed images are then passed to the Image Data Enhancement module, where a Residual Diffusion Denoising Model (RDDM) is employed for defect diversity modeling and synthetic data generation. This enhances the model's generalization capability and defect coverage under limited-sample conditions.

The enhanced image data is subsequently fed into a deep object detection network for defect identification and localization. The detection results are also fed back into a front-end security monitoring module, enabling output-based anomaly detection. For instance, abrupt changes in the number of bounding boxes or unusual clustering of defect categories can trigger an alarm or pause the model response, thus forming a closed-loop industrial vision security chain with perception, diagnosis, and response capabilities.

Overall, the proposed workflow not only ensures high detection accuracy but also integrates a three-stage defense pipeline—pre-processing, mid-processing, and post-processing—offering comprehensive protection against adversarial perturbations, transparent padding, and real-world industrial interference. This design ensures strong industrial adaptability and controllable system security.

2.1. Image Acquisition

As the first and foundational stage in the internal and external thread defect detection pipeline, the quality, viewpoint completeness, and spatial accuracy of thread image acquisition directly determine the upper performance limits of subsequent feature extraction and object recognition algorithms. To obtain high-fidelity, full-coverage, and unobstructed image inputs, this study designs a unified image acquisition system tailored for industrial field applications, capable of handling both internal and external threads.

In conventional industrial inspection systems, internal and external threads are typically imaged using separate devices and workflows due to their distinct structural positions: external threads are usually captured via multi-camera setups arranged around the object, while internal threads require endoscopic probes to access deep cavities. These differences in installation, illumination strategies, and imaging paths result in complex hardware configurations, high switching costs, and low efficiency in batch inspections.

To address these challenges, we propose an integrated image acquisition architecture for industrial thread defect detection, as illustrated in Figure 2. The system is constructed around the multi-angle structural characteristics of threaded components, incorporating independent subsystems for internal and external thread image capture. These subsystems are synchronized using stepper motors, transmission mechanisms, and an embedded control platform to achieve precise, coordinated acquisition of dynamic thread targets.

The workpiece is fixed on a central fixture driven by a lead screw mechanism powered by a stepper motor, enabling linear axial movement. The displacement of the motor and the image acquisition signals are orchestrated by a Raspberry Pi-based control unit, ensuring closed-loop synchronization of image triggering, motion control, and data transmission.

doi:10.20944/preprints202508.1070.v1

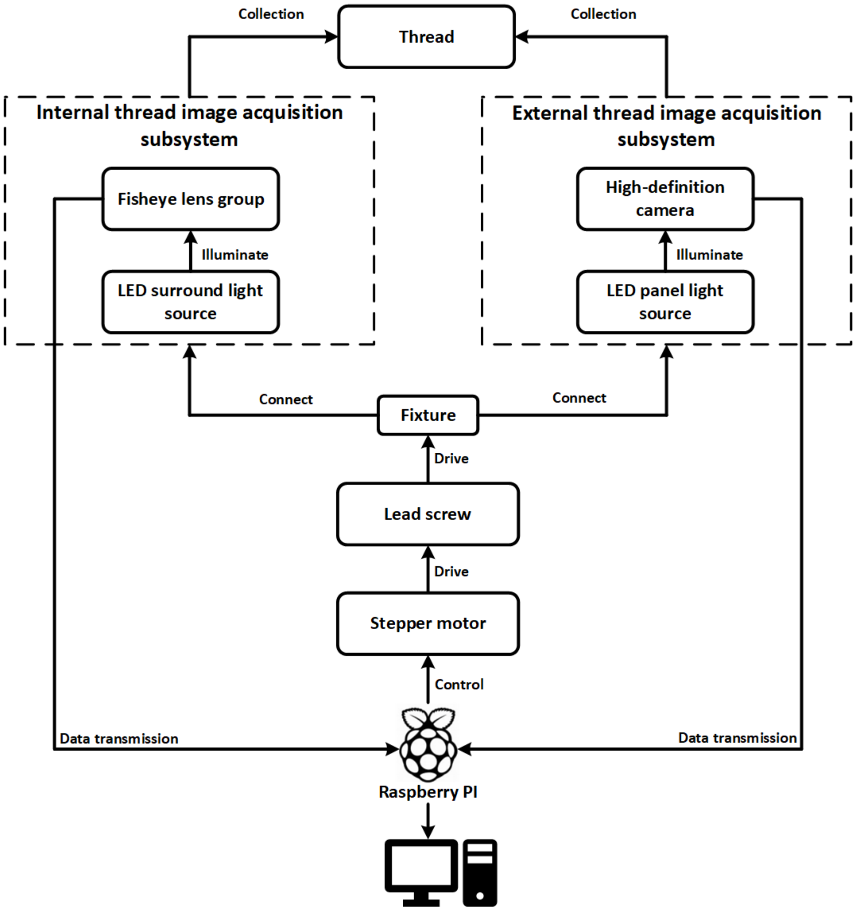


Figure 2. System architecture of the industrial internal and external thread image acquisition platform.

For internal thread imaging, the system employs a fisheye lens group combined with an LED ring light source, enabling wide-angle imaging and uniform circumferential illumination within the cavity. This setup effectively mitigates the challenges of light-shadow blind spots and angle occlusion along the thread's inner wall. For external thread imaging, a high-definition industrial camera coupled with an LED panel light source is used to achieve full circumferential coverage of the outer surface with parallel illumination, suitable for rod-like components such as screws and spindles.

Both subsystems are connected to the main control platform via the fixture linkage structure. The acquired images are transmitted in real time through the Raspberry Pi to a backend detection host, where the defect recognition network operates. Thanks to this dual-subsystem collaborative design, the platform supports unified, adjustable, and multi-angle thread image capture, forming a stable data foundation for high-precision vision-based defect detection. The hardware structure is depicted in Figure 3.

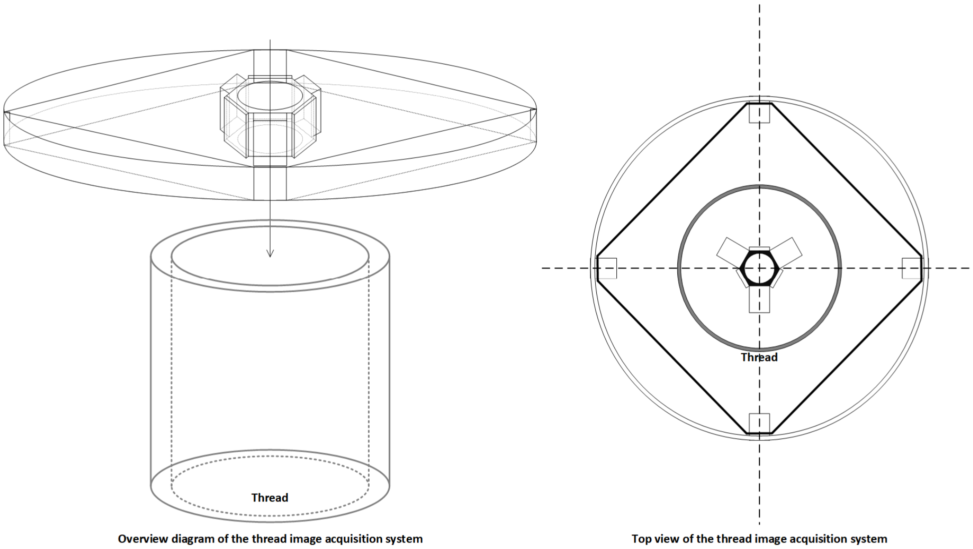

Thread
Thread
Overview diagram of the thread image acquisition system
Top view of the thread image acquisition system

Figure 3. Internal and external thread imaging hardware design.

The combination of dual imaging subsystems with a central lead-screw lifting platform and multi-angle mounting modules allows for the simultaneous and integrated acquisition of both internal and external wall images on a single device. This approach not only improves assembly consistency and reduces platform complexity, but also ensures spatial-temporal alignment and structural consistency of the images. As a result, the system provides standardized input sources for downstream defect detection models, significantly enhancing overall detection accuracy, system stability, and industrial deployability. Detailed component illustrations are shown in Figure 4.

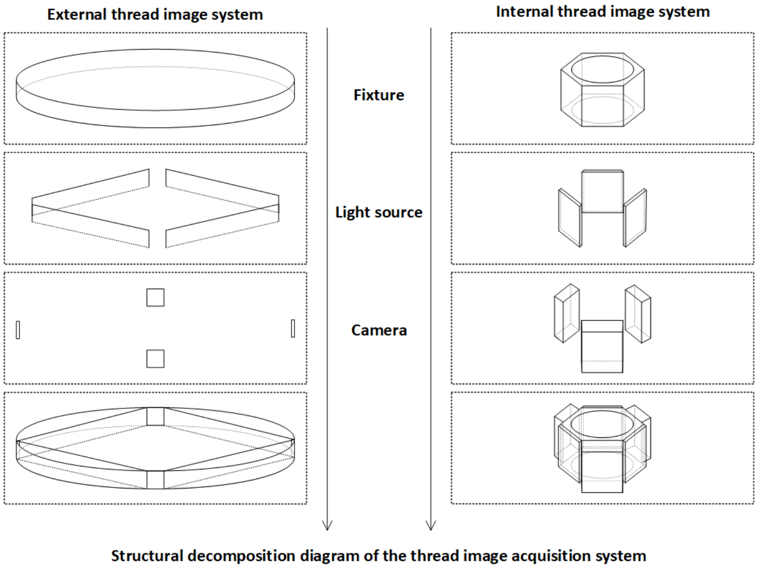

External thread image system
Internal thread image system
Fixture
Light source
Camera
Structural decomposition diagram of the thread image acquisition system

Figure 4. System structural breakdown and design details.

2.2. Dynamic Deblurring

In industrial visual inspection scenarios, image blur is a prevalent issue—particularly in regions with complex geometries such as metallic threads and tubular cavities. This type of degradation often arises from equipment vibration, insufficient exposure, or motion-induced defocus, leading to pronounced directional motion blur. Such blur is typically accompanied by attenuation of high-frequency textures, edge smearing, and structural distortion, which severely compromises image clarity and limits its usability in defect detection, depth estimation, and 3D modeling tasks. To address this, a dynamic deblurring module is introduced at the image preprocessing stage, forming a comprehensive image enhancement pipeline in conjunction with the illumination normalization module.

We adopt MLWNet (Multi-scale Network with Learnable Wavelet Transform)[32] as the backbone of the dynamic deblurring module. MLWNet incorporates learnable two-dimensional discrete wavelet transform (2D-LDWT) and a multi-scale semantic fusion mechanism to effectively capture blur features across different scales and orientations. Unlike conventional spatial-domain networks, MLWNet introduces frequency modeling during the feature extraction stage, with a specific focus on restoring high-frequency details and edge structures that are severely degraded by motion blur.

