Research on Diagnosis and Localization of Moles for Robotic Mole Removal Surgery Based on YOLO

Yanhan Sun

Sino-European School of Technology of Shanghai, Shanghai University, Shanghai, 200444, China

1056298647@shu.edu.cn

Abstract. Surgical treatment is often considered the first choice for mole removal. Treatments such as laser mole removal are widely performed, but intraoperative navigation remains a significant challenge, relying heavily on the experience of surgeons. Additionally, when multiple moles are required to be removed, a more repetitive workload can occur in the surgical procedures. To address these issues, this study aims to achieve precise intraoperative navigation and to achieve modes of automatic diagnosis and localization for robotic mole removal surgery. By training the YOLO model on an open-source categorical mole dataset, this study achieves the automatic diagnosis and localization of the moles during surgeries. In the evaluation of the study, a precision of 0.839 and an average precision mAP of 0.862 which are at a high level are observed. This study verifies that the YOLO can play a critical role in enabling the automatic diagnosis and localization of robotic mole removal surgery.

Keywords: YOLO, mole removal, robotic surgery, object detection, diagnosis and localization.

1. Introduction

Commonly known as moles or nevi, these skin lesions may not only be benign pigmented nevi but also can be malignant melanocytic nevi. Statistics indicate that pigmented nevi affect a large population worldwide [1], while some malignant skin tumors rank as the 5th most common cancer globally [2]. Mole removal surgeries, such as laser surgeries, have undergone development and exploration since the advent of various laser technologies in the mid-20th century [3]. Although the relevant techniques for mole removal during surgery are well-established, the localization of the moles still largely depends on the surgeon's experience. Moreover, removing multiple moles inevitably leads to a repetitive workload. Thus, integrating non-traditional automatic localization technologies, such as augmented reality (AR) and artificial intelligence (AI), with robotic surgery, into mole removal surgeries has significance and potential.

Currently, technologies such as AR and AI have been widely researched and applied in surgeries such as laparoscopy [4] and breast cancer surgery [5], but their application in skin tumor surgeries and mole removal surgery remains limited [6]. Research based on these technologies has primarily focused on the classification of the moles. Yang et al. [7] applied deep learning techniques to study dermoscopic image classification of pigmented nevus. Besides, robotic surgery has not yet made significant advancements in the field of mole removal.

This study is based on the You Only Look Once (YOLO) object detection model, firstly introduced in 2015 by Joseph Redmon and Ali Farhadi from the University of Washington. YOLO has been widely welcomed for its high speed and high accuracy. YOLOv8, the eighth generation of YOLO developed by Ultralytics, introduces various AI tasks, including detection, segmentation, pose estimation, tracking, and classification. In this study, YOLOv8 is trained on an open-source categorical mole dataset and a mole detection weight is obtained. This study explores the diagnosis and localization of the targets during robotic mole removal surgery.

2. Methods

2.1. Dataset

The dataset selected for this study is an open-sourced dataset from the Roboflow website under the CC BY 4.0 license. The dataset selects various skin conditions that visually resemble commonly known moles or nevi and are categorized into two types: benign and malignant. The dataset contains 1,197 images, with 855 images (71%) for the training set, 230 images (19%) for validation set, and 112 images (9%) for the testing set. The dataset has been preprocessed, including applying Auto-Orient and stretching to 800x800 pixels. Data augmentation was also applied, including horizontal flipping and random brightness adjustments within a range of -45% to +45% [8], which can enhance the robustness of the final model weight. The weighs obtained from training the dataset can be used for mole localization and for diagnosing whether the mole is benign or malignant. If the diagnosis indicates a benign mole, the patient can choose whether to have a removal surgery. If the diagnosis indicates a malignant mole, removal surgery will be performed under further pathological classification.

