



CRGF-YOLO: An Optimized Multi-Scale Feature Fusion Model Based on YOLOv5 for Detection of Steel Surface Defects

Tao Yu^{1,2} · Xu Luo^{1,2} · Qiang Li^{1,2} · Lei Li^{1,2}

Received: 5 November 2023 / Accepted: 6 June 2024
© The Author(s) 2024

Abstract

The identification of imperfections on steel surfaces is vital for ensuring the quality of industrial products. It requires the capability of real-time detection with high accuracy. This paper proposes the CRGF-YOLO (Contextual Reparameterized Generalized Feature) model based on YOLOv5. In the network, BottleneckCSP structures and depthwise separable convolutions utilizing the structural reparameterization are introduced to reduce the model size and improve performance. In addition, contextual transformer modules are employed as self-attention mechanisms to improve feature representations by capturing long-range dependencies, outperforming conventional convolutional networks. Furthermore, the simplified generalized feature pyramid network is embedded to aggregate multi-scale feature maps and enhance the network's robustness. Finally, four prediction heads with different sizes are employed to predict defects, which are supported by prior bounding boxes generated using k-means clustering algorithm. The Focal-EIOU (Exponential Intersection over Union) loss function is introduced to improve detection accuracy and expedite model convergence. The improved model achieves a mean average precision (mAP) of 82.2% on the NEU-DET dataset, outperforming the baseline YOLOv5s by 7.7% mAP while maintaining real-time speeds. Comparative evaluations demonstrate CRGF-YOLO's superior performance over previous state-of-the-art methods like Faster R-CNN (77.4% mAP), YOLOv3 (77.4% mAP), YOLOv7s (72.1% mAP), and YOLOv8s (78.7% mAP) for steel surface defect detection. Overall, this study provides valuable insights and practical guidance for the advancement of defect detection technology.

Keywords Defect detection · YOLO · Generalized feature pyramid network · Contextual transformer · Structural reparameterization

1 Introduction

In the process of manufacturing, metal products are vulnerable to defects such as surface oxidation, fractures, depressions, and scratches. These flaws have detrimental effects on the product's quality and durability, leading to potential financial loss. The need to ensure manufacturing quality has made it imperative to detect flaws on metal surfaces. As a result, there has been a surge of interest among researchers in developing rapid and precise inspection technologies.

Researchers employ the methods of vision-based defect detection since the 1980s. A typical traditional detection algorithm includes image preprocessing, Regions of Interest (ROI) detection, image segmentation, feature extraction and defect classification [1]. Traditional algorithms categorize defects using manual extraction of features to implement detection, such as HOG (Histogram of Oriented Gradient) [2], Local Binary Pattern (LBP) [3],

✉ Lei Li
nudtyt@ldy.edu.rs
Tao Yu
yt2129536@163.com
Xu Luo
luoxu2009@163.com
Qiang Li
lq1766500568@163.com

¹ College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

² Laboratory of Science and Technology on Integrated Logistics Support, National University of Defense Technology, Changsha 410073, China

Fourier transform [4], Gabor filter [5], SVM (Support Vector Machine) [6] and Random Forest [7]. However, there are limitations of algorithms above: the artificial feature extraction cannot express information adequately under complex circumstances, defect performance is highly affected by human-generated feature extractors, problems like huge computations, locations and sizes of defects are still difficult to solve.

Since 2012, Convolutional Neural Networks have become more and more important in computer vision (CV) tasks [8]. For the purpose of defect classification and location regression, the two-stage deep-learning algorithms represented by Faster-R-CNN [9] generate candidate boxes using Region Proposal Network (RPN) to predict regional images, finally features of merged Regions of Interest (ROI) can be extracted. He et al. [10] proposed a strip surface defect detection network based on the improved Faster R-CNN, and multiple hierarchical features were combined into one feature in the Resnet backbone network, which made good use of global semantic information and reduced feature loss. Xie et al. [11] proposed the improved Faster-R-CNN combined with the feature pyramid network (FPN) and ROI-Align algorithms, which reduced the information loss of small objects and errors caused by quantization, obtaining 98.5% mean Average Precision (mAP) on the PCB dataset. In comparison to two-stage algorithms, one-stage target detection algorithms such as You Only Look Once (YOLO) [12] and SSD (Single Shot Detector) [13] integrate classification and positioning tasks using regression techniques to directly compute classification results and position coordinates for multiple targets. Thanks to less complexity and faster detection time, the one-stage detectors have received much concern to meet real-time detection needs. For instance, Li et al. [14] proposed a fully convolutional YOLO detection network, which improved the prediction of the locations and sizes of surface defects on steel strips. This method offers support for real-time surface defect detection. Xie et al. [15] utilized six feature maps to generate predicted classification results and integrated the Fully Convolutional Squeeze-and-Excitation (FCSE) block into SSD to enhance the detection accuracy on the TILDA dataset.

While Convolutional Neural Networks (CNNs) are effective at extracting features for defect detection tasks, their performance can degrade when dealing with smaller targets as the network depth increases. As the depth of the CNN model grows, the receptive field of the convolutional filters also increases, making it challenging to capture intricate details and extract discriminative features from more minor defects or targets. This is because the larger receptive fields tend to capture more contextual information, which can be beneficial for understanding larger objects but may cause smaller objects or defects to be overlooked or misrepresented. In addition, CNNs face

difficulties in accurately detecting and classifying surface defects due to background interference and various shapes of the defects.

With the emergence of Vision Transformer (ViT) [16], transformer-based models have become a hot issue in the field of CV. ViT divides original images into several patches, and compiles the patch information into transformer encoder module, at last images are classified by the fully connected layer. DETR [17] is an innovative network that integrates CNN and Transformer architectures to perform object detection by treating it as an image sequence prediction task. Based on the self-attention mechanism, the transformer encoder module extracts feature information and incorporates position information, then the encoded information is sent to the decoder module to obtain global information in feature extraction and fusion. Compared to CNNs, which are limited by receptive fields of convolution kernels, transformer performs well in obtaining global context by collecting and learning relative information from adjacent regions.

As the analysis mentioned above, this paper introduces a novel approach for defect detection using an optimized version of YOLOv5. The improved model is designed to achieve accurate defect detection on steel surfaces, while simultaneously ensuring a high detection speed. The main contributions of the study are as follows:

1. BottleneckCSP architectures and standard convolution layers are replaced with reparameterized CSP (Cross Stage Partial) structures (Rep-CSP) and reparameterized depthwise separable convolutions (Rep-DSC) to lightweight the backbone part and enhance the ability of extracting features.
2. Combined CNN and transformer, the contextual transformer module is utilized to enhance feature expressions from different layers and transfer effective information to the neck part.
3. A simplified generalized FPN CoT-GFPN is designed to fuse features from different scales and enhance network generalization ability.
4. Four detection heads are set to predict defects with different sizes and the k-means clustering algorithm is utilized to generate prior bounding boxes. Moreover, the model convergence and detection precision are further improved using Focal-EIOU as the loss function.

The remainder of this paper is organized as follows: Section 2 will present related work. Section 3 will display the structure of CRGF-YOLO and proposed methods. Section 4 will demonstrate experiments evaluated on the NEU-DET dataset and verify the effectiveness of improvements. Section 5 will discuss limitations and future research. Finally, Sect. 6 will summarize the conclusion.

2 Related Work

2.1 YOLOv5

As the one-stage detector, YOLO is designed to perform object detection as a regression problem, which is achieved through the use of the convolutional neural network architecture, predicting the class probabilities of detected images and regress the coordinates of the bounding boxes. YOLOv5 is the fifth version of the YOLO series, which has received extensive attention because of its high accuracy and ability to perform real-time detection. There are three parts in the network, namely backbone, neck, and head. The CSPDarknet53 structure is constructed in the backbone part, including Focus, BottleneckCSP, Spatial Pyramid Pooling Fusion (SPPF) to extract features, and uses the stacking strategy of several residual architectures to avoid the vanishing gradient problem. The FPN facilitates the transmission of high-level semantic features in a downward direction. In addition, the Path Aggregation Network (PANet) incorporates an extra bottom-up pathway to aggregate feature maps, conveying both high-level semantic information from deeper layers and low-level locational information from shallower layers into the neck part. YOLOv5 utilizes the k-means clustering algorithm along with a genetic algorithm to determine the sizes and aspect ratios of anchor boxes, which serve as prior bounding boxes for object detection. The k-means algorithm groups the ground truth bounding box dimensions into k clusters, with each cluster represented by a single anchor box dimension. The genetic algorithm further optimizes these anchor box dimensions to improve their fitness for the specific dataset. The loss function is a critical metric for assessing the performance of a model as it measures the extent of variation between the predicted values and the actual values. A well-designed loss function is essential for achieving accurate predictions.

