A Convolutional Block Attention and Bidirectional Pyramid Network for Gear Defect Detection in Challenging Industrial Datasets

Ling Bai[1], Tom Shumka[2], Jason Shumka[2], and Zheng Liu[1]

[1] University of British Columbia, Kelowna BC V1V1V7, Canada
[2] Cleansolv International, Madison, Wisconsin 53703, United States
{ling.bai; zheng.liu}@ubc.ca

Abstract. Gear defect detection is an important task for condition monitoring and predictive maintenance in industrial transmission systems. However, practical gear inspection remains challenging because defect regions are often small, visually subtle, and affected by complex industrial imaging conditions, such as background noise, uneven illumination, oil contamination, and variations in viewing angles. In addition, gear defect datasets usually suffer from class imbalance, where some failure modes appear more frequently than others, making robust multi-class defect detection difficult. To address these challenges, this paper proposes CABiP-YOLO, a convolutional block attention and bidirectional pyramid network for vision-based gear defect detection. The proposed framework is developed based on a YOLO detection architecture and is designed to detect and classify seven representative gear failure modes, including wear, scuffing, plastic deformation, Hertzian fatigue, cracking, fracture, and bending fatigue. A convolutional block attention mechanism is integrated to enhance spatial and channel-wise defect-related features, allowing the network to focus on critical damaged regions while suppressing irrelevant background information. Meanwhile, a bidirectional pyramid network is incorporated to strengthen multi-scale feature fusion by combining low-level texture details and high-level semantic information. Experimental results demonstrate that the proposed framework can localize and classify multiple types of gear defects under challenging imaging conditions. The analysis further shows that CABiP-YOLO has practical potential for automated visual inspection and gear condition monitoring, while class imbalance and subtle defect appearance remain important research for future improvement.

Keywords: Gear defect detection · challenging industrial dataset · convolutional block attention · bidirectional pyramid network.

1 Introduction

Gears are critical components in mechanical transmission systems and are widely used for power and motion transmission in industrial equipment. The reliability of gear systems directly affects operational efficiency, production continuity, and

equipment safety. In practical industrial environments, gear defects such as wear, scuffing, plastic deformation, Hertzian fatigue, cracking, fracture, and bending fatigue may gradually degrade transmission performance and eventually lead to severe mechanical failure [1]. If these defects are not detected at an early stage, they can result in unexpected downtime, increased maintenance costs, and potential safety risks. Therefore, accurate and timely gear defect detection is an important task for condition monitoring and predictive maintenance.

Existing gear fault diagnosis methods mainly rely on manual inspection, vibration analysis, acoustic emission, and other signal-processing-based techniques. These methods have been widely studied for gear diagnostics and prognostics, especially in rotating machinery monitoring and vibration-based gear wear analysis. However, vibration-based methods may face limitations when defect signatures are weak, surface damage evolves gradually, or the relationship between visual surface degradation and vibration response is complex. Recent reviews also indicate that gear diagnosis and prognosis remain challenging because gear failure mechanisms are diverse and operating conditions vary across industrial systems [2]. In addition, manual visual inspection is labor-intensive, subjective, and difficult to apply continuously in production environments. Image-based deep learning methods provide a promising alternative because they can directly localize visible surface defects and classify defect types from gear images [3, 4]. YOLO-style object detectors are particularly suitable for such applications due to their single-stage detection pipeline and real-time inference capability [5–7]. YOLO series further improves object detection through modern feature extraction, multi-scale feature aggregation, and an anchor-free detection head. Nevertheless, gear defect detection still faces several practical challenges [8, 9]: defect regions are often small and subtle; different failure modes may share similar visual patterns; industrial images may contain noise, oil stains, uneven illumination, and background clutter; and the distribution of defect categories is often imbalanced, which may reduce the model's ability to generalize to rare defect types.

To address these challenges, this paper proposes a vision-based gear defect detection framework named CABiP-YOLO, developed based on the YOLO detection architecture. The proposed framework is designed to detect and classify seven representative gear failure modes from gear surface images. First, a convolutional block attention mechanism is introduced to enhance defect-related spatial and channel-wise features, allowing the network to focus on critical regions while suppressing irrelevant background information. This is particularly important for detecting subtle surface defects and small damaged areas. Second, a bidirectional pyramid feature fusion structure is incorporated to improve multi-scale representation learning. Bidirectional feature fusion enables information flow across different feature levels in both top-down and bottom-up directions. By combining attention-guided feature enhancement and bidirectional multi-scale fusion, CABiP-YOLO aims to improve the robustness and accuracy of gear defect localization and classification under practical industrial imaging conditions.