2.2. YOLO model

2.2.1. Brief introduction. YOLO model is a single-stage object detection model known for its high speed and high accuracy, initially introduced by Joseph Redmon and Ali Farhadi in 2015. As of October 4, 2024, YOLO has been upgraded to its eleventh generation, YOLO11. The YOLOv8 model, used in this study, is the eighth generation of YOLO, developed by Ultralytics, and features enhanced performance and efficiency. This research utilizes the detection module of the YOLOv8n model, part of the YOLOv8 series. Compared to other models in the same series, the smallest-scaled YOLOv8n model has 8.7B FLOPs, the lowest computational complexity, and the fastest prediction speed, which make it suitable for the real-time localization and classification of targets during surgery with a robotic arm actuation end and for the experimental environment of this study. The YOLOv8n model used in this study owns 225 layers, 3157200 parameters, 3157184 gradients and 8.9 GFLOPs.

2.2.2. Network structure. Unlike the two-stage object detection models which require predicting the coordinates of bounding boxes and classifying the selected regions, YOLO, as a single-stage model, can analyze the entire image in one pass. This leads to a faster speed of prediction while the model also maintains high accuracy. These characteristics make the YOLO model potential for real-time detection of the moles during surgery, which is why it was selected as the object detection model of this study.

In January 2023, Ultralytics open-sourced YOLOv8 based on YOLOv5, improving both speed and accuracy and making it more user-friendly. The network structure of YOLOv8 consists of three parts: Backbone, Neck, and Head, as shown in Figure 1. The Backbone, pre-trained on ImageNet, is responsible for extracting features from the input image. The Neck serves as an intermediate part to aggregate the features, passing them from the Backbone to the Head. The Head is responsible for regression and classification of the bounding boxes [9].

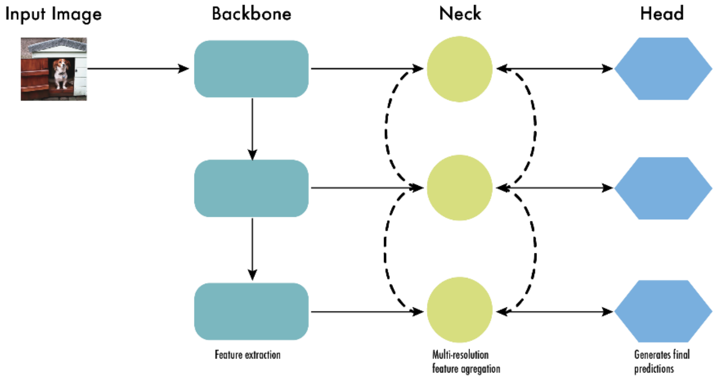
Input Image
Backbone
Neck
Head
Feature extraction
Multi-resolution feature agregation
Generates final predictions

Figure 1. Network structure of YOLOv8:Backbone, Neck and Head [10].

The input image is first split into n × n grid cells by YOLO. The Backbone processes the input image with multiple convolutional layers and residual learning modules to extract features at different scales. These features are then processed through Spatial Pyramid Pooling Fast (SPPF) which applies pooling for downsampling. The Neck facilitates the transfer of semantic and localization features from the Backbone to the Head. The Head predicts for each grid cell and returns b bounding boxes with their confidence levels. This process involves loss calculation and target detection box filtering. The detection results are evaluated with Intersection over Union (IoU), which is the ratio of the area of overlap between the predicted box and ground truth to the area of their union. Typically, an object is considered detected when the IoU exceeds 0.5 [11].

$$IoU = \frac{A \cap B}{A \cup B} = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{1}$$

