Enhanced Chest CT Detection of Pulmonary Nodules Based on YOLOv8

Yifan Chai, Xiaoxia Zhang

Abstract—**The early identification of pulmonary nodules is essential for enhancing lung cancer survival rates, with computed tomography (CT) serving as the primary diagnostic tool. However, the increasing volume of CT data poses significant challenges, particularly in detecting small and irregularly shaped nodules. To tackle this problem, we developed the Small Object Detection-YOLOv8, an extension of YOLOv8n, designed to enhance the detection of small nodules. To mitigate overfitting in limited sample scenarios, a multi-level prediction header was introduced alongside the Multi-scale Contextual Attention (MCA) mechanism to reduce noise and improve feature extraction. Moreover, the Complete Intersection over Union (CIoU) loss function was substituted with the Modified Partial Distance Intersection over Union (MPDIoU) to achieve further performance improvements. Comprehensive evaluations on the LIDC-IDRI datasets and LUNA16 demonstrated that the proposed model achieved mAP@0.5 scores of 0.769 and 0.775, representing improvements of 0.055 and 0.051 over the YOLOv8n model, respectively. These results validate the effectiveness of the proposed method in improving the accuracy of pulmonary nodule detection.**

Index Terms—**YOLOv8n ; Computer vision; Pulmonary nodule detection**

I. Introduction

The challenge of annotating lung nodule datasets complicates the determination of the presence of pulmonary nodules on chest X-rays. Typically, only medical experts in this field possess the requisite expertise to accurately label pulmonary nodules. This limitation hinders the creation of high-quality datasets for pulmonary nodule detection, thereby affecting the accuracy of detection algorithms. Additionally, continuous interpretation of X-rays can lead to fatigue among radiologists, increasing the likelihood of diagnostic errors. Therefore, in the context of pulmonary nodule detection, the implementation of a medical assistant system capable of identifying pneumonia-lesions in patients' X-ray images could be highly beneficial in clinical practice.Additionally, it could alleviate doctors' workload. Research on the automatic detection of pulmonary nodules dates back to the 1950s, when Turing, a young computer prodigy, proposed the theory of automation, laying the groundwork for artificial intelligence. Since then, extensive exploration of computer artificial intelligence has taken place from various perspectives. As computing power advanced, detection algorithms gained traction for lung nodule detection, yielding promising results. Target detection, a critical element of computer vision, has significant applications in medical diagnostics, industrial and agricultural product inspections, autonomous driving systems, and other practical domains. Object detection tasks involve identifying the locations of objects within an image and subsequently classifying these objects. In the domain of medical detection, two well-established target detection approaches are frequently utilized: single-stage target detection methods and dual-stage target detection methods. One-stage target detection algorithms include the YOLO series [1] and SSD [2], whereas two-stage target detection algorithms include Faster R-CNN [3] and RCNN [4]. In two-stage target detection algorithms, a single network is typically utilized for detecting and extracting the object region, followed by another network for classifying and recognizing the object region. Although the two-stage target detection algorithm offers high accuracy, it often suffers from slow processing speeds, which may not meet the demands of rapid target detection in medical applications.

Released in 2023, YOLOv8 is widely regarded as the most efficient and fastest algorithm developed to date. It can simultaneously classify and locate targets using a single neural network, thereby significantly enhancing computational efficiency and improving detection speed. The YOLOv8 series comprises five models, with the YOLOv8 Nano (YOLOv8n) being the smallest and fastest in terms of detection speed. In the current medical detection landscape, where efficiency is paramount, this paper selects YOLOv8n as the improved baseline algorithm. Despite its proficiency in detecting full-size targets, YOLO demonstrates suboptimal accuracy in detecting small objects. The original hybrid convolution network based on YOLOv8n achieves a 72.4% accuracy rate for lung nodule image detection. To address these limitations in medical image detection, it is essential to further optimize the performance of the YOLOv8n algorithm.

This study integrates the MCA attention mechanism into the original algorithm, allowing the network to prioritize diverse channel information through multi-head attention, thereby improving its ability to capture key features and enhancing overall target detection performance. To further optimize small object detection, we refined the multi-detection head, specifically boosting the prediction capabilities for small, multi-level features, such as pulmonary nodules. Evaluation on the LIDC-IDRI and LUNA16 datasets demonstrated that these optimizations significantly enhanced the network's representation capabilities.

