WMD: A Lightweight YOLO11 Extension for Efficient Complexity Detection in Mechanical Part Inspection

1st Xiaofeng Yue
College of Software
Taiyuan University of Technology
Taiyuan, China
https://orcid.org/0009-0003-2858-3690

2nd Jiaqi Yang
College of Software
Taiyuan University of Technology
Taiyuan,China
https://orcid.org/0009-0009-9322-8808

3rd Zeyu Qi
College of Software
Taiyuan University of Technology
Taiyuan, China
https://orcid.org/0009-0009-9165-394X

4th Ruochen Cao
College of Software
Taiyuan University of Technology
Taiyuan, China
https://orcid.org/0000-0002-0312-8355

5th Rui Cao
College of Software
Taiyuan University of Technology
Taiyuan, China
https://orcid.org/0000-0003-4906-0453

6th Xin Wen*
College of Software
Taiyuan University of Technology
Taiyuan, China
https://orcid.org/0000-0003-2363-1190

***Abstract*—A key feature of Industry 4.0 is intelligent manufacturing, which calls for the deep integration of advanced technologies in industrial production. In current manufacturing practices, the complexity detection of mechanical parts still heavily relies on manual inspection, leading to high labor costs and low efficiency. To address this challenge, this study proposes an efficient improved object detection algorithm based on YOLO11, incorporating Wavelet Convolution (WTConv), Manhattan Self-Attention (MaSA), and Distance Intersection over Union (DIoU), collectively referred to as the WMD model. The WMD model integrates WTConv and MaSA into the backbone and neck of the YOLO11 architecture to enhance feature extraction capabilities, while DIoU replaces the original bounding box regression loss function to improve localization accuracy. Experimental results demonstrate that the proposed WMD model achieves significant improvements over the baseline model: precision increases by 3.6%, mAP@0.5 improves by 1.5%, model parameters decrease by 0.042×10^6, and computational cost reduces by 0.1 GFLOPs. Overall, the WMD model effectively balances detection performance and computational efficiency, making it a promising lightweight solution for automated mechanical part complexity detection in smart manufacturing.**

***Index Terms*—YOLO11, Lightweight, Parts Complexity Detection, Wavelet Convolutions.**

I. Introduction

Recent industrial progress is characterized by the ongoing incorporation of advanced technologies to improve operational intelligence and efficiency. Under initiatives like Industry 4.0 [1], analyzing the complexity of mechanical components using technical drawings greatly accelerates manufacturing intelligence and, in turn, markedly enhances the effectiveness of production planning. Nevertheless, current methods for assessing the complexity of mechanical parts rely on manual inspection, which incurs high costs and exhibits limited efficiency, underscoring the pressing need for automated solutions.

Mechanical part detection primarily serves quality inspection and defect identification, while complexity assessment specifically informs production planning. Recent advances in this domain have predominantly enhanced YOLO-based detection through three methodological approaches: Structural optimizations, including modified pooling layers that generate specialized architectures like SPC modules [2]; Integration of attention mechanisms accelerated by hardware implementations [3]; Data enhancement techniques employing both augmentation protocols and super-resolution neural networks for refined feature extraction [4], [5]. Besides, [6]–[8] similar methods are also used to complete their work.These approaches collectively demonstrate that performance improvements are achieved through network refinement, attention-based feature weighting, and data quality enhancement.

Automatic complexity assessment of mechanical parts faces four primary challenges: (1) multi-view discrepancies in part representation, (2) recognition of orthographic projections, (3) background interference (red box, Fig. 1), and (4) demanding requirements for fine-grained feature extraction. The angular variations in technical drawings are exemplified by multi-perspective representations (blue box, Fig. 1). These factors fundamentally limit the applicability of conventional CNN [9] due to their translation-invariance characteristics. Consequently, domain-adapted YOLO [10] architectures emerge as superior solutions for this specialized task.

In this study, we propose an end-to-end detection method for assessing the complexity of process drawings in mechanical parts manufacturing, based on the aforementioned characteristics. The proposed method aims to enable a comprehensive analysis and objective evaluation of the complexity associated

* is the corresponding author.

