2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
June 28-30, 2023. Seattle, Washington, USA

BiSPD-YOLO: Surface Defect Detection Method for Small Features and Low-resolution Images

Sixu Yan, Gaoming Chen, Ao Gao, Chao Liu and Zhenhua Xiong

***Abstract*—At present, deep learning objective detection method based on learning features suffer from low detection rates and poor accuracy rates in metal surface defect detection. This is primarily due to the fact that the detected images are mostly gray images with small features and low resolution, which makes the model inefficient to train and slow to converge. This paper proposes a BiSPD-YOLO metal surface defect detection model based on YOLOv5 to solve these problems. Firstly, this model uses SDP-Conv module to replace the traditional strided convolution and pooling to enhance the training of the network for low-resolution images; BiFPN is then used to replace PANet for multi-scale feature fusion. In this way, small features in the images can be better extracted; Finally, the original loss function of YOLOv5 is improved, and the SIOU function is used to optimize the training model. The testing results on the NEU-DET dataset after data augmentation indicate that the improved model mAP achieves 97.2%, which is 4.1% higher than the original model, and is superior to other mainstream models. Compared to the original model, the detection speed is basically unchanged, and can quickly and accurately detect metal surface defects in real time.**

***Index Terms*—Surface defect detection, YOLOv5, SPD-Conv, BiFPN, SIOU.**

I. Introduction

At present, deep learning-based objective detection methods have been commonly applied to metal surface defect detection and can be divided into two categories: Two-stage objective detection methods based on candidate regions, such as R-CNN [1], SPP-Net [2], Fast R-CNN [3] and Faster R-CNN [4], and one-stage objective detection methods based on regression, such as SSD [5], RetinaNet [6] and YOLO series. However, due to a variety of difficulties in actual metal defect detection, such as noise, fuzzy defect boundary, complex background and different defect types, the classical objective detection models are plagued with problems such as slow detection speed and low detection accuracy. To solve these problems, many researchers have started to improve the classical model.

In recent years, many researchers have improved the classical two-stage objective detection methods to make them more applicable to metal surface defect detection. In 2019, Han et al. [7] designed a new metal defect detection model based on the encoder-decoder residual networks (EDR-Net), which overcame the difficulties of large amount of noise and background blur in metal defect detection, and verified its detection accuracy on the SD-saliency-900 dataset. In 2020, He et al. [8] proposed an end-to-end steel surface defect detection model based on multi-layer feature fusion. This model is based on Faster R-CNN, combines multiple hierarchical features into one feature through multi-layer feature fusion network (MFN), and achieves higher detection accuracy on NEU-DET dataset. In 2022, Li et al. [9] put forward a two-stage defect detection model. In the first stage, the improved YOLOv5 was utilized to optimise the feature extraction network to improve the network's capability to extract image features; in the second stage, Inception-ResnetV2 is adopted and CBAM attention mechanism module is embedded to achieve accurate positioning of steel surface defects.

Although the two-stage objective detection method has a higher detection accuracy than the one-stage objective detection method, its detection speed is relatively slow, making it unsuitable for real industrial scenarios. Therefore, in order to make real-time detection of metal defects a reality, many improved one-step objective detection methods have been proposed. In 2018, Li et al. [10] proposed a fully convolved improved YOLO model, which was evaluated on six different defects and achieved good mAP and high recall. In 2021, Kou et al. [11] proposed an improved YOLOv3 model using an anchor-free selection scheme and a high-density convolution module designed in the network. The model achieves high recognition accuracy on the GC10-DET as well as the NEU-DET datasets. In the same year, Lv et al. [12] proposed an EDDN defect detection model based on SSD to be applied to metal surface defect detection at different scales, and proved that the model has good robustness and detection accuracy on GC10-DET dataset. In 2022, Guo et al. [13] designed the MSFT-YOLO one-stage detection model in combination with Transformer and YOLOv5. This model designs a TRANS module on the basis of Transformer, adds it to the backbone and neck of YOLOv5, and integrates a weighted BiFPN for feature fusion of different scales. Its performance is verified on the NEU-DET dataset.

