Attention Mechanism and Novel Inspection Heads for Steel Surface Defect Detection

Zhen Yang

***Abstract*—In modern industrial production, steel is one of the most widely used basic materials, with its quality being directly related to the performance and safety of many industries. A method for detecting steel surface defects based on an improved You Only Look Once 11 (YOLO11) approach is proposed to address issues of low detection accuracy, slow detection speed, and inadequate feature extraction found in traditional methods. Firstly, to enhance the stability of model training and improve feature extraction, we substitute the self-attention module of the Inverted Residual Mobile Block (iRMB) with Cascaded Group Attention (CGA). Secondly, the Large Kernel Attention Design (LSKA) mechanism is introduced to optimize the backbone network and enhance its capability to capture multi-scale features. Finally, the L-Head inspection head is designed to enhance bounding box localization accuracy and improve small defect detection accuracy by implementing the concept of Distribution Focal Loss (DFL) regression. Experimental results on the NEU-DET defective dataset indicate that the improved model achieves an accuracy of 77.1%, a mAP50 of 78.2%, and an FPS of 312.5. In comparison to other models, the improved model strikes a good balance between detection accuracy, computational efficiency, and inference speed, successfully meeting real-time speed requirements.**

***Index Terms*—steel surface defect detection, YOLO11, attention mechanism, inspection head.**

I. Introduction

DETECTING defects on steel surfaces is essential for ensuring the quality of steel. Traditional inspection methods have significant limitations, while the YOLO algorithm, based on deep learning, offers new opportunities in this area. Defects on steel surfaces affect not only the visual quality of the material but can also lead to serious issues during subsequent processing and use. These defects can reduce structural strength, cause fatigue cracks to expand, and ultimately threaten the reliability of the entire engineering structure. Therefore, efficient and accurate detection of defects on steel surfaces is crucial for ensuring steel quality, enhancing production efficiency, and maintaining safety in industrial production.

Traditional methods for detecting surface defects in steel, including manual visual inspection, eddy current detection, and ultrasonic testing, have been effective. However, they possess significant limitations. Manual visual inspection is prone to subjective factors, which can lead to low efficiency and the risk of overlooking issues. Eddy current and ultrasonic detection technologies require highly skilled inspectors, and the equipment can be quite expensive. Additionally, these methods may not be effective when detecting complex shapes or small defects. With the rapid advancement of computer technology and artificial intelligence algorithms, deep learning-based target detection technology offers new solutions for detecting defects on steel surfaces. Among the algorithms, the YOLO algorithm has emerged as a research hotspot in this field due to its fast detection speed and high accuracy.

However, most existing research focuses on applying the YOLO algorithm in various fields, while there are relatively few optimizations and improvements specifically for detecting steel surface defects. In addition, the defects on steel surfaces come in various forms and are influenced by light, noise, and other factors, which impose greater demands on the detection capabilities of the YOLO algorithm. Therefore, exploring the application of the YOLO algorithm for detecting steel surface defects has both theoretical value and practical significance. Additionally, proposing an improvement plan for addressing the existing issues is essential. For example, Zhao et al. [1] proposed a YOLO model. It uses Res2Net blocks to expand the receptive field and extract multi-scale features, as different-sized defects can thus be better detected. Additionally, a network is integrated, which not only deepens the network but also reuses low-level features to generate rich feature representations for identifying small or subtle defects. Moreover, the model separates regression and classification tasks via decoupled heads, enabling independent optimization and leading to a more accurate detection of steel surface defects. Wang et al. [2] proposed a network. In order to efficiently acquire the multi-scale information associated with surface defects, the inspection network features a specially designed multi-scale exploration module that enhances detection performance. Additionally, a spatial attention mechanism has been introduced to help the network focus more on defect information. Ma et al. [3] developed a model for detecting defects in steel. This model employs a distinctive shunt feature network architecture and a self-rectifying transmission allocation approach to boost its performance. The network is specifically designed to handle classification and localization tasks based on varying computational needs. Additionally, it employs a self-correcting criterion that incorporates adaptive sampling and dynamic label assignment. This method makes the most of high-quality samples. It aids in regulating the data distribution and fine - tuning the training process. Lu et al. [4] proposed a WSS-YOLO model, which is based on YOLOv8. This model employs a dynamic non-monotonic focusing mechanism that utilizes WIoU loss to concentrate on anchor frames of average quality, ultimately enhancing the overall performance of the detector. Additionally, they designed a C2f-DSC module that incorporates dynamic serpentine convolution, allowing the model to adaptively adjust its sensory field. Xia et al. [5] introduced an YOLOv5s model. A reparameterized

Manuscript received March 18, 2025; revised July 20, 2025.

This work was supported by the Science and Technology Research Project of Jiangxi Provincial Department of Education (Grant No. GJJ2402903), and the Teaching Reform and Research Project of Jiangxi University of Engineering (Grant No. 2024-JGJG-56).

Zhen Yang is a lecturer of College of Smart Industry, Jiangxi University of Engineering, Xinyu 338000, PR China (e-mail: yangzhenccs@163.com).

large convolutional kernel C3 module was put forward with the goal of expanding the model's effective receptive field and strengthening its feature extraction capabilities in the face of complex texture interference. Moreover, a feature fusion architecture incorporating a multi-path spatial pyramid pooling module was constructed to accommodate the scale changes of steel surface defects. Yuan et al. [6] proposed for the detection of multi-category steel defects. This model builds upon and improves the YOLOv8n architecture. It integrates the DCNV2 module to attain an adaptive receptive field. Additionally, in the C2f module, a channel attention mechanism is included. This mechanism emphasizes valuable features and, at the same time, reduces the quantity of parameters. Li et al. [7] introduced a simulation teaching approach that integrates deep learning. This approach spans the entire spectrum, including data pre-treatment, model training, validation assessment, and innovative refinement. The objective is to bolster students' grasp of AI technology and elevate their proficiency in applying it to real-life situations. Li et al. [8] put forward an efficient and highly precise algorithm named DEW-YOLO for detecting surface defects in strip steel. This algorithm capitalizes on the merits of deformable convolutional networks (DCNs). It makes an innovation to the C2F module in YOLOv8 by introducing the C2f_DCN module. This newly-introduced module is capable of flexibly sampling features, thereby enhancing the model's capacity to learn and represent the characteristics of defects with different sizes and shapes. Li et al. [9] developed a dataset containing six types of surface defects in cold-rolled steel strips specifically for defect detection. In order to prevent overfitting, they augmented this dataset. Moreover, they optimized the YOLO network to convert it into a fully convolutional network. The improved network is composed of 27 convolutional layers, providing a complete end-to-end approach for detecting surface defects in steel strips.