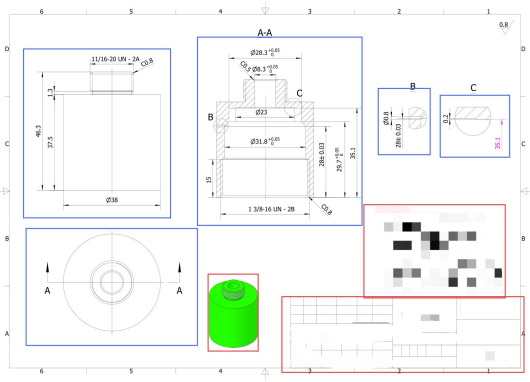

Fig. 1. An example of the Mechanical part drawing.(Privacy information has been masked)

with multi-view part drawings while effectively filtering out irrelevant interfering factors. Furthermore, it incorporates a fault-tolerant correction mechanism to enhance robustness and reliability in complexity assessment.

II. Related Work

A. Wavelet Convolution

Wavelet Convolution (WTconv), a convolutional layer based on wavelet operations, was introduced by Shahaf E. Finger et al. [11] A standard strategy for enlarging the receptive field is to increase the size of the convolutional kernel; for instance, one of the enhancements in [12] involves using larger convolution kernels. However, this strategy eventually reaches a limit: once the kernel is sufficiently large, the receptive field can no longer be effectively expanded [13]. Moreover, employing excessively large kernels can lead to over-parameterization of the network.

WTconv leverages the Wavelet Transform [14] to propose a layer that can be used as a drop-in replacement in existing architectures. For the WTconv layer, by using the Haar wavelet transform, which is efficient and straightforward [15]–[17], the low-frequency contour and high-frequency details (including vertical details, horizontal details, and diagonal details) decomposition of the input image are achieved, and convolution operations are performed at different frequencies. Finally, the output feature map is obtained through the inverse wavelet transform. WT(·) denotes the wavelet transform and IWT(·) denotes the inverse wavelet transform. **W** represents the weight tensor of a k × k deep convolution kernel. The WTconv process can be concisely expressed as (1).

$$\mathbf{Y} = \mathrm{IWT}(\mathrm{Conv}(\mathbf{W}, \mathrm{WT}(\mathbf{X}))) \tag{1}$$

One of the key benefits of employing WTconv is that it increases the number of linear parameters while simultaneously achieving an exponential expansion of the effective receptive field [18]. In addition, the frequency-based decomposition of images enables more effective feature extraction. Thus, the model could deliver improved performance on multi-angle planar graphs.

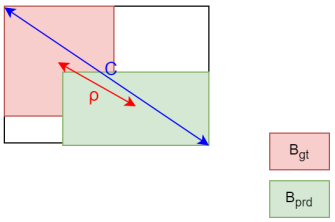

Fig. 2. The schematic diagram of the parameters required for DIoU calculation.

B. Manhattan Self-Attention

Manhattan Self-Attention (MaSA) [19] is a self-attention variant that explicitly incorporates spatial priors into the Vision Transformer [20] architecture. MaSA, learning from [21]–[27], adapts RetNet's [28] one-dimensional, unidirectional temporal decay into a two-dimensional, bidirectional spatial decay. The equation for MaSA is shown as (2) and (3).

$$D^{2d}_{nm} = \gamma^{|x_n - x_m| + |y_n - y_m|} \tag{2}$$

$$\mathrm{MaSA}(X) = (\mathrm{Softmax}(QK^\top) \odot D^{2d})V \tag{3}$$

For a pair of tokens located at coordinates (x_n, y_n) and (x_m, y_m) , the attention score is modulated by a spatial decay matrix D^{2d} , whose elements are defined as (2) where $\gamma \in (0, 1)$ is a decay coefficient. The final attention output is computed as (3) where $\odot$ denotes the Hadamard product. This formulation directly embeds the geometric distance between tokens into the attention computation, endowing the model with explicit awareness of spatial layout.

MaSA introduces a rich, explicit spatial prior by leveraging the Manhattan distance. Besides, MaSA is highly efficient, which aligns with our work on feature extraction and lightweight features.

C. Distance IoU

Distance-IoU (DIoU), introduced by Zheng et al. [29], is an enhanced variant of the Intersection over Union (IoU) metric designed specifically for bounding box regression in object detection. The calculation parameters and process are shown in Fig. 2 and (4).

$$DIoU = IoU - \frac{\rho^2\left(B_{gtc}, B_{prdc}\right)}{C^2} \tag{4}$$

where $\rho^2(\cdot)$ denotes the squared Euclidean distance between the centroids B_{gtc} (ground-truth center) and B_{prdc} (predicted center), and C represents the length of the diagonal of the smallest enclosing box that covers both bounding boxes.