2.2 Self-attention Mechanism

The self-attention mechanism has been well extensively utilized in the domain of natural language processing (NLP), which achieves self-attention by evaluating the similarity of input data and acquiring knowledge of the interdependence between information by means of selective weighted calculations. In CV tasks, the self-attention mechanism is utilized to model the relationships of feature vectors across various spatial regions, enabling the extraction of global connections within images. By employing the global multi-head self-attention mechanism, Wang et al. [18] developed a non-local operation, and gaussian

transformation is introduced to access the global similarity of information and generate corresponding weights to improve the network performance. During the processing of images, the self-attention mechanism expands the picture pixels into sequences, and the computation and memory usage will increase sharply if the sequence length is large. To solve the problem, researchers have improved transformer models by utilizing local attention mechanism that concentrates on a portion of the input sequence to decrease computation and memory requirements.

With the development of deep learning, the combination of CNN and Transformer has emerged as a prominent trend. Bello et al. [19] enhanced network performance by incorporating self-attention mechanisms into CNNs. BoT (Bottleneck Transformers) [20] is improved upon the ViT model by utilizing a bottleneck structure and replacing the 3×3 convolutional layers in ResNet with the self-attention mechanism of transformers. This results in a more efficient multi-branch model design. To use contextual information efficiently, Contextual Transformer Networks (CoT) [21] combines the self-attention mechanism with the idea of non-local blocks. This approach enables adaptive modeling of contextual information from different regions and enhance the feature expressions of the model. In general, the fusion of CNNs and Transformers can strengthen the convolutional operation of CNNs. By integrating the positional encoding of feature information, the expression of semantic information can be enhanced. This innovative combination provides an effective solution for the study of deep networks.

2.3 Multi-Scale Fusion Network

During convolution operations, extracted features from different layers have varying scales due to the high-level network containing more semantic information and the low-level network containing more geometric information. To enhance target detection ability, multi-scale features from different layers can be fused. The FPN exhibits a robust capability of equilibrating feature maps from diverse layers through its up-sampling and down-sampling operations. Based on FPN, PANet adds the route enhancement and adaptive feature pooling operations to shorten information paths and ease information propagations. Moreover, the improved FPN models like PRB-FPN [22], NAS-FPN [23], DFPN [24] have been developed to improve the performance of multi-scale target detection. EfficientDet [25] proposed a more efficient multi-scale fusion method BiFPN, which integrated bidirectional cross-layer connections and fast normalized fusion. By introducing weights, feature information at different scales was well balanced. However, previous networks only focus on feature fusion and lack internal information connections. GFPN [26] developed a novel pathway fusion approach that incorporates skip-layer connections

and cross-scale connections. The skip connections have a shorter distance between feature layers, allowing for more efficient information flow. Two connections named dense-link and \log_2n -link are introduced to reduce gradient vanish in the network and provide more effective transmission from previous nodes to next. In addition, a new cross-scale fusion named Queen-fusion is proposed to enhance feature interactions from same levels and neighbor levels. The GFPN model is able to extract rich and high-quality multi-scale features from various levels of feature maps, resulting in superior performance when compared to other enhanced FPN models.

2.4 Structural Reparameterization

Multi-branch network architectures such as ResNet and Inception significantly improve the performance at the training stage. However, inference time and memory consumption are increased because of complex structures, and it is not friendly to achieve the real-time detection. Structural reparameterization is a technique used to decouple the training-time and inference-time network, which results in a network that has both high performance and low complexity. Structural reparameterization refers to constructing the multi-branch structures in the training phase and the parameters of the structure can be converted into another inference-time plain architecture with another set of parameters. RepVGG proposed by Ding et al. [27] realized the idea. Unlike Inception models designed multi-branch architectures, RepVGG utilizes a straightforward topology and extensively employs 3×3 convolution operations, resulting in a simple, fast and flexible model. At the training stage, RepVGG is a multi-branch model that constructs the information flow of shortcut branches, and the model weights of identity branch, 1×1 branch can be added into the 3×3 convolution layer. During the inference time, structural reparameterization outperforms plain models while maintaining a lower complexity structure than multi-branch architectures.

3 Methods

3.1 CRGF-YOLO Structure

For industrial surface defect detection, there are two challenges: (1) There are various types of sample sizes in complex circumstances, which have no obvious unified division of defect boundaries; (2) Due to uneven textures of samples under weak conditions, the intra-class difference of defect images change sharply. Although the convolution operation has limitations in obtaining contextual information of targets, the transformer modules' self-attention mechanism can improve contextual semantic information and extract more

spatial information, thereby enhancing target detection ability. In addition, GFPN has an efficient multi-scale feature fusion ability that combines feature information from both high-level and low-level layers., which have good performance in the defect detections.

From the state above, this paper proposed the CRGF-YOLO (CoT-Rep-GFPN) network, and the structure of CRGF-YOLO is shown in Fig. 1.

In Fig. 1, there are three parts: (1) The backbone part is a top-down information flow made of the reparameterized depthwise separable convolution modules (Rep-DSC) and the reparameterized BottleneckCSP structures (Rep-CSP). Spatial Pyramid Pooling Fusion (SPPF) module from YOLOv5 pools the input feature map and combines the results of different scales into a final feature vector. This allows for the extraction of richer feature information on different scales. In addition, the CoT module is utilized with self-attention mechanism to enhance semantic information expression and extract effective contextual information from different feature layers. (2) The neck part is of the simplified GFPN network to fuse information from same layers and adjacent layers to upgrade feature expressions. Up-sampling modules and Rep-DSC modules have the function of adjusting channel numbers through up-sampling and down-sampling operations. (3) The prediction part has four prediction heads, which are deployed to output different sizes of feature maps. "Conv" modules are used to change the number of channels for outputting detecting information. The modules proposed above are depicted in Fig. 2.

3.2 Rep-DSC and Rep-CSP

When delving into convolutional operations, it is important to consider the potential issues of gradient vanishing and gradient exploding that influence effective training results. By utilizing the multi-branch structure's ability to converge well and create nonlinear diversified connections, the performance of these structures can be significantly improved. However, the multi-branch structure cannot effectively achieve parallel acceleration, and greatly increases the memory consumption. For the purposes of building the network with high performance and low complexity, the structural reparameterization is introduced. During the training phase, a multi-branch structure is utilized to enhance the capacity of the model, while the multi-branch architecture is consolidated into a single-path structure during the inference phase. This allows for the extraction of more feature information without increasing the computational load required for inference.

In Fig. 2, the 3×3 depthwise convolution with stride two and the 1×1 pointwise convolution are reparameterized in the Rep-DSC module, and SiLU (Sigmoid Linear Unit) activation function is used after Conv-BN (Batch Normalization)