The YOLO algorithm, a well-known deep learning model for target detection, revolutionizes the traditional detection process by integrating classification and localization into an end-to-end, real-time target detection system. In the context of steel surface defect detection, images of steel surfaces are often significantly affected by lighting, contrast, and other factors. As a result, conventional normalization methods struggle to respond effectively and fail to meet actual detection needs. To address this issue, this paper introduces an innovative steel surface defect detection method based on an improved version of YOLO. For the backbone network, this paper replaces the self-attentive module of iRMB in [10] with the CGA module from [11], and provides a significant enhancement to the C3K2 module. It effectively improves the stability of the model during the training process, enabling it to learn and extract features more robustly. To improve the model's ability to extract multi-scale features, this paper skillfully introduces the LSKA attention mechanism into the optimized Spatial Pyramid Pooling-Fast (SPPF) module. This enhancement allows the model to more effectively capture steel surface defect features across different scales, significantly increasing the accuracy and comprehensiveness of the detection process. This paper addresses the issue of low detection accuracy in existing models by elaborately designing the L-Head inspection head. This detection head integrates a regression design concept based on the DFL, which enhances its ability to accurately position bounding boxes. It excels particularly in detecting small defects on the surface of steel, significantly improving the detection accuracy for these minor issues. This advancement provides strong support for the precise control of steel surface quality.

In summary, the main contributions of this study are summarized below:

(1) There are significant differences in lighting conditions, contrast, and other factors in the images. To address this, we replaced the self-attention module of the iRMB with CGA, which enhances the model's stability during training. This change allows the model to learn and extract features more effectively.

(2) To improve the model's ability to extract features at multiple scales, we introduce the LSKA attention mechanism into SPPF. LSKA addresses the limitations of SPPF in capturing multi-scale information for complex defects. It accurately adjusts the feature weights for defects of various sizes, allowing the model to effectively retain key details. This is especially beneficial when detecting tiny defects, as the enhanced small-scale feature weights enhance the model's performance.

(3) This paper addresses the issue of low detection accuracy in existing models by designing the L-Head inspection head. This detection head integrates the DFL-based regression design concept, enabling precise localization of steel surface defects. It is particularly effective in detecting small defects, significantly improving detection accuracy for these types of imperfections.

II. Related work

The technology for detecting defects on steel surfaces has progressed alongside industrial advancements. Initially, inspections relied heavily on manual visual evaluation. Over time, techniques based on traditional image processing and machine learning began to emerge. More recently, the rise of deep learning technology has led to significant breakthroughs in this area, particularly with the increasing application of the YOLO algorithm for steel surface defect detection.

A. Traditional methods

In the past [12], detecting defects on steel surfaces primarily depended on manual visual inspection and traditional image processing methods, including machine learning techniques. Manual inspection involves workers using their eyes to directly examine the steel surface for defects such as scratches, holes, and cracks. While this approach is flexible, it is also inefficient and heavily influenced by the subjectivity of the inspectors. Fatigue from long hours of work can lead to oversight and errors in detection. For example, traditional image processing techniques such as gray-scale transformation, filtering, and edge detection are commonly employed to preprocess steel surface images and extract features. For instance, Gaussian filtering is employed to eliminate image noise, while the Canny operator is used to detect edges and highlight defect contours. The identified defective regions are subsequently isolated and recognized through morphological operations. However, these methods struggle in complex backgrounds and with tiny defects. Additionally, feature extraction often depends on numerous manual designs, which can lack strong generalization capabilities.

Traditional machine learning methods rely on image processing techniques that involve manually extracting features such as texture and shape. These features are then used with classifiers like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) for defect classification. For instance, when using SVM, it is crucial to carefully choose the appropriate kernel function and parameters in order to construct a hyperplane that can effectively differentiate between normal and defective samples. Feature engineering is challenging and struggles to address the various and complex defects found on steel surfaces. Additionally, model performance can be limited by the quality of features. For example, Tang et al. [13] provided a comprehensive overview of image processing algorithms used for detecting surface defects in steel products. This includes steps such as image preprocessing, region of interest (ROI) detection, segmentation of ROI images, feature extraction and selection, and classification of defects.

B. Detection method based on the YOLO algorithm

The YOLO algorithm is a well-established deep learning model used for object detection. Unlike traditional detection methods that execute classification and localization in separate steps, YOLO achieves real-time object detection through an end-to-end approach [14]. This ability has attracted considerable interest in the field of steel surface defect detection. Numerous scholars have enhanced the YOLO algorithm to satisfy the requirements of steel surface defect detection. Li et al. [15] put forward an optimized network featuring high-speed and high-precision performance without augmenting the overall model size. Generally, adjustments made to enhance the feature extraction capabilities of shallow networks tend to have a negative impact on the model's inference speed. To overcome this challenge and maintain an equilibrium between detection accuracy and speed, they integrated an enhanced Fusion-Faster module into the YOLOv7 backbone network. The module utilizes partial convolution (PConv) as its core operator. By doing so, it not only improves the feature extraction ability of the shallow network but also ensures that the inference speed remains unaffected. Guo et al. [16] presented an enhanced MSFT-YOLO model derived from a single-stage detector. In this model, a TRANS module, crafted based on the Transformer architecture, is embedded in both the backbone network and the detection head. Thanks to this, the model can blend local features with global information. Additionally, it utilizes a multi-scale feature fusion architecture that merges features at different scales. This significantly enhances the detector's adaptability to targets of diverse sizes.

III. Steel surface defect detection improvement methods

YOLO is a cutting-edge target detection algorithm that has gained popularity in the field of computer vision due to its speed and efficiency. YOLOv5, YOLOv8, and YOLO11 are part of the same technological development system. Compared to the YOLOv8 model, YOLO11 features a significant adjustment at the module level, as it replaces the C2F module with the C3K2 module. The C3K2 module has a distinctive configuration of convolutional kernels and connections, allowing the network to capture and integrate feature information with greater accuracy. This characteristic is extremely beneficial for subsequent target detection tasks and other vision applications, and it is anticipated to further enhance detection performance and the effectiveness of visual task processing. An improved overall structure is illustrated in Fig. 1.

A. C3k2-iRMB_CGA module

In the task of detecting defects on steel surfaces, there are significant variations in lighting conditions and image contrast, among other factors. Implementing a normalization layer can help the model adapt to these variations more effectively and enhance its generalization capability. One core idea of the inverted residual block, iRMB, is to integrate a lightweight Convolutional Neural Network (CNN) architecture with a modeling structure based on the attention mechanism, in order to create efficient mobile networks. The iRMB builds upon the iRB for CNNs by adapting its principles for attention-based models. It reevaluates the key components of the Inverted Residual Block (IRB) and the Transformer, aiming to create a unified approach. This design focuses on maximizing the efficient use of computational resources while achieving high accuracy, all while keeping the model lightweight.