The main contributions of this paper are summarized as follows:

- **Real and challenging industrial dataset.** This paper organizes a gear defect detection dataset collected from real industrial scenarios. The dataset covers seven common failure modes, including wear, scuffing, plastic deformation, Hertzian fatigue, cracking, fracture, and bending fatigue. Due to complex backgrounds, illumination variations, oil stains, and subtle defect patterns, the dataset is challenging for visual defect detection.
- **Convolutional block attention-enhanced architecture.** The CABiP-YOLO architecture integrates the convolutional block attention module into the YOLO-based architecture. The attention module enhances spatial and channel-wise defect features, improving the model's ability to focus on critical defect regions and reduce interference from irrelevant background information.
- **Bidirectional pyramid feature fusion.** A bidirectional pyramid network is adopted to replace conventional one-directional feature aggregation. This design strengthens multi-scale feature fusion by combining low-level detailed features and high-level semantic features, improving the detection of gear defects with different sizes, shapes, and visual appearances.

Experimental results show that CABiP-YOLO can effectively detect multiple gear defect types from challenging industrial images. The results also indicate that robust feature learning is important for handling small, subtle, and visually similar defects in this dataset.

2 Dataset

This section describes the gear defect dataset used in this study, including the considered failure modes, data sources, image preparation procedures, and annotation format. The dataset is designed to support vision-based detection and classification of common gear surface defects for condition monitoring and predictive maintenance.

2.1 Gear Failure Modes

Gear defects can be categorized into several typical failure modes according to their physical characteristics and degradation mechanisms. In this study, seven representative gear defect types are considered according to standards such as AGMA 1010-F14 [10]. These categories cover both surface-level degradation and structural failure patterns commonly observed in industrial gear systems. As shown in the Figure 1, the seven types of defects are displayed. Here is a detailed overview of each gear defect type:

1. **Wear:** Wear refers to material loss caused by friction between gear tooth surfaces, which gradually degrades the functional performance of the gear.

2. **Scuffing:** Scuffing is a severe form of wear that usually occurs under high load or insufficient lubrication, where heat generation and material transfer appear between contact surfaces.
3. **Plastic Deformation:** Plastic deformation occurs when the gear tooth is subjected to loads exceeding the material yield strength, resulting in permanent shape changes.
4. **Hertzian Fatigue:** Hertzian fatigue is caused by repeated contact stress cycles and may lead to surface fatigue and crack initiation.
5. **Cracking:** Cracking refers to the formation and propagation of cracks due to stress concentration in the gear material.
6. **Fracture:** Fracture represents the complete separation of gear components, often following crack propagation to a critical size.
7. **Bending Fatigue:** Bending fatigue occurs under cyclic stress on gear teeth and may eventually cause tooth breakage.

These seven failure modes provide a practical classification basis for gear defect detection. They also reflect different visual characteristics, ranging from subtle surface changes to large structural damage, making the dataset suitable for evaluating the robustness of detection models under various defect appearances.

2.2 Data Sources and Preparation

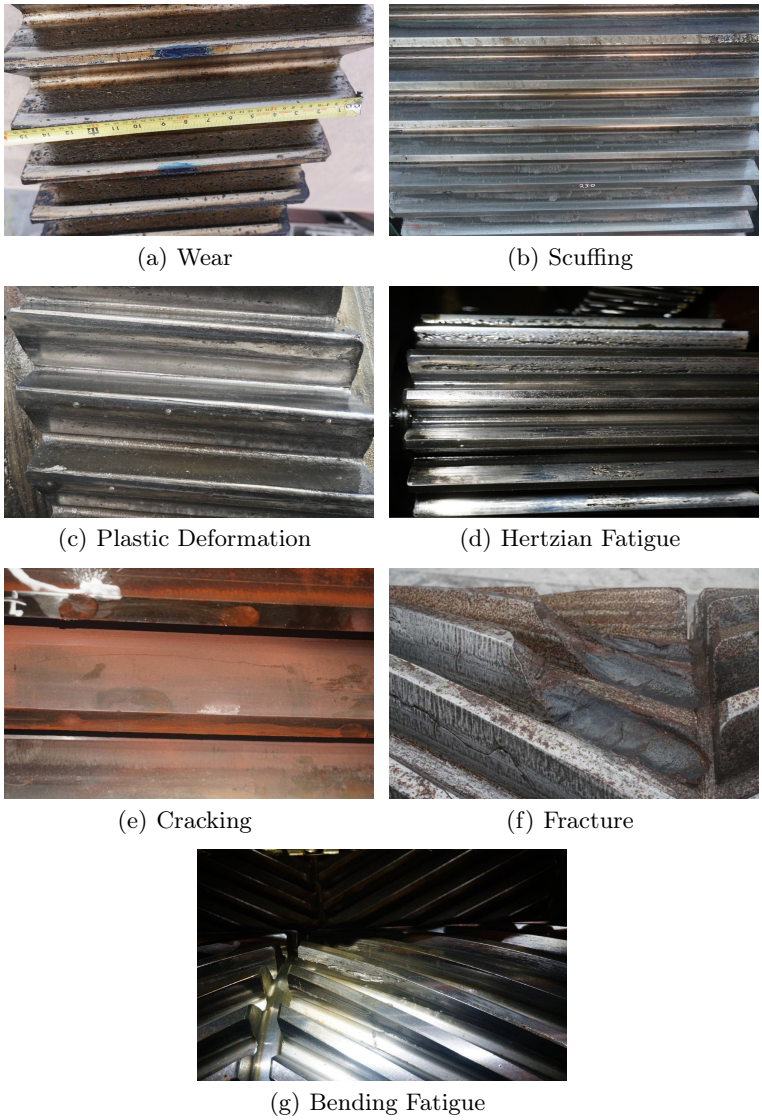
The dataset used in this study is constructed from multiple sources to improve the diversity of gear defect samples. In particular, images collected from actual gear equipment in industrial environments are included to represent practical inspection scenarios. These images contain real surface conditions and operational variations, such as background noise, illumination changes, oil stains, and different viewing angles. As shown in Figure 2, the dataset includes gear images captured from real industrial settings.

