International Journal of Innovative Computing, Information and Control
Volume **22**, Number **3**, June **2026**

ICIC International ©2026 ISSN 1349-4198
pp. **653–671**

DDC-YOLO11: A LIGHTWEIGHT METHOD FOR DETECTING MULTIPLE TYPES OF TEARS IN BELTS

Yunzheng Tao[1], Jianyun Ma[1], Zhizhen Wu[1], Shouming Cun[1], Xiaohui Guo[1]
Xiaopan Wang[2] and Shuting Wan[2,*]

[1]Guoneng Yangzonghai Power Generation Co., Ltd.
Tangchi Town, Yiliang County, Kunming 652103, P. R. China
{ 12012407; 12092957; 12088456; 12003354 }@ceic.com; 474468232@qq.com

[2]Hebei Key Laboratory of Electric Machinery Health Maintenance and Failure Prevention
North China Electric Power University
No. 619, Yonghua North Street, Baoding 071003, P. R. China
442594791@qq.com; *Corresponding author: 52450809@ncepu.edu.cn

Received July 2025; revised November 2025

Abstract. *Belt tears can lead to equipment failure, fire, and other accidents. To address the diverse forms of belt tears and the interference caused by harsh working conditions on identification accuracy, this study introduces a lightweight target detection algorithm, DDC-YOLO11, to improve the robustness of tear detection in complex scenarios. Firstly, the Dilation-wise Residual and Dilated Reparam Block (DWB) module is embedded in the C3k2 module of the YOLO11 backbone network. It utilizes a complex feature extraction and fusion mechanism to enhance the model's ability to detect targets of different scales and locations in the image. Additionally, the attention module in the C2SPA module is replaced with a variable self-attention mechanism (DAttention), which boosts the model's capacity to extract image features through adaptive adjustment of the attention area. Furthermore, the Context Anchor Attention–High-level Screening–Feature Fusion Pyramid Network (CAA-HSFPN) is adopted in the neck network. It uses adaptive pooling technology to reduce the feature map dimension and significantly decrease the number of parameters. Images acquired from the production site are preprocessed using the Adaptive Enhancement method (AdaEnhance) to improve recognition performance. Experimental results show that the DDC-YOLO11 model achieves 92.5% mAP50, with 4.2M weight parameters, 1.957 × 10^6 total parameters, and a detection speed of 100.8 FPS.*

Keywords: Belt tearing, Multiple tear types, DDC-YOLO11, Feature fusion, Attention mechanism, Complex scenarios, Lightweight algorithm

1. **Introduction.** The conveyor belt is a key component of the coal-handling system in power plants [1]. Its primary function is to continuously and stably transport coal from the coal storage yard (or other sources) to the boiler's raw-coal hopper, thereby ensuring an uninterrupted fuel supply for power generation. The belt's performance directly affects the operating efficiency and energy consumption of the coal-handling system, and its reliable operation is vital to the safe production of the plant. Once tears or other faults occur, serious safety hazards can arise during normal plant operation [2,3].

To enable timely detection of belt tears, the academic community has leveraged the high accuracy and efficiency of machine vision, leading to its deep integration into industrial tear detection scenarios as the mainstream approach [4,5]. In [6], the authors present a detection method using an auxiliary classifier with spectral normalization, which effectively

DOI: 10.24507/ijicic.22.03.653

identifies longitudinal tears through Wasserstein distance and spectral normalization strategies. [7] introduces an Adaptive Deep Convolutional Network (ADCN) that extracts multi-scale features from visible-light images of belt damage. Additionally, [8] proposes a lightweight damage detection method by integrating MobileNet with YOLOv4, significantly improving computational efficiency.

The method above effectively extracts target features from longitudinal tearing dataset samples; real-world belt operating environments present additional challenges. These include diverse tear types, significant dust accumulation, and bright light interference, especially in open-air settings. These factors complicate the accurate identification of belt tears. To address these issues, several studies have proposed innovative solutions. For instance, [9] tackles the problem of image degradation during belt conveyor operation by employing Wiener filtering to restore degraded images. It also uses the CamShift algorithm to track and capture fast-moving belt crack sequences and applies the Canny operator for robust edge detection of belt tears. [10] utilizes an enhanced Cutmix technique for dataset enrichment and adopts the CSPDarknet53 deep convolutional network to extract and integrate multi-scale tear features. [11] introduces an Improved Grayscale-Gravity Centroid Method (IGGM) to extract the centerline of a laser line, using this feature to detect longitudinal tears. [12] introduces a computer vision detection method utilizing multiple lasers, enabling precise segmentation of laser stripe regions and identification of tear areas on conveyor belts via sophisticated image processing techniques. [13] by integrating the Haar function with a modified dark channel fog removal algorithm, fog interference in belt images can be effectively eliminated. Subsequently, the classifier is trained and enhanced using the AdaBoost algorithm and further optimized through cascading with the Cascade algorithm to improve recognition performance.

While the methods above address environmental interferences, they overlook the complexity of diverse tear types in conveyor belts [14,15]. In [16], a lightweight dual-layer partial-convolution module was embedded in both the encoder and the decoder of DFPA-Net, improving its accuracy in detecting cracks with various geometries; however, the impact of adverse environmental conditions on tear-image acquisition was not considered. Most existing studies do not jointly account for environmental degradation and the diversity of tear types. Moreover, coal-conveyor belts operate outdoors year-round; consequently, they suffer not only from variable illumination and dust, but also from extreme weather such as rain, snow, and fog. Although the cited methods yield promising results, their reliance on expensive laser devices and their degraded performance in strong sunlight remain open issues. Hence, a cost-effective solution that remains robust under harsh outdoor conditions is urgently needed.

Focusing on conveyor belts in coal-handling systems of power plants, this work proposes DDC-YOLO11 – a novel tear-detection algorithm designed to handle multiple tear types under realistic operating conditions. First, tear images are acquired from both industrial sites and a controlled test rig. Next, the AdaEnhance preprocessing module augments the data by simulating rain, snow, and noise artefacts. Finally, the proposed DDC-YOLO11 network extracts discriminative features for tear localization. The main contributions of this paper are summarized below.

1) The AdaEnhance image enhancement algorithm is introduced to simulate the operating environment of the power plant belt. This enriches the dataset's image information and reduces the risk of overfitting.

2) Incorporation of the DWB module into the C3k2 backbone architecture enhances multi-scale detection performance via refined feature extraction and fusion capabilities.

3) The C2SPA module incorporates a dynamic attention mechanism that enables the model to adaptively adjust the positions of sampling points using learnable offsets, based

on the input tear data. This feature enhances the model's ability to accommodate variations in the target's shape and orientation.

4) To improve the neck network's performance, a novel feature pyramid structure called CAA-HSFPN is implemented. This architecture enhances semantic representation by selectively integrating multi-level features, thereby improving the detection of fine details in belt-tear images. Additionally, through optimized feature map compression and parameter reduction, the model remains computationally efficient while maintaining high accuracy.

The enhancements to the backbone and neck architectures, along with optimized image preprocessing, enable the target detection algorithm to overcome previous limitations in accuracy and class recognition for conveyor belt imagery. This approach exhibits consistently better performance than alternative methods.

The paper is structured as follows: Section 1 is the introduction, which presents the research background and limitations of existing methods; Section 2 elaborates on the principles and implementation of the AdaEnhance image enhancement algorithm; Section 3 proposes the belt tear detection algorithm based on DDC-YOLO11, including the design of the DWB module, deformable self-attention mechanism (DAttention), and the CAA-HSFPN feature pyramid structure; Section 4 validates the performance of the proposed method through ablation experiments and comparative experiments; Section 5 provides a conclusive summary of the study. Finally, potential directions for further investigation are outlined.

2. AdaEnhance Image Enhancement Algorithm.

During model training, the size of the dataset plays a crucial role in determining the model's performance and generalization ability. However, limited dataset size often becomes a key factor restricting model accuracy. Moreover, obtaining diverse environmental image data is essential for robust training. Image enhancement techniques serve as an effective solution to address these challenges. They can grow the dataset and strengthen the model's ability to generalize, thereby optimizing model performance even when sample sizes are limited.