2.2.3. Whole structure of YOLOv8.

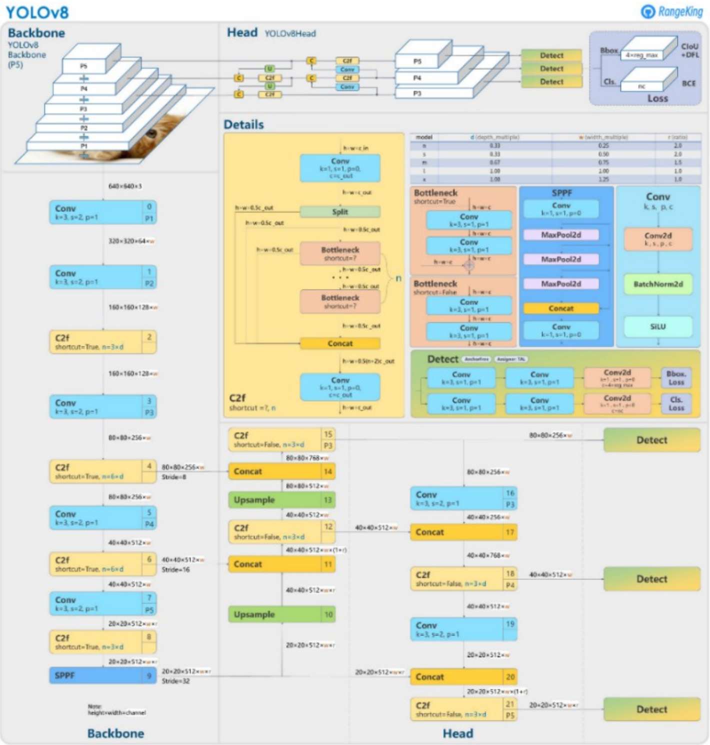

Figure 2. Whole Implemented Structure of YOLOv8 [12].

The overall implementation structure of YOLOv8 is shown in Figure 2.

2.3. Assessment methods

2.3.1. True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). TP, FN, FP, and TN reflect the relationship between predicted values and true values, classified into Positive and Negative classes. This results in all four categories: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). Table 1 presents the definitions of TP, FN, FP, and TN.

Table 1. Definition of TP, FN, FP, and TN.

TP, FN, FP, and TN		Predicted Values	
		Positive	Negative
True Values	Positive	TP[a]	FN[b]
	Negative	FP[c]	TN[d]

[a] predicting positives as positives (true prediction)
[b] predicting positives as negatives (false prediction)
[c] predicting negatives as positives (false prediction)
[d] predicting negatives as negatives (true prediction)

TP, FN, FP, and TN count the quantities for each category, but merely analyzing these values does not clearly reflect the accuracy of the prediction. Precision and Recall, calculated with TP, FN, FP, and TN, can effectively evaluate the prediction results.

2.3.2. Precision and Recall. Precision is defined as the proportion of true predictions among all the results predicted as positive. A high precision indicates that predictions are mostly accurate. Therefore, precision is an important standard for evaluating the correctness of the prediction of the weight. The formula of Precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is defined as the proportion of true predictions among all actual Positive results. Recall reflects the weight’s capability to identify all objects in the image. A high recall indicates that the weight is capable of detecting more actual objects. Therefore, recall is a critical standard for assessing the weight’s comprehensive coverage of objects. The formula of Recall is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

In general, Precision and Recall are negatively correlated. Since a good weight owns both high precision and high recall, finding the balance between them is the key to optimize the weight. In the following article, mAP and F1-score are introduced as indicators to evaluate the balance between precision and recall.

2.3.3. AP and mAP. To assess the balance between precision and recall, this study plots the Precision-Recall (PR) curve. The PR curve shows the relationship between precision and recall at different confidence levels. The area under the curve, AP (Average Precision), identifies the confidence interval which offers the best balance between precision and recall. The highest AP indicates the prime balance between precision and recall.

The mean Average Precision (mAP) is the average AP across all categories, reflecting the weight’s precision and recall for each class. Therefore, mAP is one of the key standards for evaluating the weight's overall performance. The formula of mAP is as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

2.3.4. F1-score. Similar to mAP, the F1-score evaluates the balance between precision and recall. It is calculated as the harmonic mean of precision and recall. A high F1-score indicates that the model can effectively detect all the objects while minimizing both missed detections and false positives. The F1-score is one of the key indicators for evaluating the weight's overall performance. The formula of the F1-score is as follows:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

3. Results

3.1. Hyperparameters setting

Before training, hyperparameters were set based on the dataset: batch size was set to 4 (batch=4); the initial learning rate was 0.01 (lr0=0.01); the number of training epochs was 1000 (epochs=1000); and the image size was set to 800 (imgsz=800) to match the image's size of 800x800. All other parameters were kept default. After these hyperparameters were set, the training began.

3.2. Results of training

The training finished at the 155th epoch, earlier than 1000 epochs initially set, because no significant improvement was observed in the final 50 epochs. The best weight was obtained at the 105th epoch. After training, the weight predicted the validation set, and the predicted values were compared with the ground truth, as shown in Table 2.

Table 2. Results of prediction of validation set[a].