DIoU loss can directly minimize the distance between the predicted box and the real box, and the regression efficiency will be very high when the predicted box and the real box are aligned in both horizontal and vertical directions. This makes DIoU sensitive to target positioning errors and can achieve a balance between speed and accuracy in a lightweight design that aligns with our work.

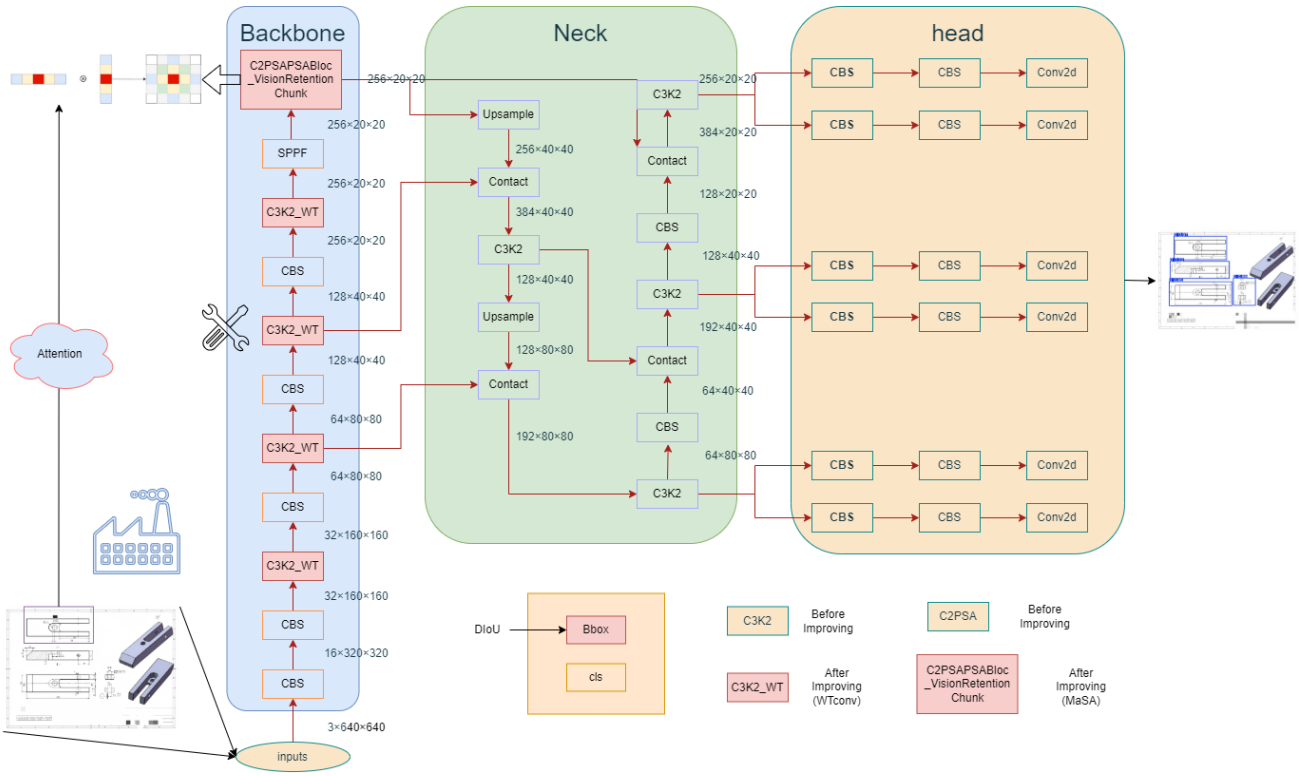

Fig. 3. The overall architecture of WMD model. It replaces the C3K2 block of YOLO11 with the C3K2_WT block fused with WTconv. The C2PSA-SABlock_VisionRetention Chunk block which substitutes for the C2PSA block integrates the MaSA mechanism. In addition, the DIoU loss function is applied in the bounding box regression head.