Several classical image processing techniques – including Histogram Equalization (HE) [17], wavelet transform [18], partial differential equation methods [19], and Retinex algorithm [20] – have demonstrated limitations when applied to belt tear detection. These conventional approaches often fail to adequately address the diverse enhancement requirements and struggle to accurately represent the complex visual characteristics of belt surfaces in challenging industrial environments. To overcome these constraints, our research introduces an Adaptive Enhancement (AdaEnhance) technique specifically designed to improve belt tear detection performance by addressing both dataset scarcity and environmental complexity issues.

The AdaEnhance algorithm integrates diverse image processing techniques, including affine transformations, brightness and contrast adjustment, perspective transformations, convert to grayscale and effects that simulate natural environmental conditions. This integrated approach aims to enrich the diversity of image datasets through the reproduction of real-world variations, and the following enhancement processing formulas are at its core:

1) Affine transformations

Scale:

$$I' = scale \times I \tag{1}$$

Translate:

$$I'(x, y) = I(x + \Delta x, y + \Delta y) \tag{2}$$

Rotate:

$$I'(x,y) = I(x\cos\theta - y\sin\theta, x\sin\theta + y\sin\theta) \tag{3}$$

Shear:

$$I'(x,y) = I(x + y \cdot shear, y) \tag{4}$$

where *scale* is the scaling factor, Δx and Δy are the translations along the width and height, θ is the angle of rotation, and *shear* is the shear strength.

2) Brightness and contrast

$$\text{Brightness adjustment: } I' = I + \Delta brightness \tag{5}$$

$$\text{Contrast adjustment: } I' = \gamma + \Delta contrast \tag{6}$$

where $\Delta brightness$ is the amount of luminance change and γ is the contrast factor. $\Delta constrast$ is the contrast factor.

3) Convert to grayscale

Convert RGB image to grayscale image:

$$I = 0.299 \times R + 0.597 \times G + 0.114 \times B \tag{7}$$

4) Perspective transformation

The perspective transform simulates a perspective effect by randomly selecting four points. The formula is

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{8}$$

(x, y) represents the coordinates of the original image point, while (x', y') denotes the coordinates of the transformed point. a controls scaling and rotation in the x-direction to adjust width; b represents the effect of the y-direction on the x-coordinate, controlling horizontal skew; c is the translation in the x-direction, controlling the horizontal position; d represents the effect of the x-direction on the y-coordinate, controlling vertical skew; e is the scaling and rotation in the y-direction, controlling height changes; f is the translation in the y-direction, controlling the vertical position; g is the perspective parameter in the x-direction, controlling the horizontal vanishing point; h is the y-direction perspective parameter, used to control the vertical vanishing point.

5) Gaussian noise

$$I(x,y) = I(x,y) + N(0,\sigma) \tag{9}$$

The image is augmented with Gaussian noise that has a mean of zero and a variance of σ.

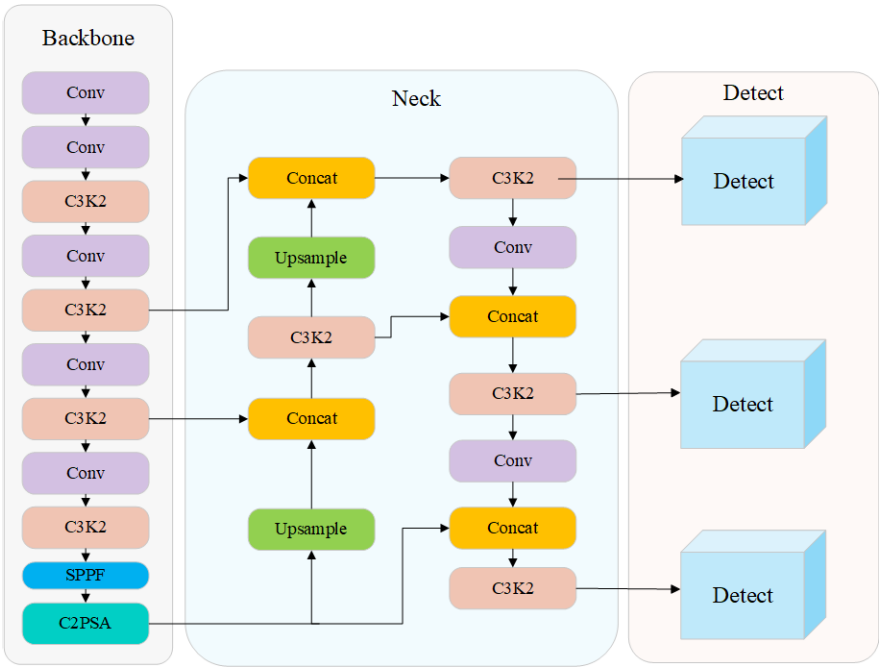
3. **Tear Detection Algorithm.**

3.1. **Fundamentals of the YOLO11 algorithm.**

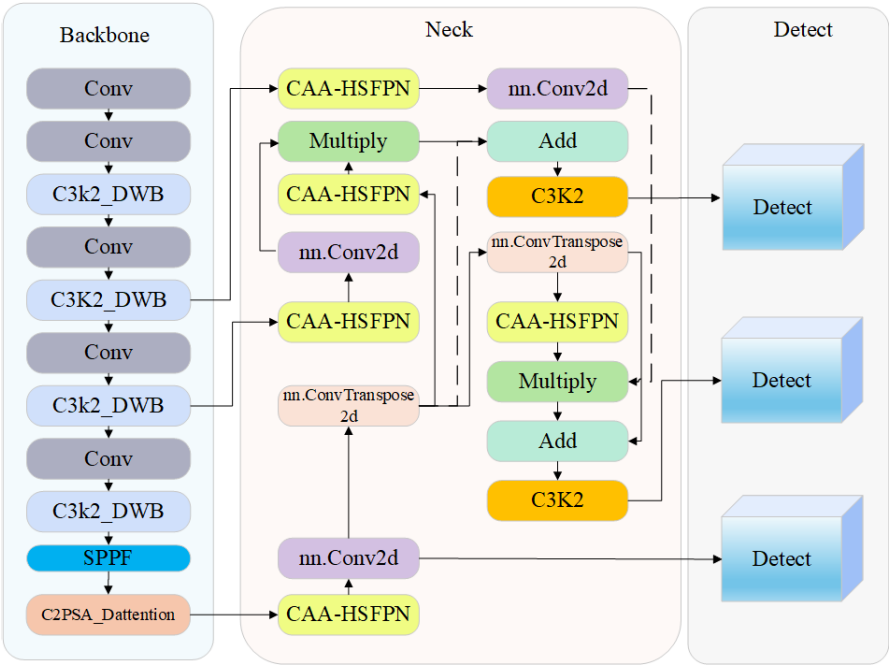
YOLO11 comprises three main components: Backbone, Neck, and Head. It offers faster, more accurate, and more efficient detection capabilities in target detection technology. YOLO11 introduces several innovative components, including the C3k2 (Cross Stage Partial with kernel size 2) module, the SPPF (Spatial Pyramid Pooling – Fast) module, and the C2PSA (Convolutional block with Parallel Spatial Attention) component. These innovations significantly enhance feature extraction and overall model performance. The structure of YOLO11 is depicted in Figure 1(a).

Initially, YOLO11 receives an input image and partitions it into a 640×640 grid for feature extraction. Utilizing a series of convolutional layers and residual concatenations, it extracts features associated with the positions and shapes of target objects within the image. YOLO11 then predicts bounding boxes and a confidence score for each grid cell,

where the confidence score represents the likelihood that a bounding box contains a target object. Additionally, each bounding box predicts a set of category probabilities, representing the likelihood that the target object belongs to each predefined class. Since multiple grid cells can detect the same object, YOLO11 applies Non-Maximum Suppression (NMS) to eliminating redundant bounding boxes, thereby reducing redundancy and improving detection accuracy. Finally, YOLO11 outputs a set of bounding boxes, each containing the target's category, location, and confidence score, which can be directly applied in subsequent tasks.