The improved models have improved the detection accuracy and recognition accuracy compared with the original models, but there are still problems with low detection efficiency, high false positives and misses when processing low resolution detection images. The main reason for this is that researchers focus on extracting small feature information in the images. They ignore the effect of image quality on detection. Existing research shows that with the reduction of the resolution of the detected image, the detection accuracy and classification

This work was supported in part by the National Natural Science Foundation of China under Grant 52175479 and the Major Science and Technology Projects for Self-Innovation of FAW (20210301032GX).

Sixu Yan, Gaoming Chen, Ao Gao, Chao Liu and Zhenhua Xiong are with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China (e-mail: yansixu_sjtu@sjtu.edu.cn; cgm1015@sjtu.edu.cn; gagagaga152@sjtu.edu.cn; aalon@sjtu.edu.cn; mexiong@sjtu.edu.cn)

**Corresponding author: Zhenhua Xiong.*

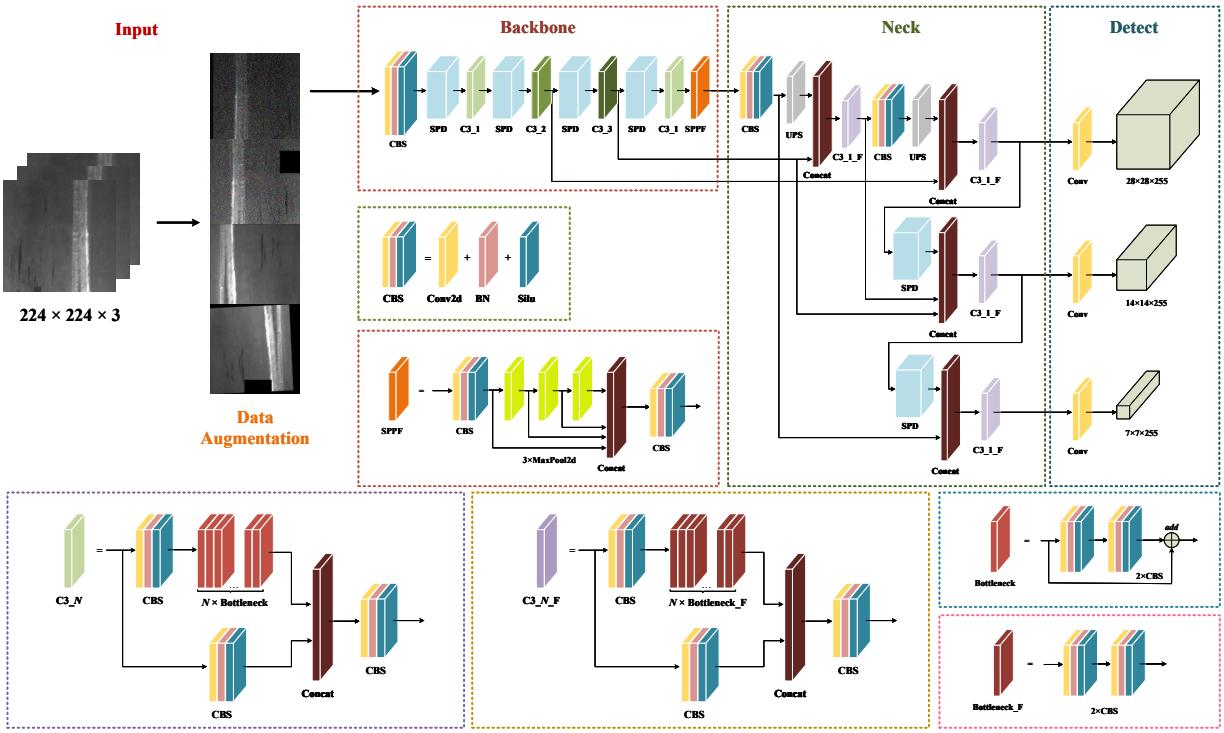

Fig. 1. Overall Structure of the BiSPD-YOLO.