III. Method

A. Dataset

The training data consists of a proprietary dataset containing annotated images of 3,639 simple mechanical parts and 1,318 complex parts. It is divided into training, validation, and test subsets using an 8:1:1 split. Due to confidentiality requirements, the dataset provider did not disclose a detailed definition of part complexity. However, all part drawings in the dataset are explicitly labeled as either "simple" or "complex".

To determine effective strategies for improving the model, we begin by analyzing the structural and semantic characteristics of the dataset in relation to domain-specific requirements. As illustrated in Fig. 1, the images display a highly organized layout, with clearly separated components in each drawing. However, due to subtle geometric distinctions among similar parts, there can be substantial differences in complexity even between drawings that appear visually alike. This underscores the necessity for robust approaches to assess complexity in planar multi-angle engineering diagrams and further highlights the need to strengthen the model's feature extraction capabilities.

Building on the observations above, the model should be designed to expand its receptive field and enhance the quality of its feature representations. Moreover, the proposed improvements need to be well-adapted to handling highly structured technical drawings and specifically optimized for detecting complexity in mechanical part illustrations. It is also important to develop a more lightweight architecture to boost computational efficiency without sacrificing performance.

B. Strategy for Network Structure Improvement

The overall architecture of the enhanced YOLO11-based model is shown in Fig. 3. In industrial scenarios, models are expected to be lightweight while providing efficient feature extraction. To meet these requirements, and to both expand the model's receptive field and decrease the number of parameters, wavelet convolution is incorporated into the C3K2 block, resulting in the C3K2_WT block.

Building upon the C2PSA block, we refine the module that integrates the MaSA mechanism to further enhance spatial feature representation and suppress irrelevant background information. This design strengthens the model's robustness, especially against background interference, which is crucial in industrial environments, while simultaneously lowering computational cost.

Considering the structured feature characteristics of the input part drawings, we employ the DIoU loss as the bounding box regression function. This modification makes the model more sensitive to target localization errors. In addition, we perform comparative experiments with several bounding box regression loss functions, and the results are summarized in Table I. Among these, DIoU offers an advantageous balance between inference speed and localization precision under a

TABLE I
COMPARATIVE RESULTS OF BOUNDING BOX REGRESSION LOSS FUNCTION. BEST VALUES ARE HIGHLIGHTED IN BOLD.

Model	P(%)	R(%)	mAP@0.5(%)	FPS
YOLO11n+GIoU	**77.5**	80.9	83.9	322.6
YOLO11n+DIoU	73.8	83.5	**85.6**	312.5
YOLO11n+CIoU	77.2	83.0	84.8	322.6
YOLO11n+EIoU	76.2	83.1	85.4	322.6
YOLO11n+SIoU	75.4	**86.1**	84.7	232.6
YOLO11n+WIoU	77.2	81.1	84.8	**333.3**

lightweight configuration, making it well suited to industrial application requirements.

As described, the improved model, integrating WTConv, MaSA, and DIoU, is named WMD, which stands for Wavelet-Multi-scale Spatial Attention-DIoU.

IV. EXPERIMENTS

A. Ablation Study

To validate the effectiveness of each individual improvement in the proposed WMD model, an ablation study was conducted by systematically adding or removing each enhancement component (WTConv, MaSA, and DIoU). This experimental setup allows for a clear assessment of the contribution of each module to the overall detection performance.

All training and experiments for the proposed model were conducted on a single NVIDIA RTX 3060 GPU. The model was trained using the Stochastic Gradient Descent (SGD) optimizer for 150 epochs, with an initial learning rate of 0.01 and a weight decay of 0.0005.

The results of the ablation study are summarized in Table II, demonstrating the incremental improvements achieved by integrating the proposed techniques. These findings provide strong evidence of the efficacy of the WMD architecture in addressing the complexity detection task in mechanical part drawings.

Incorporating WTConv into Model 1 results in a substantial drop in the parameter count (62,000) as well as a reduction in the overall number of floating-point operations (FLOPs). WTConv effectively contributes to model light-weighting without compromising detection performance. In Model 2, where MaSA is introduced, although there is a slight increase in the parameter count, the model achieves improved Recall (84.5%) and mAP@0.5 (85.6%), demonstrating the effectiveness of MaSA in enhancing feature representation and improving detection accuracy for complex mechanical parts. Furthermore, Model 3, which employs the DIoU loss function for bounding box regression, exhibits strong adaptability to the specific task. It preserves consistent mAP@0.5 performance, suggesting that DIoU is well-suited to handling the localization difficulties encountered in complexity detection for engineering drawings.