(a) Illustration of the YOLO11 model structure



(b) Illustration of the DDC-YOLO11 model structure

FIGURE 1. Comparison of YOLO11 and DDC-YOLO11 model structures

3.2. **DDC-YOLO11 algorithm.** Despite YOLO11's high detection accuracy in target detection, further improvements in its design are still possible [21-23]. Here, YOLO11n, the lightest YOLO11 version, is chosen as the basic framework. It retains high accuracy with minimal parameters and computation, making it perfect for resource-limited settings. To boost its belt tear detection performance, this paper optimizes YOLO11n into the DDC-YOLO11 algorithm, with the optimized structure shown in Figure 1(b).

To address the challenges of identifying belt tear targets in harsh operating environments – such as interference from dust, rain, and snow, as well as diverse belt tear types – the backbone network integrates the DWB module into C3k2. Combined with standard and dilated convolutional layers, this integration improves feature diversity and richness. It enhances the model's segmentation performance on instances and effectively reduces external environmental interference in identifying real belt tear targets.

For different tear sizes and belt types, the attention mechanism in C2PSA is replaced with a Dynamic Attention mechanism (DAttention). This mechanism dynamically acquires sampling point locations based on varying tear sizes and shapes, thereby improving belt tear detection accuracy.

To tackle the issues of many model parameters and the interference from scratches and scuffs caused by coal on the belt, a lightweight high-level feature pyramid module (CAA-HSFPN) is proposed. This module combines context anchor attention and spatial attention mechanisms to enhance the model's ability to recognize features from different regions of the image. As a result, it can more efficiently capture the exact location and contour of belt tears.

3.2.1. *Dilation-wise residual and dilated reparam block.* The backbone network of YOLO11 primarily relies on standard convolution when handling objects of varying sizes. This limits its ability to capture contextual information for multi-sized objects, especially in complex scenes like belt tears. This limitation also constrains its ability to express features effectively, leading to an excessive number of model parameters and high computational demands, which pose significant challenges for real-time belt tear detection. To tackle these challenges, this paper proposes a lightweight multi-scale convolutional method, named DWB (Dilation-wise Residual and Dilated Reparam Block) [24], and it is employed to substitute the residual block in the C3k2 module of the backbone network. The improved structure of the C3k2 module incorporating DWB is depicted in Figure 2(a).

DWB consists of two primary elements: DWR (Dilation-Wise Residual) and DRB (Dilated Reparam Block). Figure 2(b) visualizes the DWR module architecture. First, a 3×3 standard convolutional layer reduces the number of channels of the feature map from dim to dim/2, lowering the dimensionality and computational load while extracting local features. The output of the 3×3 convolution is then fed into three branches.

The first branch uses a standard convolutional layer to restore the feature map's channel count to dim, maintaining spatial resolution with a stride of 1.

The second and third branches utilize dilated convolution layers (DRB-3 and DRB-5) with dilation rates of 3 and 5, respectively. These layers counteract context loss induced by harsh environmental conditions. Dilated convolutions enlarge the receptive field to capture broader context without increasing parameters. This approach enhances the model's capability to extract broader contextual features.

By integrating these components, DWB effectively balances computational efficiency and feature extraction capability, rendering it highly appropriate for real-time belt tear detection tasks.

Among them, the DRB (Dilated Reparam Block) module effectively addresses the issue of missing belt-tearing feature information caused by external environmental interference. Initially, the input feature map undergoes multi-level feature extraction and transformation through multiple convolutional layers. Each convolutional layer learns the feature representation at its respective level. Subsequently, multiple connection layers repeatedly fuse feature information from different convolutional layers. This process enhances feature diversity and richness, thereby improving the model's understanding of the input data and its segmentation performance on instances. The structure of the DRB module is shown in Figure 2(c).

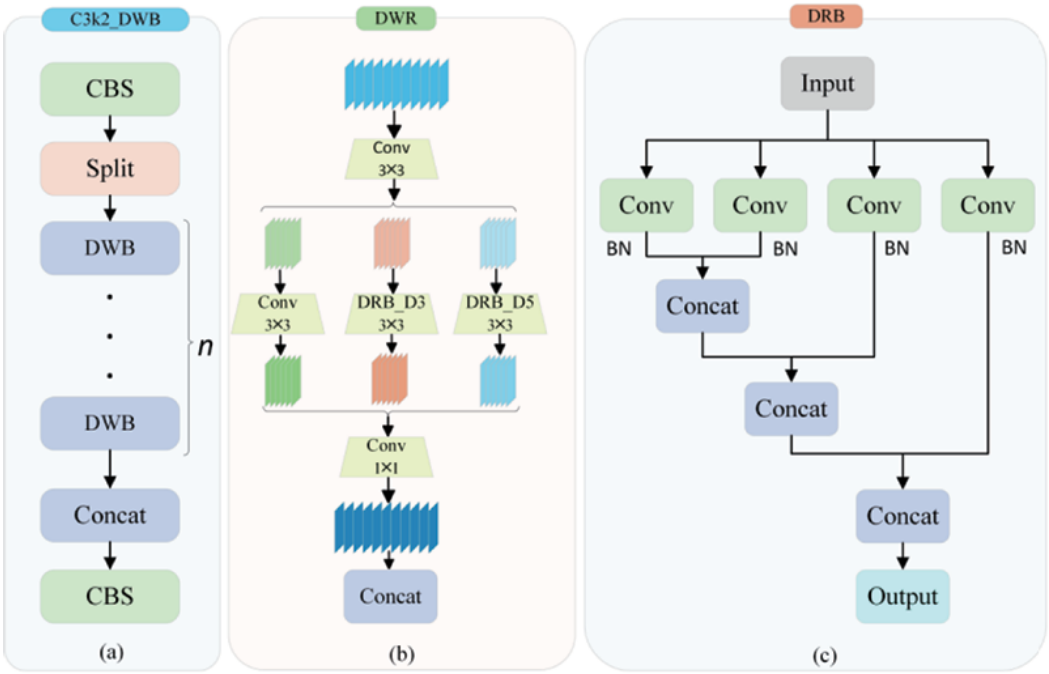


FIGURE 2. Overall structure of the C3k2 module

The feature maps produced by the three branches are concatenated along the channel dimension to create a single, consolidated representation. This fusion integrates features from diverse scales and receptive fields, thereby improving the model's capability to detect objects of various sizes. A 1×1 convolutional layer then refines the merged feature map, which adjusts the number of channels from dim/2 back to dim. This operation compresses the dimensionality of the feature map, preparing it for subsequent stages. The processed features are then fused with the initial input through a residual connection. This architecture alleviates the vanishing gradient problem during the training of deep networks and facilitates identity mapping, thereby strengthening the model's overall representational power.

DWB boosts multi-scale, multi-location detection through precise feature extraction and fusion, yielding richer representations that jointly elevate accuracy and robustness.

3.2.2. *Deformable self-attention mechanism.* In the C2SPA module, the standard self-attention mechanism relies on fixed sampling points. However, when dealing with belt-torn targets that exhibit complex shapes and varying attitudes, these fixed sampling points often fail to accurately capture the key features of the target. For objects with different perspectives, fixed sampling may overlook important parts, resulting in incomplete feature extraction.

To overcome this limitation, the C2SPA module integrates the Deformable Attention mechanism (DAttention) [25]. DAttention utilizes learnable offsets to dynamically adjust the positions of sampling points in response to the input data. This adaptive strategy allows the model to more effectively handle variations in the target's shape and pose, capture long-range dependencies, and achieve superior feature extraction capabilities compared to the conventional self-attention mechanism. The structure of the module is illustrated in Figure 3.