accuracy of the image will also decline [14]. In order to better identify metal surface defects, this paper comprehensively considers the impact of small features and low-resolution images on the detection results, and proposes a new metal surface defect method BiSPD-YOLO based on YOLOv5. The improvements to the YOLOv5 model in this paper are as follows:

- BiFPN is used as a replacement for the PANet for feature fusion in the neck network;
- Use SPD-Conv module instead of traditional strided convolutional and pooling layers for image feature extraction, reducing the loss of fine-grained information in the image extraction process;
- The bounding box loss function of YOLOv5 is refined and the SIOU function is used to optimise the training model.

II. Methodology

At present, YOLOv5 has five releases, namely 5n, 5s, 5m, 5l and 5x. The BiSPD-YOLO model proposed in this paper is based on YOLOv5s. Based on the original model, BiFPN is used to replace PANet for multi-scale feature fusion; SPD-Conv module is employed to implement the strided convolution and pooling layer in the original network; SIOU loss function is adopted to substitute CIOU loss function.

A. The Structure of BiSPD-YOLO

The overall structure of BiSPD-YOLO is depicted in Fig. 1. The input dataset is augmented at the network input end to enrich the detection data and increase the discrimination between different defect categories. The backbone network is mainly responsible for acquiring the features of images at different scales. The input images are extracted in SPD-Conv and C3 modules and then enter SPPF for downsampling to reduce the network dimension and parameters. The neck network uses BiFPN to fuse multi-scale features of the input image to reduce the redundant information generated by the correlation between different features and improve the detection accuracy. SIOU loss function can accelerate network convergence, improve network training speed and enhance model robustness. Finally, the detector will output three different scale features of 7×7, 14×14 and 28×28 images as the prediction branch.

B. Improve FPN With BiFPN

Multi-scale feature fusion refers to the fusion of feature maps from different scales in a deep learning model to improve the performance and robustness of the model. Its effects include improving the perceptual field of the model, improving the robustness of the model, improving the classification accuracy of the model and accelerating the training and inference of the model. In YOLOv5, PANet is used as a multi-scale feature fusion network. To enhance the model's ability to detect small objective defects to a greater extent, this paper chooses BiFPN to replace PANet for image feature fusion in multi-scale. BiFPN is an improved version of PANet. It simplifies the network structure of PANet, and introduces weights in the fusion process to balance feature information of different

scales. The structure of PANet and BiFPN is illustrated in Fig. 2.

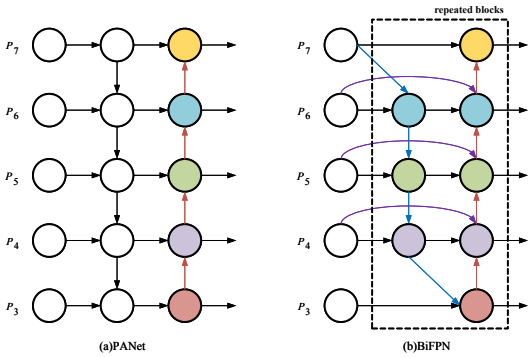

Fig. 2. PANet and BiFPN Structure. (a) PANet. The feature fusion network in YOLOv5. (b) BiFPN. The feature fusion network in BiSPD-YOLO.

Fig. 2(a) shows the structure of PANet [15], which includes two feature fusion paths: top-down and bottom-up. However, model complexity increases and network training efficiency decreases as the parameters increases. The various feature fusion result of node 6 in PANet is as follows:

$$P_6^{td} = Conv\left[P_6^{in} + Resize\left(P_7^{td}\right)\right] \tag{1}$$

$$P_6^{out} = Conv\left[P_6^{td} + Resize\left(P_5^{out}\right)\right] \tag{2}$$

where P_i^{in} represents the input characteristics of layer i, P_i^{td} represents the top-down intermediate characteristics of layer i, and P_i^{out} represents the bottom-up output characteristics of layer i.