The combined use of the two improvement modules yields different levels of effectiveness. Model 4 (WTConv + MaSA) enhances Recall and mAP@0.5 while simultaneously lowering both parameter count and FLOPs. Nonetheless, its performance remains slightly inferior to Models 1 and 2, suggesting potential for further refinement. Model 5 (WTConv + DIoU) likewise delivers higher accuracy and further validates the complementary benefits of structural improvements with advanced regression. Although FPS declines because of the added computational burden, the model still achieves a lightweight design. In contrast, Model 6 (MaSA + DIoU) shows weak performance, likely because the two modules are not well aligned in the absence of WTConv. This indicates that WTConv is crucial for facilitating effective integration of multiple components.

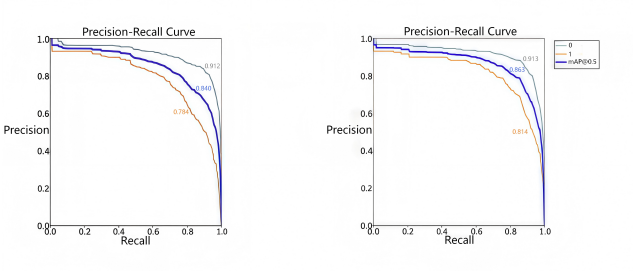

Fig. 4. The left figure illustrates the Precision-Recall (PR) curve of the YOLO11 model, while the right figure presents the PR curve of the proposed WMD model. The PR curve of YOLO11 model is almost entirely enclosed by the WMD's one, which indicates that the performance of the WMD model is superior.

The WMD model, which combines all three components (WTConv, MaSA, and DIoU), delivers the best overall results. Relative to the baseline, it increases precision by 3.6% and mAP@0.5 by 1.5%, while recall decreases only slightly (by 0.2%). In addition, it cuts the number of parameters by 0.042×10^6 and lowers the computational cost by 0.1 GFLOPs, although there is a moderate reduction in inference speed. These compromises are considered acceptable in light of the substantial improvements in detection accuracy and efficiency.

To validate the effectiveness of the improved model, we conducted repeated experiments on the WMD model. Across three independent trials, the model achieved a mean precision of 80.33% with a sample standard deviation of 0.81, a mean Recall of 81.67% with a sample standard deviation of 1.1, and a mean mAP@0.5 of 86.37% with a sample standard deviation of 0.21. These statistical results demonstrate the model's consistent and superior performance across multiple runs, confirming the efficacy of our proposed design.

B. Visualization Analysis

As illustrated in Fig. 4, the area under the PR curve for the WMD model is substantially larger than that of the baseline model, and in some intervals, the WMD curve completely encloses the baseline curve. This indicates that the improved model attains more favorable precision–recall trade-offs, demonstrating an overall improvement in detection performance for mechanical part complexity tasks. In addition, the PR curve of the WMD model is smoother than that of the baseline model, indicating greater stability and reliability over the entire detection process. As presented in Table III, WMD delivers generally better or on-par performance compared to the baseline, aside from a minor decrease in recall for

TABLE II
ABLATION STUDY OF EACH COMPONENT IN WMD. BEST VALUES ARE HIGHLIGHTED IN BOLD. THE SYMBOL 'Y' INDICATES THE ADDITION OF THE IMPROVEMENT.

Model	WTconv	MaSA	DIoU	P(%)	R(%)	mAP@0.5(%)	FPS	Params(10^6)	GFLOPS
YOLO11				77.2	83.0	84.8	322.6	2.583	6.3
Model1	Y			73.2	83.0	84.7	322.6	**2.521**	**6.2**
Model2		Y		76.6	**84.5**	85.6	312.5	2.602	6.3
Model3			Y	73.8	83.5	85.6	312.5	2.583	6.3
Model4	Y	Y		73.7	84.3	85.0	303.0	2.541	**6.2**
Model5	Y		Y	76.5	83.3	85.5	285.7	**2.521**	**6.2**
Model6		Y	Y	77.3	83.7	84.8	217.4	2.602	6.3
WMD	Y	Y	Y	**80.8**	82.8	**86.3**	222.2	2.541	**6.2**

TABLE III
COMPARISON RESULTS OF YOLO11 AND WMD.