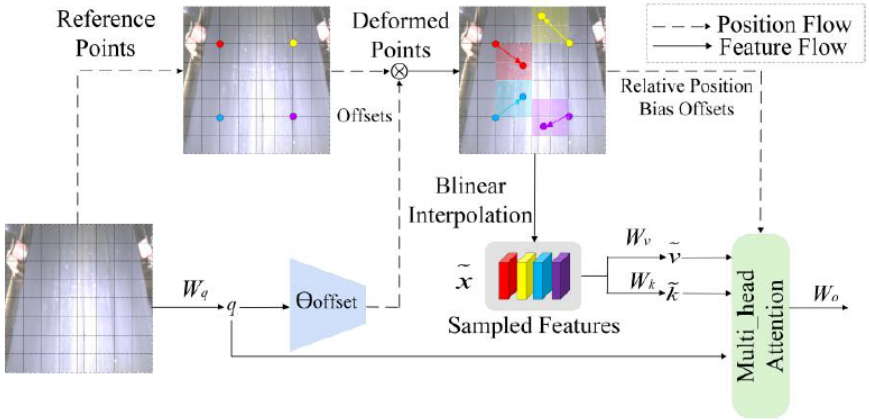


FIGURE 3. DAttention module detection schematic

As shown in Figure 3, given a feature map $x \in R^{H\times W\times C}$, first generate a uniform grid composed of $p \in R^{HG\times WG\times 2}$ as reference points.

$$HG = \frac{H}{r}, \quad WG = \frac{W}{r} \tag{10}$$

where r is the downsampling factor. The values of the reference points are normalized to the range $[-1, 1]$ based on the grid shape $HG \times WG$, which is used for subsequent attention localization calculations. Subsequently, the displacement of each reference point is obtained by projecting the feature map into query space via the learned matrix W_q, yielding

$$q = xW_q \tag{11}$$

where q is the query marker. Then, a specific offset θ_{offset} is used to generate the offset Δp in the network. These offsets dynamically adjust the position of the reference point to focus on more critical feature regions. To prevent model instability caused by excessive offsets, a fixed scale factor 's' modulates offset amplitudes, i.e.,

$$\Delta p = s \cdot \tanh(\Delta p) \tag{12}$$

After obtaining the deformation points, feature sampling is performed at the corresponding positions on the original feature map x using the bilinear interpolation method to obtain the corresponding Key and Value vectors. The formula is as follows:

$$q = xW_q,\ \tilde{k} = \tilde{x}W_k,\ \tilde{v} = \tilde{x}W_v,\ \Delta p = \theta_{offset} \tag{13}$$

$$\tilde{x} = \varphi(x, p + \Delta p) \tag{14}$$

Among them, $\tilde{k}$ and $\tilde{v}$ obtain the corresponding Token Key and Value through interpolation based on the reference point after movement, and $\varphi(;)$ is bilinear interpolation. Finally, multi-head attention is performed on the obtained q, $\tilde{k}$, and $\tilde{v}$, and the corresponding position offset R is used, and then a linear projection layer W_o is used to obtain the final output z. The output formula of the attention head is

$$z^m = \sigma\left(q^{(m)}\tilde{k}^{mT}\Big/\sqrt{d} + \varphi(B;R)\right)\tilde{v}^m \tag{15}$$

where z^m is the embedded output calculated by the mth attention head, d is the feature dimension that each head should process, and $\sigma()$ is a Softmax function.

In summary, the DAttention module generates offsets via deformable convolution, enabling it to dynamically adjust the region of attention and better adapt to local feature changes in the image. Meanwhile, positional encoding and residual concatenation are incorporated into the module to enhance its capacity for capturing spatial details. Additionally, grouped convolution and bilinear interpolation are employed to reduce computational complexity and improve efficiency significantly.

3.2.3. *CAA-HSFPN space pyramid structure.*

3.2.3.1. *HSFPN model.* HS-FPN (High-level Screening-feature Fusion Pyramid Networks) [26] is primarily designed to address multi-scale challenges in target detection. Its fundamentals include two key components: the feature selection module and the feature fusion module.

1) Feature Selection Module: By integrating cross-channel attention with dimensional adaptation, the module dynamically filters multi-scale spatial features through parallel global pooling pathways, generating optimized channel weights that enhance informative feature selection without compromising processing speed.

2) Feature Fusion Module: Through an intelligent fusion process, this module combines processed low-level features with upsampled high-level features (using bilinear interpolation or transposed convolution), boosting the model's performance in identifying belt tear patterns.

3.2.3.2. *Context anchor attention model.* For belt tear detection, precise localization and morphological characterization of defects are essential. However, real-world scenarios frequently encounter difficulties including multi-scale targets, uneven lighting conditions, and partial obstructions. To overcome these limitations, we developed the Context Anchor Attention (CAA) mechanism, which combines channel-wise and spatial attention principles to replace conventional channel attention in the Hierarchical Selective Feature Pyramid Network (HSFPN). The proposed CAA module significantly improves regional feature discrimination, enabling more accurate tear localization and boundary delineation. Additionally, the enhanced architecture demonstrates improved adaptability to viewpoint changes and environmental variations. By effectively combining tear-specific features with contextual information, it strengthens the feature selection and integration performance of HSFPN, leading to superior detection precision. Unlike conventional channel attention or computationally intensive attention variants, our anchor attention employs an efficient architecture that simplifies network integration while significantly reducing both model complexity and computational requirements. These advantages make the solution particularly suitable for industrial belt tear inspection systems and enable practical field deployment.

The architecture of the Contextual Anchor Attention (CAA) module is depicted in Figure 4. Initially, the feature map undergoes processing via an average pooling layer (avg_pool), which serves to decrease spatial dimensionality and extract local features.

$$F^{pool} = Conv_{1\times1}\left(P_{avg}\left(X^{(1)}\right)\right) \tag{16}$$

where P_{avg} denotes the average pooling operation and the output features are F^{pool}. Two deep strip convolutions are then applied as an approximation to the standard large kernel deep convolution:

$$F^w = DWConv_{1\times k_b}\left(F^{pool}\right) \tag{17}$$

$$F^h = DWConv_{k_{b\times 1}}(F^w) \tag{18}$$

F^w is the width output feature and F^h is the height output feature.

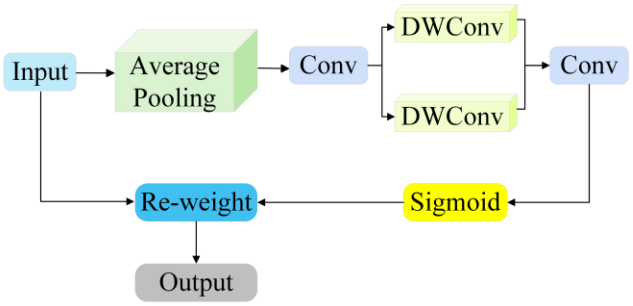


FIGURE 4. Schematic diagram of CAA structure

The deep strip convolution is selected based on two primary considerations. First, strip convolution is highly lightweight. Compared to conventional 2D depthwise convolution, CAA can achieve similar performance using only a few 1D depthwise kernels while reducing the number of parameters by half. Second, strip convolution is particularly effective for recognizing and extracting features from objects with elongated shapes, such as belt tears. It enhances the ability of advanced feature pyramids to establish relationships between distant pixels, all without significantly increasing computational costs due to its strip-based design.

Ultimately, the CAA module produces an attention weight A, which is subsequently employed to further refine the output of the HSFPN module.

$$A = Sigmoid\left(Conv_{1\times 1}\left(F^h\right)\right) \tag{19}$$

$$F^{attn} = (A \odot P) \oplus P \tag{20}$$

The Sigmoid function ensures that the attention graphs A is in the range $(0, 1)$, $P \in R^{\frac{C_l}{2}\times H_l\times W_l}$ denote the output features, $\odot$ denotes the multiplicity of assignments in the elemental directions, $\oplus$ denotes the sum of the elemental directions, and $F^{attn} \in R^{\frac{C_l}{2}\times H_l\times W_l}$ is the augmented feature. The output of the nth feature at stage l is obtained by the following equation:

$$X = Conv_{1\times 1}(F^{attn}) \tag{21}$$

X denotes the output of the last feature.