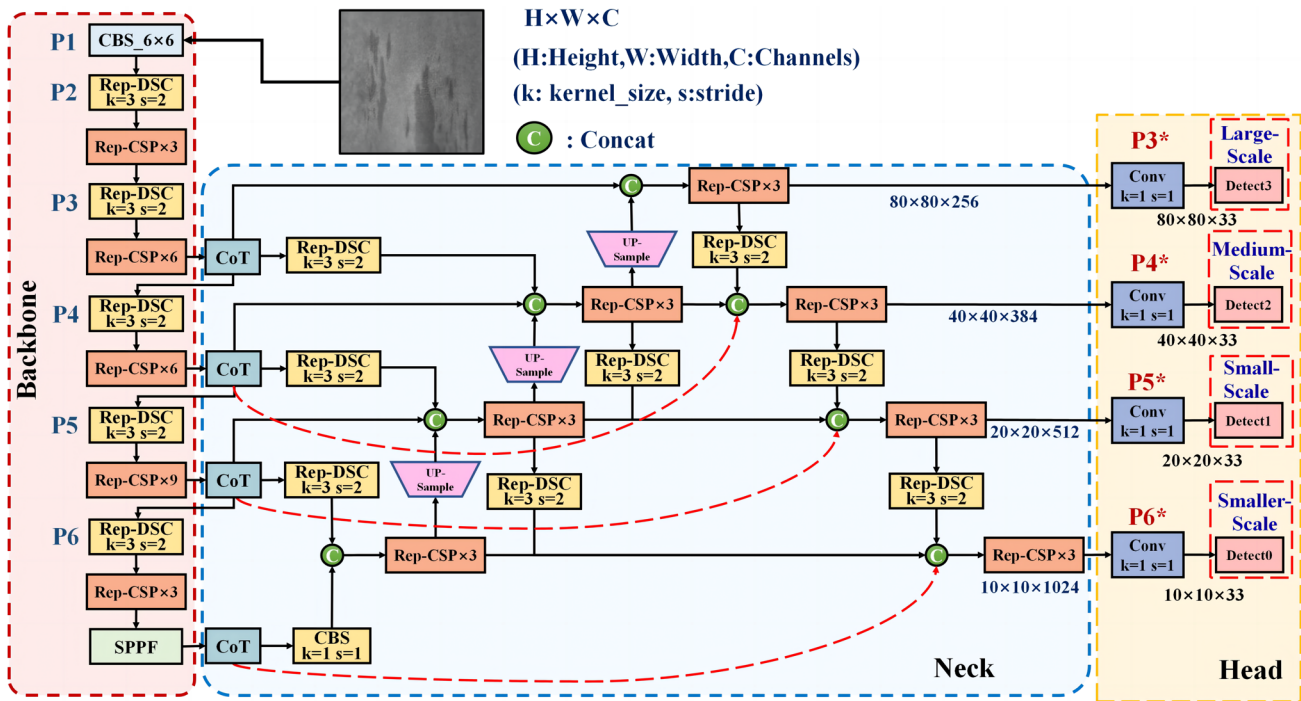


Fig. 1 The structure of CRGF-YOLO

modules. The Rep-DSC module offers the advantage of having fewer parameters of depthwise separable convolution [28], and the architecture of the multi-branch fusion diversifies the network. The Rep-CSP module can be split into two branches, one of which has the benefits of a multi-branch structure at training time, as well as a stacking strategy of bottlenecks with residual architecture to enhance performance. The other branch performs concatenation with this. By utilizing structural reparameterization, the inference-phase results in richer features compared to the BottleneckCSP module.

Taking the structural reparameterization of the CBS module as an example. At the training stage, the standard 3×3 convolution is decoupled into a multi-branch architecture composed of the 3×3 branch, the 1×1 branch and the identity branch. By combining the convolution without bias and BN, it is beneficial to improve the convergence speed of the network and control the occurrence of overfitting. Then the size of each branch convolution kernel is converted into 3×3 , and finally the multi-branches are fused. The process of structural reparameterization is demonstrated in Fig. 3.

The formula of the convolution operation (w_k means the weight coefficient):

$$Conv(X) = X' = w_k * X. \tag{1}$$

BN includes five parameters: mean (μ), variance (σ), weight (w_b), bias (b) and a minimum value set to prevent the denominator from falling to zero (ϵ):

$$BN = w_b \times \frac{X' - \mu}{\sqrt{\sigma^2 + \epsilon}} + b. \tag{2}$$

The combination of Conv and BN is:

$$BN(Conv(X)) = \frac{w_b \times w_k}{\sqrt{\sigma^2 + \epsilon}} * X + b - w_b \times \frac{\mu}{\sqrt{\sigma^2 + \epsilon}}. \tag{3}$$

The Conv-BN module not only accelerates the training speed, but also realizes the additivity on the multi-branch structure. The convolution kernel size is adjusted to 3×3 , the white square represents the element with zero in the convolution kernel. For a convolution kernel with a size of 1×1 , a circle of zero elements is filled to change the size of the convolution kernel to 3×3 . Because there is no convolution operation in the identity branch, it is necessary to construct a 3×3 convolution kernel automatically so that the output of the input data does not change.

The structure consists of the 3×3 branch, the 1×1 branch and the identity branch, corresponding to the transfer function $f(x)$, $g(x)$ and x . By adjusting the branch dimension, the final output of the multi-branch structure is:

$$F(x) = f(x) + g(x) + x. \tag{4}$$

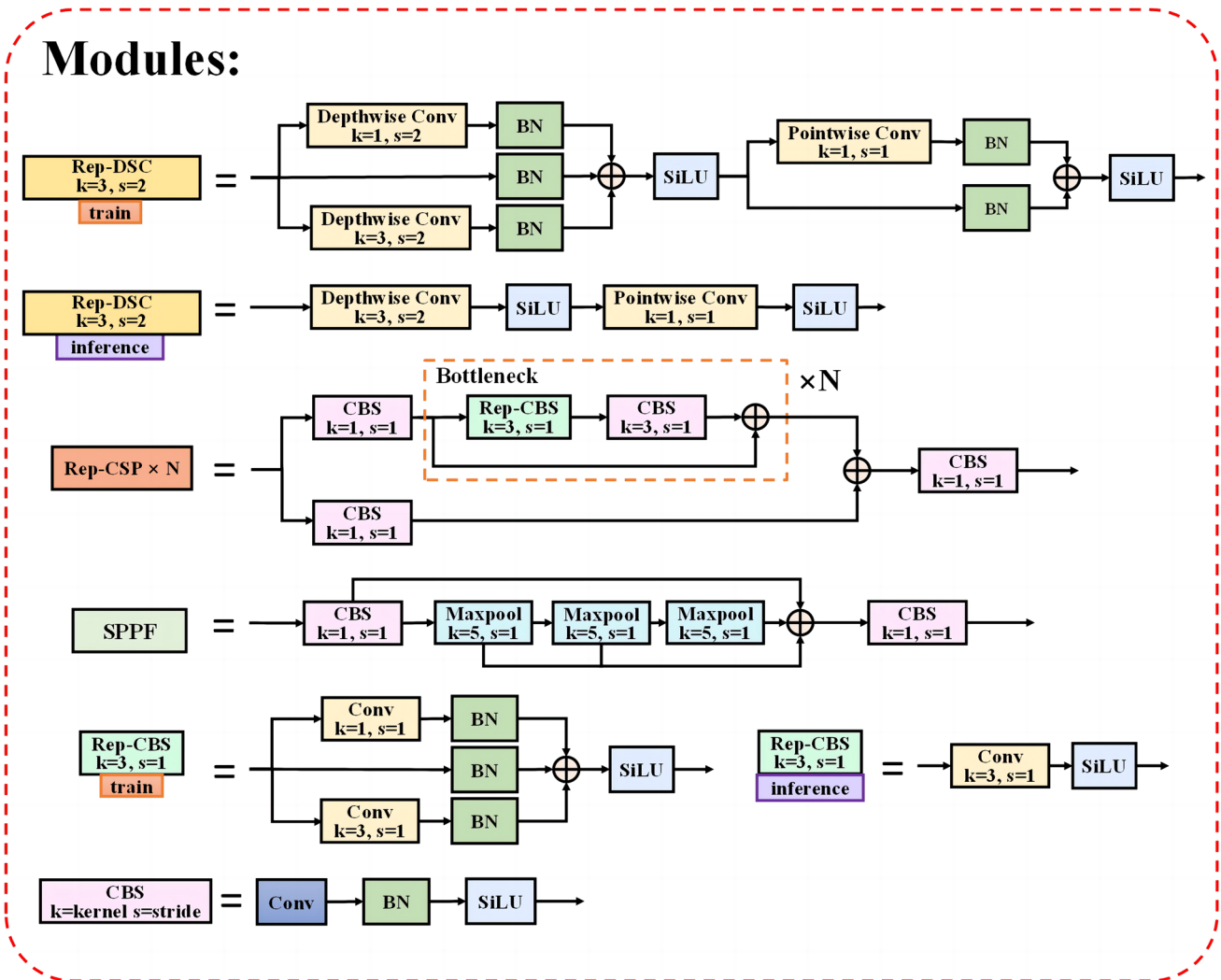


Fig. 2 Modules in CRGF-YOLO

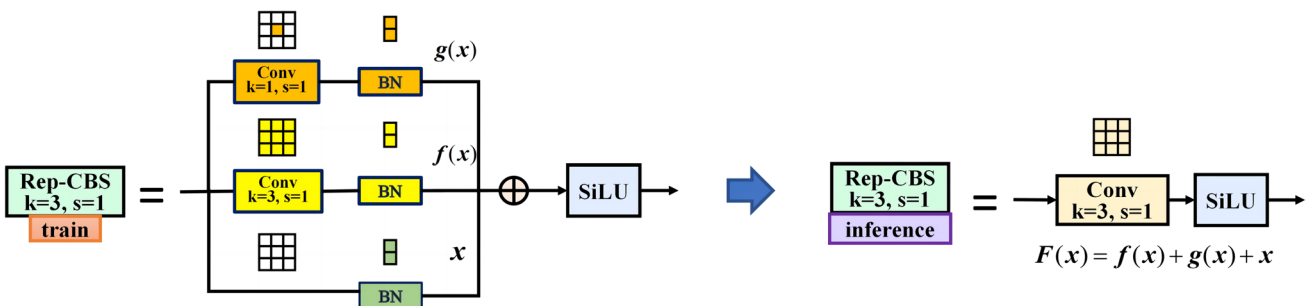


Fig. 3 The process of structural reparameterization

3.3 CoT Module

CNN models effectively combine visual features and semantic information, with low-level features such as edge, texture, angular providing feature information, while high-level

