Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37347

Deep Learning-Based Visual Detection Technology for Metal Gear Surface Defects

Jie Meng

International Department, Hefei 168 Middle School, Hefei, China
mj168dp@outlook.com

Abstract. Due to complex background interference, a lack of real defect samples, and inconsistent evaluation of real-time deployment in metal gear surface defect detection, four representative studies were selected for comparison and four technical approaches were analyzed: lightweight detection, interference resistance, few-sample augmentation, and production-line application. The results show that lightweight networks can increase the speed of detection; attention mechanisms and multi-scale feature fusion help suppress interference from oil contamination, shadows, etc.; and generative models can address the shortage of defect samples. Insufficient coverage of real samples still affects the model's generalization. Although the actual production-line system meets the cycle-time requirements, its false-positive rate and stability in the face of changes have not yet been optimized. Based on the above analysis, this paper proposes expanding the collection of real samples across different gear models and production batches, strengthening tests under complex imaging conditions, standardizing the evaluation of complete inspection time, and conducting comprehensive assessment using both false-positive and false-negative rates. The above recommendations provide a reference for the engineering application of visual inspection systems for metal gear surface defects.

Keywords: metal gear surface defects, machine vision, deep learning, few-sample detection, real-time deployment

1. Introduction

Gears are the basic components of a mechanical transmission system, and the surface quality directly affects the smoothness, stability and service life of transmission. During the production of automobiles and machinery, gear surfaces may have defects such as scratches, impact damage, missing teeth, wear and surface roughness. The previous method of manual visual inspection relied on the inspector's experience and was relatively inefficient and inconsistent. With the development of intelligent manufacturing, machine vision has gradually been used to inspect the quality of gears because it is non-contact, high-efficiency and easy to automate. Previous studies have shown that complex backgrounds, illumination variations and small defects all increase the difficulty of visual inspection; therefore, improving the robustness of inspection systems is of great practical value [1-3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37347

Recently, deep learning has been applied to the large-scale problem of gear surface defect detection. Research has been continuously conducted on lightweight networks, attention mechanisms, multi-scale feature fusion and gear-tooth surface defect segmentation to provide new technical paths for improving the detection accuracy, speed and adaptability to complex scenarios [4-6]. In addition, targeted studies have explored lightweight detection, interference immunity, few-shot augmentation and practical production-line applications. Four typical studies have been selected for this paper.

Based on the above studies, this paper compares the features of different methods in three areas: interference immunity, few-sample detection and real-time deployment. It investigates the deficiencies of the present methods in terms of data sources, defect categories, detection speed, false positives and adaptability to actual production lines. Based on the above, the following directions for improvement are proposed: expand the coverage of real samples, strengthen testing under complex imaging conditions, standardize the evaluation of complete inspection time, and improve the assessment of false positives and false negatives; thus, provide a reference for the engineering application of future visual inspection systems for metal gear surface defects.

2. Research on gear defect detection

Machine vision continuously acquires images and automatically detects defects on the gear surface without manual inspection. The Detection Performance is limited by image quality, sample size and model speed. Yan and others have reduced the detection speed by using a light-weight YOLOv5 model; Zhang and others have solved the interference problem caused by oil contamination and shadows; Zhou and others have added generated images to supplement real defect samples; and Idzik and others have applied a detection model to a practical helical-gear inspection process. The four studies are four technical paths: lightweight detection, interference immunity, sample augmentation and production-line application.

2.1. Case 1: study by Yan et al.

Yan et al. have introduced lightweight modifications to YOLOv5 to improve the speed of inference without sacrificing detection accuracy [7]. ShuffleNetv2 was selected as the lighter-weight model in this study, and an attention module and a new loss function were added for fracture, missing-defect and scratch detection. On a dataset of 3,000 gear images, the model reached an mAP@0.5 of 98.6% and had a detection speed of 130.4 FPS. Although the lightweight model had good accuracy, only three defect types were used in the experiment and data from a single dataset were employed; thus, its performance on other gear types and in workshop environments has not been verified.

2.2. Case 2: study by Zhang et al.

Zhang et al. improved YOLOv5 to address the problem that oil contamination and shadows in gear images can easily be confused with actual defects [8]. The study introduced an attention module and multi-scale feature fusion into the network to enhance feature representation in defect regions, increasing mAP@0.5 by 12 percentage points over the original model. However, detection speed decreased from 33 FPS to 25 FPS, and the 1,200 images were labeled with only one defect category, "gear impact damage." The method improved recognition under complex backgrounds, but the reduction in speed and the use of a single defect category limit the applicability of the results.

DOI: 10.54254/2755-2721/2026.37347

2.3. Case 3: study by Zhou et al.

Zhou et al. s have conducted some few-sample defect detection for non-standard gears, focusing on sand inclusion, surface roughness and sand expansion defects. First, BAGAN was used to augment the samples, and then an improved YOLOv8 model was applied for detection [9]. The model has an mAP50 of 97.30% on the dataset. However, 460 of the 690 images were generated images, and the proportion of real samples was relatively small. The paper found that surface roughness defects with considerable shape changes were still easily misidentified. Therefore, the generated images can reduce the problem of sample scarcity to some extent, but they cannot replace full-scale validation in practice.