CGA enhances diversity in features fed to the attention head, as illustrated in Fig. 2. CGA differs from traditional self-attention by offering unique input segmentations for each attention head and cascading the output features across these heads. This method reduces computational redundancy in multi-head attention and enhances the model's capacity by increasing network depth, leading to a more efficient and powerful model. The CGA specifically divides the input features into different sections, These are then allocated to each attention head. Every head computes its respective self-attention mapping. Subsequently, the outputs of all heads are aggregated and projected back to the input dimension via a linear layer. CGA enhances the model's computational efficiency without introducing extra parameters. Furthermore, each head's output is incorporated into the next head's input through a sequential process, progressively refining the feature representation.

In this study, we enhance the C3k2 module by substituting the self-attention module of iRMB with the CGA module, as illustrated in Fig. 3. The improved model offers several advantages: it enhances stability. Normalization accelerates convergence, reduces internal covariate bias, and promotes training stability. In the scenario of detecting defects on steel surfaces, the normalization layer can help the model better handle variations in lighting, contrast, and other factors, thereby enhancing its generalization ability. Secondly, involves optimizing nonlinear feature extraction. This process allows the model to learn more intricate feature representations. Because steel surface defect features often exhibit nonlinear characteristics, the activation function plays a crucial role in helping the model capture these complex features, ultimately enhancing the accuracy of defect detection.

The input data first passes through a 1×1 convolutional layer, which adjusts the number of channels and integrates

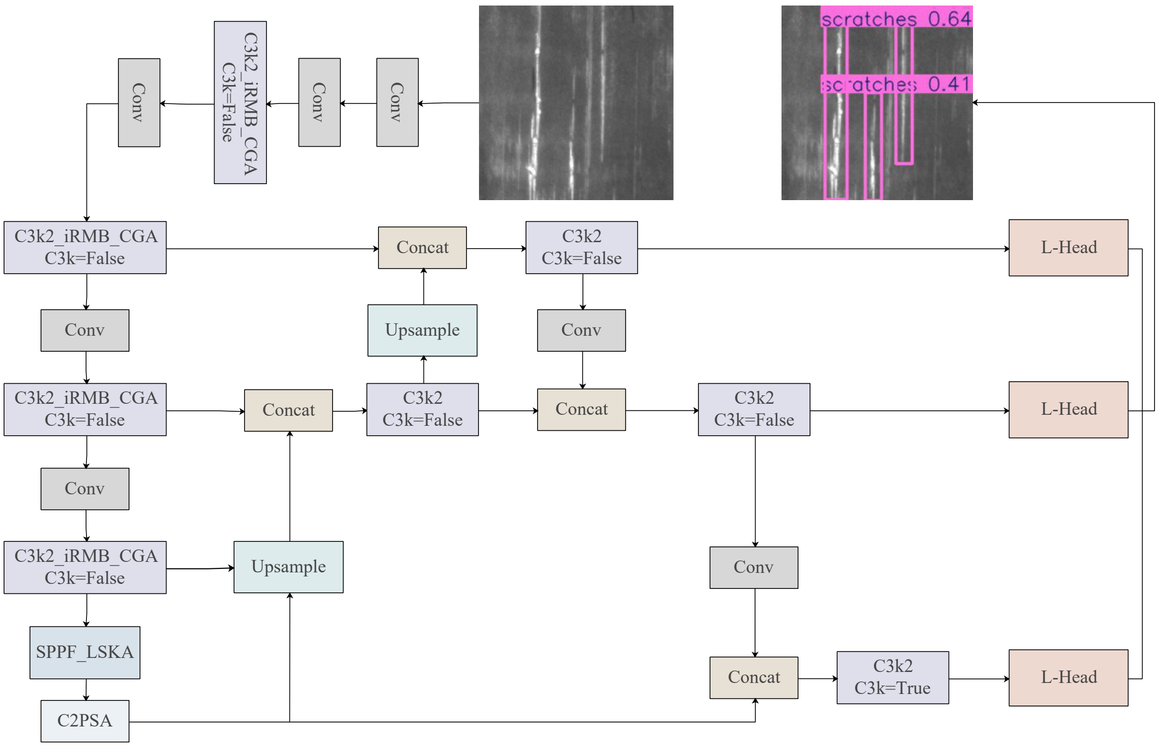

Fig. 1: The improved YOLO11 algorithm.

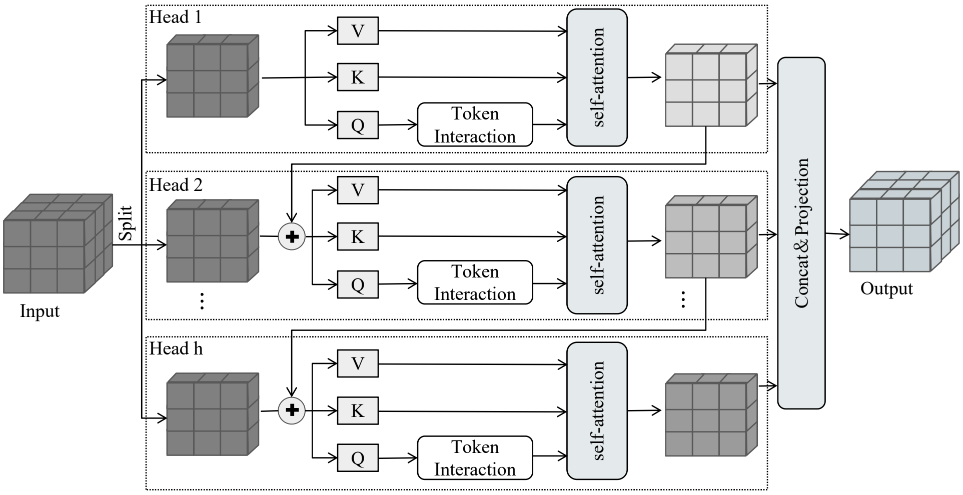

Fig. 2: Cascaded Group Attention.

information across channels to reduce subsequent computations. Next, the data enters a 3×3 deep convolutional layer, where local features are efficiently extracted by convolving each channel independently. Finally, the data goes through another 1×1 convolutional layer to further adjust the channel dimensions and integrate the features. After that, the output from the second 1×1 convolutional layer is combined with the output from the first 1×1 convolutional layer through a skip connection. This approach helps to fuse different levels of features and addresses the issue of gradient vanishing. Finally, the fused features are processed through the CGA cascade group attention module, where the attention mechanism assigns weights based on the significance of the features. This enables the model to focus on the most important features, thereby enhancing its expressive capabilities.