Before training, the collected images are processed to improve data quality and reduce irrelevant information. The preparation process includes image cropping, background removal, normalization, and data augmentation. Image cropping is used to focus on gear surfaces and potential defect regions, while background removal reduces interference from complex industrial scenes. Normalization improves image consistency under different acquisition conditions. Data augmentation, including rotation, flipping, and noise perturbation, is applied to improve model robustness and generalization.

These preparation steps help reduce the influence of background clutter and imaging inconsistency, making the dataset more suitable for training the proposed gear defect detection model.

2.3 Dataset Characteristics and Challenges

The real gear defect dataset presents several challenges for visual detection. First, many defect regions are small and subtle, especially early-stage wear, cracks, and fatigue-related damage. These defects may occupy only a small portion of the



(a) Wear (b) Scuffing

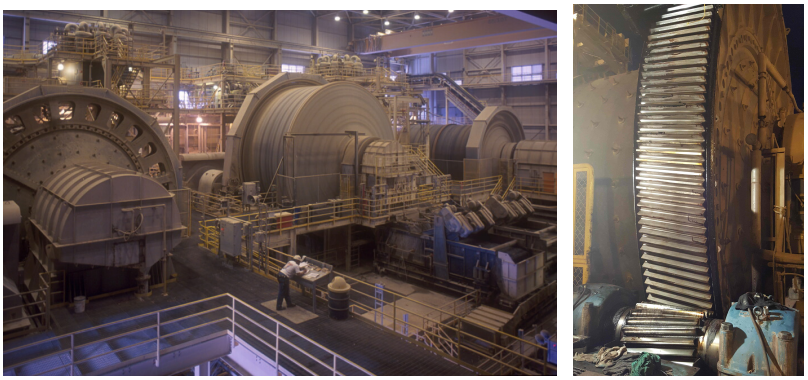
(c) Plastic Deformation (d) Hertzian Fatigue

(e) Cracking (f) Fracture

(g) Bending Fatigue

Fig. 1. Seven Types of Defects.

image and can be difficult to distinguish from normal surface texture. Second, different defect types may have similar visual appearances. For example, wear, scuffing, and Hertzian fatigue can all appear as surface degradation patterns, which increases the difficulty of classification. Third, real industrial images may contain noise, uneven lighting, oil contamination, and complex backgrounds. These factors can affect both localization accuracy and classification reliability.



(a) (b)

Fig. 2. Gears from Real Industrial Settings.

Another important challenge is class imbalance. Some defect categories, such as plastic deformation, may appear more frequently in the dataset, while other categories may have fewer annotated instances. This imbalance can bias the model toward frequent classes and reduce its ability to recognize rare defect types. Therefore, data augmentation, balanced sampling, and further collection of diverse real-world samples are important for improving model robustness.

Overall, the dataset provides a practical foundation for developing and evaluating gear defect detection models. By covering seven representative failure modes and incorporating images from public, synthetic, and real industrial sources, it supports the study of automated visual inspection methods for gear condition monitoring.

3 CABiP-YOLO

This section presents the proposed CABiP-YOLO framework for gear defect detection and classification. The model is developed based on the YOLO detection architecture and is designed to identify seven representative gear defect types, including wear, scuffing, plastic deformation, Hertzian fatigue, cracking, fracture, and bending fatigue. The proposed framework introduces two main structural improvements: a convolutional block attention mechanism for enhancing defect-related features and a bidirectional pyramid network for multi-scale feature fusion.