semantic information is interpreted in terms of the low-level features. In addition, transformers are better suited for processing high-level semantic information. The CoT module enhances feature extraction from feature maps by combining the convolution operation with transformer’s self-attention

mechanism and integrated contextual information between adjacent keys through self-attention learning.

In CV tasks, the structure of transformer comprises of three main components: Image Patch Embedding, encoder, and Multi-Layer Perceptron (MLP). The Image Patch Embedding is responsible for extracting image features and transforming each image patch into a tensor form that can be processed by the encoder for target detection. The encoder, on the other hand, encodes the input image tensor and mines global information using Multi-Head Self-Attention (MSA). In addition, MLP maps features to higher dimensions. MSA obtains the weighted values of the attention distribution between features by querying the matrix Q and the key matrix K . The output is then obtained by multiplying the feature matrix V by the softmax function.

As is shown in Fig. 4, the input feature is X ($X = H \times W \times C$, H : Height, W : Width, C : Channels). The CoT module utilizes the MSA mechanism, through the 3×3 convolution operation, all adjacent keys (K) in the kernel are contextually encoded to obtain the static contextual information K_1 . The local static context information keys K_1 performs the concatenation with Q , and then attention matrix A will be obtained through two consecutive 1×1 convolutions.

The matrix A is expressed as:

$$A = [K_1, Q] \cdot W_\alpha \cdot W_\beta \tag{5}$$

A is applied for aggregating with V (self-attention mechanism), and the dynamic contextual information K_2 is received after the concatenation of aggregated feature maps.

$$K_2 = V \otimes A \tag{6}$$

Finally, the extraction of feature information is enhanced by fusing the static contextual information K_1 and the dynamic contextual information K_2 . In this paper, we find that placing the CoT module with self-attention mechanism after Rep-CSP module can enhance the ability of feature expression and transmit effective information at different levels. To control the number of channels and create a deeper nonlinear level, the 1×1 CBS module is embedded before the CoT model. In Sect. 4, comparative experiments are conducted to evaluate the effectiveness of the CoT module.

3.4 CoT-GFPN

The FPN is designed to effectively merge information from multiple feature layers in the backbone network, which has been proved to be effective with limited parameters. A conventional FPN model typically combines multi-scale features in a top-down manner. However, this method has limitations in terms of one-way information flow. To address this issue, GFPN has been developed to serve as a new “neck” that integrates global contextual information into the feature pyramid, allowing for a more holistic understanding of information. This helps in capturing long-range dependencies and improving the overall performance. However, GFPN significantly increases the model size, which makes it unfriendly for real-time detection applications.

The CoT-GFPN is displayed in Fig. 5. In this network, four prediction heads for detecting different sizes are employed firstly, and the feature maps of different scales are extracted from the CoT modules. There are two information flows: one path is the skip-layer connection of same level layers, and the other one is the queen-fusion integrating high-level semantic information and low-level geometric

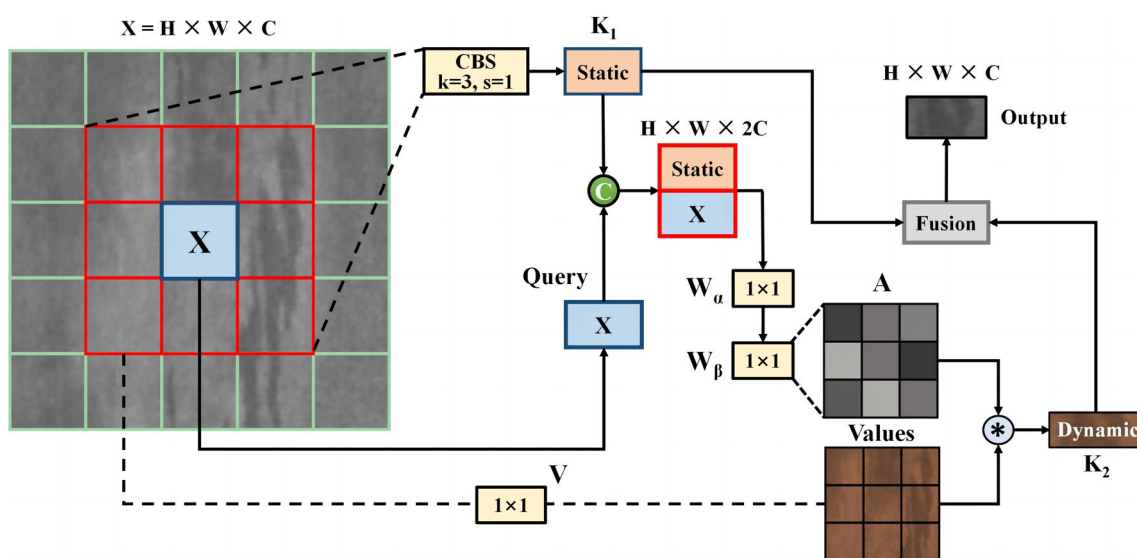
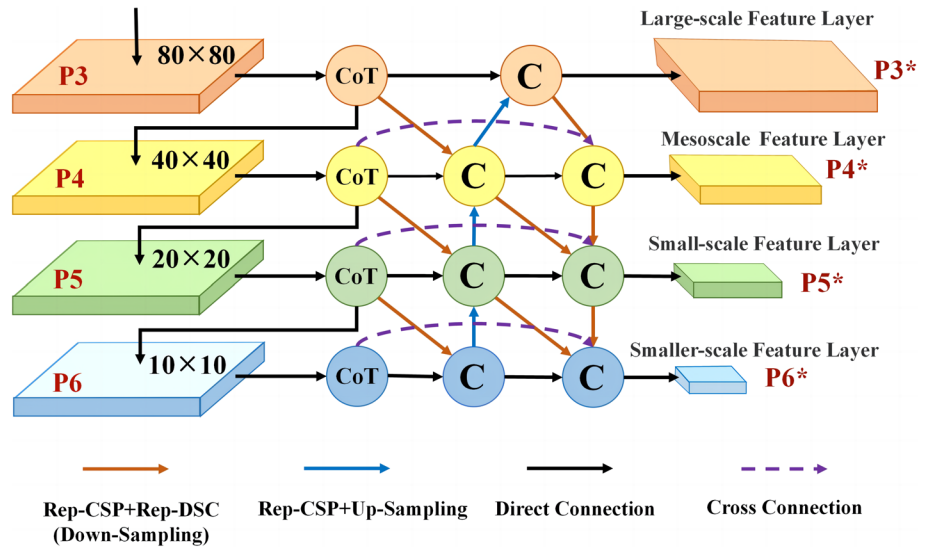


Fig. 4 The structure of the CoT module

Fig. 5 The design of CoT-GFPN network. Here the CoT modules are utilized to extract effective information and transfer to the neck part. The simplified GFPN in the neck part has the cross-scale connection of same level layers, and the queen-fusion connection of different level layers



information. The skip-layer connection allows for the direct transfer of information between feature maps from different layers that share same dimensions. This contributes to enhancing the transmission of intricate details present in the low-level feature maps, while also integrating them into the high-level feature maps. Queen-fusion adopts an adaptive weight strategy during the fusion process by assigning different weights to feature maps from the same level and neighboring levels. This allows the model to selectively emphasize more informative features while suppressing less useful ones, resulting in more effective multi-scale feature fusion. Moreover, this connection is performed in a sequential manner, starting from low resolution and gradually progressing to high resolution. This approach allows for the effective utilization of feature information at various scales, thereby enhancing the network’s capability to perceive multi-scale targets.