At the modeling level, wavelet transform decomposes an image $f(t)$ into a low-frequency approximation term and multiple high-frequency directional components:

$$f(t) = \sum_{j>j_0} \sum_k d_{j,k}\, \psi_{j,k}(t) + \sum_k c_{j_0,k}\, \phi_{j_0,k}(t) \tag{1}$$

where $d_{j,k}$ represents the high-frequency detail coefficients, and $c_{j_0,k}$ denotes the low-frequency approximation coefficients. This decomposition provides multi-scale frequency resolution capabilities.

In the network, high-pass and low-pass filters $\vec{a}_1$ and $\vec{a}_0$ are used to perform recursive convolution operations, resulting in four sets of 2D filtered wavelet components—LL, LH, HL, and HH—which are concatenated to form a four-channel wavelet convolution kernel K_w.

To ensure reversibility and energy conservation during both the forward and inverse wavelet transformations, Perfect Reconstruction Constraints are introduced:

$$A_0(-z)S_0(z) + A_1(-z)S_1(z) = 0,\ \ A_0(z)S_0(z) + A_1(z)S_1(z) = 2 \tag{2}$$

In terms of architecture, the input image is first processed through multiple Simple Encoder Blocks (SEBs) to extract shallow features and perform multi-scale downsampling. The central module, Wavelet Fusion Block (WFB), employs Learnable Wavelet Nodes (LWNs) to conduct forward wavelet transformation and directional detail modeling. Frequency-domain features are extracted using depthwise separable convolutions and channel reconstruction, and are then fused back into the spatial domain through residual connections. The decoding phase uses several Wavelet Head Blocks (WHBs) for progressive feature upsampling and image clarity restoration, ultimately producing a high-resolution deblurred image.

For the training strategy, the network is optimized using two types of loss functions: the multi-scale loss $\mathscr{L}_{\text{multi}}$, which supervises the pixel-wise discrepancies between outputs at different scales and the ground truth (GT); and the wavelet reconstruction loss $\mathscr{L}_{\text{wavelet}}$, which ensures consistent frequency-domain modeling by the Learnable Wavelet Node (LWN) module. The final total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{multi}} + \lambda \cdot \mathcal{L}_{\text{wavelet}} \tag{3}$$

2.3. Illumination Normalization

In real-world industrial inspection environments, the surfaces of threaded metallic components frequently exhibit strong shadows, specular highlights, and non-uniform exposure due to the periodic geometry, high reflectivity of materials, and significant variations in ambient lighting. These conditions impose substantial challenges to vision-based defect detection models, often leading to unstable or inaccurate predictions.

To enhance the consistency of input images and improve illumination robustness, this study introduces an image-level illumination normalization module during the data preprocessing stage.

This study adopts an illumination processing framework based on the Retinex theory[33], which models the observed image $\mathrm{I}(x,y)$ as the product of a reflectance component $\mathrm{R}(x,y)$ and an illumination component $\mathrm{L}(x,y)$:

$$\mathrm{I}(x,y) = \mathrm{R}(x,y) \cdot \mathrm{L}(x,y) \tag{4}$$

In the logarithmic domain, this multiplicative relationship is transformed into an additive model for easier processing:

$$\log \mathrm{I}\,(x,y) = \log \mathrm{R}\,(x,y) + \log \mathrm{L}\,(x,y) \tag{5}$$

To extract reflectance components that are more sensitive to subtle structural defects, we propose a local brightness-constrained enhancement strategy, which combines local contrast amplification with gamma compression into a unified normalization framework. This approach performs spatial mean filtering in the brightness channel to suppress low-frequency illumination artifacts while adaptively adjusting the contrast range of the image. It enhances the visibility of fine textures and edge-related features that are critical for defect detection.

For practical implementation, we adopt the fast and deployable Retinex by Adaptive Filtering (RAF) method as the primary algorithm, combined with gamma compression:

$$\mathrm{I}_{\mathrm{norm}}(x,y) = \left(\frac{\mathrm{I}(x,y)}{\mathrm{G}_{\sigma}(\mathrm{I})}\right)^{\gamma} \tag{6}$$

Here, $\mathrm{G}_{\sigma}(\mathrm{I})$ denotes the Gaussian-smoothed output of the image brightness channel, and γ controls the non-linear compression of brightness. This method suppresses overexposure in locally bright areas and enhances contrast in low-illumination or occluded regions. It is particularly effective for inner surfaces of metallic threads, where reflective lighting often causes pseudo-defect patterns due to structural highlights.

The network architecture, as illustrated in Figure 5, adopts a dual-stage cooperative design aimed at simultaneously addressing low-light enhancement and image deblurring. The overall network consists of two distinct stages: the Low-Light Enhancement Stage (LOL Stage) and the Deblurring Stage (Deblur Stage). Both stages utilize a symmetric encoder–decoder architecture with multi-scale feature extraction capabilities, and are connected via skip connections to ensure efficient transmission and fusion of feature information.

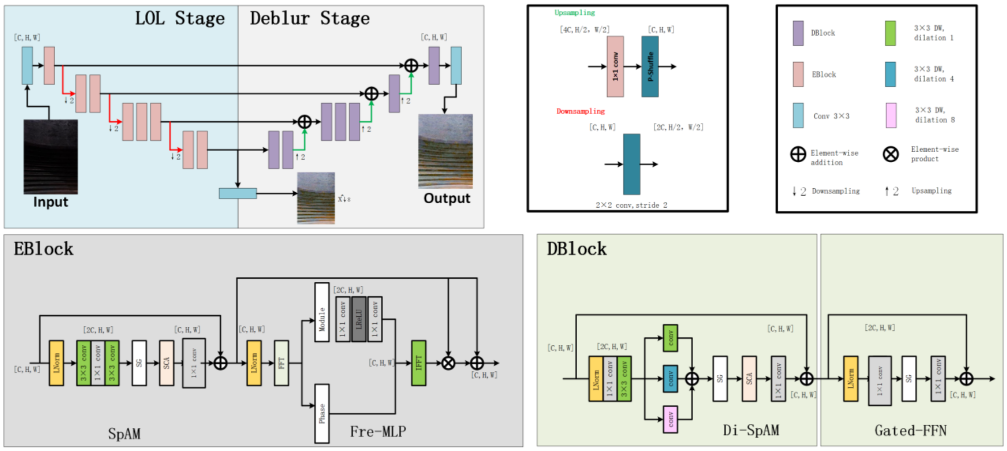


Figure 5. Architecture of the proposed dual-stage DarkIR network for low-light enhancement and image deblurring.

The LOL Stage focuses on restoring brightness and enhancing fundamental details in low-illumination input images, thereby improving overall visibility. Building on the enhanced outputs, the Deblur Stage further strengthens edge structures and restores texture details, effectively compensating for blur caused by low lighting or acquisition jitter.

In terms of module design, DarkIR integrates an EBBlock (Enhancement Block) into the LOL Stage. This block contains two key submodules:

(1) The SpAM (Spatial Attention Module) enhances local responses via spatial attention mechanisms, improving brightness expression under uneven lighting conditions;

(2) The Fre-MLP (Frequency-aware MLP) module centers on frequency-domain modeling, leveraging frequency information to preserve fine details and reduce noise—especially suited for handling high-frequency regions such as industrial surface textures.

The EBBlock output is fused with the main feature stream via residual connections, ensuring stability throughout the enhancement process.

In the Deblur Stage, DarkIR incorporates the DBlock module for high-quality restoration of blurred regions. DBlock consists of:

(1) Di-SpAM (Dilated Spatial Attention Module), which uses dilated convolutions to enlarge the receptive field and capture edge cues in low-contrast backgrounds;

(2) Gated-FFN (Gated Feed-Forward Network), which enables discriminative modeling between blurred and sharp regions during information propagation, thus better preserving structural integrity and suppressing artifacts.

The entire network employs standard strided convolutions and transposed convolutions for downsampling and upsampling, respectively. Additionally, skip connections between multiple scales enable the flow of semantic and fine-grained visual information across layers, further enhancing the network's multi-scale perceptual capability.

2.4. Image Data Augmentation

In industrial internal and external thread defect detection tasks, the acquisition of high-quality and representative image samples is often constrained by factors such as complex spatial structures, reflective metallic surfaces, and occluded viewpoints. These limitations result in a scarcity of annotated data, which restricts the robustness and generalization capability of deep learning-based detection models.

To address the limited-data problem, this study introduces a high-fidelity defect image generation framework based on diffusion modeling during the data augmentation phase. Specifically, a Residual Diffusion Denoising Model (RDDM) is employed to perform conditional sampling on original thread images, thereby simulating a broader distribution of diverse and representative defect types.

The RDDM method follows the classical forward–reverse diffusion modeling paradigm. In the forward process, the original image x_0 is progressively injected with Gaussian noise via a Markov chain, expressed as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I) \tag{6}$$

Where β_t is the diffusion coefficient at time step t, controlling the noise injection intensity. The reverse generation process is guided by a residual-conditioned denoising predictor to iteratively reconstruct the original signal, defined by the target distribution:

$$p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c, t), \Sigma_\theta(x_t, c, t)) \tag{7}$$

Here, c denotes the defect category label, and μ_θ and $\Sigma_\theta$are the conditional mean and variance estimated by the learned model.

Unlike traditional DDPM approaches, RDDM introduces a residual prediction strategy, which does not directly predict the original image x_0, but instead predicts the residual information:

$$\hat{r}_\theta(x_t, c, t) = x_t - \mu_\theta(x_t, c, t) \tag{8}$$

This residual-based formulation effectively mitigates issues such as edge blurring, texture degradation, and is particularly suitable for enhancing fine-grained defects like thread breaks, burrs, and contamination in industrial images.

In this study, we utilize real-world internal and external thread defect samples as priors. The RDDM is conditioned on defect labels to perform diffusion-based sampling, generating high-fidelity defect images that not only retain geometric consistency with real samples but can also simulate diverse defect types across different sampling iterations.

2.5. Defect Detection

In this study, the YOLO (You Only Look Once)[34,35] series is adopted as the foundational framework for metallic surface defect detection due to its high inference efficiency as a single-stage object detector. Among them, YOLOv8[36] significantly enhances feature extraction and multi-scale fusion capabilities by introducing the C2f module and BiFPN structure. However, challenges remain in accurately detecting small-scale defects under complex industrial backgrounds. The baseline network architecture is illustrated in Figure 6.