3.1 Overall Architecture

The overall architecture of CABiP-YOLO consists of three main components: backbone, neck, and detection head. The backbone extracts hierarchical visual features from the input gear image. The convolutional block attention mechanism

is then applied to refine the extracted features by emphasizing informative spatial regions and channels. The neck aggregates multi-scale feature maps through spatial pyramid pooling and bidirectional feature fusion. Finally, the detection head predicts bounding boxes, confidence scores, and class probabilities for gear defects. Figure 3 illustrates the overall structure of the proposed CABiP-YOLO framework. Compared with the baseline YOLO architecture, the proposed model enhances feature representation by incorporating convolutional block attention after feature extraction and bidirectional pyramid fusion in the neck.

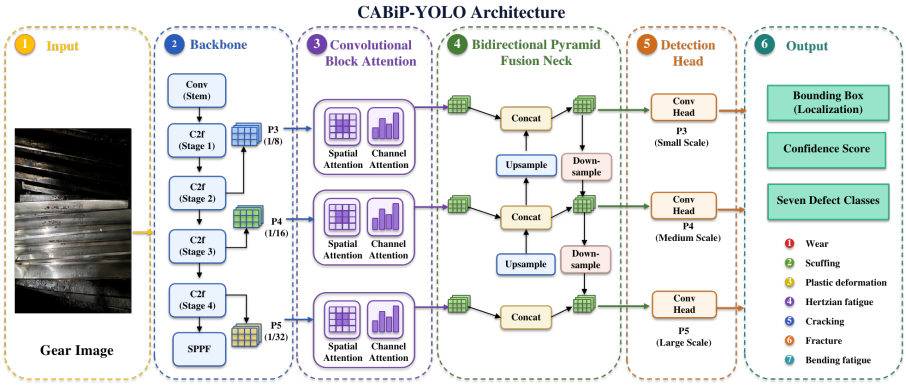


Fig. 3. The proposed CABiP-YOLO architecture.

Given an input gear image $I \in R^{H\times W\times 3}$, the proposed model learns a mapping function:

$$\mathcal{F}_\theta : I \rightarrow \{B_i, c_i, p_i\}_{i=1}^{N}, \tag{1}$$

where B_i denotes the predicted bounding box of the i-th detected defect, c_i denotes the confidence score, p_i denotes the predicted class probability distribution, N is the number of detected defect instances, and θ represents the learnable parameters of the model.

The predicted bounding box is represented as:

$$B_i = (x_i, y_i, w_i, h_i), \tag{2}$$

where (x_i, y_i) denotes the center coordinate of the bounding box, and w_i and h_i denote its width and height, respectively.

3.2 Backbone Feature Extraction

The backbone of CABiP-YOLO is responsible for extracting hierarchical feature representations from gear surface images. It transforms raw image pixels into multi-level feature maps containing low-level texture information and high-level

semantic information. These features are essential for detecting gear defects with different visual appearances and spatial scales.

The basic operation in the backbone is convolution. For an input feature map X, a convolutional layer applies a kernel W to produce an output feature Y:

$$Y(u,v) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} W(i,j) \cdot X(u+i, v+j), \tag{3}$$

where k is the kernel size, $W(i,j)$ represents the convolution kernel weight, and $X(u+i, v+j)$ is the local input feature value.

After convolution, a nonlinear activation function is applied. In this work, the SiLU activation function is used:

$$\mathrm{SiLU}(x) = x \cdot \sigma(x), \tag{4}$$

where $\sigma(x)$ is the sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{5}$$

To improve feature extraction efficiency, the backbone adopts cross-stage partial feature learning. Given an input feature map F, the feature map is divided into two parts, F_1 and F_2. One part is transformed through convolutional operations, while the other part is preserved and later fused:

$$F_{\mathrm{out}} = \mathrm{Concat}(F_2, H(F_1)), \tag{6}$$

where $H(\cdot)$ denotes a sequence of convolution, normalization, and activation operations. This design reduces redundant feature computation and improves gradient flow.

In addition, the C2f module is used to enhance feature propagation across shallow and deep layers. The C2f module can be formulated as:

$$F_{\mathrm{C2f}} = \mathrm{Concat}(F_2, H_1(F_1), H_2(H_1(F_1)), \ldots, H_m(\cdot)), \tag{7}$$

where $H_1, H_2, \ldots, H_m$ represent multiple feature transformation blocks. This structure improves feature reuse and maintains computational efficiency.

3.3 Convolutional Block Attention Mechanism

Gear defects are often small, subtle, and visually similar to normal surface textures. In industrial images, defect regions may also be affected by background clutter, uneven illumination, oil stains, or noise. To enhance the model's ability to focus on critical defect regions, CABiP-YOLO introduces a convolutional block attention module (CBAM) [11] after backbone feature extraction.

The convolutional block attention module refines the feature map by combining spatial attention and channel attention. Given a feature map $F \in R^{C\times H\times W}$, the attention-refined feature map is defined as:

$$F_{\mathrm{att}} = M_c(F) \otimes M_s(F) \otimes F, \tag{8}$$

where $M_c(F)$ is the channel attention map, $M_s(F)$ is the spatial attention map, and $\otimes$ denotes element-wise multiplication with broadcasting.