Model	Label	P(%)	R(%)	mAP@0.5(%)
YOLO11	0	83.3	90.0	91.2
	1	71.1	76.1	78.4
WMD	0	84.5	88.8	91.3
	1	77.2	76.8	81.4

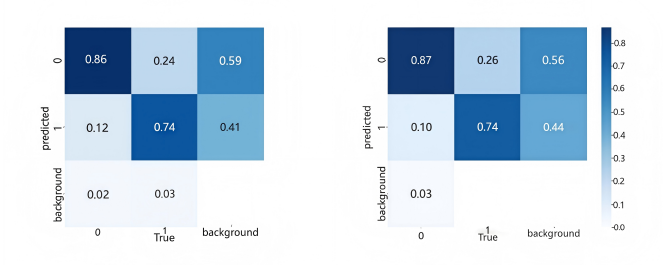

Fig. 5. The confusion matrices of the YOLO11 and WMD models are displayed in the left and right, respectively. Comparing whether diagonal elements are larger and non diagonal elements are smaller, it indicates that the performance of the WMD model is superior.

simple parts. Importantly, it substantially boosts precision on complex parts, directly tackling the primary shortcoming of the baseline. Precision for simple parts also rises, reflecting improved overall consistency. Collectively, these findings confirm the effectiveness of WMD in enhancing detection accuracy and robustness. By examining the confusion matrices in Fig. 5, it becomes apparent that the WMD model achieves marginally better performance along the diagonal compared to the baseline model. In general, the WMD model yields fewer misclassifications, which makes it more appropriate for real-world applications.

C. YOLO Series Algorithm Comparison Experiment

To evaluate the effectiveness and practical utility of our proposed improvements, we designed and conducted a series of comparative experiments. The results are reported in Table IV.

In the experiment, the WMD model is compared with other models trained using previous YOLO series algorithms. According to the results, the WMD model achieves the best

TABLE IV
COMPARISON OF YOLO SERIES MODEL EFFICIENCY. BEST VALUES ARE HIGHLIGHTED IN BOLD.

Model	P	R	mAP@0.5	FPS	Params	GFLOPs
yolov5n6u	68.7	**84.8**	81.8	**333.3**	**2.503**	7.1
yolov8	73.5	80.7	81.7	312.5	3.006	8.1
yolov10	76.5	81.0	84.5	294.1	2.695	8.2
YOLO11	77.2	83.0	84.8	322.6	2.583	6.3
WMD	**80.8**	82.8	**86.3**	222.2	2.541	**6.2**

overall performance, demonstrating superior precision (80.8%) and mAP@0.5 (86.3%). It also shows competitive performance in terms of recall, model parameters, and computational cost. These results validate the effectiveness of the proposed improvements. The detection outcomes of the model are presented accordingly.

V. LIMITATION

Although the proposed design improves the efficiency of object detection, certain limitations remain. For instance, the model may misidentify global assembly views as individual parts or inconsistently label different sections of the same part drawing with varying complexity levels. Future research will focus on enhancing the model's robustness alongside its performance to achieve more reliable results.

VI. CONCLUSION

For the automatic complexity evaluation of mechanical parts from engineering drawings, YOLO11 is adopted as the baseline model, as it satisfies the demands for fast detection, high accuracy, multi-angle analysis, robust anti-interference performance, and fault-tolerant correction. To strengthen feature extraction, enlarge the receptive field, and maintain a lightweight architecture, the WTConv, MaSA, and DIoU modules are integrated into the network. Following these enhancements and subsequent training, the WMD model is obtained. The effectiveness of the proposed modifications is demonstrated through ablation studies and comparative experiments. The detection outcomes are visually illustrated in Fig. 6 and Fig. 7.

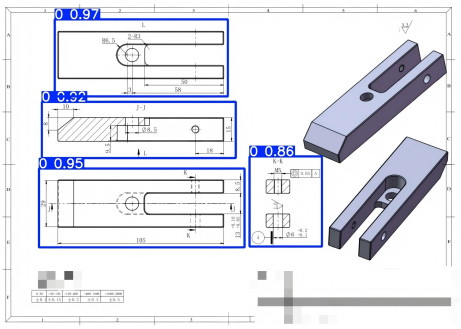

Fig. 6. An example of the model detecting simple mechanical parts. The '0' on the box is the label (The part is the simple one), and the number after it represents the confidence level.