4. Experimental Outcomes.

4.1. **Setup for experimentation.** The experimental setup is illustrated in Figure 5. The belt used in the experiment is an NN-300(L) nylon cord core rubber belt with a tensile strength of 1200 N/mm, a total length of 253 m, a width of 50 cm, and a thickness of 8 mm. The cover rubber has a thickness of 1.5 mm and a tensile strength of 15 MPa. A 1330M-A-I1 camera is employed for image acquisition, with a resolution of 2304×1296 pixels.

The experiments were performed on a computer with Windows 11. The hardware setup features an AMD Ryzen 7 7840H processor equipped with integrated Radeon 780M Graphics, along with an NVIDIA GeForce RTX 4060 GPU that has 16GB of dedicated video memory. The software environment includes PyTorch 2.1.0, CUDA 11.8, and Python 3.8.

FIGURE 5. Experiment building platform

The model is trained on images with a resolution of 640 × 640 pixels, using an initial learning rate of 0.01 and Stochastic Gradient Descent (SGD) as the optimizer. A batch size of 16 is applied throughout 300 epochs of training. The model is initialized with pre-trained weights from YOLO11n.pt. The testing procedure is presented in Figure 6.

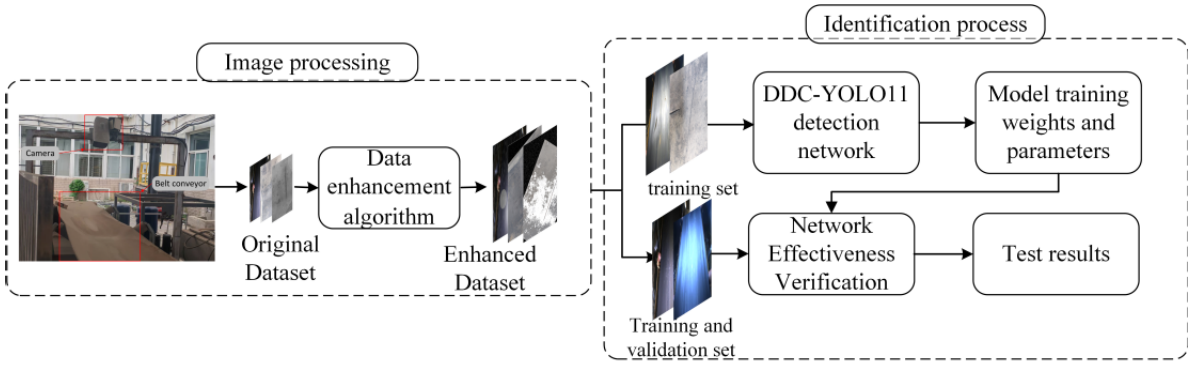
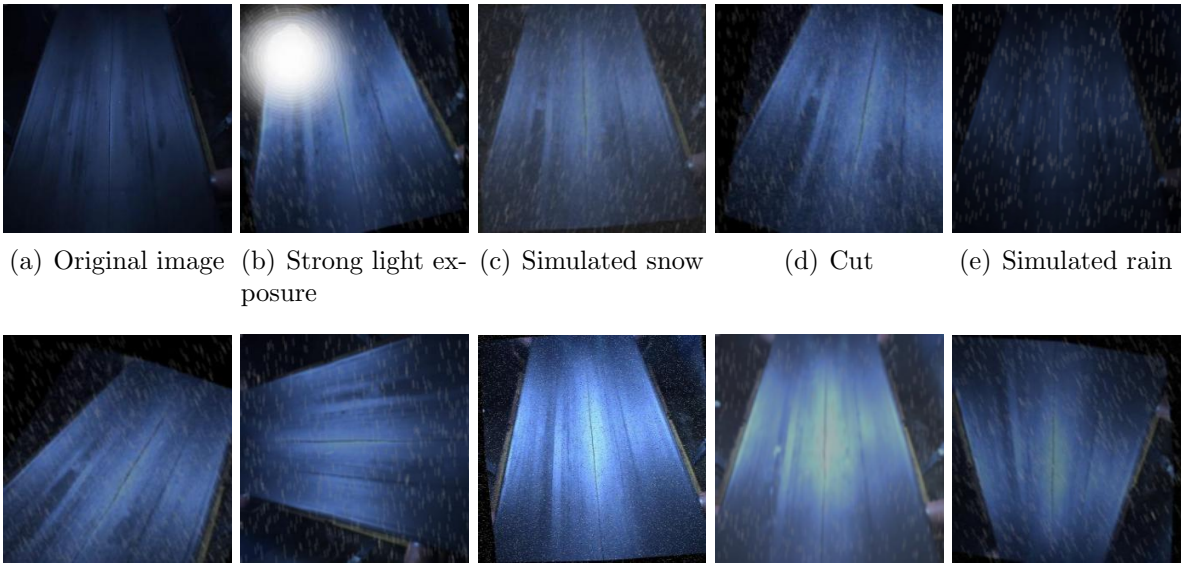


FIGURE 6. Test flow chart

4.2. **Dataset generation and preprocessing.** The dataset used in this study comprises three distinct components: actual images obtained from the Yunnan Guoneng Yangzonghai Power Plant, simulated images generated by the experimental platform, and images sourced from online repositories. The laboratory experimental bench was located outdoors and designed to replicate the open-air environment of the power plant, eliminating the need for additional lighting equipment. In total, 334 images of scratches and tears were collected. However, given the scarcity of belt tear samples, the dataset's size and quality had a direct influence on the deep learning model's training performance, thereby increasing the likelihood of overfitting.

The dataset was augmented and the belt's imaging effects in different environments were simulated using the AdaEnhance algorithm. This expanded the dataset of belt tear images from 334 to 1200. (Figure 7 shows examples of processed images, where (a) is the original image and the rest are images enhanced by the AdaEnhance algorithm). This enhancement improved the model's capability to precisely locate the target in the image and efficiently conduct feature extraction.

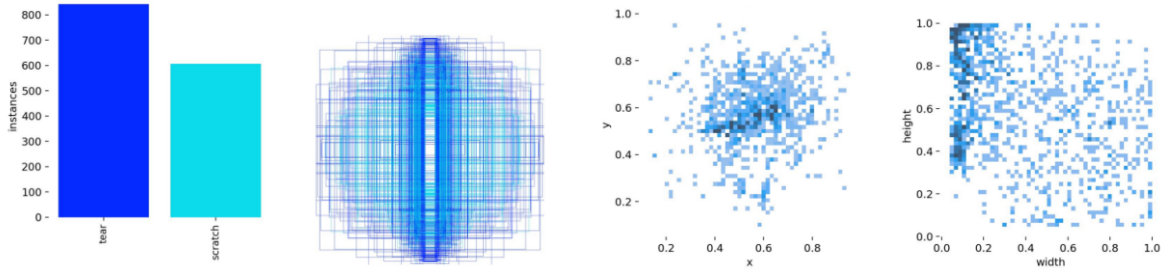
Considering that minor scratches and scuffs on the belt, which do not significantly affect its normal operation, may interfere with the model's accurate identification of tears, we categorized these minor imperfections uniformly as "scratch" while reserving the term "tear" for actual belt tears. The dataset was divided into training, validation, and test sets in an 8 : 1 : 1 ratio. The labeling statistics of the dataset are shown in Figure



(a) Original image (b) Strong light exposure (c) Simulated snow (d) Cut (e) Simulated rain

(f) Affine transformation (g) Rotating operation (h) Gaussian noise (i) Brightness enhancement (j) Turn upside-down