Fig. 2(b) shows the structure of BiFPN. Based on PANet, BiFPN [16] has made the following improvements: removing the node of single input side, simplifying the network structure, and reducing the effort of parameter calculation, such as the intermediate node of layer 7 in Fig. 2(a); an edge is added from input node to output node of each layer to fuse more features with less cost and enhance the feature representation ability; fast normalized fusion method is adopted to adjust the proportion of features of different scales in the fusion, so as to increase the percentage of useful features.

The fast normalized fusion method is as follows:

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \cdot I_i \tag{3}$$

where I_i and w_j represent the input characteristics and corresponding weights. Here, ε is taken as 0.0001 to avoid instability in numerical calculation.

The feature fusion results of node 6 in BiFPN are as follows:

$$P_6^{td} = Conv\left[\frac{w_1 \cdot P_6 + w_2 \cdot Resize\left(P_7^{in}\right)}{w_1 + w_2 + \varepsilon}\right] \tag{4}$$

$$P_6^{out} = Conv\left[\frac{w_1\prime \cdot P_6^{in} + w_2\prime \cdot P_6^{td} + w_3\prime \cdot Resize\left(P_5^{out}\right)}{w_1\prime + w_2\prime + w_3\prime + \varepsilon}\right] \tag{5}$$

BiFPN simplifies the network structure of PANet and incorporates more features of different scales. In addition, BiFPN also gives each feature a learnable weight, which greatly enhances the generalization ability of the model. Since BiSPD-YOLO has stronger feature extraction capability and extracted features have better classification performance.

C. SPD-Conv Module

When the image resolution is low or the detection objective is small, the performance of the YOLO model will decrease rapidly. This is due to the layout of strided convolution layers and pooling layers for downsampling in the detection network, which leads to the loss of fine-grained information and thus reduces the model's ability to extract features. To deal with this problem, Raja Sunkara et al. [17] proposed the SPD-Conv module, which does not lose the original feature information of the image during downsampling.

The SPD-Conv module consists of a space-to-depth (SPD) layer and a non-strided convolution layer. After the SPD layer downsamples the image, it preserves all the information in the original dimension of the channel and does not cause any loss of information. Learnable parameters in the non-strided convolution layer are used for reducing the number of channels increased by the SPD layer, so that the image is downsampled with as little loss of original information as possible.

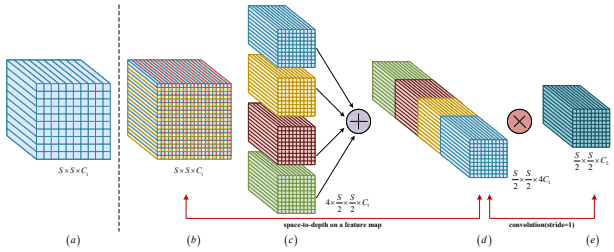

Fig. 3. The SPD-Conv module ($scale$ = 2).

Now consider the intermediate feature map X of size $S \times S \times C_1$. The processing of the feature map by SPD-Conv can be divided into the following steps:

1) Slice the intermediate feature map X with size $S \times S \times C_1$ according to the given $scale$. After slicing, feature submaps with number of $scale$ and size $\frac{S}{scale} \times \frac{S}{scale} \times C_1$ will be obtained. The slicing method refers to Eq. (6), where $i, j \in [0 : scale - 1]$. Fig. 3(a)(b)(c) shows the slicing results when $scale = 2$. After slicing the intermediate feature map X, four submaps are obtained, which are $f_{0,0}$, $f_{1,0}$, $f_{0,1}$ and $f_{1,1}$ respectively, and the size of each one is $\frac{S}{2} \times \frac{S}{2} \times C_1$;

$$f_{i,j} = X\left[i : S : scale, j : S : scale\right] \tag{6}$$

2) Connect all submaps along the channel dimension to obtain a new feature map X'. Compared with the original feature map X, its spatial dimension is reduced by $scale$ and the channel dimension is increased by $scale^2$. The function of the SPD layer is to convert the feature map