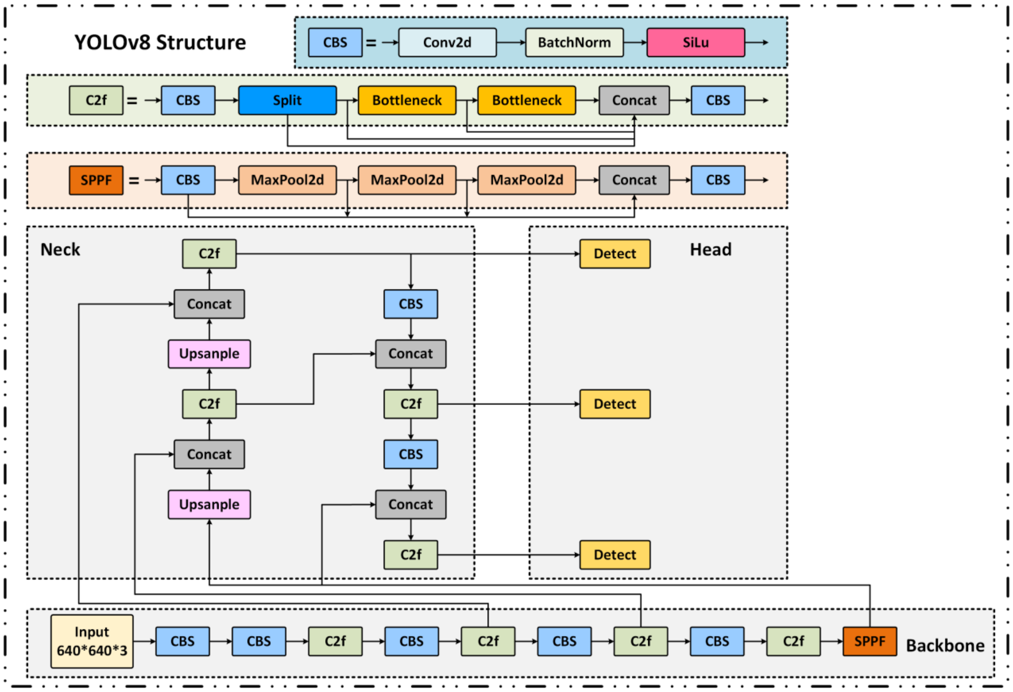

YOLOv8 Structure
CBS = Conv2d BatchNorm SiLu
C2f = CBS Split Bottleneck Bottleneck Concat CBS
SPPF = CBS MaxPool2d MaxPool2d MaxPool2d Concat CBS
Neck
C2f
Concat
Upsanple
C2f
Concat
Upsanple
CBS
Concat
C2f
CBS
Concat
C2f
Head
Detect
Detect
Detect
Input
640*640*3
CBS CBS C2f CBS C2f CBS C2f CBS C2f SPPF
Backbone

Figure 6. Architecture of the baseline YOLOv8 detection network.

To overcome these limitations, we propose a lightweight enhanced architecture named SLF-YOLO, which integrates three key components:

(1) the SC_C2f module for improved channel-wise feature fusion;
(2) the Light-SSF_Neck structure for efficient multi-scale aggregation;
(3) a novel loss function termed FIMetal-IoU, designed to optimize bounding box regression under industrial constraints.

The overall structure of SLF-YOLO is shown in Figure 7. In the backbone, the SC_C2f module incorporates a Star Block for enhanced feature interaction and leverages a Channel-Gated Linear Unit (CGLU) activation to enable fine-grained dynamic channel selection.

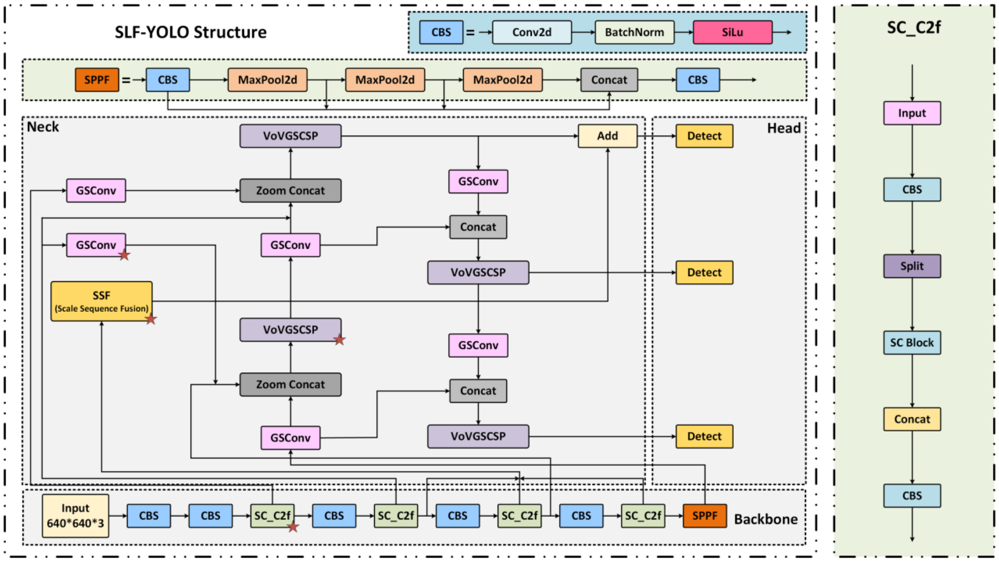


Figure 7. Architecture of the proposed SLF-YOLO detection network with modular enhancements.

The key computation formulas are defined as follows:

To optimize information flow and multi-scale feature fusion in the Neck stage, we adopt the Light-SSF_Neck structure. Its core component, the GSConv module, fuses Standard Convolution (SC) and Depthwise Separable Convolution (DSC) to enhance channel-wise information exchange via dual-path computation:

$$\mathrm{Y_{GSConv}} = \mathrm{Shuffle}\big(\mathrm{Concat}(\mathrm{Y_{SC}}, \mathrm{Y_{DSC}})\big) \tag{9}$$

Additionally, the Scale-Sequence Fusion (SSF) module extracts multi-scale features from P3, P4, and P5, and applies convolution, upsampling, and 3D convolutional fusion to construct cross-scale contextual representations:

$$\mathrm{Y_{fusion}} = \mathrm{Conv3D}\big(\mathrm{Concat}(\mathrm{Y_{P3}}, \mathrm{Y_{P4\rightarrow P3}}, \mathrm{Y_{P5\rightarrow P4}})\big) \tag{10}$$

To further enhance localization accuracy, we propose a novel loss function called FIMetal-IoU, which introduces an auxiliary box mechanism and piecewise weighting scheme into the IoU computation (as illustrated in Figure 9). The auxiliary box IoU is defined as:

$$\mathrm{IoU_{Inner}} = \frac{\mathrm{inter_{inner}}}{\mathrm{union_{inner}}} \tag{11}$$

Based on this, we apply piecewise weighting to different IoU intervals, and the final loss function is expressed as:

$$|\mathrm{FIMetal} - \mathrm{IoU}| = \begin{cases} 0, & \mathrm{IoU_{Inner}} < \mathrm{d} \\ \frac{\mathrm{inter_{inner}} - \mathrm{union_{inner}} \cdot \mathrm{d}}{\mathrm{union_{inner}}}(\mathrm{u} - \mathrm{d}), & \mathrm{d} < \mathrm{IoU_{Inner}} < \mathrm{u} \\ 1, & \mathrm{IoU_{Inner}} > \mathrm{u} \end{cases} \tag{12}$$

2.6. Adversarial Attacks on Image-Based Systems

Adversarial attacks on images represent a major security threat to deep learning-based vision systems. Their core objective is to induce incorrect predictions or outputs by introducing subtle but intentionally crafted perturbations to the input image, thereby compromising the system's robustness and trustworthiness.

Based on their implementation methods and attack effects, adversarial attacks can be categorized into various types. Among them, Alpha attacks, CCP (Color Channel Perturbation), and Patch attacks are three representative methods, as summarized in Table 1.

Table 1. Comparison of Representative Adversarial Attack Types on Visual Systems.

Attack Type	Perceptible	Implementation Complexity	Attack Strength	Applicability	Engineering Deployment Risk	Attack Type
Alpha	No	Medium	Very High	High	High	Alpha
CCP	Yes	Low	Medium	Medium	Medium	CCP
Patch	Yes	Low	Medium	High	Medium	Patch

Alpha attacks combine high imperceptibility with extremely strong attack capability, making them one of the most severe threats to current image recognition systems. Therefore, it is essential to develop high-sensitivity defense mechanisms specifically targeting this form of attack.

Although CCP and Patch attacks pose relatively lower threats, they still introduce practical security risks—especially in large-scale deployments of vision systems, where adversaries may exploit their low complexity to achieve rapid system compromise.

Accordingly, the design of image-level security defense strategies should be based on a multi-layered security framework, incorporating:

(1) robust model architecture design;
(2) adversarial training techniques;
(3) multimodal detection methods.

These measures collectively enhance the model's adversarial robustness and resistance to diverse threat vectors in real-world industrial environments.

Alpha channel attack is a stealthy adversarial method based on the transparency dimension of image representation. In recent years, it has emerged as a highly concealed and engineering-feasible input-level threat in security-sensitive industrial visual inspection systems. This method exploits the structural vulnerabilities in image processing pipelines by embedding adversarial perturbations into the alpha (transparency) channel of standard image formats (e.g., PNG), which are typically not perceived by human vision systems.

While alpha channel manipulations are generally unsupported in human-viewing libraries, they remain invisible yet processable in industrial vision pipelines that rely on image pre-processing frameworks such as OpenCV, TensorRT, PIL, or PyTorch. As a result, these perturbations bypass typical input validation and are treated as valid tensors by deep neural networks, allowing attackers to create stealthy adversarial samples without altering pixel color or brightness. The attack is formulated as:

$$x_{adv} = (1 - \alpha) \odot x_{orig} + \alpha \odot \delta \tag{13}$$

Here, x_{orig} is the original industrial image; δ is the adversarial perturbation map; α is the transparency mask controlling the blend intensity.This operation introduces controllable perturbations through the Alpha channel without altering the color and brightness distribution of the image pixels. It effectively interferes with the responses of the model in the convolution feature extraction of the previous layer, especially having a significant interference effect on the periodic textures, gap edges and multi-scale concave structures in the threaded images.