Class[b]	Images[c]	Instances[d]	Box (P	R	mAP50[e]	mAP50-95[f])
All (Average)	230	248	0.839	0.796	0.862	0.544
Benign	230	92	0.819	0.836	0.88	0.572
Malignant	230	156	0.86	0.756	0.845	0.517

[a] 0.4ms preprocess, 4.2ms inference, 0.0ms loss, 1.1ms postprocess per image
[b] classes of the targets
[c] number of the images in the validation set
[d] number of the labelled images of each class
[e] mAP with IoU larger than 0.5
[f] mAP with different IoU (from 0.5 to 0.95, step = 0.05)

Table 2 indicates that the mAP50 of both classes reached a high level of 0.862, with Precision and Recall in balance. Based on these results, it can be primarily concluded that the weight is well capable of localizing the moles. A more detailed analysis is as follows, combining various charts.

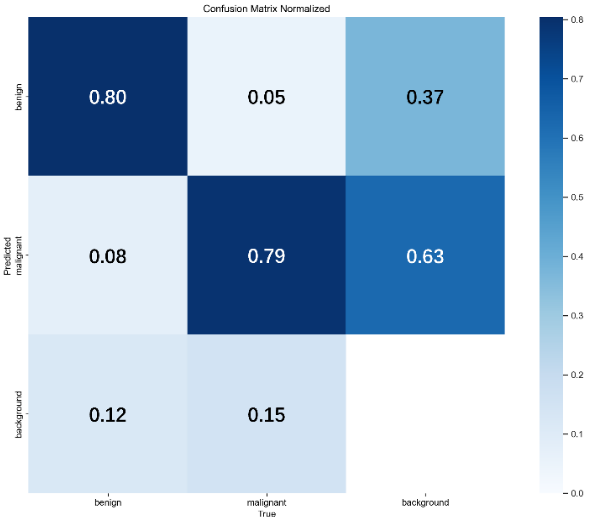

Figure 3. Normalized confusion matrix.

Figure 3 presents that the proportion of truly predicting benign as benign and malignant as malignant reached close to 0.8 in both categories. This demonstrates that the weight owns high TP and FN, effectively distinguishing between benign, malignant, and the background. The YOLO model reliably classifies targets and contributes to the preliminary diagnosis of whether a mole is benign or malignant.

After verifying the classification ability of YOLO, the next step is to evaluate YOLO's detection and localization capability based on the Precision-Recall curve and F1-score.

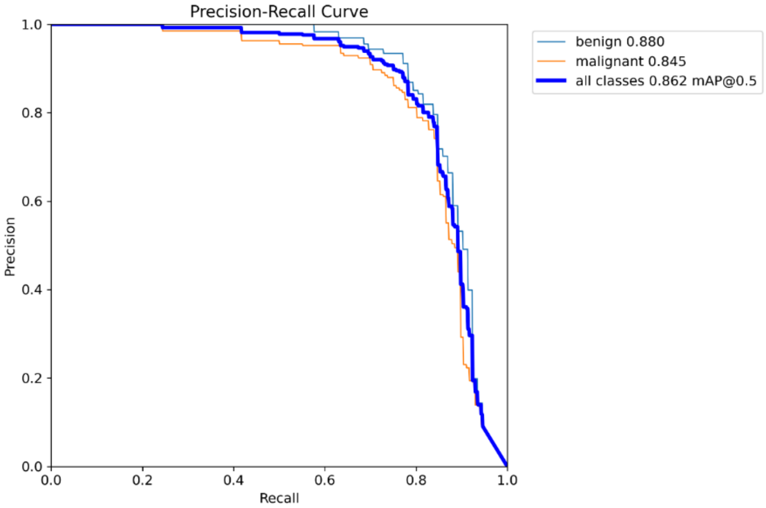

Figure 4. PR curve.

Figure 4 shows that the mAP of benign reached 0.880 and the mAP of malignant reached 0.845, with an average mAP of 0.862, demonstrating a high level close to 1. This result is merely a visualization of the data from Table 2 Next, the F1-score, another metric for evaluating the balance between precision and recall, will be used to further assess the weight's detection and localization performance.