3.5 Prediction Network

Four prediction heads with sizes of $10 \times 10 \times 1024$, $20 \times 20 \times 512$, $40 \times 40 \times 384$, $80 \times 80 \times 256$ are set to output detecting information. After the employed 1×1 convolution with stride one layers and sigmoid activation function, the channel will be adjusted to the value of C: $3 \times (6 + 5)$. For each grid, three bounding boxes are designed to output three types of parameters: (1) the box parameter with four values, namely the center coordinates of the box (x, y) and the width and height of the box (w, h); (2) confidence with values between 0 and 1; (3) the type of the dataset is a set of conditional class detecting probabilities, which is the value between 0 and 1. The final outputs of the convolution layers are three-dimensional tensors, which include: $10 \times 10 \times 33$, $20 \times 20 \times 33$, $40 \times 40 \times 33$ and $80 \times 80 \times 33$. Each branch has feature maps with varying receptive field sizes, allowing

for the detection of targets of different sizes, including large, medium, small, and smaller size of targets.

In YOLOv5s, CIOU (Complete Intersection over Union) loss function is used in the regression of the prediction box. This loss function considers the IOU, aspect ratio, and distance between the center point of the real box and the predicted box. However, it is essential to note that the CIOU loss function only considers the relative aspect ratio between the predicted and ground truth bounding boxes, which may need to be revised to capture the actual aspect ratio differences accurately, potentially leading to ambiguity or error in specific scenarios. A good approach is to use the Focal-EIOU (Focal and Efficient Intersection over Union) loss function to solve the problem. The expression of EIOU loss [29] function is as follows:

$$L_{EIOU} = 1 - IOU + \frac{\rho^2(B_{ct}, B_{ct}^{gt})}{h_c^2 + w_c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2}, \tag{7}$$

where B_{ct}^{gt} , w^{gt} and h^{gt} are the center coordinates, width, and height of the real bounding box, B_{ct} , w and h are the center coordinates, width, and height of the predicted bounding box, c_w , c_h are the width and height of the minimum bounding rectangle of the real bounding box and the predicted bounding box. $\rho(\cdot)$ is the method of Euclidean distance.

In addition, IOU loss can be expressed as follows:

$$IOU = \frac{|A \cap B|}{|A \cup B|}, \tag{8}$$

where A represents the predicted box’s area and B represents the real ground box’s area. This formula is used to determine the ratio between the overlapping area and the common area. The Loss function $L_{Focal-EIOU}$ can be expressed as:

$$L_{Focal-EIOU} = IOU^\gamma L_{EIOU}, \tag{9}$$

where γ is a hyper-parameter used to control the curvature of the curve, it is determined to have a value of 2 in this paper.

Yolov5 uses the k-means clustering algorithm to generate different sizes of prior bounding boxes, which contributes to predicting the target with different scales. Due to four detection heads employed in this paper, 12 prior bounding boxes are generated for improving the detection accuracy of steel surface defects in this paper. Table 1 is given to demonstrate a new set of prior bounding boxes serving for each detection head.

4 Experiments

This section introduces the dataset and evaluation metrics, followed by comparative experiments with different models and detectors, which are evaluated on the NEU-DET dataset. The experimental comparison shows that the performance of CRGF-YOLO is reasonable, and it achieved real-time detection in the industrial scenario.

4.1 Datasets and Experimental Environment

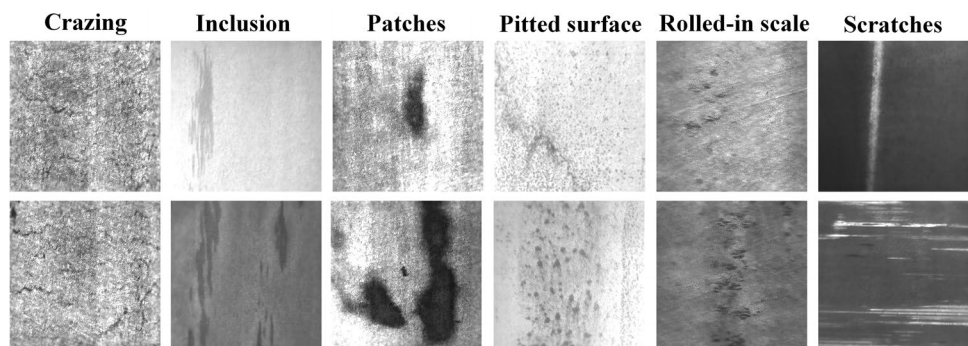
In this paper, NEU-DET dataset [30] is a steel strip surface defect database, offered by Northeast University, which includes six types, inclusion, pitted surface, crazing, rolled-in scales, patches, and scratches. There are 1800 gray-scale

Table 1 Prior bounding boxes for each detection head

Detection layer	Feature-map dimensions	Anchor boxes [w, h]
P3*	80×80	[15, 166], [19, 36], [24, 75]
P4*	40×40	[36, 168], [50, 71], [51, 36]
P5*	20×20	[74, 106], [78, 177], [103, 59]
P6*	10×10	[155, 80], [162, 26], [164, 183]

Detection layer can be seen in (Fig. 5)

Fig. 6 Presentation of six types of defects in the NEU-DET dataset



pictures with 200×200 pixels, and each type of defects has 300 labeled samples. Six types of samples are shown in Fig. 6. The dataset we used is divided into training set, validation set and test set in a ratio of 8:1:1.

The experiments are performed on the configuration with NVIDIA GeForce RTX 3060Ti GPUs, and deep-learning framework is PyTorch 1.9.0. By utilizing the image adaptive amplification function, the input image size was adjusted to 640×640 pixels. This method helps strengthen the detection ability at various scales on images, particularly for small targets. Moreover, enlarging the image can decrease the model’s sensitivity to resolution and image ratio, improving the model’s robustness and generalization ability.

Moreover, the SGD optimizer is a momentum of 0.9 and weight decay is 0.0005, and the learning rate increased from 0.0033 to 0.01. The models mentioned below are trained 300 epochs, and the batch size is 8. To enhance the robustness of the dataset, image enhancement methods was introduced at the training stage, such as Mosaic, Image Flip, and HSV Conversion.

4.2 Evaluation Metrics

To demonstrate the performance of the CRGF-YOLO algorithm in an objective manner, the following metrics are utilized:

a. Precision (P), Recall (R), and mean Average Precision (mAP) calculated as follows:

$$\text{Precision} = TP / (TP + FP),$$

$$\text{Recall} = TP / (TP + FN), \text{ mAP} = \Sigma(\text{AP}_i) / N,$$

where: TP = True Positives FP = False Positives FN = False Negatives AP_i = Average Precision for class i N = Number of classes

P measures the accuracy of model’s predictions, while R measures the model’s ability to detect positive instances. mAP denotes the average accuracy value across different categories.

b. Params (M), Inference Time, and GFLOPs

Params is the metric used to quantify the model size, while the computational volume of the model is represented by FLOPs (Floating Point Operations, 1 GFLOPs = 10^9 FLOPs), and Inference Time is used to demonstrate the detection speed.

4.3 Ablation Study

To assess and analyze the contribution of each module, ablation experiments are conducted to evaluate the following components: simplified GFPN, an additional detection head for detecting small targets, CoT module with the self-attention mechanism, structural reparameterization, and the improved loss function Focal-EIOU. Related improved models are displayed in Table 2, and evaluation metrics relevant to the corresponding cases are illustrated in Table 3.

In this section, various experiments were conducted to assess the efficiency of the one-stage detector presented, and main metrics need to be focused on evaluating the models' performance. As shown in Table 3, it can be shown that evaluation metrics have improved when adding another detection head, and prior bounding boxes are generated using k-means clustering algorithm. This method makes Model 1 improve 3.3 and 2.9% in R and mAP. Compared to the baseline model, all evaluation metrics are improved at

different levels in Model 2. The improved neck part includes additional feature information from different levels using Queen-fusion and skip-layer connections. Although this fusion increases the computational load and the structure becomes complex, it results in a robust algorithm generalization and greater robustness. The multi-scale CoT-GFPN network in Model 4 is employed to increase 5.0% in mAP, exhibiting varying degrees of enhancement in defect detection at different scales. It is noteworthy that the parameters sharply increase to 17.6 M.