To ensure imperceptibility and maintain image quality, the perturbation design is subject to the following constrained optimization:

$$\min_{\delta} \mathcal{L}\,(f(x_{adv}), y_{target}) + \lambda \cdot ||\delta||_p \tag{14}$$

Here,$f(\cdot)$ denotes the target detection model,y_{target} is the desired misclassification output,The loss function is used to guide the model's output to deviate from the original detection result, while the regularization term controls the magnitude of the perturbation to meet the perceptual constraints.This form enables attackers to deceive the model under multiple task settings, including common industrial errors such as misclassification of defect types, deviation in position regression, and decrease in the confidence level of bounding boxes.$\mathscr{L}(\cdot)$ is the loss function guiding the attack, and $\|\delta\|_p$ is the perturbation norm, regularized by λ.

To preserve perceptual quality and system integrity, the following constraints are enforced:

$$\|\delta\|_\infty < \epsilon, \quad \alpha < \tau \tag{15}$$

Where ϵ is the maximum perturbation bound and τ is the upper bound for transparency. Both are typically set below 0.1 to avoid triggering quality-based preprocessing thresholds and to ensure compatibility with image format standards.

In threaded defect detection applications, Alpha channel perturbations have been experimentally demonstrated to cause typical false detections and missed detections at the output of neural network models. Specifically, such perturbations can lead to:Positional drift in defect localization (e.g., misaligned gap detection),Destruction of edge integrity and Interference from structurally repetitive regions.

Notably, even under static image dimensions, the adversarial effect exhibits strong transferability across samples and models, indicating cross-model attack capability. Given that most industrial lightweight detection networks—such as the SLF-YOLO model proposed in this study—do not explicitly regulate or suppress four-channel inputs, Alpha-based perturbations present a realistic deployment risk, warranting serious attention during the system design phase for security reinforcement.

In conclusion, Alpha channel attacks represent a form of implicit input-level perturbation characterized by:High stealthiness,Cross-model adaptability, andEngineering feasibility.

They have become an emerging but critical security threat in industrial-grade object detection systems. This work constructs an attack modeling framework tailored to thread-structured images, and systematically uncovers the disruptive mechanisms and misleading effects of Alpha perturbations on convolutional feature responses.

Furthermore, this study highlights the necessity of removal or masking mechanisms for the Alpha channel in the image preprocessing pipeline. Combined with the proposed [XX] defense model (placeholder for your actual model name), the approach provides a practical reference for Alpha-channel risk assessment and mitigation strategies during the pre-deployment stage of industrial vision systems, aiming to:Reduce vulnerability to adversarial inputs at the source, andEnhance the overall robustness and security of the system.

3. Experiments

3.1. Dynamic Deblurring

In industrial thread defect detection tasks, image blur is one of the key interference factors that significantly affects the accuracy of detection models. This issue is particularly pronounced in external threads, which often exhibit periodic structural features and small-scale defects. Even minor motion blur can severely degrade edge sharpness and target contrast, resulting in localization deviation, reduced confidence scores, or even complete miss detections.

Figure 8 illustrates a visual comparison of typical external thread corrosion defect images before and after dynamic deblurring.

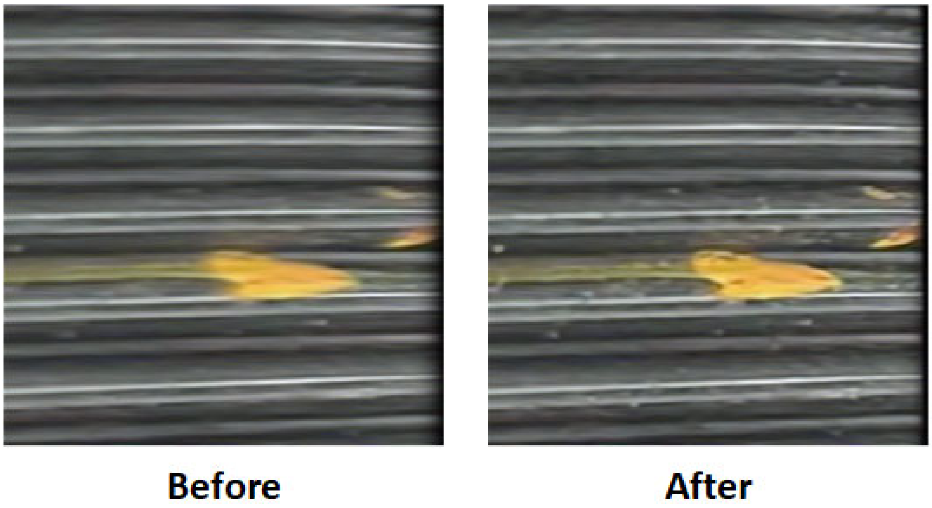


Figure 8. Comparison of thread corrosion defect images before and after dynamic deblurring.

The left image represents the original unprocessed input, where the high-speed rotation of the threaded pipe or minor camera vibration during image acquisition has introduced noticeable motion blur. This is evident in the blurred boundary of the corrosion spot (yellow area) and the streaking of background thread lines.

In contrast, the right image, processed using the proposed dynamic modeling-based deblurring enhancement method, shows sharpened defect edges, restored periodic thread patterns, and significantly improved contrast and texture clarity.

The proposed dynamic deblurring module integrates residual attention-based local blur recognition with a frequency-domain compensation mechanism, allowing for precise localization of degraded regions and adaptive restoration. While maintaining global structural consistency, it significantly enhances the separability and visual saliency of corrosion boundaries.

This method also demonstrates strong generalizability to common local blur issues in industrial thread imagery, such as:exposure trailing caused by illumination variation, and misalignment between motion speed and sampling rate.

It thus provides a robust and adaptable solution to blur-related challenges in real-world industrial inspection scenarios.

3.2. Impact of Illumination Normalization on Detection Accuracy

We conducted a visual comparison on internal thread images commonly found in industrial scenarios before and after enhancement.

In the left image (originally captured under extreme low-light conditions), the thread structure is nearly completely obscured by shadows, exhibiting severe black suppression and significant illumination non-uniformity.

The right image, processed using DarkIR, shows substantial improvements in detail visibility, with the hierarchical thread structures clearly recovered and the overall dynamic brightness range significantly expanded.

As an end-to-end deep learning method tailored for ultra-low-light image restoration, DarkIR adopts a learnable nonlinear mapping structure that enhances brightness while suppressing color distortion and noise amplification, issues that are frequently observed in traditional enhancement techniques. Specifically, DarkIR leverages a multi-scale attention mechanism and a dark-region-aware feature modeling module to implement an adaptive brightness compensation strategy. This makes it especially effective in industrial surface scenarios characterized by complex geometries and low reflectivity, such as threads and pipelines.

As illustrated in Figure 9, the step structures and inner-wall textures of the threads are clearly reconstructed after enhancement. Previously invisible micro-defects become discernible, thereby significantly improving the perceptual capability of downstream defect detection models in both localization and classification tasks.

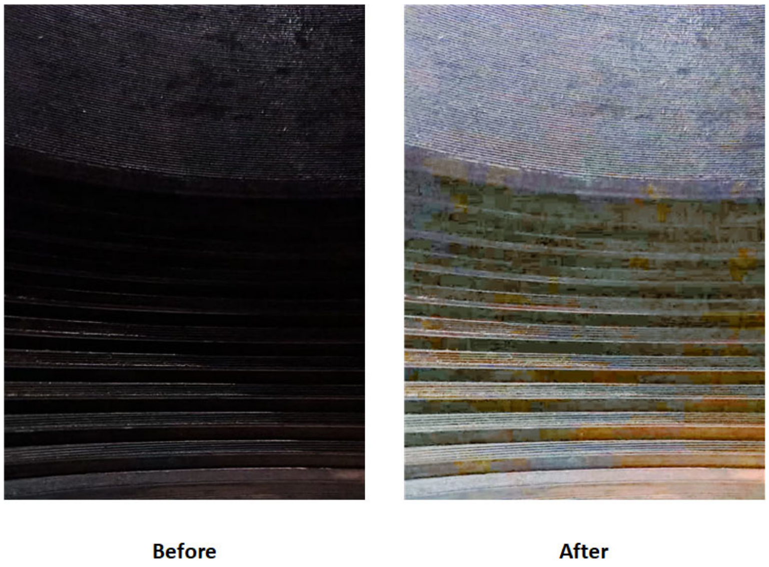


Figure 9. Internal thread image enhancement using DarkIR under low-light conditions.

Moreover, the enhanced images maintain sharp edge boundaries and structural integrity, which also provides a stable input foundation for further processing steps such as depth estimation and 3D reconstruction.

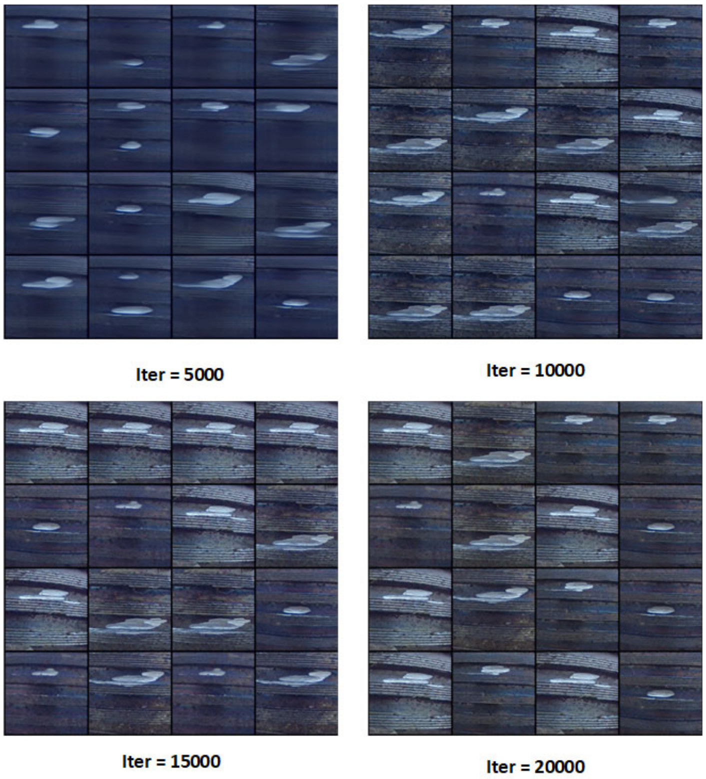
3.3. Image Data Augmentation Strategy and Model Generalization Analysis

To evaluate the adaptability and robustness of diffusion models in simulating industrial thread defects, this study conducted controllable generation experiments based on a Residual Denoising Diffusion Model (RDDM). The detailed parameter settings are listed in Table 2.

Table 2. Training configuration of the residual denoising diffusion model (RDDM) for industrial thread defect synthesis.