FIGURE 7. Diagram of the effect of dataset enhancement

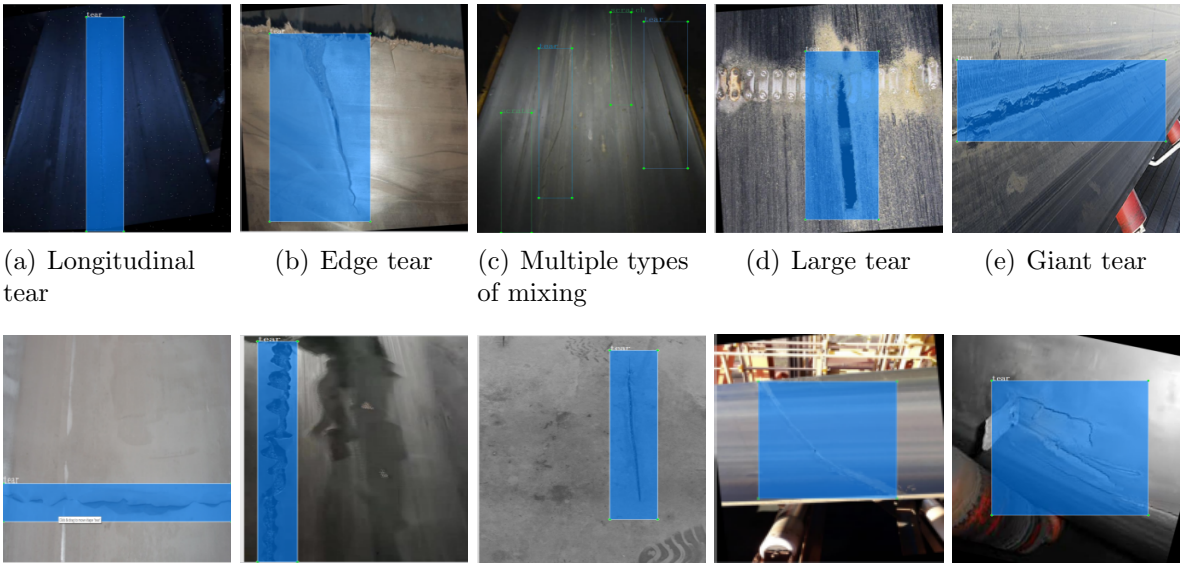


(a) (b) (c) (d)

FIGURE 8. (a) Statistics on the number of categories in the dataset; (b) statistics on the size of annotated boxes in the dataset; (c) statistics on the position of the center point of annotated boxes in the image; (d) statistics on the width-to-height ratio of annotated boxes relative to the entire image

8. In the dataset used in this paper, each image has been precisely annotated. From the category distribution shown in 8(a), the number of tear samples is slightly higher than that of scratch samples, reflecting the issue of sample imbalance in real-world belt operation scenarios. In 8(b), the positions and sizes of the annotated boxes in the images are relatively dispersed, indicating that the locations of belt tears are highly random, which closely aligns with the real-world conditions of belt tears in industrial environments and effectively tests the model's generalization ability. In 8(c), the distribution analysis of the annotations shows that the center points of the annotated boxes are relatively dispersed. In 8(d), the positions and sizes of the annotations are also relatively dispersed, fully demonstrating the superiority of the data augmentation techniques employed in this study.

Utilize the LabelImg tool to annotate tear images of various types, enabling the model to learn tear characteristics comprehensively and thereby achieve higher detection accuracy. The annotated images are presented in Figure 9.



(a) Longitudinal tear (b) Edge tear (c) Multiple types of mixing (d) Large tear (e) Giant tear

(f) Warped tear (g) Zigzag tear (h) Small laceration (i) Horizontal tear (j) Extensive wear and tear

FIGURE 9. Labeling annotation schematic

4.3. **Performance metrics.** We evaluate the model using Precision, Recall, mAP, parameter count, model size, and inference speed (FPS). Specifically, mAP is calculated as the average of mAP@0.50 (%), representing the mean precision AP across all categories when the Intersection over Union (IoU) threshold is set at 0.5. Since mAP is derived from the Precision-Recall curve, it provides a comprehensive measure of the algorithm's accuracy and robustness. The number of parameters is one of the core metrics for measuring model size and complexity. The greater the number of parameters, the larger the model file and the higher the storage requirements. Furthermore, parameter count directly governs the computational cost of forward propagation, which in turn impacts inference speed. FPS (Frames Per Second) is a core metric for measuring model real-time performance [27]. The formulas for calculating detection accuracy (P), recall (R), and mAP are as follows:

$$P = \frac{TP}{TP + FP} \tag{22}$$

$$R = \frac{TP}{TP + FN} \tag{23}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{24}$$

where TP is the number of positive samples predicted correctly; FN is the number of positive samples predicted incorrectly; FP is the number of negative samples predicted incorrectly; AP_i is the average precision of category i; n is the total number of categories.

These parameters evaluate the model from three dimensions: detection accuracy, processing speed, and model complexity, enabling a reasonable and accurate assessment of the model's detection performance.

4.4. **Experimentation and analytical evaluation.**

4.4.1. *Ablation experiment.* Ablation experiments were conducted to evaluate the individual and combined effects of the proposed modules, with results summarized in Table 1.

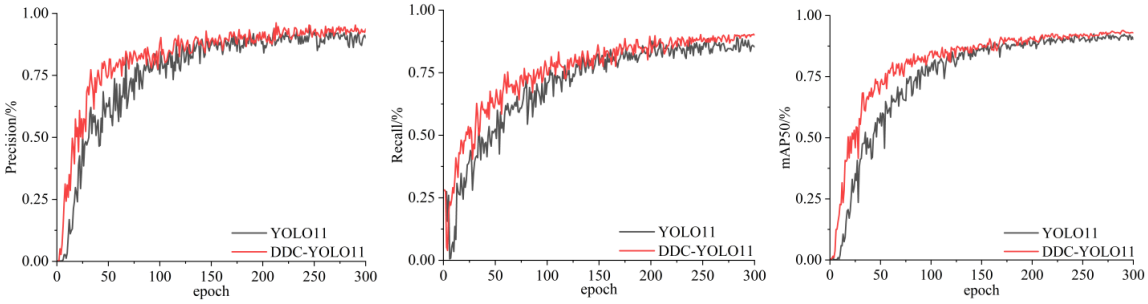
TABLE 1. Ablation experiments

Model	YOLO11	DWR	DWB	CAA-HSFPN	DAttention	Precision /%	mAP50 /%	Recall /%	Size (M)	Params /10^6	FPS
1	√					90.1	89.9	81.9	5.3	2.582	114.9
2	√	√				89	91.7	88.7	5.3	2.565	122.4
3	√		√			94	92.1	86.6	5.2	2.490	90.4
4	√		√	√		90.0	91.5	87.5	4.1	1.940	93.3
5	√		√	√	√	92.5	92.5	88.8	4.2	1.957	100.8

The outcomes of the experiment indicate that the three strategies have enhanced the model's overall performance in different ways. Before the DWB was fused with the DRB (i.e., when only the DWR was added in the table), the mAP50 and R were enhanced, but the impact on other parameters was relatively small. After the DWR and DRB were integrated into the DWB, all evaluation metrics of the model were improved.

To further optimize the model, the CAA-HSFPN was introduced, which achieved the goal of reducing the number of parameters. However, this had a negative impact on detection accuracy. To counteract the adverse effects caused by the reduction in parameters, the DAttention module was incorporated. By dynamically adjusting the detection of tear contours, it managed to restore the detection accuracy to its highest level with minimal impact on the number of parameters. Therefore, the coordinated interaction of the three modules is the most effective for the model.

To provide a more in-depth analysis of the performance variations during the model detection process, and to present the comparative results in a more intuitive and clear manner, this paper plots the change curves of the three key detection metrics – P, mAP, and R – for the YOLO11 and DDC-YOLO models during the detection process, as shown in Figure 10. As shown in the figure, the improved DDC-YOLO model not only achieves higher detection accuracy but also converges faster and exhibits a smoother curve. This indicates that the DDC-YOLO model has significantly enhanced its ability to identify belt tears in complex environments.