To lightweight the network in Model 4, DSC plays an important role in reducing parameters and computational cost. Through the employment of DSC, the model size is successfully reduced by 24%. However, the use of separate filters in DSC can result in inadequate information transmitted in channels. The values of P and mAP have decreased by 2.1 and 0.3% compared with Model 4.

To reduce the number of parameters and improve the generalization ability of the model, the structural reparameterization is utilized in DSC and BottleneckCSP. Compared with Model 5, Model 6 has shown an increase of 1.0 and 2.1% in mAP and R. To accelerate convergence speed and reduce the imbalance rate of sample categories, the loss function Focal-EIOU is implemented, and the mAP value is increased by 1.9%, reaching to 82.2%. As a result,

Table 2 Cases of ablation study

	YOLOv5s	Detection head	Simplified GFPN	CoT	DSC	Reparameterization	Focal-EIOU
Baseline	P						
Model 1	P	P					
Model 2	P		P				
Model 3	P	P	P				
Model 4	P	P	P	P			
Model 5	P	P	P	P	P		
Model 6	P	P	P	P	P	P	
CRGF-YOLO	P	P	P	P	P	P	P

P means selecting the corresponding modules in the table

The significance of bold fonts indicates the optimal values under the same parameters

Table 3 Evaluation metrics about corresponding cases evaluated on the NEU-DET dataset

	Baseline	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	CRGF-YOLO
Precision (P)	72.2	72.5	74.0	74.5	75.2	73.1	74.5	76.3
Recall (R)	72.5	75.8	74.1	73.3	74.1	75.6	77.7	76.8
mAP	74.6	77.5	77.0	78.2	79.6	79.3	80.3	82.2
Params (M)	7.3	12.0	10.7	13.9	17.6	13.4	13.3	13.3
Inference Time (ms)	7.2	8.7	9.3	10.7	11.3	14.5	15.6	16.1
GFLOPs	16.8	17.1	23.3	20.6	22.9	17.9	22.7	22.7

The significance of bold fonts indicates the optimal values under the same parameters

CRGF-YOLO shows increase in mAP and P of 7.7 and 4.1% compared to YOLOv5s.

4.4 Experiments on Attention Mechanisms with Grad-CAM

It is necessary to have a better understanding of the network's capacity to extract efficient features by looking at the regions of predicting a class. Grad-CAM [31] is introduced to visualize the importance of detecting regions in convolutional layers and reflect them in the form of heat maps. Grad-CAM calculates significant weights in each channel on feature maps based on the gradient ascent strategy, creating heat maps to represent the model's region of interest. This improves the visibility and interpretability of deep neural networks and helps users understand models better.

Experiments are performed on the NEU-DET dataset and Grad-CAM is applied to the networks below. To verify that CoT module with self-attention mechanism has better effects on feature extraction in the network, CoT can be placed with SE [32], CBAM (Convolutional Block Attention Module) [33], SimAM [34], and ECA [35], then compare the visualization results of CoT-integrated network (Network + CoT) with baseline (Model 6 without CoT modules), SE-integrated network (Network + SE), CBAM-integrated network (Network + CBAM), SimAM-integrated (Simple Attention Module) network (Network + SimAM), and ECA-integrated (Efficient Channel Attention) network (Network + ECA). In every network, six images are selected with different types of defects for comparison, and the Grad-CAM helps to display the focus area and the confidence. Figure 7 demonstrates the visualization results of the network integrated with different attention mechanisms, and relevant evaluation metrics about Grad-CAM visualization results on different networks are shown in Table 4.

In Fig. 7, Grad-CAM shows attended area of various types of defects, and visualization results have been improved by adding attention mechanisms at different levels. Compared with the baseline network, CoT-integrated network has best performance of extracting effective information and covers largest attended areas in target regions. It can be seen that complex background textures have negative impacts on extracting efficient features, which is particularly evident in the pictures of crazing, pitted surface, and rolled-in scale. The CoT module enhances feature expression by establishing the contextual relationship between information to effectively locate and enlarge the attended area. This improves the network's performance and reduces interference from background factors.

Integrated with data from Table 4, it can be shown that mAP values of different networks have been increased with attention mechanisms. Although the number of parameters has increased by 40%, the CoT-integrated network outperforms,

improving 2.4%, 4.5%, and 3.8% in P, R and mAP compared with Network. It can be seen that complex background textures have negative impacts on extracting efficient features, which is particularly evident in the pictures of crazing, pitted surface, and rolled-in scale. The CoT module enhances feature expression by establishing the contextual relationship between information to effectively locate and enlarge the attended area. This proves that CoT module capitalizes on the contextual information to strengthen the capacity of visual representation and transfers useful information to the network.

4.5 Comparison of Different Models

Table 5 shows the results of different detectors evaluated on the NEU-DET dataset. To measure the effectiveness of defect detection, it is important to consider both the detecting speed and the accuracy rate. Metrics such as mAP, Inference Time, and Params are important indicators when evaluating these capabilities. Under the experimental environment and setting parameters in this paper, six representative existing methods are selected for comparative experiments: Faster R-CNN, YOLOv3_spp [36], YOLOv5s, YOLOv7s [37], YOLOX_s [38], and YOLOv8s [39].

As shown in Table 5, Faster R-CNN achieves satisfactory mAP values in Patches and Scratches, but its model size is the largest and Inference Time is the longest, making it unsuitable for lightweight deployment. In comparison to Faster R-CNN, the YOLO series has less Inference Time and smaller model sizes, making it a more efficient option. As a comparative algorithm, YOLOv5s has the smallest model size and fastest detection speed, but mAP value is not satisfactory. On the other hand, our FPS is slower 46% than YOLOv5s because of the increased model size. Influenced by complex textures, Crazing has obscure boundaries and its performance in terms of mAP values is significantly lower than other types. Depending on the CoT module with the self-attention mechanism and multi-scale feature fusion from different layers, the ability of feature expressions is enhanced. As a result, the performance of Crazing is increased by 17.2% compared to YOLOv5s. In Addition, mAPs of other types have also been improved at different levels, which demonstrates CRGF-YOLO has strong ability to detect defections with different scales, and FPS is improved than that of Faster R-CNN. Some visualization results from CRGF-YOLO on the NEU-DET dataset are shown in Fig. 8.

5 Limitations and Future Research

The improved model based on YOLOv5s outperforms other algorithms. However, the interference of complex background noise can affect the effectiveness of detecting some defections. It is crucial to explore methods to minimize the