X with the size of $S \times S \times C_1$ into the feature map $X^{'}$ with the size of $\frac{S}{scale} \times \frac{S}{scale} \times scale^2 C_1$. Fig. 3(d) shows the conversion results at $scale = 2$;

3) Input the feature map $X^{'}$ of the output of the SPD layer to the non-strided convolution layer with a C_2 filter to reduce the channel dimension. The feature map $X^{'}$ with size $\frac{S}{scale} \times \frac{S}{scale} \times scale^2 C_1$ is converted into the feature map $X^{''}$ with size $\frac{S}{scale} \times \frac{S}{scale} \times C_2$ at the non-strided convolution layer, where $C_2 < scale^2 C_1$.

D. Improve Loss Function with SIOU

Surface defect detection consists of two types of tasks, a classification task, where the identified defects are classified, and a bounding box regression task, i.e. defect localization, which requires loss regression on the predicted bounding box. In the YOLOv5 model, the CIOU loss function is used to estimate the bounding box loss. The calculation of the CIOU loss is defined in Eq. (7).

$$L_{CIOU} = 1 - IOU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \tag{7}$$

In Eq. (7), $\mathbf{b}$ and $\mathbf{b}^{gt}$ represent the center points of anchor box B and target box B^{gt} respectively; $\rho(\bullet)$ is the Euclidean distance between two points; c is the diagonal length of the minimum enclosing rectangle containing the anchor box and the target box; $IOU = |B \cap B^{gt}| / |B \cup B^{gt}|$ is used to reflect the degree of overlap between the anchor box and the target box; v and α are used to measure the difference in the shape (or aspect ratio) of the anchor box and the target box.

The CIOU loss function is designed with full consideration of the Euclidean distance, overlap area and shape between the anchor box and the target box. However, it does not take into account the mismatch in direction between the anchor box and the target box. This shortcoming causes the anchor box to constantly "wander" during the training process, resulting in too slow convergence. To solve this problem, Zhora Gevorgyan [18] proposed a new loss function SIOU, which takes into account the angle between the line connecting the centres of the target and prediction anchors and the coordinate axes when calculating the loss function. The composition of the SIOU loss function is as follows:

1) Angle Loss. The definition is as follows:

$$\Lambda = 1 - 2\sin^2\left(\alpha - \frac{\pi}{4}\right) \tag{8}$$

2) Distance Loss. The definition is as follows:

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma \rho_t}\right) = 2 - e^{-\gamma \rho_x} - e^{-\gamma \rho_y} \tag{9}$$

where $\rho_x = \left(\frac{b_x^{gt} - b_x}{w^c}\right)^2, \rho_y = \left(\frac{b_y^{gt} - b_y}{h^c}\right)^2, \gamma = 2 - \Lambda$.

3) Shape Loss. The definition is as follows:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-w_t}\right)^{\theta} = \left(1 - e^{-w_w}\right)^{\theta} + \left(1 - e^{-w_h}\right)^{\theta} \tag{10}$$

where $w_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, w_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})}, \theta \in [2, 6]$.

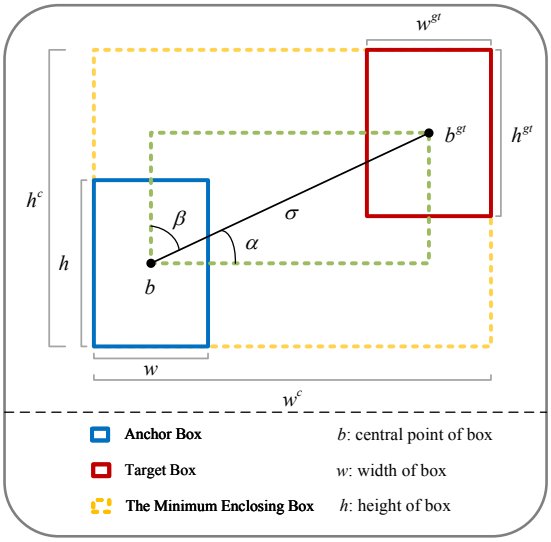

Fig. 4. SIOU Loss Calculation Scheme.