Manuscript received April 13, 2024; revised October 24, 2024.

Y. F. Chai is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 2397789262@qq.com).

X. X. Zhang is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (corresponding author, phone: 86-0412-5929812; e-mail: aszhangxx@163.com).

II. Materials And Method

A. Overall Structure

The YOLO series was initially proposed by Joseph Redmon in 2015 as a material measurement method based on convolutional neural networks [5]. Each component of YOLOv1 requires separate training, making the training process intricate. In response, Ultralytics, a small start-up, has developed and maintained the YOLOv8 algorithm to support image classification, detection, and instance segmentation. As depicted in Figure 1, the YOLOv8 network architecture comprises four main components: the Input module, the Backbone (neural network structure), the Neck (feature aggregation layer), and the Head (prediction network layer).

B. Component Structure

The input module mainly includes Mosaic image augmentation, adaptive anchor box computation, and adaptive image scaling. Mosaic Image Enhancement, introduced by YOLOv4 [6], involves the random splicing of four images into one during training to enrich the dataset for pulmonary nodule detection. Adaptive Anchor Frame Calculation automatically computes the most suitable anchor frame parameters for the input image through learning prior to network training, enhancing target detection accuracy and robustness without requiring manual configuration. However, the image aliasing enhancement technology proposed by YOLOv4 may impact training accuracy when enabled throughout the entire training process. YOLOv8 addresses this by deactivating the image aliasing enhancement technology during the later stages of training, thereby improving overall training efficacy.

The backbone network is a crucial component of the overall network, designed to extract image features. Its comprehensive structure comprises ConvBiSiLU (CBS), shortcut (C2F), Spatial Pyramid Pooling Fast (SPPF), and other essential modules. ConvBiSiLU (CBS) performs convolution operations on input images and aids C2F (shortcut) in feature extraction, while Spatial Pyramid Pooling Fast (SPPF) achieves adaptive size output. This section is mainly tasked with extracting feature representations from the target.

Among these modules, CBS and SPPF are adopted from YOLOv5, whereas the C2F (shortcut) module is influenced by the ELAN design concept introduced in YOLOv7 [7], effectively replacing the original C3 module with the C2F (shortcut) module. The C3 module within the YOLOv5 architecture improves the network's depth, expands the receptive field, and strengthens its feature extraction capabilities. The structural diagram of the C2F (shortcut) module in YOLOv8 is presented in Figure 2, while the structural diagram of the C3 module in YOLOv5 is depicted in Figure 3. The bottleneck represents a specialized residual structure.