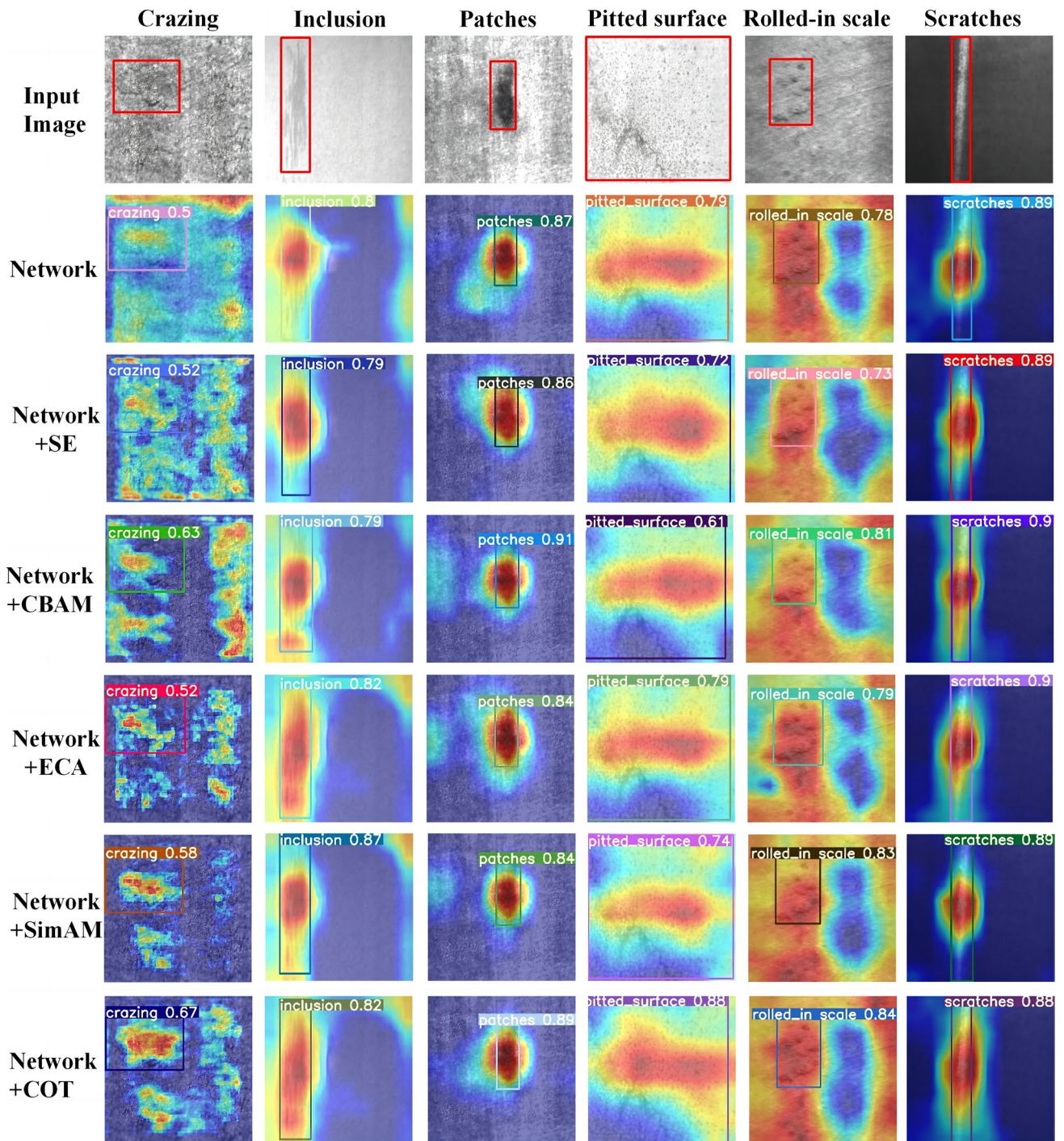


Fig. 7 Grad-CAM visualization results of networks with different mechanisms

impact of background noise on defect targets and enhance inspection outcomes. In addition, for categories that exhibit minimal differences within classes, conducting research on edge detections can significantly enhance the classification accuracy of similar categories. Furthermore, the deep-learning framework encompasses numerous hyperparameters,

necessitating further experiments to identify the optimal detection parameters and avoid falling into the local optimal dilemma.

To enhance the performance in defect detection, future work can explore the use of Generative Adversarial Networks (GAN) and its variants to expand the dataset. In

Table 4 Evaluation metrics about Grad-CAM visualization results on networks with different attention mechanisms

Metrics	Network	Network +SE	Network +CBAM	Network +ECA	Network +SimAM	Network +CoT
P	72.1	71.3	70.7	73.1	72.6	74.5
R	73.2	73.3	73.8	71.3	76.1	77.7
mAP	76.5	76.7	77.3	76.8	78.6	80.3
Params (M)	9.5	10.1	13.4	9.9	9.9	13.3
GFLOPs	13.7	19.0	22.9	18.9	18.9	22.7

The significance of bold fonts indicates the optimal values under the same parameters

Table 5 The results of different networks evaluated on the NEU-DET dataset

Methods	Faster R-CNN	YOLOv3	YOLOv5s	YOLOv7s	YOLOX_s	YOLOv8s	CRGF-YOLO
Crazing	47.4	50.8	42.3	36.8	39.0	50.1	59.5
Inclusion	78.6	87.0	86.5	86.5	80.3	89.1	90.0
Pitted surface	86.9	92.9	93.0	96.1	78.6	78.1	94.5
Patches	95.2	76.2	78.5	76.6	77.0	93.1	89.6
Scratches	95.7	70.7	59.6	53.3	62.0	89.2	70.9
Rolled-in scale	60.5	86.9	86.9	83.1	86.1	72.5	88.5
mAP	77.4	77.4	74.6	72.1	68.8	78.7	82.2
Params (M)	137.1	9.6	7.3	9.3	8.9	11.1	13.3
GFLOPs	20.2	23.6	17.0	26.7	18.2	28.4	22.7
Inference Time (ms)	58.9	7.3	7.2	14.2	7.8	14.5	16.1

The significance of bold fonts indicates the optimal values under the same parameters

addition, improving the preprocessing of images by suppressing background noise and enhancing local feature information can help improve the accuracy and robustness of the model. It may not be the optimal one to choose YOLO as the basic model. Other CNNs or even building custom networks can be taken into consideration, as well as more improved methods, such as dilation convolution and other transformer variants, which benefits the network to increase the receptive fields of feature extraction. Although different networks may perform well in validation on datasets, they often necessitate on-site testing and fine-tuning in various industrial scenarios to meet requirements.

6 Conclusions

As the research on CNNs for steel defect detection progresses, researchers have started evaluating on real-time collected defect datasets to train models and improve detection capabilities. This study specifically focuses on making improvements to YOLOv5s, realizing exceptional outcomes on the NEU-DET dataset.

In this paper, an enhanced YOLO model called CRGF-YOLO is designed for detecting steel surface defects. Based on YOLOv5s, four detection heads are employed to detect targets with different scales, and the performance

of detecting small targets is improved. The CoT module utilizes the self-attention mechanism to improve feature expression by connecting contextual information from different feature maps, resulting in more efficient information transfer to the neck part. On the other hand, the simplified GFPN facilitates multi-scale feature fusion and enables effective information exchange between different feature maps. By utilizing Grad-CAM for visualizing the results, it can be found that the CoT-GFPN network outperformed the others and covered the largest attended areas in detecting regions. Furthermore, Rep-DSC and Rep-CSP implement the structural reparameterization to create nonlinear diversified connections and improve the generalization ability of the network, maintaining high detection accuracy and a compact model. Finally, the loss function Focal-EIOU is utilized to accelerate model convergence and improve the imbalance of sample categories. Based on the experimental results and evaluations, it is concluded that CRGF-YOLO outperforms at the training and validation stage, with the mAP value of 82.2% compared to other detectors, which indicates that the improved model has both high detection accuracy and meet real-time requirements.

This study makes significant contributions by offering valuable insights and practical guidance for enhancing defect detection technology in the steel industry.