4) IOU Loss. The definition is as follows:

$$L_{IOU} = 1 - IOU \tag{11}$$

The meanings of other parameters are shown in Fig. 4. The complete loss function of the SIOU can be expressed as

$$L_{SIOU} = 1 - IOU + \frac{\Delta + \Omega}{2} \tag{12}$$

Since the SIOU loss function adds to the calculation of the loss in the relative direction from the anchor box to the target box, the problem that YOLOv5 converges slowly when using the CIOU loss function is solved, and accuracy of locating defects is improved.

III. Experiment

A. Experimental Settings

BiSPD-YOLO model is build on the Pytorch deep learning framework and runs on NVIDIA GeForce RTX 3060. Python is used for programming and CUDAv11.7 and CUDNNv8.6 for GPU acceleration. Some parameters during model training are given in TABLE I.

TABLE I
Model Partial Hyperparameters Setting

weight decay	batch size	learning rate	momentum	epoch
0.0005	16	0.01	0.937	300

B. Experimental Dataset and Data Augmentation

This experiment uses the public NEU-DET [8] dataset as training data. This dataset contains 6 common surface defects of hot-rolled steel, including cracks (CR), inclusions (IN), patches (PA), pitted surface (PS), rolled scale (RS), and

scratches (SC). The dataset consists of 1800 grey-scale images of 200×200 size, 300 images for each defect type, and is the most authoritative and representative dataset in steel surface defect detection.

During this experiment, some data augmentation operations including cropping, rotation, translation, brightness adjustment and adding salt and pepper noise, etc. are used to ensure that the original defect types, number and relative positions in the image remain unchanged after the operation is applied to the image. The dataset after data augmentation is 5 times of the original dataset, with a total of 9000 pictures. The data augmentation results of patch and scratch defects are shown in Fig. 5. The training set, verification set and test set are divided in the ratio 8:1:1.

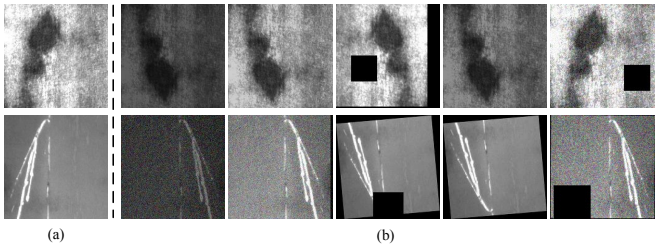

Fig. 5. Data Augmentation Results for Patch and Scratch Defects. (a) Original Images. Original images of patch and scratch defects. (b) Images After Data Augmentation. Images of patch and scratch defects after data augmentation.

C. Model Training

To verify the capability of BiSPD-YOLO for defect detection on small features and low resolution images, it is compared with other typical defect detection methods. It is easy to see from TABLE II that YOLOv5s and BiSPD-YOLO are lightweighted network models compared to two-stage Faster R-CNN and one-stage SSD, Retinanet, YOLOv3 and YOLOv4. Despite the computational complexity of the BiSPD-YOLO model being twice that of YOLOv5s, the average accuracy is 4.1% higher than YOLOv5s, which is much higher than YOLOv3, YOLOv4, Retinanet and SSD. In addition, the parameters of BiSPD-YOLO are much lighter than those of YOLOv3 and YOLOv4, and will not increase the detection speed too much. It can be observed that the BiSPD-YOLO model effectively identifies defects on the steel surface, which is superior to the comparative objective detection method.