Parameter	Value
Image size	640 × 640
Timesteps	1000
Training steps	800
Optimizer	AdamW
Learning rate	1e-4
Batch size	16
Loss function	L1 + LPIPS + Residual penalty
Augmentations	Rotation, crop, contrast jitter

As shown in Figure 10, the model's progressive reconstruction results of internal and external thread defect images are visualized at different training iterations (Iter = 5000, 10000, 15000, 20000), clearly demonstrating the evolution from blurred structures to highly detailed and realistic defect images.



Iter = 5000

Iter = 10000

Iter = 15000

Iter = 20000

Figure 10. Progressive reconstruction of thread defect images using RDDM at different training stages.

At iteration 5000, the generated images remain in the high-noise reverse diffusion phase, with blurry defect contours, limited texture, and vague geometric structures.

By iteration 10000, the thread contours become more defined, and the metallic surface textures along with spatial coherence of the defect areas begin to emerge, indicating the model has preliminarily learned the semantic features of industrial thread structures.

At iteration 15000, the model is capable of synthesizing high-quality images with typical damage characteristics, such as localized wear, burr edges, and corrosion spots—reflecting its ability to accurately simulate mid-scale structural degradations.

By the final iteration 20000, the generated images achieve high photorealism, with well-reconstructed surface textures, illumination reflections, and fine-grained defect details. These images exhibit complexity and discriminative features comparable to real-world inspection data, making them highly suitable for enhancing model generalization under limited data conditions.

3.4. Performance Evaluation of the Defect Detection Model

To comprehensively evaluate the adaptability and generalization performance of the proposed SLF-YOLO model in real-world industrial inspection scenarios, we conducted visual analyses of its detection performance on a variety of typical thread surface defect images. The results are presented in Figure 11.

For “scratch”-type defects, the model accurately identifies linear scratches of varying lengths and textures. It demonstrates the ability to distinguish between clear boundaries and partially blurred edges. Despite some scratch edges blending into the background due to color similarity, the model consistently provides high-confidence predictions (score ≥ 0.9), indicating strong sensitivity to linear texture features. Moreover, it offers low-confidence indications in uncertain regions, which can support manual verification or multi-model ensemble processing.

In the case of “broken”-type defects, the model shows exceptional stability, especially in detecting vertically distributed multi-point damage, achieving high-confidence multi-object predictions (confidence ≥ 0.9). These results suggest that SLF-YOLO has a high detection rate and localization consistency for small-scale, discrete defects, making it highly suitable for automated inspection tasks involving thread wear and microcracks.

For “corrosion”-type defects, the model successfully identifies irregularly shaped, blurred-boundary corrosion regions, assigning relatively high confidence scores despite the indistinct edges.

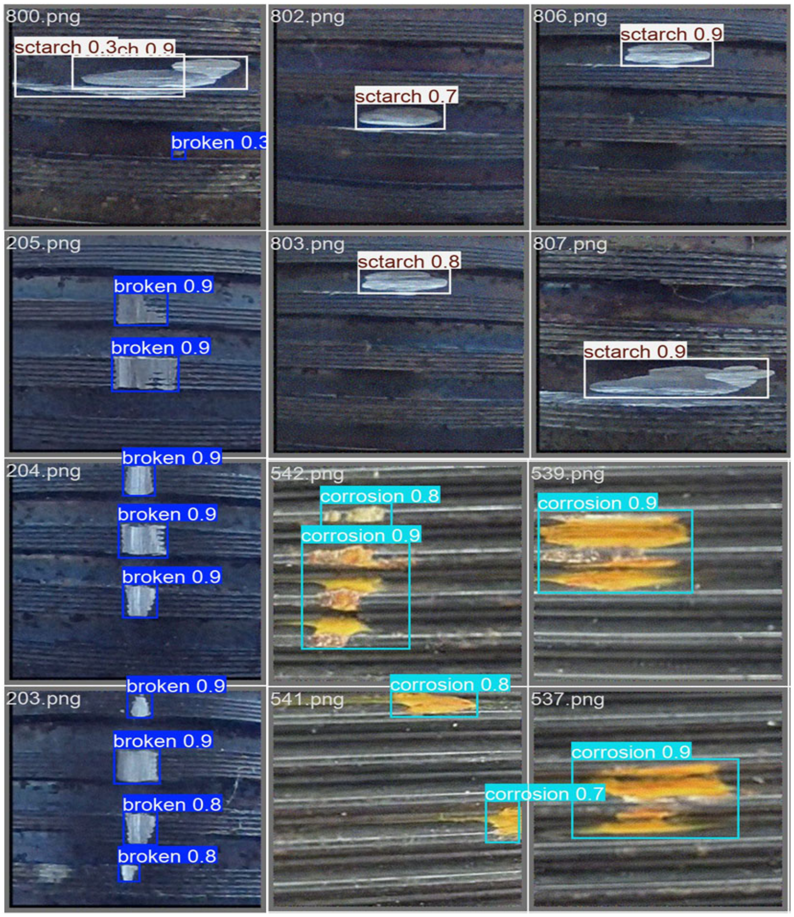


Figure 11. Detection results of SLF-YOLO on various real-world thread defect types: scratch, broken, and corrosion.

Overall, SLF-YOLO demonstrates excellent performance in detecting scratches, fractures, and corrosion on threaded surfaces. It exhibits strong defect type discrimination, multi-scale adaptability, and robustness under complex backgrounds. The high confidence, consistent multi-target detection, and stability across varied defect scenarios validate the effectiveness and industrial applicability of the proposed structural improvements—namely, the Channel-Gated Linear Unit (CGLU) mechanism and the SlimNeck lightweight fusion structure.

To further assess the training stability and convergence behavior of the proposed model, we visualized the variation trends of key loss functions and performance metrics throughout training, as shown in Figure 12.

The box regression loss, classification loss, and distribution focal loss (DFL) rapidly decreased during the first 100 epochs and gradually stabilized thereafter. This trend indicates good convergence behavior without noticeable overfitting. The consistency between training and validation loss curves further confirms the model’s generalization capability in thread defect detection.

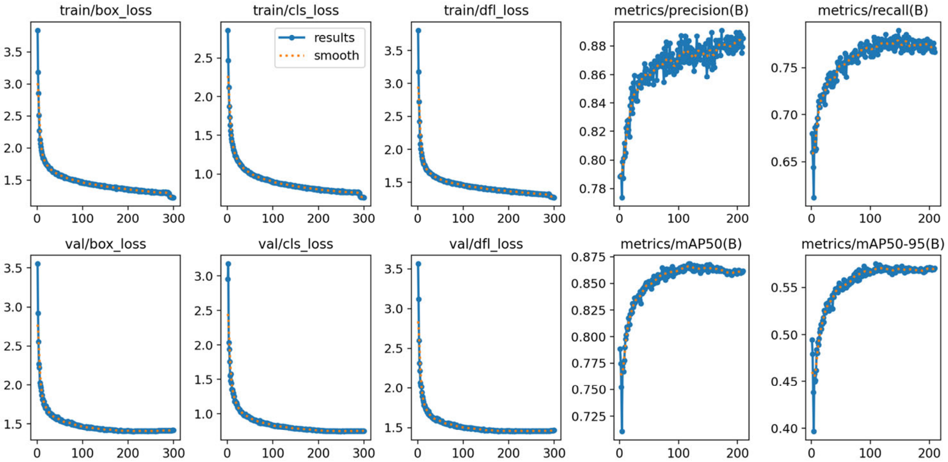


Figure 12. Training and validation loss curves and performance metric trends of the SLF-YOLO model.

In parallel, key performance metrics showed continuous improvement:

(1) Precision increased steadily from ~0.78 to 0.88, indicating a significant reduction in false positives;

(2) Recall improved from 0.68 to 0.78, reflecting a substantial drop in missed detections;

(3) The final mAP@0.5 reached 0.88, while the more stringent mAP@0.5:0.95 reached 0.56, demonstrating the model's strong detection capability across varying object sizes and IoU thresholds.

3.5. Analysis and Visualization of Adversarial Perturbation Effects

Alpha channel attacks represent a form of adversarial input manipulation based on the transparency dimension of image data. These attacks exhibit extremely high stealthiness and practical feasibility. From a visual perspective, Alpha attacks do not alter the image's color, brightness, or structural content. However, by embedding subtle perturbations into the alpha (transparency) channel, they can significantly disrupt feature extraction and inference pathways in deep neural networks—posing a substantial threat to defect detection tasks that rely heavily on edge clarity and texture consistency, as is common in industrial imagery.

In this study, we specifically investigate whether lightweight industrial defect detection models such as YOLO are vulnerable to Alpha channel exposure, particularly under default conditions where RGBA-format images are accepted without channel masking. Furthermore, we evaluate how such perturbations impact the detection accuracy and output stability of the model. Through this experiment, we aim to uncover the potential risks posed by input-level implicit attacks and provide a quantitative foundation and technical reference for the design of robust defense mechanisms in industrial vision systems.

To systematically evaluate the impact of Alpha channel attacks and other input perturbation methods, we conducted a series of comparative experiments based on the trained SLF-YOLO model. The objective is to assess how three representative adversarial attack types affect the model's accuracy, stability, and false detection risk under different mechanisms, as illustrated in Figure 13.

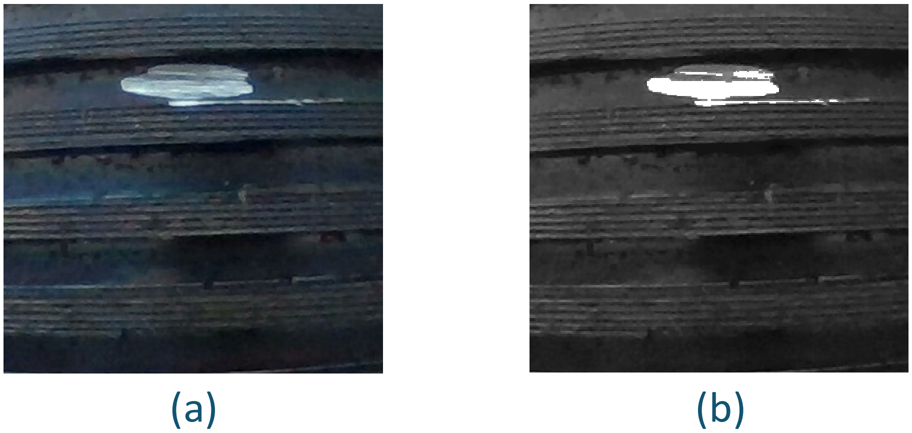


Figure 13. Visual and statistical impact comparison of Alpha, CCP, and Patch adversarial attacks on thread defect detection using SLF-YOLO.