Channel Attention Channel attention determines which feature channels are more informative for defect detection. It emphasizes useful channels and suppresses less relevant or redundant channels. The channel attention map [11] is computed by applying global average pooling and global max pooling, followed by shared fully connected layers:

$$M_c(F) = \sigma\left(W_1\left(W_0(\mathrm{GAP}(F))\right) + W_1\left(W_0(\mathrm{GMP}(F))\right)\right), \tag{9}$$

where $\mathrm{GAP}(\cdot)$ and $\mathrm{GMP}(\cdot)$ denote global average pooling and global max pooling, respectively. W_0 and W_1 are learnable weights in the shared fully connected layers.

Spatial Attention Spatial attention determines where the model should focus in the image. It highlights important spatial regions that are more likely to contain defects. The spatial attention map [11] is generated by applying average pooling and max pooling along the channel dimension, followed by a convolution operation:

$$M_s(F) = \sigma\left(f^{7\times7}\left([\mathrm{AvgPool}_c(F); \mathrm{MaxPool}_c(F)]\right)\right), \tag{10}$$

where $\mathrm{AvgPool}_c(\cdot)$ and $\mathrm{MaxPool}_c(\cdot)$ represent channel-wise average pooling and max pooling, $[\cdot;\cdot]$ denotes concatenation, $f^{7\times7}(\cdot)$ denotes a convolution operation with a 7×7 kernel, and $\sigma(\cdot)$ is the sigmoid activation function.

The attention mechanism acts as a feature selection module. By applying both spatial and channel attention, the network can focus on defect-related regions and informative feature channels, which is beneficial for detecting small cracks, wear patterns, fatigue marks, and other subtle surface defects.

3.4 Neck with Bidirectional Pyramid Feature Fusion

The neck of CABiP-YOLO bridges the backbone and the detection head. Its main function is to aggregate feature maps from different levels and generate multi-scale representations for defect localization and classification. This is important because gear defects may appear at different scales, from small cracks to larger deformations or fractures.

The neck contains two main components: Spatial Pyramid Pooling-Fast (SPPF) and a bidirectional pyramid network.

Spatial Pyramid Pooling-Fast The SPPF module is used to capture contextual information from different receptive fields without significantly increasing computational cost. Given an input feature map F, SPPF applies max pooling operations at multiple spatial scales and concatenates the resulting features:

$$F_{\mathrm{SPPF}} = \mathrm{Concat}\left(F, \mathrm{MaxPool}_{k_1}(F), \mathrm{MaxPool}_{k_2}(F), \mathrm{MaxPool}_{k_3}(F)\right), \tag{11}$$

where k_1, k_2, and k_3 denote the kernel sizes of different max pooling operations. This module increases the receptive field and improves the model's ability to capture contextual information around defect regions.

Bidirectional Pyramid Network Conventional feature aggregation mainly transfers semantic information from high-level features to low-level features in a top-down direction. However, for gear defect detection, both low-level detailed features and high-level semantic features are important. Low-level features preserve edges, textures, and small defect patterns, while high-level features provide semantic information for defect classification.

To improve multi-scale feature representation, CABiP-YOLO incorporates a bidirectional pyramid network, BiFPN [12]. This structure allows feature information to flow in both top-down and bottom-up directions. Let $\{P_3, P_4, P_5\}$ denote multi-scale feature maps extracted from different network levels. The top-down feature fusion can be expressed as:

$$P_l^{td} = \phi\left(P_l + \mathrm{Up}(P_{l+1}^{td})\right), \tag{12}$$

where P_l^{td} is the top-down fused feature at level l, $\mathrm{Up}(\cdot)$ denotes upsampling, and $\phi(\cdot)$ denotes convolutional refinement.

The bottom-up fusion is then formulated as:

$$P_l^{out} = \phi\left(P_l^{td} + \mathrm{Down}(P_{l-1}^{out})\right), \tag{13}$$

where $\mathrm{Down}(\cdot)$ denotes downsampling, and P_l^{out} is the final fused feature map at level l.

To further improve adaptive feature fusion, learnable weights are introduced to control the contribution of different input features. The weighted feature fusion is defined as:

$$F_{\text{fusion}} = \frac{\sum_{i=1}^{n} \alpha_i F_i}{\epsilon + \sum_{i=1}^{n} \alpha_i}, \quad \alpha_i \geq 0, \tag{14}$$

where F_i denotes the input feature map from the i-th level, α_i is a learnable fusion weight, and ϵ is a small constant used for numerical stability.

Through bidirectional feature fusion, the model can combine detailed local information and global semantic information more effectively. This improves the detection of gear defects with different sizes, shapes, and visual appearances.

3.5 Detection Head

The detection head generates the final predictions for gear defect localization and classification. It receives the fused multi-scale feature maps from the neck and predicts bounding box coordinates, confidence scores, and class probabilities.