TABLE II
COMPARE EXPERIMENTAL RESULTS

Algorithm	Backbone	mAP	Parameters(M)	FPS	GFLOPs
SSD	VGG-16	0.773	24.83	12.7	67.49
Retinanet	ResNet-50	0.846	36.17	12.6	40.31
Faster R-CNN	ResNet-50	0.913	41.14	10.3	51.7
YOLOv3	Darknet53	0.829	61.51	15.1	154.6
YOLOv4	CSPDarknet53	0.897	64.36	23.4	29.95
YOLOv5s	CSPDarknet53	0.931	**7.02**	**33.3**	**15.8**
BiSPD-YOLO	CSPDarknet53	**0.972**	8.53	30.2	33.1

For the purpose of comparison of model detection speed, this paper tests the detection speed of each model separately. The experiments are conducted on the same GPU. The detection speed of BiSPD-YOLO model decreases by 2.1FPS compared with that before the improvement, and increases by 17.5FPS, 17.6FPS, 19.9FPS, 15.1FPS and 6.8FPS compared with SSD, Retinanet, Faster R-CNN, YOLOv3 and YOLOv4, respectively, which also confirms the effectiveness of the improved model, it can ensure the rapid and accurate detection of defects on the surface of the steel.

Fig. 6 shows the defects detection results of YOLOv5 and BiSPD-YOLO. It can be seen from the results that the BiSPD-YOLO is more adequate for the detection of defects, as shown in the Fig. 6 for the detection of cracking defects, the YOLOv5 did not detect defects at adjacent locations. Moreover, for the same defect, BiSPD-YOLO has a better detection effect, and the confidence of the detection results is higher than that of YOLOv5.

D. Ablation Experiment on BiSPD-YOLO

To better investigate the contribution of each improvement to the performance of BiSPD-YOLO, an ablation experiment was conducted. As shown in TABLE III, SPD-Conv module contributes the most to mAP improvement, which increases 2.1% compared with YOLOv5s. The use of BiFPN increased the YOLOv5s model mAP by 1.4%. The improvement of the loss function with SIOU has a significant impact on the improvement of the training speed and the acceleration of the model convergence, which is 2.2FPS higher than the original model, but has no significant impact on the detection accuracy. The combination of BiFPN and SPD-Conv increases the mAP of the original model by 3.6%, but the increase in the number of parameters reduced the detection speed by 4.7FPS. Other combination methods also optimized the overall performance of YOLOv5 to varying degrees, and combining the three methods proved to be the best at improving the model's detection accuracy.

TABLE III
ABLATION EXPERIMENT RESULT

BiFPN	SPD-Conv	SIOU	AP						mAP	FPS
			CR	IN	PA	PS	RS	SC		
-	-	-	0.872	0.922	0.991	0.935	0.892	0.972	0.931	33.3
√	-	-	0.902	0.939	0.993	0.943	0.917	0.975	0.945	30.1
-	√	-	0.919	0.943	0.993	0.956	0.927	0.977	0.952	30.8
-	-	√	0.871	0.923	0.991	0.936	0.884	0.975	0.930	**35.5**
√	√	-	0.964	0.953	**0.994**	0.967	0.940	0.940	0.967	28.6
√	-	√	0.936	0.941	**0.994**	0.956	0.939	0.980	0.958	32.0
-	√	√	0.915	0.942	0.993	0.958	0.916	0.977	0.950	32.3
√	√	√	**0.968**	**0.955**	**0.994**	**0.973**	**0.958**	**0.984**	**0.972**	30.2

IV. CONCLUSION

Aiming at the problems of metal surface defect detection, YOLOv5 model is improved in this paper to solve the challenge faced by existing models when dealing with small features and low-resolution images. BiSPD-YOLO uses BiFPN in feature