B. The SPPF-LSKA module

The SPPF module primarily achieves feature fusion through maximum pooling operations at multiple scales. However, steel surface defects exhibit a wide range of intricate and diverse features. This complexity can make it challenging for the module to effectively capture and fuse the various types of multi-scale information present in complex defect scenarios. When it comes to tiny defects, large-scale pooling operations often result in the loss of important details. On the other hand, for larger defects, small-scale pooling struggles to effectively capture the overall features. As a consequence, the final fused features fail to accurately represent complex defect scenarios, which negatively impacts detection accuracy.

Unlike traditional methods, the LSKA attention mechanism applied to multi-scale feature integration, as illustrated

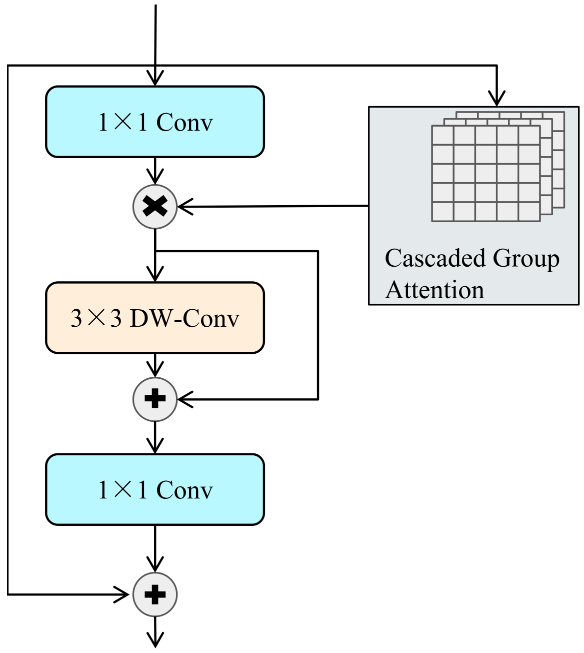

Fig. 3: The iRMB_CGA module.

in Fig. 4, can adaptively assign different weights to features at various scales. This capability effectively enhances the feature information that is crucial for defect detection. For instance, when addressing small-sized defects, LSKA increases the emphasis on small-scale features, prompting the model to concentrate more on these details. Conversely, when handling large-sized defects, it highlights large-scale features, enabling the model to make more efficient and accurate judgments when dealing with defects of varying sizes.

As shown in Fig. 5, the LSKA module is integrated into the SPPF and constructed as the SPPF-LSKA module. In this module, the original K×K convolution operation is re-disassembled. The process begins with a decomposition into a $(2d-1)$×$(2d-1)$ depth convolution, a K/d×K/d cavity depth convolution, and a 1×1 convolution. Next, both the 2D depth convolution and the cavity convolution are further refined and broken down into 1D horizontal and vertical convolutions. Finally, these decomposed convolution kernels are cascaded sequentially to complete the entire computational process. This module can significantly enhance the effectiveness of multi-scale feature fusion. LSKA can compensate for the shortcomings of SPPF in capturing multi-scale information related to complex defects. It accurately adjusts the feature weights for defects of different sizes, allowing the model to effectively retain key details when detecting small defects by enhancing the importance of small-scale features. Conversely, when identifying large defects, LSKA emphasizes large-scale features to capture the overall characteristics in a comprehensive manner. This fusion greatly enhances the model's capability to identify complex defects on steel surfaces. As a result, it significantly improves detection accuracy and ensures the efficient and precise execution of steel surface defect detection tasks. This advancement provides more reliable technical support for quality control in industrial steel production.

C. L-Head module

Small surface defects on steel can often be overlooked during inspection. To address this issue, we designed the L-Head inspection head using a DFL regression design. The structure of the L-Head inspection head is depicted in Fig. 6. Firstly, the L-Head allows for more precise bounding box positioning. The shapes and sizes of defects on the steel surface exhibit diverse characteristics, and accurate bounding box positioning is crucial for properly assessing the severity of these defects. DFL redefines the bounding box regression problem as a distribution learning problem. This approach enables more precise localization of bounding boxes by predicting the probability distribution for each potential location. In comparison to traditional bounding box regression methods, DFL offers improved accuracy in identifying the boundaries of steel surface defects. The L-Head plays a crucial role in enhancing the detection accuracy of small defects. Its distribution modeling approach, known as DFL, improves the model's ability to identify small targets. When it comes to detecting minor imperfections on steel surfaces, the L-Head can more accurately predict both the location and size of these defects, thanks to DFL. This significantly boosts the overall detection accuracy for small defects.

IV. Tests and analysis

A. Datasets source

The dataset chosen for this study is the NEU-DET dataset, an open-source resource provided by Northeastern University. This dataset is essential for researching surface defect detection algorithms for steel plates. It encompasses a variety of complex situations that are commonly encountered in industrial production environments. This includes typical types of steel surface defects such as Scratches, Patches, Inclusions, Rolled-in_scale, Pitted_surface, and Cracking. As illustrated in Fig. 7.

B. Test environment and parameter configuration

We implemented our environment with VSCODE, all running on Ubuntu. The software stack utilized included Python 3.10, PyTorch 2.2. The hyperparameter configurations for the model are presented in Table I.

TABLE I: The configuration of our model.

Hyperparameter	Value
optimizer	sgd
batch Size	32
epoch	200
weight Decay	0.005
learning rate	0.01

C. Model evaluation indexes

In this paper, we select several metrics for model evaluation, including Precision (P), Recall (R), mean Average Precision (mAP), Frames Per Second (FPS), GFLPOs, and Parameters. The optimal values for each metric are highlighted in bold in the table. Precision measures the proportion of positive cases that the model predicts correctly, as indicated in Eq. (1). Recall, assesses how well the model predicts

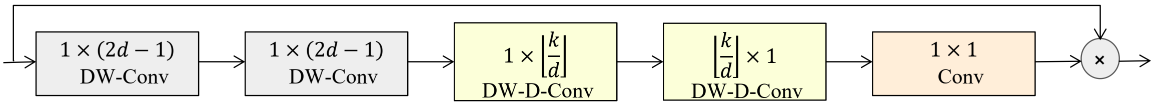

Fig. 4: The LSKA module.

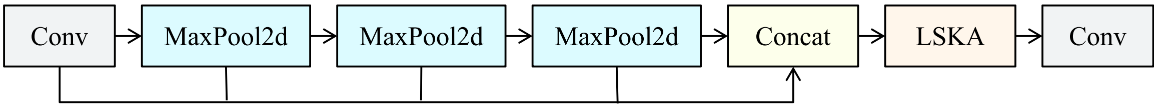

Fig. 5: The SPPF-LSKA module.