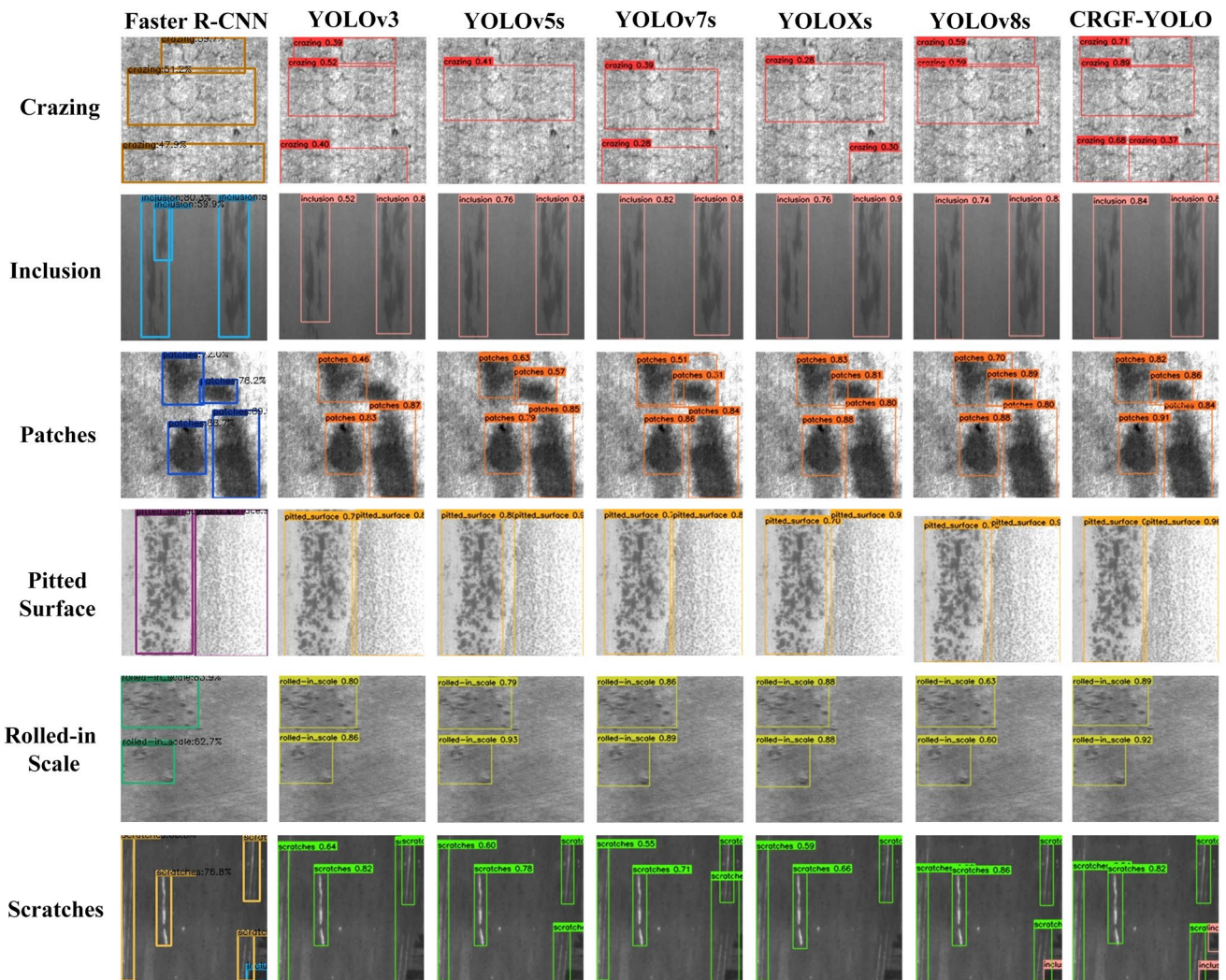


Fig. 8 Comparison of visualization results from different models on the NEU-DET dataset

By overcoming the drawbacks of conventional methods and harnessing innovative techniques, this research sets the stage for more robust and effective defect inspection systems. Ultimately, it aids in strengthening product quality and minimizing economic losses within steel manufacturing.

Author Contributions Conceptualization, T.Y.; methodology, T.Y., Q.L.; investigation, T.Y.; software, T.Y.; writing—original draft preparation, T.Y.; writing—review and editing, T.Y., X.L. and Q.L.; visualization, X.L. and L.L.; supervision, X.L. and L.L.; project administration, L.L. All authors have read and agreed to the published version of the manuscript.

Funding This research was supported by the National Natural Science Foundation of China (Grant No. 52075539).

Data Availability The data presented in this paper is: www.kaggle.com/datasets/ousmanesangary/neu-det.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Informed Consent Not applicable.

Institutional Review Board Statement Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Tang, B., Chen, L., Sun, W., Lin, Z.K.: Review of surface defect detection of steel products based on machine vision. *Iet Image Process.* **17**, 303–322 (2023)
- Wang, Y.L., Xia, H.B., Yuan, X.F., Li, L., Sun, B.: Distributed defect recognition on steel surfaces using an improved random forest algorithm with optimal multi-feature-set fusion. *Multimed. Tools Appl.* **77**, 16741–16770 (2018)
- Liu, Y., Xu, K., Xu, J.W.: An improved MB-LBP defect recognition approach for the surface of steel plates. *Appl. Sci. Basel* (2019). <https://doi.org/10.3390/app9204222>
- Park, C., Choi, S., Won, S.: Vision-based inspection for periodic defects in steel wire rod production. *Opt. Eng.* (2010). <https://doi.org/10.1117/1.3284779>
- Choi, D.C., Jeon, Y.J., Yun, J.P., Kim, S.W.: Pinhole detection in steel slab images using Gabor filter and morphological features. *Appl. Optics* **50**, 5122–5129 (2011)
- Hou, J., Xia, K., Fan, Y., Aslam, N.: Surface defects detection of strip steel based on semi-definite programming SVM. *J. Inf. Comput. Sci.* **12**, 3137–3145 (2015)
- Zhu, C.X., Yuan, H.T., Ma, G.H.: An active visual monitoring method for GMAW weld surface defects based on random forest model. *Mater. Res. Express* (2022). <https://doi.org/10.1088/2053-1591/ac5a38>
- Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. Acn* **60**, 84–90 (2017)
- Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. *Ieee Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017)
- He, Y., Song, K.C., Meng, Q.G., Yan, Y.H.: An End-to-end steel surface defect detection approach via fusing multiple hierarchical features. *Ieee Trans. Instrum. Meas.* **69**, 1493–1504 (2020)
- Xie, Y., Dong, Y., He, H.: Defect detection of printed circuit board based on small target recognition network. In: the 23rd International Conference on electronic packaging technology (ICEPT), pp. 1–5. 10–13 August 2022
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 779–788. Las Vegas, NV, USA, 27–30 June 2016
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: the European Conference on computer vision, Amsterdam, Netherlands, pp. 21–37. (2016)
- Li, J.Y., Su, Z.F., Geng, J.H., Yin, Y.X.: Real-time detection of steel strip surface defects based on improved YOLO detection network. *Ifac Papersonline* **51**, 76–81 (2018)
- Xie, H.S., Zhang, Y.F., Wu, Z.S.: An improved fabric defect detection method based on SSD. *Aatcc J. Res.* **8**, 182–191 (2021)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: transformers for image recognition at scale. In: the 9th International Conference on learning representations, Virtual, Online, 3–7 May 2021
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. *arXiv*, arXiv: 2005.12872 (2020)
- Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: the 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp. 7794–7803. 18–23 June 2018
- Bello, I., Zoph, B., Le, Q., Vaswani, A., Shlens, J.: Attention augmented convolutional networks. In: the 2019 IEEE/CVF International Conference on computer vision (ICCV), pp. 3285–3294. Seoul, Korea (South), 27 October–2 November 2019
- Srinivas, A., Lin, T., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: the 2021 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 16514–16524. 20–25 June 2021
- Li, Y., Yao, T., Pan, Y., Mei, T.: Contextual transformer networks for visual recognition. *Ieee Trans. Pattern Anal. Mach. Intell.* **45**, 1489–1500 (2023)
- Chen, P.Y., Chang, M.C., Hsieh, J.W., Chen, Y.S.: Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *Ieee Trans. Image Process.* **30**, 9099–9111 (2021)
- Ghiasi, G., Lin, T., Le, Q.V.: NAS-FPN: learning scalable feature pyramid architecture for object detection. In: the 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 7029–7038. 15–20 June 2019
- Yilmaz, M.A., Tekalp, A.M.: DFPN: deformable frame prediction network. In: the 2021 IEEE International Conference on image processing (ICIP), pp. 1944–1948. Anchorage, AK, United States, 19–22 September 2021
- Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: the 2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 10778–10787. 13–19 June 2020
- Jiang, Y., Tan, Z., Wang, J., Sun, X., Lin, M., Li, H.: GiraffeDet: a heavy-neck paradigm for object detection. *arXiv*, arXiv: 2202.04256 (2022)
- Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: RepVgg: making VGG-style ConvNets great again. In: United states, pp. 13728–13737. (2021)
- Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv*, arXiv: 1704.04861 (2017)
- Zhang, Y.F., Ren, W.Q., Zhang, Z., Jia, Z., Wang, L., Tan, T.N.: Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **506**, 146–157 (2022)
- Song, K., Yan, Y.: A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **285**, 858–864 (2013)
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020)
- Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: the 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp. 7132–7141. 18–23 June 2018
- Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: the 15th European Conference on computer vision (ECCV), pp. 3–19. Munich, Germany (2018)
- Yang, L.X., Zhang, R.Y., Li, L.D., Xie, X.H.: SimAM: a simple, parameter-free attention module for convolutional neural networks. In: the International Conference on machine learning (ICML), vol. 139, Electr Network, 18–24 July 2021

35. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-net: efficient channel attention for deep convolutional neural networks. In: the 2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 11531–11539. 13–19 June 2020
36. Redmon, J., Pag, A.F.: YOLOv3: An incremental improvement. *arXiv*, [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
37. Wang, C., Bochkovskiy, A., Liao, H.M.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*, [arXiv: 2207.02696](https://arxiv.org/abs/2207.02696) (2022)
38. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., YOLOX: exceeding YOLO series in 2021. *arXiv*, [arXiv: 2107.08430](https://arxiv.org/abs/2107.08430) (2021)
39. Reis, D., Kupec, J., Hong, J., Daoudi, A.: Real-time flying object detection with YOLOv8. *arXiv*, [arXiv:2305.09972](https://arxiv.org/abs/2305.09972) (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.