As summarized in Table 3, the Alpha channel attack caused a drastic degradation in SLF-YOLO's detection performance.

Table 3. Detection performance metrics of SLF-YOLO under Alpha channel attack.

Type	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5–0.95
No Attack	0.893	0.958	0.969	0.717
Alpha	0.017	0.128	0.027	0.0045

Under the no attack condition, the model demonstrated strong performance:
Precision (P) = 0.893
Recall (R) = 0.958
mAP@0.5 = 0.969
mAP@0.5:0.95 = 0.717
These results reflect the model's high sensitivity to subtle defects on thread surfaces.
However, under Alpha perturbation, performance dropped catastrophically:
Precision = 0.017
Recall = 0.128
mAP@0.5 = 0.027
mAP@0.5:0.95 = 0.0045

This stark contrast reveals that although Alpha perturbations are visually imperceptible, they can fatally disrupt the model's discriminative mechanisms, rendering the detection task almost entirely ineffective. The attack bypasses traditional pixel-based anomaly detectors and directly targets the front end of the perception pipeline, highlighting its extreme stealth and destructive potential.

4. Discussion

4.1. Contribution Analysis of the Preprocessing Module to Model Performance

Image preprocessing plays a critical role in enhancing the performance of deep learning models, particularly in tasks such as motion deblurring and low-light enhancement. To quantify the impact of preprocessing on final image quality, we compared the performance of various algorithms using two standard metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The detailed results are summarized in Table 4.

Table 4. Quantitative comparison of different preprocessing algorithms in motion deblurring and low-light enhancement tasks.

Task	Algorithm	PSNR（dB）↑	SSIM↑
motion deblurring	MLWNet-B	30.3	0.940
	DeblurGAN-v2	27.6	0.903
	SRN	28.7	0.910
Low-light enhancement	DarkIR	26.4	0.945
	HWMNet	27.0	0.922
	FLOL	25.9	0.920

In the motion deblurring task, MLWNet-B achieves the best performance, with a PSNR of 30.3 dB and SSIM of 0.940, significantly outperforming DeblurGAN-v2 (27.6 dB, 0.903) and SRN (28.7 dB, 0.910). These results demonstrate MLWNet-B's superior capability in blur modeling and multi-scale detail recovery. Notably, in cases where the blur kernel is non-estimable, MLWNet-B's adaptive wavelet-based feature extraction mechanism effectively enhances image clarity and perceptual quality while preserving structural consistency.

In the low-light enhancement task, DarkIR attains the highest SSIM (0.945), indicating excellent structural preservation during enhancement. However, its PSNR (26.4 dB) is slightly lower than that of HWMNet (27.0 dB), suggesting marginally weaker noise suppression. Overall, HWMNet achieves a balanced trade-off between brightness enhancement and detail preservation. FLOL, while slightly lower in both PSNR (25.9 dB) and SSIM (0.920), delivers stable performance, demonstrating good generalization in illumination modeling.

These results affirm that effective preprocessing modules significantly mitigate image degradation, which is common in complex industrial settings. By enhancing edge sharpness, texture contrast, and structural separability, preprocessing serves as a supportive foundation for downstream tasks such as object detection and classification.

4.2. PerformanceApplicability of Image Data Augmentation in Thread Defect Detection

In the task of synthetic generation of thread defect images, we evaluated the Fréchet Inception Distance (FID) trend of the Residual Denoising Diffusion Model (RDDM) under different sampling steps. As illustrated in Figure 14, the results clearly demonstrate the model's capability to progressively enhance image quality during the generation process. As the number of sampling steps increases from 5 to 100, the FID value drops significantly from 69.6 to 24.92, exhibiting a distinct nonlinear decreasing trend.

Under low sampling steps, RDDM is already capable of rapidly generating defect images with coherent global structure and basic surface morphology. These samples exhibit sufficient structural similarity and discriminative features, making them suitable for real-time pseudo-sample generation and online data augmentation, especially where computational efficiency is critical.

As the number of sampling steps increases, the model gains more capacity to refine texture, lighting, and boundary details, enabling the synthesis of high-fidelity samples that more closely resemble real-world defects. These refined images are particularly beneficial for improving the robustness of defect detection models, especially in multi-type defect scenarios involving corrosion, scratches, and fractures.

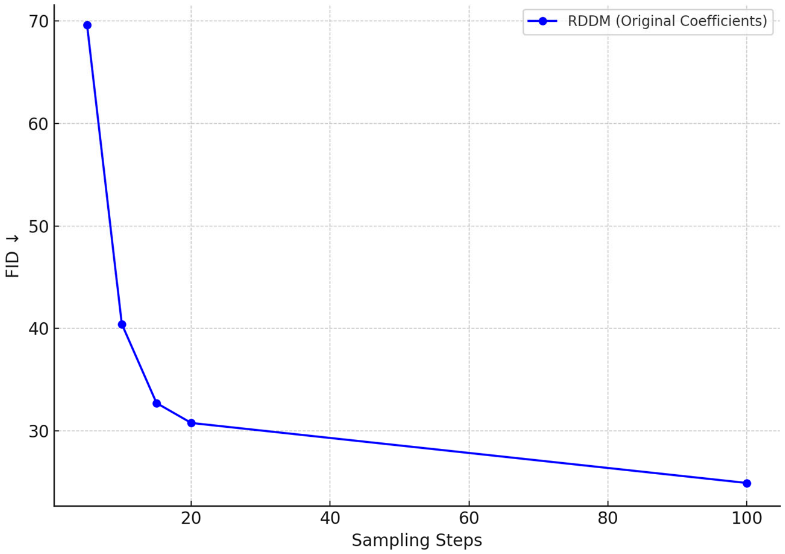


Figure 14. FID variation of RDDM-generated thread defect images under different sampling steps.

RDDM achieves this by leveraging a decoupled residual and noise diffusion mechanism, which enables the model to capture global structural patterns while incrementally enriching fine-grained details. This significantly enhances the structural expressiveness of defect regions. Furthermore, the high quality of generated samples serves as a stable input foundation for subsequent lightweight detection networks, improving system robustness against complex backgrounds, lighting variations, and sample distribution shifts.

The FID curve validates that RDDM is capable of delivering high-quality image synthesis at relatively low sampling costs, making it highly applicable for a wide range of industrial vision tasks, including multi-source defect modeling, pseudo-sample generation, and adversarial sample construction for defense training. Therefore, RDDM serves as a critical generative component in enabling intelligent thread defect detection systems.

4.3. Ablation Study of Detection Module Components

To systematically evaluate the contribution of each structural improvement to the performance of the proposed thread defect detection model, a series of ablation experiments were conducted. The comparison results are shown in Table 4, analyzing performance across multiple dimensions including Precision, Recall, mean Average Precision (mAP@0.5 and mAP@0.5:0.95), and model complexity (number of parameters and FLOPs).

Using the Baseline model (with no structural enhancements) as the reference, it achieved a Precision of 0.821 and mAP@0.5 of 0.759, serving as a performance benchmark. With the introduction of the SC_C2f module, Precision increased to 0.842 and mAP@0.5 rose to 0.781, indicating that this shallow attention mechanism improves feature expressiveness.

Similarly, the Light-SSF_Neck module showed stronger spatial feature aggregation capabilities. Although its Recall (0.691) was slightly lower, its Precision improved to 0.841, still outperforming the Baseline. In terms of loss function, integrating the FIMetal-IoU led to more precise bounding box supervision. While the improvement in mAP@0.5 was modest (from 0.759 to 0.774), the mAP@0.5:0.95 significantly increased to 0.449, highlighting its strength in refining fine-grained localization.

The combination of SC + Neck achieved a balanced improvement across metrics, raising Recall to 0.784 and mAP@0.5 to 0.793. Configurations SC + IoU and Neck + IoU yielded Precision values of 0.864 and 0.868, respectively, validating the complementarity between feature enhancement and localization optimization modules. However, Neck + IoU showed slightly lower Recall, possibly due to reduced robustness in handling low-quality proposals.

doi:10.20944/preprints202508.1070.v1

The complete model, SLF-YOLO (All), which integrates all proposed modules, achieved the best overall performance:

Precision: 0.881

Recall: 0.794

mAP@0.5: 0.813

mAP@0.5:0.95: 0.521

Compared to the Baseline, these represent respective improvements of 6.0%, 7.6%, 7.1%, and 8.9%. Importantly, this performance gain comes with manageable complexity:

Parameters: 9.65M

FLOPs: 24.6G

Indicating strong potential for real-world deployment in industrial settings.

Table 4. Ablation study on structural components of SLF-YOLO.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Params (M)	FLOPs (G)
Baseline	0.821	0.718	0.759	0.411	11.12	28.4
SC_C2f	0.842	0.742	0.781	0.445	10.16	25.2
Light-SSF_Neck	0.841	0.691	0.776	0.445	10.33	26.3
FIMetal-IoU	0.855	0.741	0.774	0.449	11.12	28.4
SC + Neck	0.847	0.784	0.793	0.462	9.65	24.6
SC + IoU	0.864	0.755	0.785	0.449	10.16	25.2
Neck + IoU	0.868	0.665	0.785	0.458	10.33	26.3
All	0.881	0.794	0.813	0.521	9.65	24.6

To further validate the overall performance of SLF-YOLO in industrial defect detection tasks, we compared it with leading lightweight detection models: YOLOv5s, YOLOv8s, YOLOv9s, and YOLOv10s, using the same dataset and training strategy. As shown in Table 5, SLF-YOLO achieved the highest Precision (0.881) among all models, outperforming YOLOv5s (0.862), YOLOv10s (0.850), and slightly surpassing YOLOv9s (0.870).

Importantly, SLF-YOLO also achieved the highest Recall (0.794), indicating better resistance to missed detections. While mAP@0.5 (0.813) was slightly lower than YOLOv8s (0.832) and YOLOv9s (0.829), SLF-YOLO demonstrated a more balanced trade-off between accuracy and recall, confirming its detection stability and generalization capacity.

Table 5. Comparison of SLF-YOLO with state-of-the-art lightweight YOLO models.

Models	Precision	Recall	mAP@0.5
YOLOv5s	0.862	0.629	0.725
YOLOv8s	0.869	0.732	0.832
YOLOv9s	0.870	0.729	0.829
YOLOv10s	0.850	0.703	0.817
Ours	0.881	0.794	0.813

In conclusion, the proposed SLF-YOLO architecture achieves significant gains in both detection accuracy and stability while maintaining a lightweight structure, outperforming current mainstream YOLO variants. These results validate the effectiveness and advancement of the proposed model in practical industrial defect detection applications.