(a) Precision comparison curve (b) Recall comparison curve (c) mAP50 comparison curve

FIGURE 10. Dynamic progression of evaluation criteria

4.4.2. *Comparison experiment.* To fully demonstrate the inherent advantages of the YOLO11 model and enable it to perform at its best in the belt tear detection task, this paper designs a series of comparative experiments. Not only are performance comparisons made with other similar models, but also various improvement methods for the YOLO11 model itself are explored. The specific experimental settings are shown in Table 2.

Table 2. Comparative analysis of model training performance

Model	R/%	mAP50/%	mAP50-95/%	Size (M)	Params/10^6
DDC-YOLO11	88.8	92.5	59	4.2	1.9
YOLO8n.pt	80.7	89.2	55.1	5.9	3.0
YOLO10n.pt	83	89.3	56.6	5.5	2.6
YOLO12n.pt	84.9	89.6	57.8	5.5	2.6
SDC-YOLO11	86.4	86.4	54.6	25.2	13.1
FocalModulation-YOLO11	86	89.7	56	4.8	2.3
EMSC-YOLO11	89.2	92.3	57.2	5.2	2.5
FDPN-YOLO11	89.5	93	60	5.6	2.7
BIMAFPN-YOLO11	84.9	92.5	57.3	4.8	2.3
FeaturePyramid-SharedConv-YOLO11	85	93.1	58.2	5.2	2.5

Table 2 shows that our method surpasses YOLOv8, YOLOv10 and even the newest YOLOv12 in both detection accuracy and parameter efficiency.

Based on the method described in [28], the SDC-YOLO11 model incorporates the improved modules DWB and CAA-HSFPN introduced in this paper, along with a multi-scale attention (SEAM) mechanism detection head. This mechanism aggregates information from each channel, enhancing the connectivity between channels. The results show that the addition of the detection head has a negative impact on all metrics.

Drawing from [29], this variant substitutes YOLO11's feature pyramid with a FocalModulation network. Despite achieving parameter reduction, the configuration suffers from considerable accuracy degradation, rendering it unsuitable for our application requirements.

In accordance with the method detailed in [30], the EMSC-YOLO11 model integrates a multi-scale convolutional EMSC (Efficient Multi-Scale Convolutional) module into the C3k2 component. The results show that this improvement has little beneficial effect on the model and still does not meet the requirements of this paper.

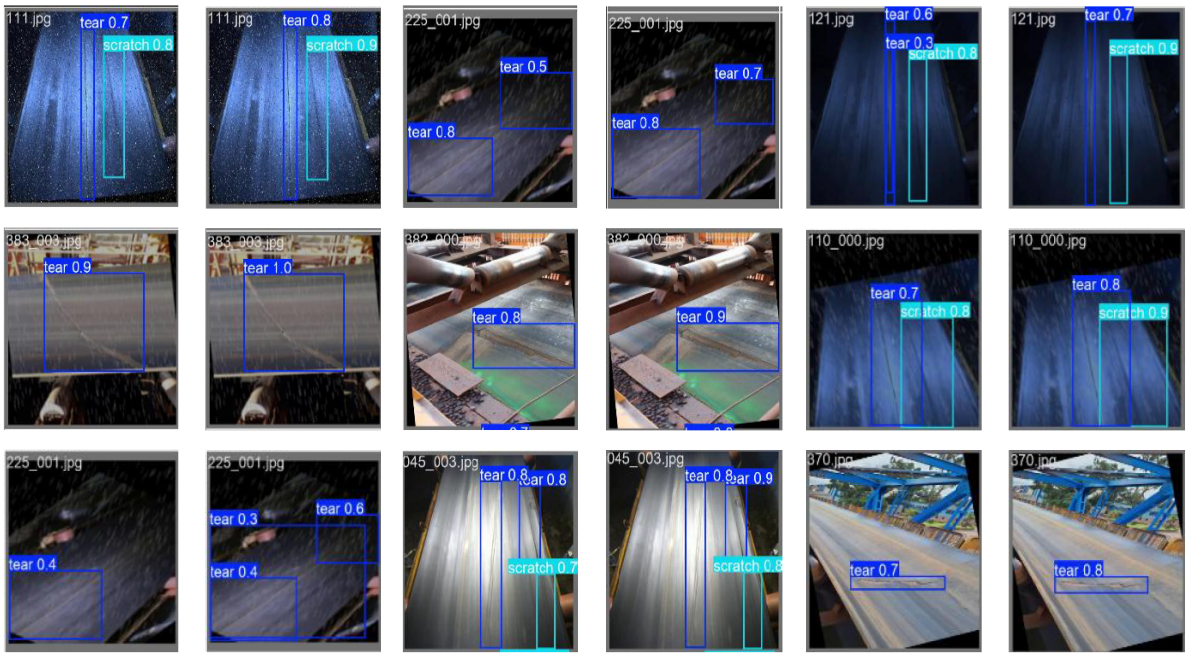
Drawing on the method outlined in [31], the FDPN-YOLO11 model incorporates a Feature Focusing Diffusion Pyramid Network (FDPN) into YOLO11. The model enriches each feature scale with detailed contextual information. However, the results indicate that the parameter count has increased.

Drawing on the work presented in [32], the BIMAFPN-YOLO11 model incorporates a Weighted Bidirectional Feature Pyramid Network (BIMAFPNA). This network assigns varying weights to different input features based on their priority and iteratively employs this structure to strengthen feature fusion. The experimental results highlight significant differences in accuracy and parameter size compared to the model introduced in this paper.

Finally, the FeaturePyramidSharedConv-YOLO11 model replaces the spatial pyramid in YOLO11 with the multi-scale feature extraction FeaturePyramidSharedConv module. By using convolutional layers with different expansion rates, the module can extract features of different scales, which is conducive to capturing information of different sizes and contexts in images. The experimental results show that its performance metrics are inferior to those of the DDC-YOLO11 model.

In summary, the improved method described in this paper demonstrates significant superiority in terms of detection accuracy and lightweight design.

To visually highlight how the proposed method outperforms the baseline in detection accuracy on the same image, a comparison chart of detection accuracy was generated. Figure 11 presents a side-by-side comparison of two adjacent images. Specifically, the baseline model YOLO11's detection accuracy is illustrated in 11(a), 11(c), and 11(e), while 11(b), 11(d), and 11(f) display the results of the proposed improved method. As evident from the figure, the improved algorithm significantly enhances the detection accuracy of belt tears and effectively reduces the occurrence of missed detections.



(a) YOLO11 test results (b) DDC-YOLO11 test results (c) YOLO11 test results (d) DDC-YOLO11 test results (e) YOLO11 test results (f) DDC-YOLO11 test results

FIGURE 11. Comparison of YOLO11 model and DDC-YOLO11 model detection effect

5. **Conclusion.** This paper focuses on detecting various types of belt tears in harsh environments and proposes a lightweight object detection algorithm based on the DDC-YOLO11 model.

1) Propose the AdaEnhance image preprocessing method, which effectively simulates belt running images in harsh environments and solves the problem of insufficient data sets.

2) Improved optimization of the backbone and neck networks increases the detection accuracy of multiple tear types in harsh environments while reducing the number of parameters.

3) The experimental results demonstrate that DDC-YOLO11 achieves notable improvements in training precision without compromising computational efficiency, making it particularly suitable for real-world implementation of visual belt tear monitoring solutions. Future work will focus on expanding dataset diversity and enhancing detection performance while pursuing more compact model architectures.

Acknowledgment. This study is supported by the National Natural Science Foundation of China (Grant No. 52275109); the Natural Science Foundation of Hebei Province (Grant No. E2022502007).

REFERENCES

[1] L. Zhou, Y. Liu, X. Zeng, H. Yao and Z. Wu, Lightweight design of truss structure of pipe belt conveyor based on doe analysis, *International Journal of Innovative Computing, Information and Control*, vol.19, no.3, pp.959-971, 2023.