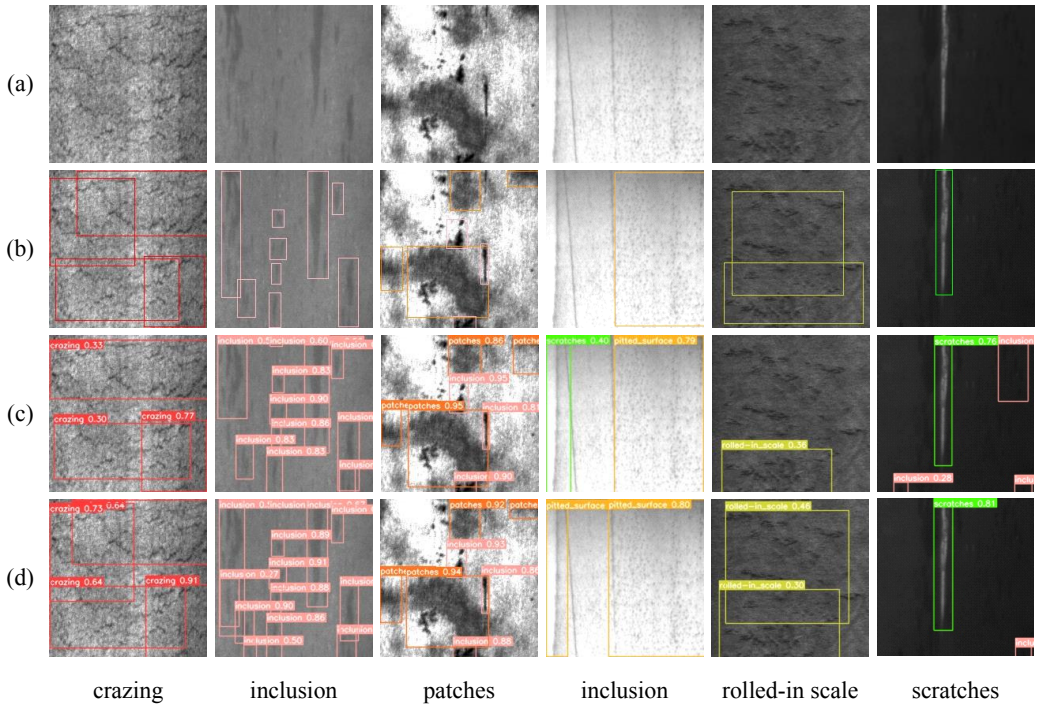

Fig. 6. Detection Results. (a) Original Images. (b) Labelled Images. (c) YOLOv5 Detection Results. (d) BiSPD-YOLO Detection Results.

fusion, a more efficient bi-directional cross-scale and weighted feature fusion approach that fuses more features at less cost and can enhance the characterisation of the fused features; The SPD-Conv maximises the retention of fine-grain information in images; The original CIOU loss function is replaced by the SIOU function, which speeds up convergence and improves the robustness of the model. The method complies with real-time requirements and has a low computational cost. In the experiment, the detection method is matched with other deep learning objective detection methods, and the accuracy rate reaches 97.2%. This can effectively improve the detection performance of metal surface defects.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[3] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[7] C. Han, G. Li, and Z. Liu, "Two-stage edge reuse network for salient object detection of strip steel surface defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[8] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1493–1504, 2020.

[9] Z. Li, X. Tian, X. Liu, Y. Liu, and X. Shi, "A two-stage industrial defect detection framework based on improved-yolov5 and optimized-inception-resnetv2 models," *Applied Sciences*, vol. 12, no. 2, 2022.

[10] 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, 5th IFAC Workshop on Mining, Mineral and Metal Processing MMM 2018.

[11] X. Kou, S. Liu, K. Cheng, and Y. Qian, "Development of a yolo-v3-based model for detecting defects on steel strip surface," *Measurement*, vol. 182, p. 109454, 2021.

[12] X. Lv, F. Duan, J.-J. Jiang, X. Fu, and L. Gan, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors (Basel, Switzerland)*, vol. 20, no. 6, p. E1562, March 2020.

[13] 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.

[14] M. Koziarski and B. Cyganek, "Impact of low resolution on image recognition with deep neural networks: An experimental study," *International Journal of Applied Mathematics and Computer Science*, vol. 28, no. 4, pp. 735–744, 2018.

[15] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[16] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[17] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," *arXiv preprint arXiv:2208.03641*, 2022.

[18] Z. Gevorgyan, "Siou loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022.