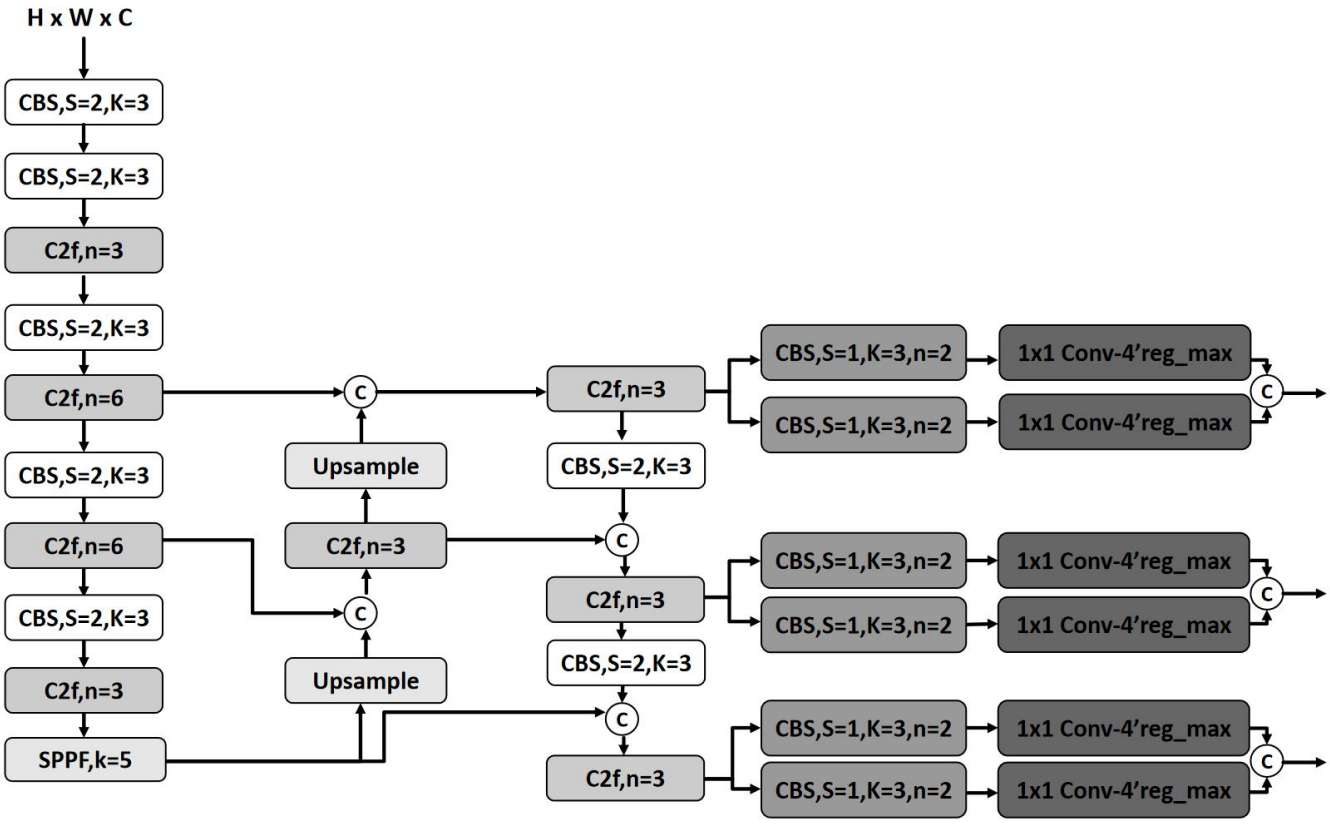


Fig. 1. Network structure of YOLOv8

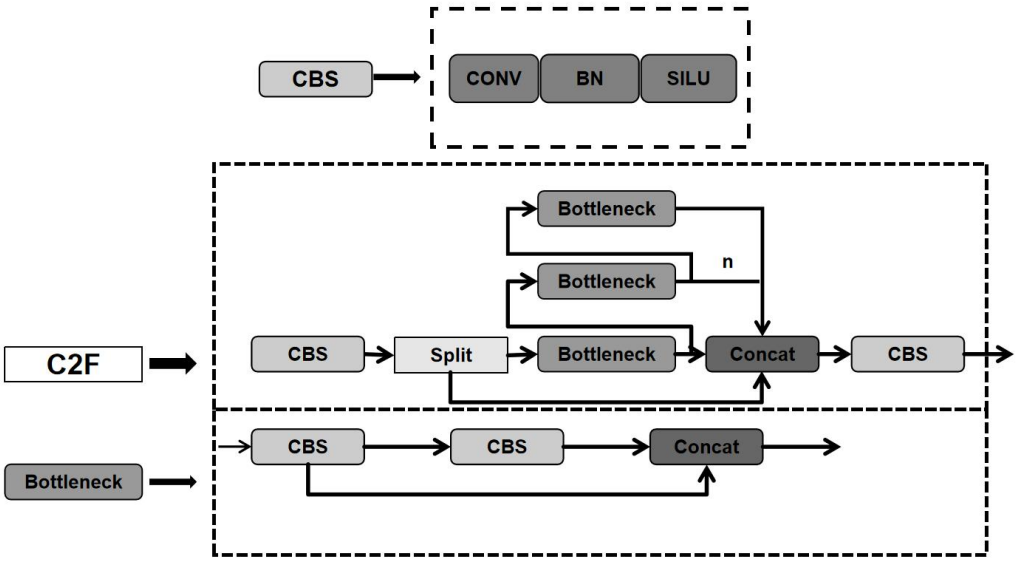


Fig. 2. C2F module structure of YOLOv8

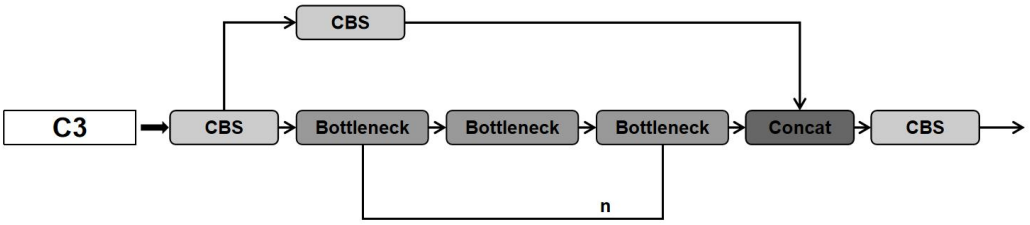


Fig. 3. C3 module structure of YOLOv5

The Neck structure, as depicted in Figure 4, is a sophisticated network layer that consists of a convolutional layer and a C2F module. This architecture strategically integrates the Path Aggregation Network (PAN) [8] and Feature Pyramid Network (FPN) [9] frameworks to enhance multi-scale feature fusion. The primary goal of this design is to effectively transfer image features to the prediction layer, ensuring robust performance across varying image scales.

In the left section of Figure 4, the PAN structure is illustrated. The PAN architecture employs down-sampling to link low-resolution feature maps with high-resolution ones, facilitating effective feature integration. This creates an interconnected pathway that allows for the fusion of information between adjacent layers of the feature map. Consequently, feature maps across various scales are enriched with both semantic and visual details, enhancing prediction accuracy regardless of the input image size.