[2] W. Liu et al., YOLO-STOD: An industrial conveyor belt tear detection model based on Yolov5 algorithm, *Scientific Reports*, vol.15, no.1, 1659, 2025.

[3] X. Guo, X. Liu, H. Zhou et al., Belt tear detection for coal mining conveyors, *Micromachines*, vol.13, no.3, 449, 2022.

[4] Y. Wang et al., Longitudinal tear detection of conveyor belt based on improved YOLOv7, *IEEE Access*, vol.12, pp.24453-24464, 2024.

[5] Z. Wang and X. Li, A mining conveyor belt tear detection method based on improved STDC, *2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT)*, 2024.

[6] G. Wang et al., AC-SNGAN: Multi-class data augmentation for damage detection of conveyor belt surface using improved ACGAN, *Measurement*, vol.224, 113814, 2024.

[7] M. Zhang et al., A new paradigm for intelligent status detection of belt conveyors based on deep learning, *Measurement*, vol.213, 112735, 2023.

[8] M. Zhang et al., Application of lightweight convolutional neural network for damage detection of conveyor belt, *Applied Sciences*, vol.11, no.16, 7282, 2021.

[9] F. B. Wang, H. Sun et al., Edge-expanded belt tear support vector machine vision detection, *Chinese Mechanical Engineering*, vol.30, no.4, pp.455-460, 2019.

[10] G. Wang et al., A belt tearing detection method of YOLOv4-BELT for multi-source interference environment, *Measurement*, vol.189, 110469, 2022.

[11] Z. Lv et al., Visual detection method based on line lasers for the detection of longitudinal tears in conveyor belts, *Measurement*, vol.183, 109800, 2021.

[12] W. Li, C. Li and F. Yan, Research on belt tear detection algorithm based on multiple sets of laser line assistance, *Measurement*, vol.174, 109047, 2021.

[13] G. Wang et al., Machine vision-based conveyor belt tear detection in a harsh environment, *Measurement Science and Technology*, vol.33, no.8, 084006, 2022.

[14] X. Wang and X. Li, Causes of belt tearage in belt conveyors and preventive measures, *Baogang Science and Technology*, vol.51, no.3, pp.63-65, 2025.

[15] Z. Jin and Z. Wang, Research on belt tear detection technology for coal conveyor belts using a stereo camera, *China's New Technologies and New Products*, vol.4, pp.23-25, 2025.

[16] H. Zhang et al., DFPA-Net: A high-performance lightweight network for mining transport belt tear segmentation and degree prediction, *Knowledge-Based Systems*, 113908, 2025.

[17] S. Liu, Q. Lu and S. Dai, Adaptive histogram equalization framework based on new visual prior and optimization model, *Signal Processing: Image Communication*, vol.132, 117246, 2025.

[18] S. S. Parida, S. Bose and G. Apostolakis, Earthquake data augmentation using wavelet transform for training deep learning based surrogate models of nonlinear structures, *Structures*, vol.55, 2023.

[19] D.-K. Jang, K. Kim and H. H. Kim, Partitioned neural network approximation for partial differential equations enhanced with Lagrange multipliers and localized loss functions, *Computer Methods in Applied Mechanics and Engineering*, vol.429, 117168, 2024.

[20] L. Rong et al., Reconstruction efficiency enhancement of amplitude-type holograms by using single-scale retinex algorithm, *Optics and Lasers in Engineering*, vol.176, 108097, 2024.

[21] Y. Wang et al., Enhanced feature extraction with AL-YOLOv9s lightweight model: Application in key component recognition within highly integrated device environments, *Information Technology and Control*, vol.53, no.4, pp.1028-1041, 2024.

[22] V. Alamelu and S. Thilagamani, Lion based butterfly optimization with improved YOLO-v4 for heart disease prediction using IoMT, *Information Technology and Control*, vol.51, no.4, pp.692-703, 2022.

[23] R. Chen and Z. Yu, LSYOLO: An algorithm for linear scan PCB defect detection, *Measurement Science and Technology*, vol.36, no.1, 016040, 2024.

[24] H. Wei et al., DWRSeg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation, *arXiv Preprint*, arXiv: 2212.01173, 2022.

[25] Z. Xia et al., Vision transformer with deformable attention, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[26] Y. Chen et al., Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases, *Computers in Biology and Medicine*, vol.170, 107917, 2024.
[27] S. Luo et al., EPDD-YOLO: An efficient benchmark for pavement damage detection based on Mamba-YOLO, *Measurement*, 117638, 2025.
[28] Z. Yu et al., YOLO-FaceV2: A scale and occlusion aware face detector, *Pattern Recognition*, vol.155, 110714, 2024.
[29] J. Yang et al., Focal modulation networks, *Advances in Neural Information Processing Systems*, vol.35, pp.4203-4217, 2022.
[30] D. Li et al., YOLOv8-EMSC: A lightweight fire recognition algorithm for large spaces, *Journal of Safety Science and Resilience*, vol.5, no.4, pp.422-431, 2024.
[31] P. Yang et al., Diagnosis of obsessive-compulsive disorder via spatial similarity-aware learning and fused deep polynomial network, *Medical Image Analysis*, vol.75, 102244, 2022.
[32] M. Tan, R. Pang and Q. V. Le, EfficientDet: Scalable and efficient object detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Author Biography

Yunzheng Tao received a Bachelor of Science degree in Thermal Energy Engineering from Shanghai Jiao Tong University, China, in June 1998 and obtained a Master of Engineering degree in Power Engineering from Kunming University of Science and Technology, China, in June 2011. He joined Yunnan Yangzonghai Power Generation Co., Ltd. (now renamed Guoneng Yangzonghai Power Generation Co., Ltd.), China, in July 1998. His research areas encompass condition monitoring and fault diagnosis of power plant electrical equipment.

Jianyun Ma graduated with a bachelor's degree in Computer Science and Technology from Yunnan Open University, China, in 2024. He joined Yunnan Yangzonghai Power Generation Co., Ltd. (now renamed Guoneng Yangzonghai Power Generation Co., Ltd.), China, in July 1997. His research areas encompass condition monitoring and fault diagnosis of power plant electrical equipment.

Zhizhen Wu graduated from the Yunnan Administrative College of the Party School of the CPC Yunnan Provincial Committee with a degree in Economic Management, China, in 2005. He joined Yunnan Yangzonghai Power Generation Co., Ltd. (now renamed Guoneng Yangzonghai Power Generation Co., Ltd.), China, in November 2000. His research areas encompass condition monitoring and fault diagnosis of power plant electrical equipment.

Shouming Cun received a bachelor's degree in Thermal and Power Engineering from North China Electric Power University, China, in 2011 and joined Yunnan Yangzonghai Power Generation Co., Ltd. (now renamed Guoneng Yangzonghai Power Generation Co., Ltd.), China, in July 2011. His research areas encompass condition monitoring and fault diagnosis of power plant electrical equipment.

Xiaohui Guo received a Bachelor of Science degree in Materials Science and Engineering from Chongqing University, China, in 1997 and a Master of Business Administration degree from Yunnan University, China, in 2011. He joined Yunnan Yangzonghai Power Generation Co., Ltd. (now renamed Guoneng Yangzonghai Power Generation Co., Ltd.), China, in July 1997. His research areas encompass condition monitoring and fault diagnosis of power plant electrical equipment.

Xiaopan Wang received a Bachelor of Science degree in Mechanical Design, Manufacturing, and Automation from the School of Mechanical and Electrical Engineering at Xingtai University, China, in 2023. She is currently pursuing a master's degree at North China Electric Power University, China. Her primary research interests include fault diagnosis and deep learning.

Shuting Wan received the Ph.D. degree in Electric Machines and Electric Apparatus from North China Electric Power University, China, in 2006.

Dr. Wan joined the Department of Mechanical Engineering, North China Electric Power University, in 1994. He has also been the Head of the Hebei Key Laboratory of Electric Machinery Health Maintenance and Failure Prevention. His areas of interests include condition monitoring and fault diagnosis of power equipment.