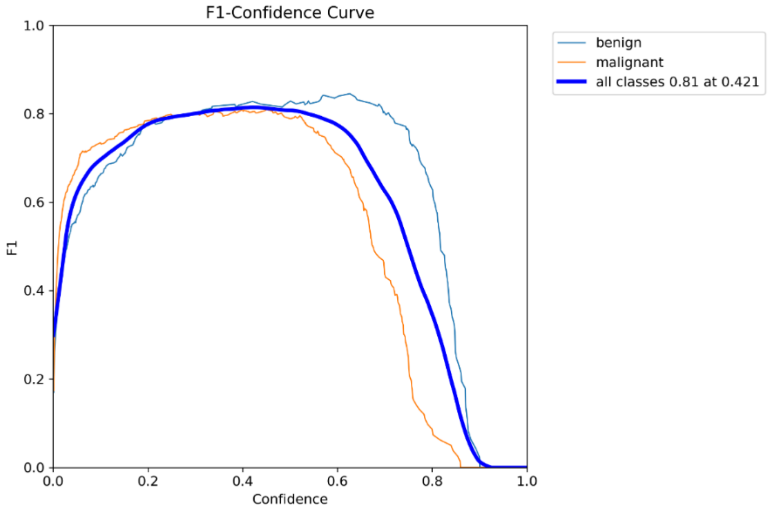

Figure 5. F1 curve.

Figure 5 shows that the F1-score reached the maximum value of 0.81 with a 0.421 confidence, indicating that the weight achieved the best balance, and the balance was at a high level. This result aligns with the findings from Figure 4, leading to the conclusion that the weight can achieve a high success rate in object detection and can identify all target categories as comprehensively as possible.

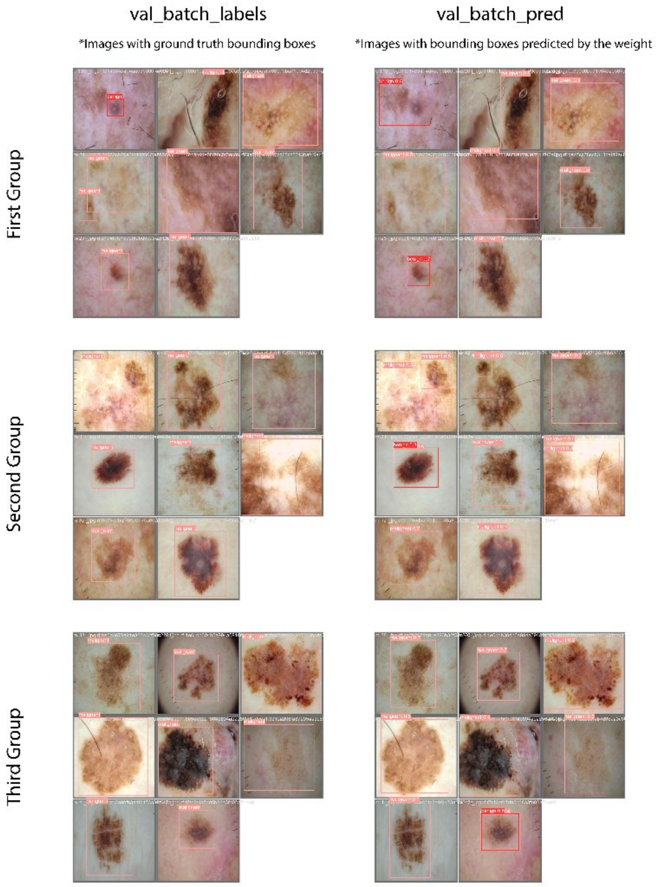
val_batch_labels
*Images with ground truth bounding boxes
val_batch_pred
*Images with bounding boxes predicted by the weight
First Group
Second Group
Third Group

Figure 6. Comparison between ground truth and predictions.

The previous quantitative analysis was accurate but lacked intuitiveness. To visually demonstrate the weight's localization and classification abilities, three sets of images were selected, with each set consisting of the ground truth labels from the validation set and the bounding boxes predicted by the weight, as shown in Figure 6. Through these images, the weight was able to classify the moles effectively, and the predicted boxes generally overlapped with the ground truth boxes. The comparison illustrates the weight's good localization and classification capabilities. However, it is still important to note that several errors were observed.