2.4. Case 4: study by Idzik et al.

Idzik et al. applied a defect detection model to a real-world inspection process for automotive helical gears and evaluated its performance in a production environment [10]. The system used two cameras to take pictures of a rotating gear and employed Faster R-CNN for tooth damage, unclean areas and edge impact damage identification. The total scan and detection cycle was 7.5 seconds; thus, it met the factory's requirement of a 10-second production cycle. However, for defect-free gears, the system had a false-positive rate of 24.2%, and this rose to 27.3% after changing the imaging environment. Although the system met the requirements for the production line, the false positives still had to be manually verified, and this problem was even more serious in all imaging environments.

To better understand the above research content, Figure 1 summarizes the four representative research works, as detailed below.

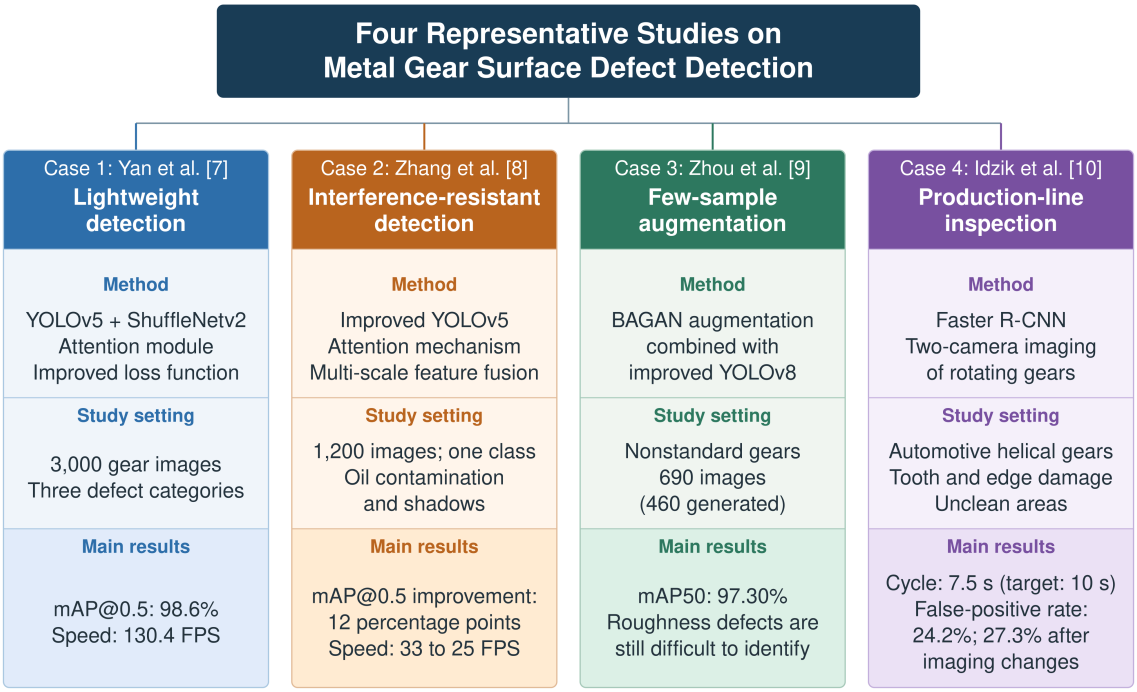

Figure 1. Overview of four representative studies on gear surface defect detection (picture credit : original)

DOI: 10.54254/2755-2721/2026.37347

3. Analysis of limitations

Although the studies above improved visual detection of metal gear surface defects from different perspectives, several limitations remain. The main deficiencies are as follows:

Table 1. Main limitations of the four representative studies

Study	Main limitations
Case 1	Data were obtained from a single dataset, and only three defect categories were included; performance of different gear types and in workshop environments has not been verified.
Case 2	Only one "gear impact damage" tag was used, and after interference-resistant modification, the detection speed decreased to 25 FPS from 33 FPS.
Case 3	Out of 690 images, 460 were generated, and surface roughness defects with significant morphological changes were still frequently misclassified.
Case 4	The false-positive rate for defect-free gears was 24.2%, and after the environment changed, it increased to 27.3%; therefore, the imaging conditions were likely to cause this rise.

As shown in Table 1, sample sources and imaging environments are common problems in many studies. If only one dataset is used for sample collection, the model's performance on other gears or under different light conditions cannot be known. Such variations in actual production lines may also result in more false positives. Therefore, mAP alone is insufficient to determine whether a system is suitable for on-site deployment.

The studies also have different detection speeds. Yan et al. and Zhang et al. report FPS as the processing speed of a single image, while the 7.5 s reported by Idzik et al. includes gear rotation, image acquisition and recognition. Therefore, the two types of measurements cannot be directly compared. To determine whether a system is suitable for production-line deployment, one must consider not only the FPS but also the total time required for a complete inspection and whether the false-positive and false-negative rates are acceptable.