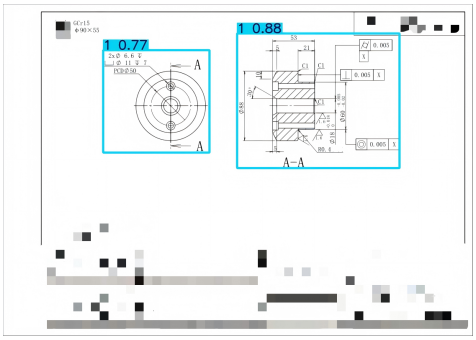

Fig. 7. An example of the model detecting simple mechanical parts. The '1' on the box is the label (The part is the complex one), and the number after it represents the confidence level.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (62206196) and the Fundamental Research Program of Shanxi Province (Jointly Supported, Transportation, 202403011222002)

REFERENCES

[1] M. Ghobakhloo, "Industry 4.0, digitization, and opportunities for sustainability," *Journal of cleaner production*, vol. 252, p. 119869, 2020.

[2] Y. Liu, S. Duan, Z. Shen, Z. He, and L. Li, "Grasp and inspection of mechanical parts based on visual image recognition technology," *Journal of Theory and Practice of Engineering Science*, vol. 3, no. 12, pp. 22–28, 2023.

[3] L. Yu, J. Zhu, Q. Zhao, and Z. Wang, "An efficient yolo algorithm with an attention mechanism for vision-based defect inspection deployed on fpga," *Micromachines*, vol. 13, no. 7, p. 1058, 2022.

[4] H. Chen, Z. He, B. Shi, and T. Zhong, "Research on recognition method of electrical components based on yolo v3," *IEEE Access*, vol. 7, pp. 157 818–157 829, 2019.

[5] H. Zhu, Z. Li, Y. Lang, Y. Li, and Y. Liu, "Research on yolo-based component detection algorithms," *Scientific Journal of Technology*, vol. 7, no. 1, pp. 88–97, 2025.

[6] L. Jiang, B. Yuan, Y. Wang, Y. Ma, J. Du, F. Wang, and J. Guo, "Ma-yolo: a method for detecting surface defects of aluminum profiles with attention guidance," *IEEE Access*, vol. 11, pp. 71 269–71 286, 2023.

[7] X. Wang, Z. Jing, L. Shi, G. Cheng, Y. Gao, and B. Hao, "Ma-yolo: A lightweight vehicle detection framework based on yolo," in *2023 International Conference on Artificial Intelligence and Automation Control (AIAC)*. IEEE, 2023, pp. 277–281.

[8] F. Mushtaq, K. Ramesh, S. Deshmukh, T. Ray, C. Parimi, P. Tandon, and P. K. Jha, "Nuts&bolts: Yolo-v5 and image processing based component identification system," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105665, 2023.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[10] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[11] S. E. Finder, R. Amoyal, E. Treister, and O. Freifeld, "Wavelet convolutions for large receptive fields," in *European Conference on Computer Vision*. Springer, 2024, pp. 363–380.

[12] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[13] M. Tan and Q. V. Le, "Mixconv: Mixed depthwise convolutional kernels," in *British Machine Vision Conference (BMVC)*, 2019.

[14] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[15] S. E. Finder, Y. Zohav, M. Ashkenazi, and E. Treister, "Wavelet feature maps compression for image-to-image cnns," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 592–20 606, 2022.

[16] R. Gal, D. C. Hochberg, A. Bermano, and D. Cohen-Or, "Swagan: A style-based wavelet-driven generative model," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–11, 2021.

[17] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1689–1697.

[18] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[19] Q. Fan, H. Huang, M. Chen, H. Liu, and R. He, "Rmt: Retentive networks meet vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 5641–5651.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[21] Q. Fan, H. Huang, J. Guan, and R. He, "Rethinking local perception in lightweight vision transformer," *arXiv preprint arXiv:2303.17803*, 2023.

[22] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 175–12 185.

[23] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[25] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 32–42.

[26] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31.

[27] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, and A. Yuille, "Lite vision transformer with enhanced self-attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 998–12 008.

[28] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei, "Retentive network: A successor to transformer for large language models," 2024. [Online]. Available: https://openreview.net/forum?id=UU9Icwbhin

[29] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.