4.4. Analysis of Adversarial Perturbation Effects on Detection Model Mechanisms

This section explores the internal interference mechanisms of adversarial perturbations against lightweight industrial defect detection models. We investigate three representative types of attacks:

doi:10.20944/preprints202508.1070.v1

(1) Alpha Channel Attack: Introduces imperceptible structural interference by embedding a low-opacity perturbation layer in the Alpha channel. This simulates the risk of unfiltered RGBA input passing through pre-processing stages unhandled.

(2) Color Channel Perturbation (CCP) Attack: Distorts RGB channel ratios via color mapping matrix adjustments, resulting in global color shifts that hinder accurate edge and texture recognition.

(3) Patch Attack: Applies high-contrast patches to critical image regions (e.g., thread junctions or edges), simulating physical occlusions or targeted adversarial triggers.

Figure 13 illustrates the impact of these attacks. Subfigure (a) shows the original defect image; (b) is the Alpha-perturbed adversarial sample; (c) shows CCP-affected input; and (d) presents the Patch-augmented sample. All inputs maintain the same resolution, brightness, and content to ensure that performance changes stem solely from the applied perturbations.

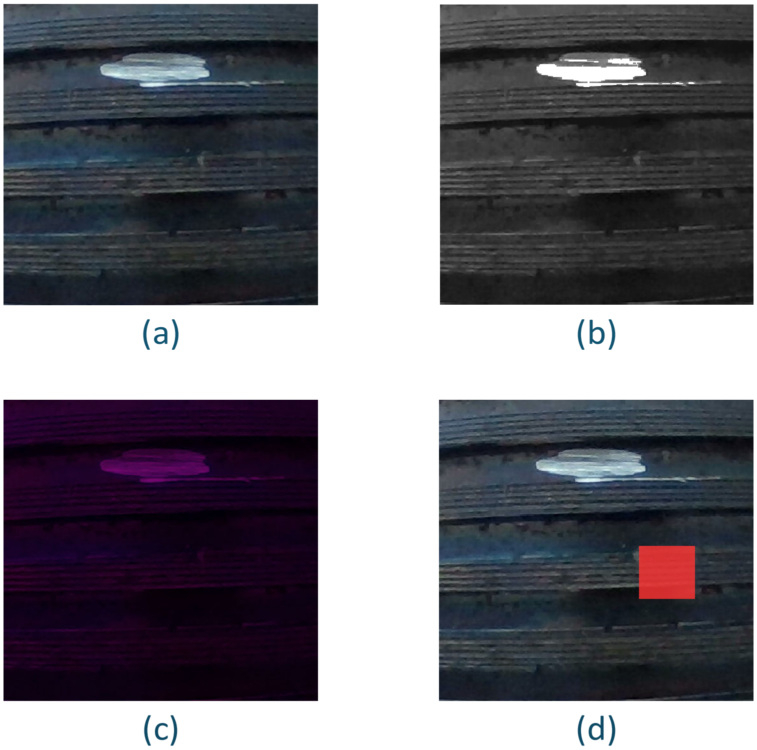


Figure 13. Visualization of adversarial perturbation effects on industrial thread defect images under three attack types: Alpha channel attack, Color Channel Perturbation (CCP), and Patch-based attack.

A unified SLF-YOLO model, with fixed structure and weights, is used to evaluate detection performance across four standard metrics: Precision, Recall, mAP@0.5, and mAP@0.5:0.95. Additionally, we assess per-class recognition rate changes and feature activation shifts to further analyze the disruption patterns caused by each perturbation mechanism.As can be seen in Table 6.

Table 6. Performance of SLF-YOLO under different adversarial attacks.

Type	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5–0.95
No Attack	0.881	0.794	0.813	0.521
Alpha	0.017	0.128	0.027	0.005
CCP	0.732	0.373	0.515	0.340
Patch	0.754	0.840	0.764	0.511

In the absence of any attacks, the model demonstrated strong performance, achieving a Precision of 0.881, Recall of 0.794, mAP@0.5 of 0.813, and mAP@0.5–0.95 of 0.521, reflecting its high accuracy and generalization capability.

Under Alpha channel perturbation, model performance collapsed severely. The mAP@0.5–0.95 dropped precipitously to 0.005, while Precision and Recall fell to 0.017 and 0.128, respectively. This confirms the catastrophic impact of Alpha perturbations, which—despite being visually imperceptible—severely disrupt the model's feature extraction and recognition logic.

In contrast, CCP attacks led to moderate degradation, with Recall dropping to 0.373 and mAP@0.5–0.95 falling to 0.340, indicating the model's sensitivity to chromatic distortions, especially in texture-based defect recognition.

Patch attacks, simulating physical occlusions, had localized effects. Although Recall remained relatively high (0.840), Precision dropped, and mAP@0.5–0.95 decreased to 0.511, suggesting a rise in false positive detections due to attention misdirection.

To further examine the perceptual behavior shifts induced by these perturbations, we utilized Grad-CAM to visualize the model's feature response under different attack conditions. As shown in Figure 14.

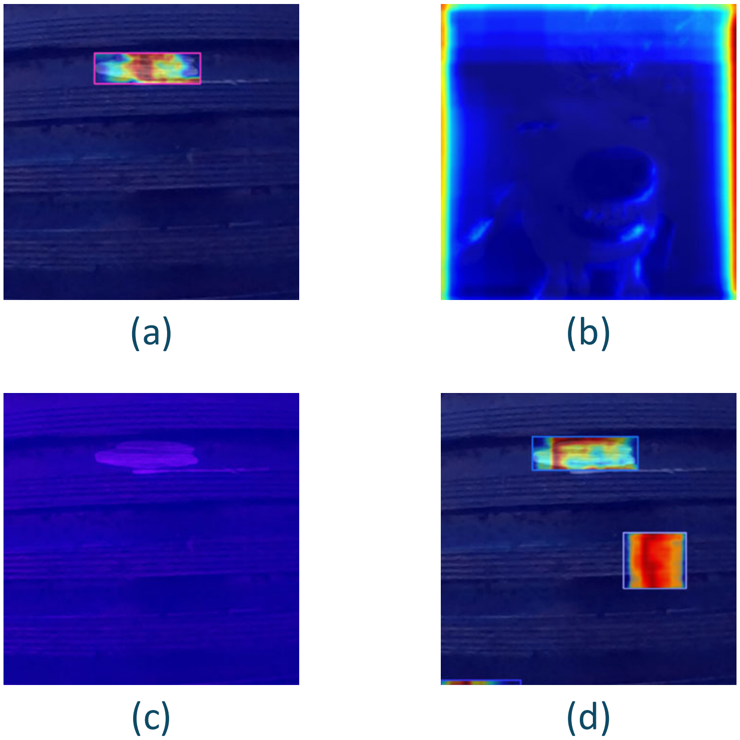


Figure 14. Grad-CAM visualizations of SLF-YOLO attention maps under different perturbation scenarios.

This analysis confirms that Alpha channel perturbations pose the most severe threat, primarily due to their invisible nature and disruptive power at the feature extraction stage. Meanwhile, CCP and Patch attacks, although visually perceptible, also warrant defense due to their practical deployment feasibility in industrial environments.

These results highlight the critical need for preprocessing modules that strip non-RGB channels and implement adversarial input screening, as well as for designing robust network architectures resilient to both implicit and explicit perturbations in visual industrial applications.

4.5. Effectiveness and Deployment Feasibility of the Defense Strategy

To address the high stealthiness and misleading nature of Alpha channel attacks in industrial vision systems, this study proposes a visual consistency analysis method based on perceptual differences. Unlike conventional defense strategies that rely on model structures or inference paths, this method operates directly at the image level, detecting latent discrepancies between AI perception and human vision to enable unsupervised identification of Alpha channel attack samples.

Specifically, we reconstruct two versions of each input image: the AI-perceived image (IAI) and the human-viewed image (IEye). The IAI is obtained by averaging the RGB channels, simulating how most vision models process inputs when the Alpha channel is implicitly discarded. In contrast, the IEye is reconstructed using the standard alpha compositing formula by blending the RGBA image with a default background (e.g., white, as in typical web displays), thereby approximating the actual visual experience of human observers. For normal images, these two representations should be nearly identical. However, under Alpha attacks, they often exhibit significant divergence in pixel distribution and structure.

To quantify such differences, we introduce two key detection metrics:

(1) Histogram Overlap, which measures the grayscale distribution similarity between IAI and IEye — lower values indicate greater discrepancy.

(2) Mean Squared Error (MSE), which captures pixel-wise deviation between the two images.

Empirical results show that Alpha-attacked images typically exhibit a histogram overlap below 0.05 and an MSE significantly higher than that of clean samples, making them robust indicators of adversarial manipulation.

As illustrated in Figure 15, the grayscale histogram comparison between IAI and IEye shows a highly divergent distribution, with an overlap of only 0.0015, indicating near-complete disassociation — a signature of Alpha-based perturbation.

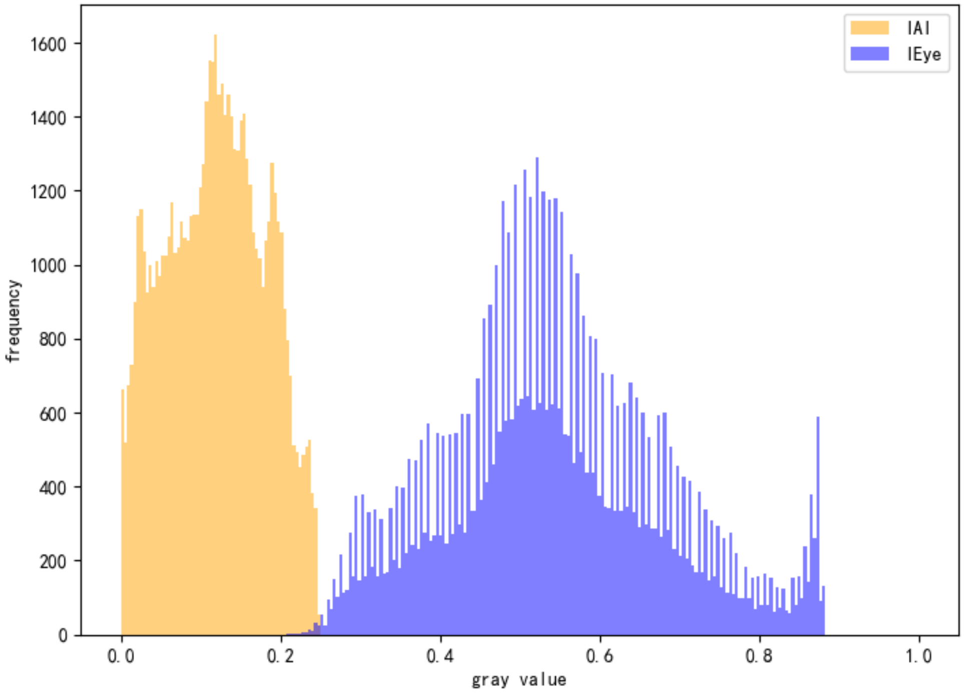


Figure 15. Grayscale Histogram Comparison between AI Perceived Image (IAI) and Human Viewed Image (IEye) under Alpha Channel Attack.

doi:10.20944/preprints202508.1070.v1

Figure 16 presents the MSE heatmap, where the orange-red regions highlight systematic pixel-level discrepancies between the original content and the adversarial overlay, with an average MSE of 0.2054, far exceeding the noise tolerance threshold of typical images.