CABiP-YOLO adopts an anchor-free detection design. Unlike anchor-based detectors, which rely on predefined anchor boxes, the anchor-free head directly predicts the position and size of bounding boxes from feature maps. This improves flexibility when detecting gear defects with irregular shapes and varying scales.

For each grid location, the predicted bounding box is formulated as:

$$\hat{x} = \sigma(t_x) + c_x, \quad \hat{y} = \sigma(t_y) + c_y, \tag{15}$$

$$\hat{w} = e^{t_w}, \quad \hat{h} = e^{t_h}, \tag{16}$$

where $(\hat{x}, \hat{y})$ denotes the predicted center coordinate, $(\hat{w}, \hat{h})$ denotes the predicted width and height, (t_x, t_y, t_w, t_h) are network outputs, and (c_x, c_y) denotes the corresponding grid cell coordinate.

The class probability is computed using the softmax function:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \tag{17}$$

where z_i is the logit of class i, and C is the number of defect categories. In this study, $C = 7$.

The detection head outputs:

$$\hat{Y} = \left\{ \hat{B}_i, \hat{s}_i, \hat{p}_i \right\}_{i=1}^{N}, \tag{18}$$

where $\hat{B}_i$ is the predicted bounding box, $\hat{s}_i$ is the confidence score, and $\hat{p}_i$ is the predicted class probability vector.

3.6 Training Objective

The training objective of CABiP-YOLO consists of bounding box regression loss, confidence loss, and classification loss. The overall loss function is defined as:

$$\mathcal{L} = \lambda_{\text{box}} \mathcal{L}_{\text{box}} + \lambda_{\text{conf}} \mathcal{L}_{\text{conf}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}, \tag{19}$$

where $\mathcal{L}_{\text{box}}$ is the bounding box localization loss, $\mathcal{L}_{\text{conf}}$ is the confidence loss, $\mathcal{L}_{\text{cls}}$ is the classification loss, and λ_{box}, λ_{conf}, and λ_{cls} are balancing coefficients.

The bounding box regression loss is based on the overlap between predicted and ground-truth boxes. The Intersection over Union (IoU) between a predicted box $\hat{B}$ and a ground-truth box B is defined as:

$$\text{IoU}(\hat{B}, B) = \frac{|\hat{B} \cap B|}{|\hat{B} \cup B|}. \tag{20}$$

The box loss can be written as:

$$\mathcal{L}_{\text{box}} = 1 - \text{IoU}(\hat{B}, B). \tag{21}$$

The confidence loss is formulated using binary cross-entropy:

$$\mathcal{L}_{\text{conf}} = -\left[y \log(\hat{s}) + (1 - y) \log(1 - \hat{s}) \right], \tag{22}$$

where y indicates whether a defect is present, and $\hat{s}$ is the predicted confidence score.

The classification loss is formulated as categorical cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{C} y_i \log(\hat{p}_i), \tag{23}$$

where y_i is the one-hot ground-truth label for class i, and $\hat{p}_i$ is the predicted probability for class i.

3.7 Non-Maximum Suppression

After the detection head generates candidate bounding boxes, Non-Maximum Suppression (NMS) is applied to remove redundant detections. NMS keeps the bounding box with the highest confidence score and suppresses other highly overlapping boxes whose IoU exceeds a predefined threshold.