The right section of Figure 4 illustrates the FPN structure. The FPN is constructed by down-sampling high-resolution feature maps and up-sampling low-resolution feature maps to create a pyramid structure. This architecture facilitates the integration of information between layers, ensuring that essential target information from high-level feature maps is preserved, while low-level background details are enhanced by the high-level features. This dual integration approach ensures comprehensive feature representation, contributing to more accurate detection and classification in diverse visual tasks.

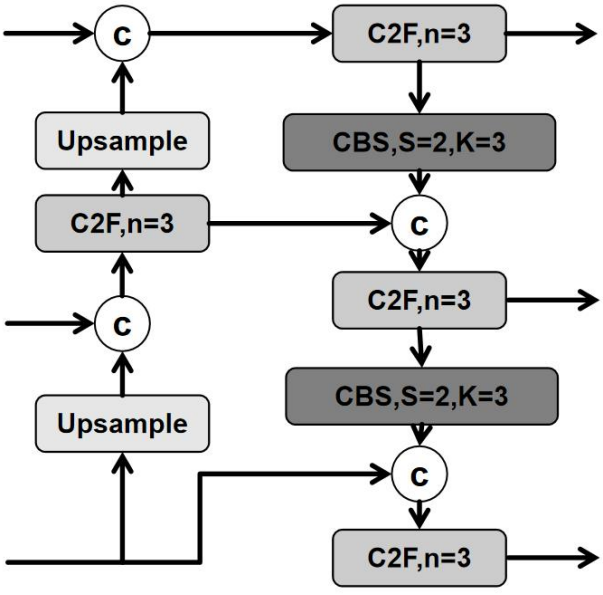


Fig.4. Neck structure: PAN-FPN

Figure 5 illustrates the structure of the head prediction layer. Due to the differing focuses on classification and positioning—where classification prioritizes texture content and positioning emphasizes edge information—YOLOv8 employs decoupled detection headers. This segregation into distinct branches for classification and detection tasks enhances overall detection effectiveness. Additionally, the channel configuration of the regression header is adjusted to further optimize performance.