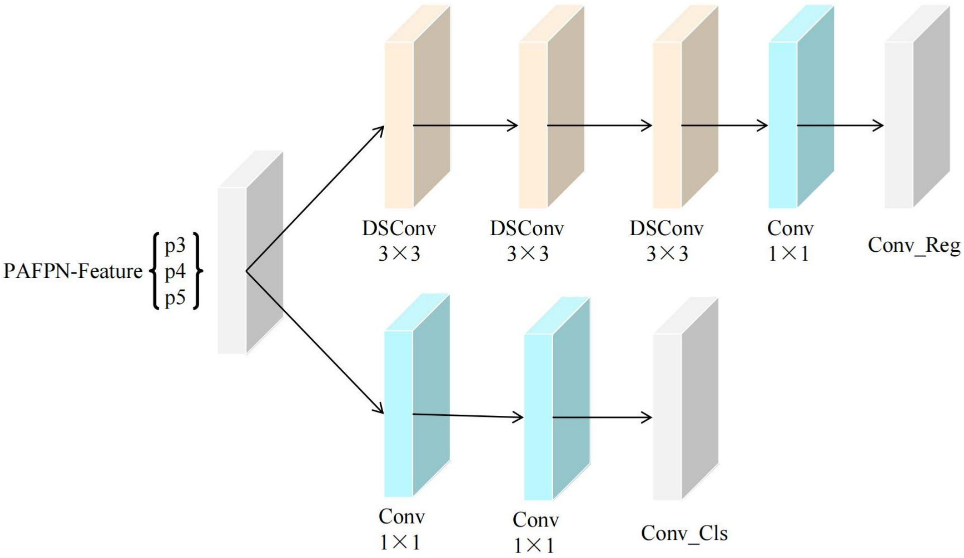

Fig. 6: The L-Head module.

actual positive cases, as shown in Eq. (2). Together, Precision and Recall provide a comprehensive assessment of both the accuracy and completeness of the model's classification. The mAP thoroughly evaluates the weighted average of precision rates across various recall rates in target detection tasks. It provides a detailed measure of the model's overall performance in each detection category. The calculation formula can be found in Eq. (3).

Frames per second (FPS) indicates the number of frames an algorithm processes in one second. It is a crucial metric for assessing the efficiency and responsiveness of algorithms in real-time applications. The calculation formula can be found in Eq. (4). Giga Floating Point Operations (GFLPOs) measures the computational load of the model during operation, while parameters (Params) represent the size of the model. Together, these two metrics offer insights into the model's complexity and resource requirements. By evaluating performance from these different perspectives, they provide valuable support for model optimization and assessment.

$$P = \frac{tp}{tp + fp} \tag{1}$$

$$R = \frac{tp}{tp + fn} \tag{2}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{3}$$

$$FPS = \frac{1}{T_{total}} = \frac{1}{T_{pre} + T_{det} + T_{post}} \tag{4}$$

D. Comparative experiments with different models

To compare the efficiency of the improved algorithms presented in this paper, we selected the two-stage algorithm Faster R-CNN and the one-stage algorithms YOLOv5s and YOLOv6n for our comparison experiments. The results of these experiments are summarized in Table II, and some visualization results are displayed in Fig. 8.

To further investigate the model's detection performance in various scenarios, we conducted a systematic evaluation of its effectiveness in identifying different types of steel surface defects. The results are detailed in Table IV. This data visualization emphasizes the model's strengths in various defect detection scenarios while clearly outlining its limitations. These findings provide crucial insights for the subsequent optimization and improvement of the model, serving as a key foundation for enhancing its performance.

Our improved model, which incorporates the C3K2-iRMB_CGA module, the SPPF-LSKA module, and the L-Head module, demonstrates strong performance in terms of accuracy. As shown in Table III, our model achieves an accuracy of 77.1%. This is significantly higher than the second-order algorithm, Faster R-CNN, which achieves an accuracy of 67.4%. It also outperforms other first-order algorithms, such as YOLOv8n, which has an accuracy of

TABLE II: Comparison experiments of different models.

Models	P	R	mAP50	mAP50-95	Params(M)	GFLOPs	FPS
Faster R-CNN	0.674	0.685	0.722	0.407	43.5	183.5	46.8
YOLOv5s	0.755	0.732	0.730	0.432	**5.9**	**2.4**	259.4
YOLOv6n	0.718	0.712	0.723	0.423	11.6	4.5	295.3
YOLOv8n	0.749	**0.758**	0.757	0.435	8.1	3.0	218.7
YOLOv10n	0.696	0.688	0.713	0.417	6.7	2.7	245.2
YOLO11	0.711	0.741	0.780	0.442	6.3	2.6	**303.0**
ours	**0.771**	0.717	**0.782**	**0.448**	7.8	2.7	**312.5**

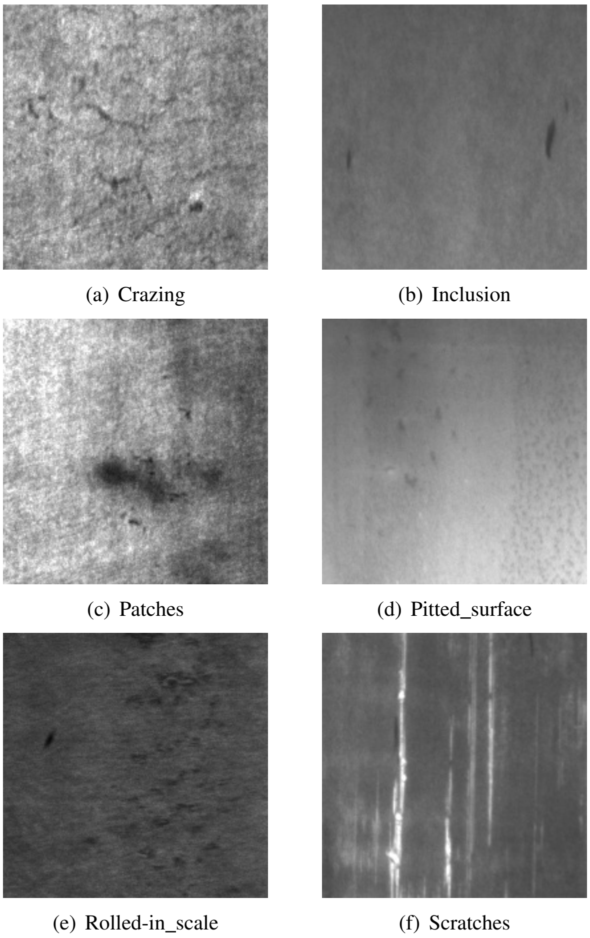
(a) Crazing
(b) Inclusion
(c) Patches
(d) Pitted_surface
(e) Rolled-in_scale
(f) Scratches

Fig. 7: Steel surface defects.