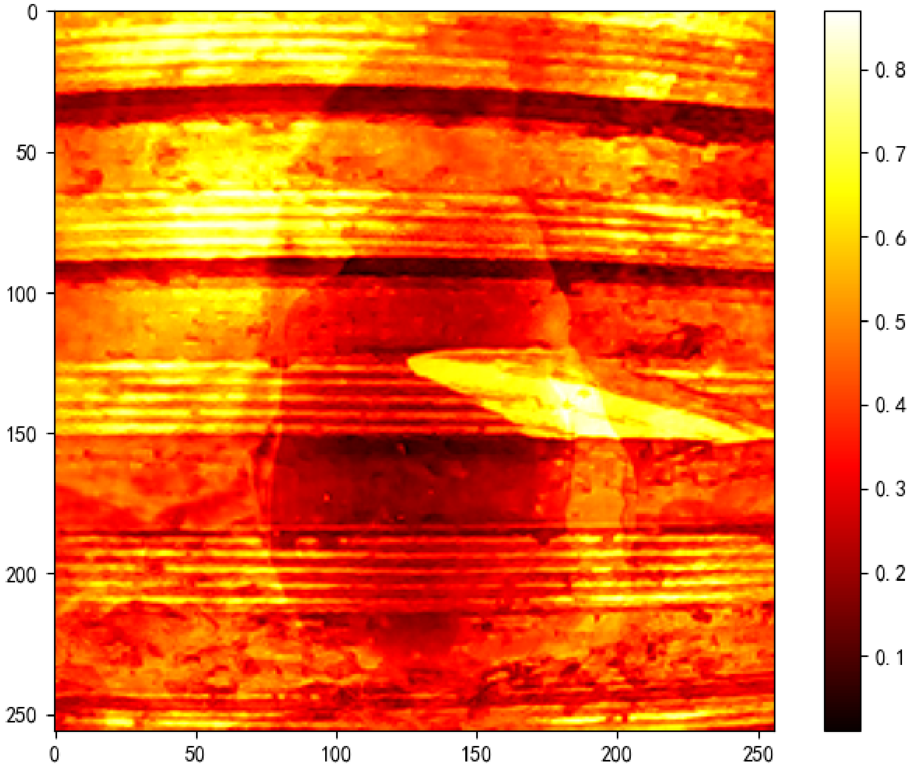


Figure 16. MSE Heatmap Highlighting Structural Deviations Caused by Alpha Channel Attack.

Experiments confirm that when the histogram overlap falls below 0.05 or the MSE exceeds 0.01, the image can be reliably flagged as a potential Alpha channel attack. The entire detection pipeline takes less than 20ms on a standard CPU, requires no dependence on deep network architectures or prior knowledge of specific attacks, and features lightweight computation, clear thresholds, and strong generalization ability — making it a practical and deployable solution for mitigating Alpha-based adversarial threats in real-world industrial vision systems.

5. Conclusions

This work presents an intelligent vision system for detecting internal and external thread defects, featuring a closed-loop design from image acquisition and enhancement to sample generation, lightweight detection, and adversarial defense. Experimental evaluations show that integrating MLWNet and DarkIR in preprocessing substantially improves input image discriminability, while high-fidelity pseudo-samples generated via RDDM enhance model generalization under limited data. The proposed SLF-YOLO achieves high detection accuracy and stability in complex industrial environments. With its dual defense strategy—combining input perturbation suppression and output anomaly detection—the system exhibits strong resilience against Alpha channel attacks. Overall, the method strikes a balanced trade-off among detection precision, security robustness, and deployment efficiency, demonstrating strong potential for industrial-scale adoption. Future work will focus on multimodal defense mechanisms and cross-device robustness optimization, driving inspection systems toward greater intelligence, security, and adaptability.

doi:10.20944/preprints202508.1070.v1

References

1. Yu, Jing, et al. "Geometric error modeling of the contact probe in a three-dimensional screw thread measuring machine." Measurement 194 (2022): 111026.
2. Yu, Jing, et al. "Design and characteristic research of contact probe for high-precision 3D thread-measuring machine." The International Journal of Advanced Manufacturing Technology 119.3 (2022): 2235-2245.
3. Yi, Guo, Ma Liting, and Su Chong. "The method of thread defect detection based on machine vision." 2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2019.
4. Dou, Xiaohan, et al. "Research on internal defect detection method based on machine vision." Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition. 2024.
5. Liu, Yanhua, et al. "Research on Deep Hole and Large Thread Defect Detection Based on Machine Vision Fusion." International Conference on Computing, Control and Industrial Engineering. Singapore: Springer Nature Singapore, 2024.
6. Liu, Lei, et al. "Review of optical detection technologies for inner-wall surface defects." Optics & Laser Technology 162 (2023): 109313.
7. Huo, Qishuo, and Zhihong Liang. "A structural scheme design of internal thread detection based on laser profile scanning." Journal of Physics: Conference Series. Vol. 2921. No. 1. IOP Publishing, 2024.
8. Zhang, Zhouqiang, et al. "Knitting needle fault detection system for hosiery machine based on laser detection and machine vision." Textile Research Journal 91.1-2 (2021): 143-151.
9. Zuo, Fengyuan, et al. "An X-ray-based automatic welding defect detection method for special equipment system." IEEE/ASME Transactions on Mechatronics 29.3 (2023): 2241-2252.
10. Rafiei, Mehdi, Jenni Raitoharju, and Alexandros Iosifidis. "Computer vision on x-ray data in industrial production and security applications: A comprehensive survey." Ieee Access 11 (2023): 2445-2477.
11. Wu, Lin, et al. "A dataset for surface defect detection on complex structured parts based on photometric stereo." Scientific Data 12.1 (2025): 276.
12. Wang, Chenwei, et al. "Defect detection method based on sparse scanning with laser ultrasonics." Scientific Reports 15.1 (2025): 13175.
13. Jin, Lei, et al. "Outer surface defect detection of steel pipes with 3D vision based on multi-line structured lights." Measurement Science and Technology 35.6 (2024): 065203.
14. Jang, Seok-Woo, Limin Yan, and Gye-Young Kim. "Deep Supervised Attention Network for Dynamic Scene Deblurring." Sensors 25.6 (2025): 1896.
15. Ren, Wenqi, et al. "Deblurring dynamic scenes via spatially varying recurrent neural networks." IEEE transactions on pattern analysis and machine intelligence 44.8 (2021): 3974-3987.
16. Gao, Hongyun, et al. "Dynamic scene deblurring with parameter selective sharing and nested skip connections." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
17. Zhang, Jiawei, et al. "Deep dynamic scene deblurring from optical flow." IEEE Transactions on Circuits and Systems for Video Technology 32.12 (2021): 8250-8260.
18. Chen, Mingju, et al. "An efficient image deblurring network with a hybrid architecture." Sensors 23.16 (2023): 7260.
19. Vasluianu, Florin-Alexandru, et al. "Towards image ambient lighting normalization." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
20. Da Cruz, Steve Dias, et al. "Illumination normalization by partially impossible encoder-decoder cost function." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.
21. Huang, Jie, et al. "Transition-constant normalization for image enhancement." Advances in Neural Information Processing Systems 36 (2023): 20562-20576.
22. Rad, Mahdi, Peter M. Roth, and Vincent Lepetit. "ALCN: Adaptive local contrast normalization." Computer Vision and Image Understanding 194 (2020): 102947.
23. Goswami, Suranjan, and Satish Kumar Singh. "A simple deep learning based image illumination correction method for paintings." Pattern Recognition Letters 138 (2020): 392-396.
24. Jiang, Zhihao, et al. "Internal Thread Defect Generation Algorithm and Detection System Based on Generative Adversarial Networks and You Only Look Once." Sensors 24.17 (2024): 5636.

25. Dou, Xiaohan, et al. "Internal thread defect detection system based on multi-vision." Plos one 19.5 (2024): e0304224.
26. Xu, Haitao, Haipeng Pan, and Junfeng Li. "Surface defect detection of bearing rings based on an improved YOLOv5 network." sensors 23.17 (2023): 7443.
27. Wu, KeZhu, et al. "RBS-YOLO: A Lightweight YOLOv5-Based Surface Defect Detection Model for Castings." IET Image Processing 19.1 (2025): e70018.
28. Patil, R. S.** (2024). A Comparative Analysis of YOLOv8 and YOLOv5 for Nut Thread Classification – Deep Learning Approach. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 12(2), 1863–1869.
29. Lang, Xianli, et al. "MR-YOLO: An improved YOLOv5 network for detecting magnetic ring surface defects." Sensors 22.24 (2022): 9897.
30. Tabernik, Domen, et al. "Segmentation-based deep-learning approach for surface-defect detection." Journal of Intelligent Manufacturing 31.3 (2020): 759-776.
31. Wang, Junpu, et al. "Defect transformer: An efficient hybrid transformer architecture for surface defect detection." Measurement 211 (2023): 112614.
32. Gao, Xin, et al. "Efficient multi-scale network with learnable discrete wavelet transform for blind motion deblurring." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
33. Rahman, Zia-ur, Daniel J. Jobson, and Glenn A. Woodell. "Retinex processing for automatic image enhancement." Journal of Electronic imaging 13.1 (2004): 100-110.
34. Terven, Juan, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas." Machine learning and knowledge extraction 5.4 (2023): 1680-1716.
35. Jiang, Peiyuan, et al. "A Review of Yolo algorithm developments." Procedia computer science 199 (2022): 1066-1073.
36. Sohan, Mupparaju, Thotakura Sai Ram, and Ch Venkata Rami Reddy. "A review on yolov8 and its advancements." International Conference on Data Intelligence and Cognitive Informatics. Springer, Singapore, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.