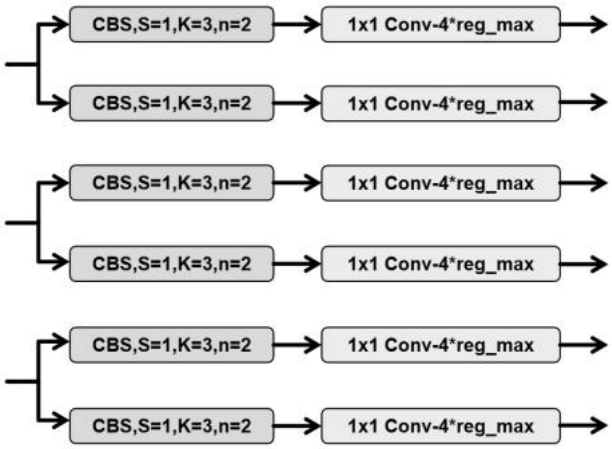


Fig.5. Head structure: Decoupled head

III. Improved Strategy

A. MCA Attention Mechanism

The attention mechanism in deep learning emulates the visual focus observed in biological systems, particularly the human eye, which selectively emphasizes specific areas with "high resolution." This technique empowers models to prioritize and discern critical feature information within an image, thereby enhancing their interpretative capabilities. The attention mechanism operates by varying the degree of attention assigned to different elements of the input, accomplished through the application of weights. Conceptually, it integrates a query matrix, key, and a weighted average to form what is known as a Multilayer Perceptron (MLP) attention mechanism, as depicted in Figure 6.

In a broader context, channel attention plays a pivotal role in substantially improving network performance. This enhancement arises from the ability of channel attention to direct the network ' s focus towards essential semantic information within the image, while simultaneously filtering out extraneous details. By doing so, channel attention effectively reduces the negative impact of noise on feature extraction from the input image, thus resulting in a notable improvement in the model's detection accuracy. This selective attention mechanism is especially valuable in scenarios where precision and accuracy are paramount, as it facilitates more refined and reliable outcomes in deep learning tasks.

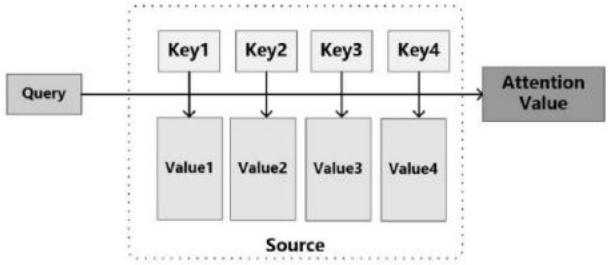


Fig.6. Principle of attention mechanism

Similarly, Multi-scale Cross-axis Attention (MCA) [10] aims to address the challenges of multi-scale information and long-range dependencies in medical image segmentation. This method effectively captures global information by employing efficient axial attention to calculate the two-way cross-attention between parallel axial attention. To accommodate significant variations in individual lesion size and organ shape, multiple strip convolutions with varying kernel sizes are utilized in each axial attention path to enhance the efficiency of spatial information encoding. This method is incorporated into the MSCAN backbone network, forming what is referred to as MCANET. The architecture of the MCA attention mechanism is depicted in Figure 7. Notably, MCANET, with only 4M+ parameters, outperforms many previous attention mechanisms in four challenging tasks: skin lesion segmentation, nuclear segmentation, abdominal multi-organ segmentation, and polyp segmentation.

The MCA Attention Mechanism enhances the axial attention mechanism by integrating strip convolution to introduce multi-scale features, thereby improving precision in localizing target areas. Concurrently, a double cross-attention mechanism is established between two spatial axial attentions to effectively leverage multi-scale features and identify ambiguous boundaries. MCANET adeptly encodes global context and accommodates diverse sizes and shapes of lesion regions or organs, thereby enhancing the accuracy of medical image detection. Furthermore, MCANET enhances the model's global context perception by integrating multiple scales of attention, thus contributing to more precise medical image detection.