74.9%. The C3K2-iRMB_CGA module enhances model stability and optimizes nonlinear feature extraction. The SPPF-LSKA module compensates for the limitations of the original SPPF module in capturing multi-scale information about complex defects and can accurately adjust feature weights for defects of various sizes. The L-Head module integrates a regression design concept based on DFL, providing strong capabilities for bounding box localization. The combined effect of these modules significantly increases the model's accuracy in predicting positive samples, thereby reducing the likelihood of misjudgment. Moreover, the R of our improved model is 71.7c, which is slightly lower than YOLOv8n's recall of 75.8%. However, it remains at a high level and is sufficient for actual detection purposes. Although the L-Head module alters feature processing to some degree, which may limit the extraction of recall-related features, the strengths of the other modules compensate for this shortcoming. As a result, the overall recall rate remains high. The mAP50 and mAP50-95 scores of 78.2% and 44.8% exceed those of other algorithms. This improvement is due to enhancement modules that ensure stable and accurate target detection, while performing well across different overlap levels.

In terms of computational metrics, the GFLPOs for Faster-RCNN reach as high as 183.5G, with the number of parameters totaling 43.5M. These figures are significantly higher than those of other models, indicating that Faster-RCNN requires substantial computational resources during operation. In contrast, our improved model achieves GFLPOs of 7.8G and has 2.7M parameters. While this is slightly more than YOLOv5s, which has GFLPOs of 5.9G and 2.4M parameters, our model demonstrates a clear advantage in computational efficiency, meeting the demands of real-time detection.

Notably, the CGA mechanism in the C3K2-iRMB_CGA module greatly enhances the model's computational efficiency without adding extra parameters. This optimization is crucial in maintaining a balance between overall computational volume and efficiency.

In summary, our enhanced model attains an excellent equilibrium between detection accuracy, computational efficiency and inference speed, and has significant advantages over other comparative models, with high application value and potential in the field of target detection.

As shown in Fig. 8, YOLO11 faces several detection challenges. It has issues with repeated detections for surface defects such as Crazing, Inclusion, and Patches. Additionally, it misses detections for the Pitted Surface defect and misidentifies defects like Rolled-in Scale and Scratches. These problems stem from YOLO11's unoptimized model structure, which struggles to accommodate the complex and varied characteristics of steel surface defects.And our enhanced algorithm successfully identified all of them accurately.

While our model exhibits lower detection accuracy specifically for the Rolled-in scale defect, it remains effective in accurately detecting the target overall. It is important to note that the accuracy for detecting other types of defects is significantly higher. This improved performance can be attributed to the synergy between the modules, which enhances the model's ability to identify and locate a variety of steel surface defects. This capability provides a strong assurance for the quality inspection of steel surfaces.

E. Ablation experiments

In target detection, enhancing model performance typically depends on the combined effectiveness of its individual components. To thoroughly examine the specific effects of the improved C3K2-iRMB_CGA module, the SPPF-LSKA, and the L-Head module on the YOLO11 model's performance, we conducted model ablation experiments. The results of these experiments are presented in Table III.

TABLE III: Ablation experiments.

Yolo11	C3K2-iRMB_CGA	SPPF-LSKA	L-Head	P	R	mAP50	mAP50-95	FPS
✓				0.711	0.741	0.780	0.442	303.0
✓	✓			0.729	0.731	0.778	0.447	**500.0**
✓		✓		0.736	**0.742**	0.764	0.444	250.0
✓			✓	0.725	0.731	0.781	0.447	208.3
✓	✓	✓		0.714	0.715	0.771	0.437	93.4
✓	✓		✓	0.677	0.720	0.769	0.431	476.1
✓		✓	✓	0.690	0.729	0.774	0.443	84.7
✓	✓	✓	✓	**0.771**	0.717	**0.782**	**0.448**	312.5

Table III presents the results from a series of experiments conducted with the YOLO11 model. The first set of experiments serves as a benchmark using the YOLO11 base model without any additional components. This benchmark achieved a P of 71.1%, a R of 74.1%, a mAP50 of 78.0%, a mAP50-95 of 44.2%, and a processing speed of 303.0 FPS.

The introduction of the iRMB_CGA module leads to a 1.8% improvement in P and a 0.2% increase in mAP50-95, despite a slight decline in R and mAP50. Notably, the FPS improved significantly by 197. This enhancement is primarily due to the CGA mechanism, which optimizes the model's computational efficiency without adding additional parameters. When the SPPF-LSKA module is added alone, the P increases by 2.5%, and the R rises by 0.1%. This improvement occurs because the LSKA module effectively compensates for the limitations of the SPPF in capturing multi-scale information related to complex steel defects. By accurately adjusting the feature weights for steel defects of different sizes, the model can better preserve key details when detecting small defects. The enhanced small-scale feature weights contribute to improved detection accuracy. In the experiment where the L-Head module was introduced alone, the recall decreased by 1%. This may be due to the fact that L-Head changes the feature processing and has some limitations in extracting features related to target detection recall. At the same time, the FPS was reduced due to the fact that L-Head increased the computational complexity of the model. However, it is noteworthy that the P, mAP50, and mAP50-95 show a small improvement.

The final set of experiments includes all the enhanced components. Although the recall rate decreased by 2.4%, the precision rate improved significantly by 6%, along with slight improvements in other indicators. In summary, the improved method proposed in this paper outperforms the YOLO11 base model in terms of overall performance. It not only enhances the model's detection accuracy but also optimizes detection speed to a certain extent, allowing for more efficient and accurate detection of steel surface defects.

As shown in Table IV, there is a significant difference in the performance of the models for detecting steel surface defects. Our improved model achieved the highest overall performance, with a mAP of 78.2% and the top accuracy in detecting Crazing, Pitted_surface, and Scratches defects, with rates of 51.6%, 83.6%, and 96.8%, respectively. This indicates a strong detection capability. The YOLO11 model follows closely, with a mAP50 of 78.0%. It demonstrates a clear advantage in detecting Inclusion, Patches, and Rolled-in_scale defects, achieving accuracies of 88.7%, 93.9%, and 67.2%, respectively. In contrast, other models like Faster R-CNN and YOLOv5s do not perform as well as these two in terms of multi-class defect detection and overall performance.