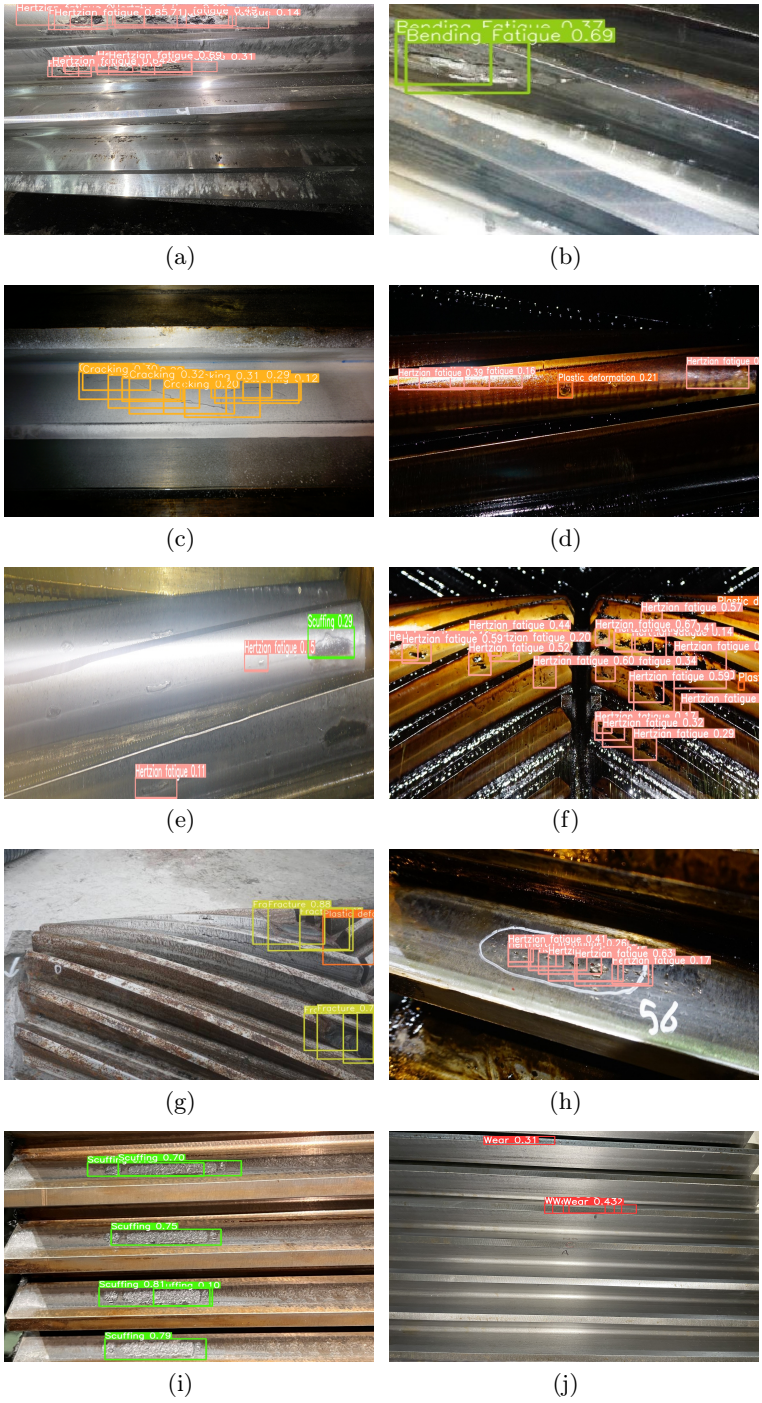
In summary, CABiP-YOLO combines attention-guided feature refinement and bidirectional multi-scale feature fusion within a YOLO-based detection framework. The convolutional block attention mechanism improves the model's focus on defect-related regions, while the bidirectional pyramid network enhances feature aggregation across different scales. These two components allow the model to detect and classify gear defects more effectively under practical industrial imaging conditions.

4 Experimental Results and Analysis

The objective of the experiments is to evaluate whether the proposed model can accurately localize and classify different types of gear defects under practical imaging conditions. All experiments were performed on an NVIDIA Tesla T4 GPU with 16 GB memory. The gear defect dataset contains 600 images and was split into training, validation, and testing sets with a 70:15:15 ratio. All images were resized to 640×640 pixels. The model was trained for 100 epochs with a batch size of 8. The evaluation focuses on both localization and classification performance.

As shown in Figure 4, the proposed CABiP-YOLO model can detect and classify different gear defect types under challenging industrial imaging conditions. The results show that the model can localize multiple failure modes, using bounding boxes with corresponding confidence scores. In particular, the model is able to detect both single and multiple defect instances within the same image, indicating its ability to handle defects with different scales and appearances. However, some predictions have relatively low confidence, especially for subtle defects or regions with strong visual interference, suggesting that class imbalance, small defect size, and ambiguous surface textures remain challenging. Overall, these qualitative results demonstrate the effectiveness and practical potential of CABiP-YOLO for multi-class gear defect detection, while also indicating the need for further improvement using more balanced and diverse industrial data.

This real industrial gear defect dataset is challenging due to small defect regions, subtle visual differences, and complex surface textures. As shown in Table 1, CABiP-YOLO achieves the highest Precision, F1-score, and mAP@0.5 among the compared methods. Compared with standard YOLO-based detectors, the proposed model integrates the convolutional block attention and bidirectional pyramid fusion, which help enhance defect-related feature responses and multi-scale representations. This makes the model more suitable for detecting small and subtle gear defects under complex industrial imaging conditions than the original YOLO baselines [13–15]. Compared with staged CNN detectors such



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 4. Classification results.

as Faster R-CNN and RetinaNet [16, 17], CABiP-YOLO provides a better balance between precision and overall detection accuracy. The transformer-based RT-DETR represents another detection paradigm, but it may require more diverse training data to fully exploit its global modeling ability [18]. It is also noted that CABiP-YOLO does not achieve the highest Recall or mAP@0.5:0.95, indicating that some small, rare, or visually ambiguous defects may still be missed and that fine-grained localization under stricter IoU thresholds remains challenging. Overall, the results show that attention-guided feature refinement and bidirectional multi-scale fusion are effective for this challenging task, while improving recall and localization robustness remains an important direction for future work.

Table 1. Comparison of detection performance across different methods.