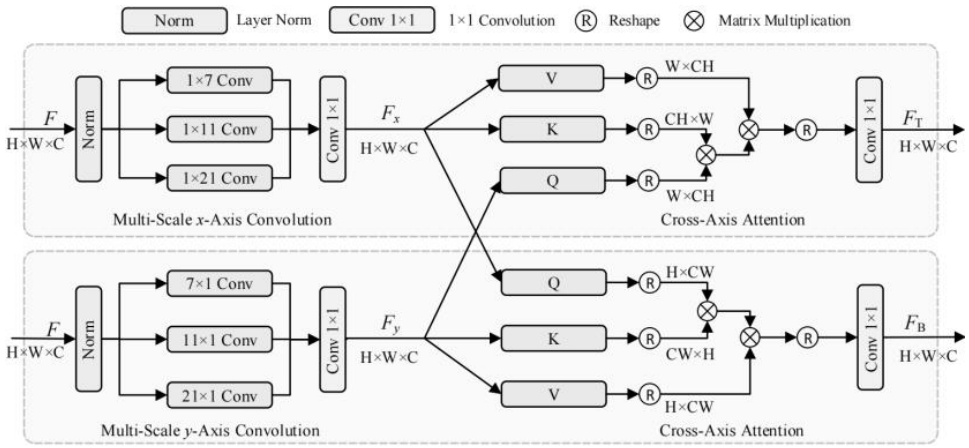


Fig.7. MCA network structure

B. Multiple Detection Head Optimization

The target detection network model utilizes established backbone networks such as VGG [11], ResNet [12], DenseNet [13], MobileNets [14], EfficientNet [15], CSPDarknet 53, and Swin Transformer [16], known for their robust feature extraction capabilities in classification tasks. Consequently, this study adopts the original YOLOv8 network architecture as the backbone. The backbone network primarily performs feature extraction, while the head uses these feature maps to detect target position and category. Detection heads are generally classified as single-stage or two-stage detectors. The RCNN series is widely acknowledged as the leading representative of two-stage detectors. In contrast, single-stage detectors concurrently predict both the bounding box and the object class, offering faster processing speeds but generally lower accuracy. Notable single-stage detectors in deep learning include the YOLO series, SSD, and RetinaNet [17]. Since the introduction of YOLOv3 [18], the YOLO detector has been enhanced with three prediction headers, integrating various detection scales and feature levels to optimize performance across diverse scenarios.

Detecting pulmonary nodules, especially small targets, remains a significant challenge in medical image detection. Enhancing the accuracy of TB detection for small targets is a primary focus in medical detection. Feature extraction in CNNs can lead to loss of crucial information, especially for small targets, complicating this issue. To address this, a prediction header is implemented to detect multi-level features of small-sized pulmonary tuberculosis targets. When combined with the other three prediction heads, this approach effectively mitigates overfitting in scenarios with small sample sizes. The enhanced small object detection-YOLOv8 network structure, illustrated in Figure 8, demonstrates significant improvements in detecting small objects without introducing additional model parameters or computational overhead.

C. Loss Function Improvement

Selecting an appropriate loss function in deep learning is vital for determining the overall effectiveness of the trained model. In YOLOv8, the employed loss function differs from that used in the YOLOv5 and YOLOv7 series, as it is divided into two components. Specifically, VFL Loss [19] is utilized for classification loss coordination, while the default loss function employed is *CIoU* Loss [20]. The calculation of the *CIoU* loss is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho(b, b^{gt})}{C^2} + av \tag{1}$$

$$IoU = \frac{A \cap B}{A \cup B} \tag{2}$$

$$v = \frac{4}{\pi^2}\left[\arctan\left(\frac{w^{gt}}{h^{gt}}\right) - \arctan\left(\frac{w}{h}\right)\right]^2 \tag{3}$$

$$a = \frac{v}{(1 - IoU) + v} \tag{4}$$

$$\begin{cases} \dfrac{\partial v}{\partial w} = \dfrac{8}{\pi^2}\left[\arctan\left(\dfrac{w^{gt}}{h^{gt}}\right) - \arctan\left(\dfrac{w}{h}\right)\right] \times \dfrac{h}{w^2 + h^2} \\ \dfrac{\partial v}{\partial h} = \dfrac{8}{\pi^2}\left[\arctan\left(\dfrac{w^{gt}}{h^{gt}}\right) - \arctan\left(\dfrac{w}{h}\right)\right] \times \dfrac{w}{w^2 + h^2} \end{cases} \tag{5}$$

Where A is the prediction box and B indicates the true box. IoU represents the intersection ratio between A and B, specifically defined as the proportion of the overlapping area of A and B to the total area of their union. A higher value of IoU indicates that the predicted box is closer to the ground truth box. However, in cases where there is no overlap between the predicted box and the ground truth box, or when they are perfectly aligned, IoU cannot be evaluated. b represents the center of the predicted box, b^{gt} represents the center of the ground truth box, and ρ refers to the Euclidean distance calculation. C denotes the diagonal length of the smallest enclosing area that can contain both the predicted and ground truth boxes. Parameter a is used to adjust the balance ratio, Parameter v is utilized to represent the similarity in aspect ratio between the predicted box and the ground truth box. When the centers coincide, v serves as an indicator to evaluate the closeness between the predicted and actual boxes.