F. Comparative experiments with different detection heads

To validate the effectiveness of the proposed detection head, L-Head, we selected several leading detection heads for comparison, including LSDECD [17], LAWDS [18], dyHead [19], and LSCD [20]. We conducted a comprehensive and rigorous comparison experiment focused on key performance indicators. The goal was to evaluate the performance of these different detection heads from multiple perspectives, including detection accuracy and inference speed. The results are shown in Table V.

In Table V of the experimental results, the L-Head (ours) achieves a P of 72.5%. This surpasses the performance of other detection heads, including LSDECD at 69.5%, LAWDS at 70.4%, LSCD at 70.5%, and dyHead at 72.3%. This indicates that the L-Head has a high accuracy in predicting positive samples. Regarding the R index, the L-Head reaches 73.1%, which is higher than other comparative detection heads such as LSDECD at 64.5% and LAWDS at 62.5%. This highlights its strong ability to recognize all positive samples effectively.

In terms of mAP50, our L-Head achieved a value of 78.1%, which is significantly higher than LSDECD and LSCD, both at 67.1%, and LAWDS at 67.8%. It is only slightly lower than dyHead, which scored 77.4%. This performance indicates that our detector head can detect targets more stably and accurately when the Intersection over Union (IoU) threshold is set at 0.5. When evaluating the mAP50-95 metric, our L-Head scored 44.7%, surpassing the performance of other detector heads. This suggests that it has superior detection capabilities across various IoU threshold ranges. Notably, our designed L-Head also achieved a detection speed of 208.3 FPS, significantly higher than that of other detector heads.

G. Comparison experiment between C3K2-iRMB_CGA and C3k2-iRMB

To thoroughly explore the advantages of our improved C3K2-iRMB_CGA module over the C3k2-iRMB module, we conducted comparative experiments. The experimental results are shown in Table VI.

From the experimental results in Table VI, our improved module, C3K2-iRMB_CGA, achieves 72.9%, 73.1%, 77.8%, and 44.7% in P, R, mAP50, and mAP50-95, respectively, which is an improvement of 2.1%, 0.5%, 0.01%, and 0.02%, respectively, compared to the C3k2-iRMB module. Notably, the C3K2-iRMB_CGA module reaches 500 FPS, a 214.3 improvement over the C3k2-iRMB's 285.7. This is because the CGA module optimizes the computational efficiency of

TABLE IV: Comparison of ours model with other typical target detection algorithms for different types of defects.

Model	Crazing	Inclusion	Patches	Pitted_surface	Rolled-in_scale	Scratches	mAP50
Faster R-CNN	0.432	0.793	0.851	0.774	0.595	0.884	0.772
YOLOv5s	0.443	0.792	0.870	0.796	0.613	0.869	0.730
YOLOv6n	0.429	0.765	0.860	0.818	0.585	0.880	0.723
YOLOv8n	0.480	0.823	0.888	0.817	0.627	0.906	0.757
YOLOv10n	0.404	0.772	0.845	0.809	0.601	0.847	0.713
YOLO11	0.502	**0.887**	**0.939**	0.724	**0.672**	0.957	0.780
ours	**0.516**	0.816	0.919	**0.836**	0.635	**0.968**	**0.782**

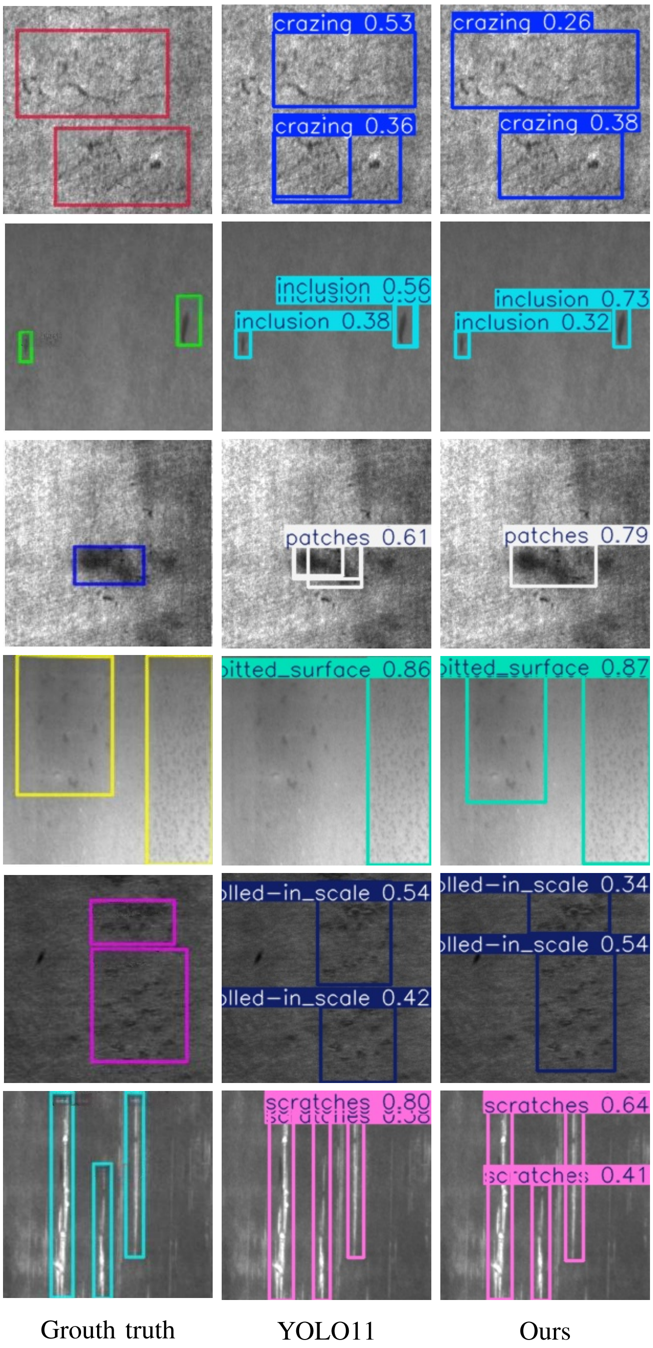

Fig. 8: Visualization results of experiments comparing different models.

TABLE V: Comparison experiments of different detection heads.

Models	P	R	mAP50	mAP50-95	FPS
LSDECD	0.695	0.645	0.671	0.378	23.4
LAWDS	0.704	0.625	0.678	0.386	44.6
dyHead	0.723	0.709	0.774	**0.448**	120.5
LSCD	0.705	**0.642**	0.671	0.375	25.1
L-Head(Ours)	**0.725**	0.731	**0.781**	0.447	**208.3**

TABLE VI: Comparison experiment between C3K2-iRMB_CGA and C3k2-iRMB.