According to the comprehensive evaluation of the prediction results, the weight's achieved a balanced performance between precision and recall, demonstrating overall good performance in classifying and localizing targets. The YOLO model is competent at assisting robotic surgery systems and handling the fundamental tasks of diagnosis and localization in mole removal surgery.

4. Discussion

Although the weight trained based on the YOLO model has good performance, there are still some limitations and optimizing space.

This study selected an open-source dataset with a relatively small number of images (1,197). The limited size of the dataset may influence the robustness of the predictions. Additionally, the study classified the target moles into two categories: benign and malignant. While this classification can provide accurate feedback to patients and doctors, helping them decide whether mole removal surgery is necessary, it remains imprecise. This study did not categorize the specific types of mole symptoms, which limits its ability to directly guide the selection of appropriate treatments.

This study employed the official YOLOv8 model without making specific optimizations for the mole dataset and for the goal of mole localization and classification in robotic surgery. As a result, the current weight and the model have space for optimization. Since moles are small-sized, the images of them were enlarged in the dataset. However, YOLO's detection performance is known to be less effective when dealing with small objects or when there is a complex background. Future research will explore the impact of small-sized moles and complex backgrounds on the detection accuracy of YOLO-based weights and optimize YOLO to better handle small targets.

5. Conclusion

This study introduces the YOLO model for the localization and classification of moles. Trained on an open-source dataset, the model achieved good performance: a precision of 0.839, a recall of 0.796, mAP50 of 0.862, and mAP50-95 of 0.544. This study confirms that the YOLO model effectively detects and classifies moles, accurately distinguishing between benign and malignant moles, providing recommendations on whether surgery is necessary, and guiding the choice of specific surgical methods. Additionally, the YOLO model offers the potential for intraoperative navigation for robotic arm actuation end, enabling surgical robots to perform mole removal operations, reducing reliance on the surgeon's experience, and minimizing repetitive tasks during mole removal procedures. The YOLO model also holds promise for diagnosing the specific pathology of various mole symptoms. Additionally, when the segmentation module of YOLO is utilized, the YOLO model can delineate the boundaries for the moles, providing more precise navigation for the robotic arm actuation end.

References

[1] Saritas S Tekin HG Høgsberg T Hölmich LR and Juel J 2022 Ugeskrift for Laeger vol 184 V10210786
[2] https://www.wcrf.org/wp-content/uploads/2021/02/skin-cancer.pdf
[3] Wheeland RG McBurney E and Geronemus RG 2000 The role of dermatologists in the evolution of laser surgery Dermatol Surg vol 26 pp 815-22
[4] Horita K Hida K Itatani Y Fujita H Hidaka Y Yamamoto G Ito M and Obama K 2024 Real-time detection of active bleeding in laparoscopic colectomy using artificial intelligence Surg Endosc vol 38 pp 3461-69

[5] Urso L Manco L Castello A Evangelista L Guidi G, Castellani M Florimonte L Cittanti C Turra A and Panareo S 2022 PET-Derived Radiomics and Artificial Intelligence in Breast Cancer A Systematic Review Int J Mol Sci vol 23 p 13409

[6] Huang K Liao J He J Lai S Peng Y Deng Q Wang H Liu Y Peng L Bai Z et al 2024 A real-time augmented reality system integrated with artificial intelligence for skin tumor surgery experimental study and case series Int J Surg vol 110 pp 3294-306

[7] https://universe.roboflow.com/surawiwat-school-suranaree-university-of-technology/skin_cancer_detection-v2

[8] Hussain M 2023 YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection Machines vol 11 p 677

[9] Terven J Córdova-Esparza D-M and Romero-González J-A 2023 A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Mach Learn Knowl Extr vol 5 pp 1680-716

[10] Wang X Gao H Jia Z and Li Z 2023 BL-YOLOv8 An Improved Road Defect Detection Model Based on YOLOv8. Sensors vol 23 p 8361

[11] https://github.com/RangeKing

[12] Guo Z Wang C Yang G Huang Z and Li G 2011 MSFT-YOLO Improved YOLOv5 Based on Transformer for Detecting Defects of Steel Surface Sensors vol 22 p 3467