YOLOv8's *CIoU* loss integrates overlap, center point distance, and aspect ratio, leading to more stable bounding box regression. but it is not perfect. Parameter v assesses aspect ratio similarity relative to the ground truth box's width and height but does not represent their actual values, which may hinder model optimization. When one of the values of w and h increases, the other must decrease, and they cannot maintain the same increase and decrease, this can result in slower convergence of the loss function and imprecise localization of the regression box. To enhance the algorithm's performance and detection accuracy, this paper utilizes *MPDIoU* loss [21] as a replacement for *CIoU* loss. The formula for calculating *MPDIoU* loss is given as follows.

$$d_i = (x_i^{gt} - x_i^{prd})^2 + (y_i^{gt} - y_i^{prd})^2 \tag{6}$$

$$MPDIoU = IoU - \frac{\sum_{i=1}^{2} d_i^2}{h^2 + w^2} \tag{7}$$

In the formula, d_i denotes the separation between the top-left and bottom-right corners of the predicted box and the actual box, respectively. As illustrated in Figure 9, the bottom-right box represents the ground truth, whereas the top-left box corresponds to the predicted box. The *MPDIoU* loss functions to minimize loss by reducing the distance between two predicted corner points. As a method based on point distance measurement, the *MPDIoU* loss effectively resolves the issue of existing loss functions in optimizing predicted boundary boxes and real boundary boxes with similar aspect ratios but vastly different length-width values. This method streamlines the calculation process and exhibits enhanced adaptability for target detection across diverse scales. Moreover, the *MPDIoU* loss is agnostic to boundary frame size, enabling effective management of scenarios involving substantial variations in target scale while bolstering localization capabilities for small targets.