4. Directions for improvement

To address insufficient samples, sensitivity to environmental changes, and inconsistent production-line evaluation metrics identified in the studies above, improvements can be made in four areas: data, models, imaging, and deployment evaluation. Specific recommendations are presented in Table 2.

Table 2. Improvement recommendations and evaluation focus

Cases	Improvement recommendation	Evaluation focus
Case 1	Expand the collection of real samples across different gear models, production batches, and defect categories. Generated images may supplement minority categories, but final validation should use real samples.	Generalization across different gears and production batches

DOI: 10.54254/2755-2721/2026.37347

Table 2. (continued)

Case 2	Add training samples acquired under complex imaging conditions; keep the light source, camera position, and imaging parameters as fixed as possible; and increase testing under environmental changes.	False-positive rate, false-negative rate, and performance after environmental changes
Case 3	In on-site testing, record the complete elapsed time from gear positioning and rotation through image acquisition to result output, rather than reporting model FPS alone.	Complete inspection time per part and production cycle time
Case 4	Include multiple defect categories and report mAP, false positives, and false negatives together, avoiding system evaluation based on a single accuracy metric.	Multi-category detection capability and overall reliability

In Table 2, future data collection should include various gear models, production batches, degrees of oil contamination, and imaging angles, and the models should be independently tested with production-line images that were not included in the training set. Generated images are suitable for supplementing defect categories with limited samples, but final model validation should still be performed on real samples.

Record the total time from gear positioning and image acquisition to result output during on-site testing, rather than just reporting model FPS, and also monitor false positives and false negatives. To reduce the impact of oil contamination, reflections and changes in illumination, corresponding scenarios can be added to the training data, and the light source, camera position and imaging parameters should be kept as consistent as possible.

5. Conclusion

Generally speaking, the four representative studies have shown that deep learning has raised the level of automation for metal gear surface defect detection significantly. The various methods have achieved good experimental results through lightweight design, interference immunity, few-shot learning augmentation and production-line application.

However, in terms of actual application, the existing studies still have some deficiencies. Some experiments use a single data source, have only a small number of defect categories, or contain a high proportion of generated samples; thus, there is insufficient evidence of the model's generalisation across different gear models and real workshop environments. There are also false alarms in a complex imaging environment. Different studies have used different definitions of detection speed, and FPS alone cannot reflect the entire production cycle from positioning and image acquisition to recognition. Therefore, the performance indicators of gear vision inspection systems should include detection accuracy, false-positive rate, false-negative rate, full inspection time and stability under environmental changes.

In the future, the goal of research should be to achieve a higher mAP or FPS on a single dataset, and data collection and independent testing in a real industrial environment need to be strengthened. Expand the range of gear models and production batches, add interference scenarios such as oil contamination and illumination variation, standardise the evaluation of the total inspection time, and jointly evaluate false positives and false negatives to determine whether a model can operate reliably on the production line more accurately. Future work will continue to promote the coordinated

DOI: 10.54254/2755-2721/2026.37347

optimization of lightweight models, few-shot learning with real data, and engineering-oriented imaging systems while maintaining detection accuracy.

References

[1] Kumar, A., Gandhi, C. P., Zhou, Y., et al. (2020). Latest developments in gear defect diagnosis and prognosis: A review. *Measurement, 158*, 107735.
[2] Yu, Z., Wang, Z., & Duan, Z. (2019). Detecting gear surface defects using background-weakening method and convolutional neural network. *Journal of Sensors, 2019*, Art. no. 3140980.
[3] Xu, Z., Wang, A., Hou, F., & Zhao, G. (2023). Defect detection of gear parts in virtual manufacturing. *Visual Computing for Industry, Biomedicine, and Art, 6*(1), Art. no. 6.
[4] Zhu, D., Cheng, G., & Yao, J. (2024). Surface defect detection of automotive gear components based on improved YOLOv5. *Computer Engineering and Applications, 60*(5), 321–327.
[5] Zhou, X., Zhang, Y., Ren, Z., et al. (2024). A Unet-inspired spatial-attention transformer model for segmenting gear tooth surface defects. *Advanced Engineering Informatics, 62*, Art. no. 102933.
[6] Shuai, Y., Lin, Z., Chen, W., et al. (2024). SF-YOLO: An evolutionary deep neural network for gear end surface defect detection. *IEEE Sensors Journal, 24*(13), 21762–21775.
[7] Yan, R., Zhang, R., Bai, J., et al. (2023). STMS-YOLOv5: A lightweight algorithm for gear surface defect detection. *Sensors, 23*(13), 5992.
[8] Zhang, S., He, M., Zhong, Z., & Zhu, D. (2023). An industrial interference-resistant gear defect detection method through improved YOLOv5 network using attention mechanism and feature fusion. *Measurement, 221*, Art. no. 113433.
[9] Zhou, F., Chao, Y., Wang, C., et al. (2023). A small sample nonstandard gear surface defect detection method. *Measurement, 221*: Art. no. 113472.
[10] Idzik, T., Veres, M., Tarry, C., & Moussa, M. (2023). A real-time inspection system for industrial helical gears. *Sensors, 23*(20), Art. no. 8541.