Method	Precision	Recall	F1-score	mAP@0.5	mAP@0.5:0.95
CABiP-YOLO	**0.8373**	0.3467	**0.4903**	**0.4072**	0.1977
YOLO11s	0.6763	0.2910	0.4069	0.3185	0.1772
YOLO11n	0.6742	0.2713	0.3869	0.2577	0.1469
RT-DETR	0.5890	0.3278	0.4212	0.3248	0.1692
YOLOv8s	0.5704	**0.3695**	0.4485	0.3638	**0.2241**
YOLOv8n	0.5547	0.2759	0.3685	0.2952	0.1436
Faster R-CNN	0.4759	0.2438	0.3225	0.2612	0.1604
RetinaNet	0.2363	0.2247	0.2303	0.1051	0.0400

5 Conclusion

This paper proposed CABiP-YOLO, a vision-based framework for gear defect detection under challenging industrial imaging conditions. The model detects and classifies seven representative gear failure modes, including wear, scuffing, plastic deformation, Hertzian fatigue, cracking, fracture, and bending fatigue. By integrating a convolutional block attention mechanism and a bidirectional pyramid feature fusion structure into a YOLO-based architecture, the proposed method enhances defect-related feature representation and improves multi-scale detection capability. Experimental results show that CABiP-YOLO can localize and classify multiple gear defect types from complex gear images. The results demonstrate its potential for automated visual inspection, gear condition monitoring, and predictive maintenance. However, several limitations remain. The dataset is imbalanced across defect categories, and some defects are small, subtle, or visually similar to normal surface textures. Image quality issues, such as oil contamination, uneven illumination, and background noise, may also affect detection performance. Future work will focus on collecting more diverse industrial data, improving class balance, enhancing model robustness, and exploring multimodal condition monitoring by combining image data with vibration signals.

References

1. Kumar, A., Gandhi, C.P., Zhou, Y., Kumar, R., Xiang, J.: Latest developments in gear defect diagnosis and prognosis: A review. Measurement **158**, 107735 (2020).
2. Kundu, P., Darpe, A.K., Kulkarni, M.S.: A review on diagnostic and prognostic approaches for gears. Structural Health Monitoring **20**(5), 2853–2893 (2021).
3. Zhang, D., Zhou, S., Zheng, Y., Xu, X.: Review on Application of Machine Vision-Based Intelligent Algorithms in Gear Defect Detection. Processes **13**(10), 3370 (2025).
4. Huang, H., Liang, Q., Wu, R., Yang, D., Wang, J., Zheng, R., Xu, Z.: Gear Target Detection and Fault Diagnosis System Based on Hierarchical Annotation Training. Machines **13**(10), 893 (2025).
5. Wang, Z., Zhao, L., Li, H., Xue, X., Liu, H.: Research on a Metal Surface Defect Detection Algorithm Based on DSL-YOLO. Sensors **24**(19), 6268 (2024).
6. Ruan, S., Zhan, C., Liu, B., Wan, Q., Song, K.: A high precision YOLO model for surface defect detection based on PyConv and CISBA. Scientific Reports **15**(1), 15841 (2025).
7. Ge, Y., Li, Z., Meng, L.: YOLO-MSD: a robust industrial surface defect detection model via multi-scale feature fusion. Applied Intelligence **55**(12), 840 (2025).
8. Zhang, S., He, M., Zhong, Z., Zhu, D.: An industrial interference-resistant gear defect detection method through improved YOLOv5 network using attention mechanism and feature fusion. Measurement **221**, 113433 (2023).
9. Zhou, F., Chao, Y., Wang, C., Zhang, X., Li, H., Song, X.: A small sample nonstandard gear surface defect detection method. Measurement **221**, 113472 (2023).
10. American Gear Manufacturers Association: ANSI/AGMA 1010-F14, Appearance of Gear Teeth – Terminology of Wear and Failure. American Gear Manufacturers Association, Alexandria, VA (2014).
11. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018).
12. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and Efficient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10781–10790 (2020).
13. Jocher, G., Qiu, J.: Ultralytics YOLO11. Version 11.0.0 (2024).
14. Khanam, R., Hussain, M.: YOLOv11: An Overview of the Key Architectural Enhancements. arXiv preprint arXiv:2410.17725 (2024).
15. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. Version 8.0.0 (2023).
16. Rane, N., Choudhary, S., Rane, J.: YOLO and Faster R-CNN Object Detection in Architecture, Engineering and Construction (AEC): Applications, Challenges, and Future Prospects. SSRN preprint (2023).
17. Kadiyam, A.T.S., Padhan, C.R., Panigrahi, S.: A Study on Object Detection using Faster R-CNN, RetinaNet, RPN_R_50_FPN, RPN_R_50_C4. In: 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT), pp. 1–6 (2024).
18. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs Beat YOLOs on Real-time Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965–16974 (2024).