Models	P	R	mAP50	mAP50-95	FPS
C3k2-iRMB	0.708	0.726	0.777	0.445	285.7
C3K2-iRMB_CGA	**0.729**	**0.731**	**0.778**	**0.447**	**500**

the model and reduces computational redundancy in the polytomous attention without adding additional parameters, resulting in significantly faster model inference.

Overall, the C3K2-iRMB_CGA module exceeds the performance of the C3k2-iRMB module in metrics such as P, R, mAP50, mAP50-95, and FPS, indicating greater potential for applications in detecting steel surface defects.

V. Conclusion

This paper explores the issue of detecting defects in steel and proposes an improved method based on the YOLO11 algorithm. Specifically, the self-attention module of the iRMB has been replaced with the CGA module, and enhancements have been made to the C3k2 module. improves the stability of the model during training, enabling it to learn and extract features more robustly. In addition, the LSKA attention mechanism has been integrated into the SPPF module to enhance its ability to capture multi-scale information related to complex defects. This integration addresses the SPPF module's limitations in managing feature weights for defects of varying sizes. As a result, the model can effectively emphasize small-scale features, allowing for better detection of tiny defects. Conversely, when larger defects are present, the mechanism highlights large-scale features, ensuring that the overall characteristics of these defects are thoroughly captured. Additionally, the L-Head inspection head has been designed with a focus on integrating the regression design concept based on DFL. This design enhances its bounding box positioning capability, allowing for accurate detection of defects on steel surfaces, particularly small defects. As a result, it significantly improves the precision in identifying these small defects and provides robust support for maintaining high steel surface quality. The experimental results indicate that the enhanced algorithm is competitive, achieving 77.1% of P, 78.2% of mAP50, and 44.8% of mAP50-90 on the NEU-DET dataset, with a frame rate of 312.5 FPS.

Although the improved model already has some advantages in computational efficiency, with the increasing demand for real-time and resource-constrained scenarios in industrial production, model lightweighting is still of great significance. In the future, we will explore more efficient lightweight convolutional structures to further reduce the number of parameters and computation of the model without significantly decreasing the detection accuracy, so that it can be operated on resource-limited edge devices and expand the application scope of the model.

REFERENCES

[1] C. Zhao, X. Shu, X. Yan, X. Zuo, and F. Zhu, "Rdd-yolo: A modified yolo for detection of steel surface defects," *Measurement*, vol. 214, p. 112776, 2023.

[2] L. Wang, X. Liu, J. Ma, W. Su, and H. Li, "Real-time steel surface defect detection with improved multi-scale yolo-v5," *Processes*, vol. 11, no. 5, p. 1357, 2023.

[3] H. Ma, Z. Zhang, and J. Zhao, "A novel st-yolo network for steel-surface-defect detection," *Sensors*, vol. 23, no. 22, p. 9152, 2023.

[4] M. Lu, W. Sheng, Y. Zou, Y. Chen, and Z. Chen, "Wss-yolo: An improved industrial defect detection network for steel surface defects," *Measurement*, vol. 236, p. 115060, 2024.

[5] K. Xia, Z. Lv, C. Zhou, G. Gu, Z. Zhao, K. Liu, and Z. Li, "Mixed receptive fields augmented yolo with multi-path spatial pyramid pooling for steel surface defect detection," *Sensors*, vol. 23, no. 11, p. 5114, 2023.

[6] Z. Yuan, H. Ning, X. Tang, and Z. Yang, "Gdcp-yolo: Enhancing steel surface defect detection using lightweight machine learning approach," *Electronics*, vol. 13, no. 7, p. 1388, 2024.

[7] L. Li, R. Zhang, T. Xie, Y. He, H. Zhou, and Y. Zhang, "Experimental design of steel surface defect detection based on msfe-yolo—an improved yolov5 algorithm with multi-scale feature extraction," *Electronics*, vol. 13, no. 18, p. 3783, 2024.

[8] J. Li and M. Chen, "Dew-yolo: An efficient algorithm for steel surface defect detection," *Applied Sciences*, vol. 14, no. 12, p. 5171, 2024.

[9] J. Li, Z. Su, J. Geng, and Y. Yin, "Real-time detection of steel strip surface defects based on improved yolo detection network," *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 76–81, 2018.

[10] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue, B. Zhang, Z. Jiang, T. Huang, Y. Wang, and C. Wang, "Rethinking mobile block for efficient attention-based models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2023, pp. 1389–1400.

[11] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 420–14 430.

[12] S. Li and W. Liu, "Small target detection model in aerial images based on yolov7x+," *Engineering Letters*, vol. 32, no. 2, pp. 436–443, 2024.

[13] D. Gai, J. Zhang, Y. Xiao, W. Min, H. Chen, Q. Wang, P. Su, and Z. Huang, "Gl-segnet: global-local representation learning net for medical image segmentation," *Frontiers in Neuroscience*, vol. 17, p. 1153356, 2023.

[14] Q. Wei and K. Yang, "Research on interpretable recommendation algorithms based on deep learning." *Engineering Letters*, vol. 32, no. 3, pp. 560–568, 2024.

[15] Y. Li, S. Xu, Z. Zhu, P. Wang, K. Li, Q. He, and Q. Zheng, "Efc-yolo: An efficient surface-defect-detection algorithm for steel strips," *Sensors*, vol. 23, no. 17, p. 7619, 2023.

[16] Z. Guo, C. Wang, G. Yang, Z. Huang, and G. Li, "Msft-yolo: Improved yolov5 based on transformer for detecting defects of steel surface," *Sensors*, vol. 22, no. 9, p. 3467, 2022.

[17] Q. Wang, F. Liu, Y. Cao, F. Ullah, and M. Zhou, "Lfir-yolo: Lightweight model for infrared vehicle and pedestrian detection," *Sensors*, vol. 24, no. 20, p. 6609, 2024.

[18] Y. Liu, M. Zhang, F. Fan, D. Yu, and J. Li, "Edasnet: efficient dynamic adaptive-scale network for infrared pedestrian detection," *Measurement Science and Technology*, vol. 35, no. 11, p. 115406, 2024.

[19] H. Zou, J. Yang, J. Sun, C. Yang, Y. Luo, and J. Chen, "Detection method of external damage hazards in transmission line corridors based on yolo-lsdw," *Energies*, vol. 17, no. 17, p. 4483, 2024.

[20] J. Wang, Y. Huang, and Y. Liu, "Low contrast stamped dates recognition for pill packaging boxes based on yolo-sfd and image fusion," *Digital Signal Processing*, vol. 153, p. 104602, 2024.