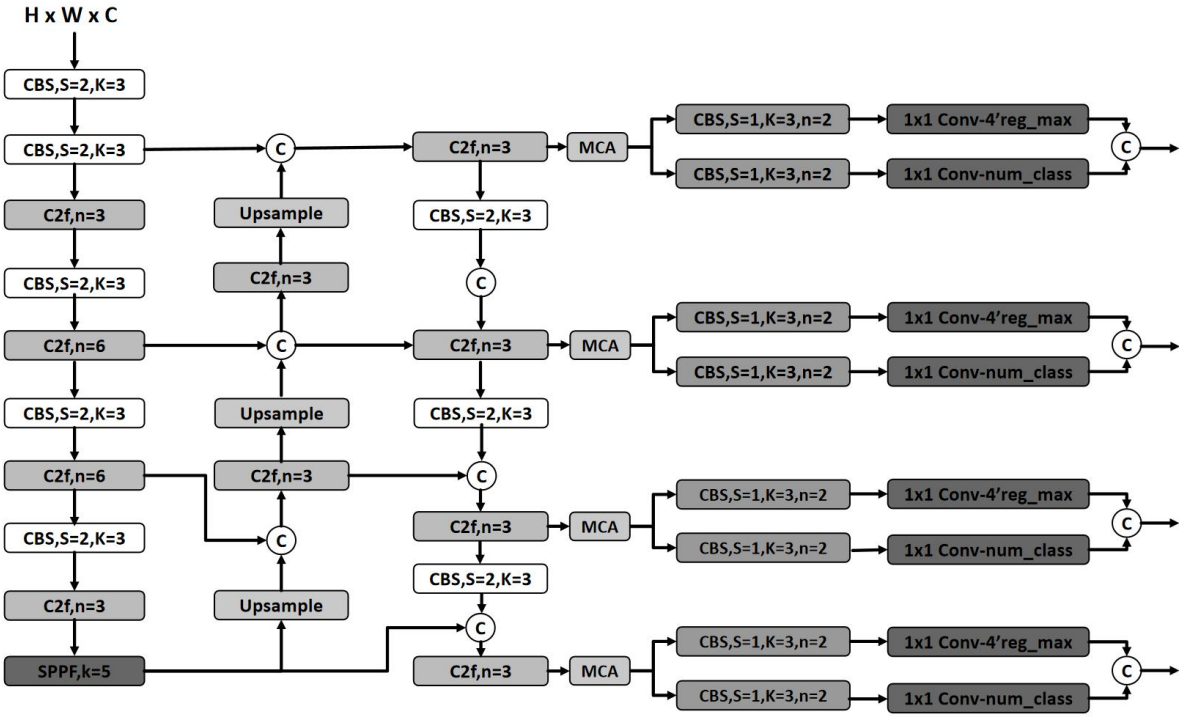


Fig. 8. Network structure of small object detection-YOLOv8

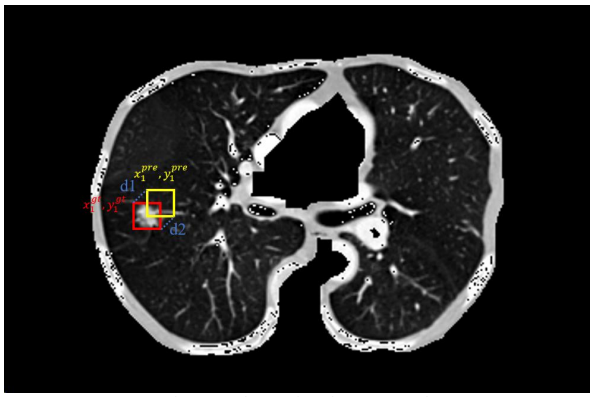


Fig. 9. Schematic of MPDIoU loss

IV. Analysis of Experiments And Results

The experimental platform consists of a hardware configuration featuring an Intel Core i7-11700 CPU alongside an NVIDIA GeForce GTX 3070 GPU. The software framework utilizes PyTorch 1.8-GPU as the primary deep learning environment, with PyCharm Community IDE employed for model design and training execution. The experiment parameters include 100 epochs, a batch size of 16, and a data division strategy where the dataset is split into training, validation, and testing sets in an 8:1:1 ratio.

A. Data Set Selection

The LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) dataset offers a comprehensive collection of 1,018 thoracic CT scans designed for developing and evaluating lung nodule detection algorithms. It includes scans with varying slice thicknesses and a wide range of nodule types. Each scan is independently annotated by four radiologists, with nodules larger than 3 mm classified as "consensus nodules" if identified by at least three radiologists. The dataset covers nodules of varying sizes and malignancy levels, providing a diverse and clinically representative challenge, making it ideal for testing model generalization across different clinical scenarios.

In contrast, the LUNA16 dataset[22], which is a curated subset of LIDC-IDRI and part of the LUng Nodule Analysis 2016 Challenge, consists of 888 thin-slice CT scans, excluding 130 scans with slice thicknesses exceeding 2.5 mm. The dataset identifies 1,186 nodules with an average diameter of 8.3 mm, each annotated with precise coordinates (XYZ) and diameter. Divided into 10 subsets for ten-fold cross-validation, LUNA16 focuses on nodules larger than 3 mm, identified by at least three of the four radiologists from LIDC-IDRI, while excluding non-nodular micronodules (<3 mm) and irrelevant findings. This makes LUNA16 an essential benchmark for evaluating nodule detection models in a more controlled yet relevant context. Figure 10 provides a partial view of the LUNA16 dataset.

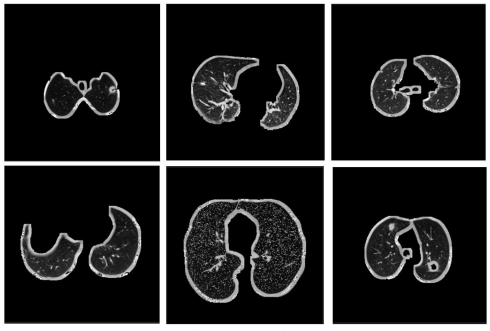


Fig. 10. A partial display of the LUNA16 dataset

B. *Experimental Evaluation Criteria*

This study employs Mean Average Precision (mAP) [23] and Rank-n as key metrics for evaluating model performance. mAP acts as a foundational measure for assessing retrieval capabilities, incorporating both Precision and Recall to provide a well-rounded evaluation of retrieval accuracy. Specifically, it represents the average precision (AP) across several queries, reflecting the model's ability to consistently deliver accurate results. In contrast, Rank-n quantifies the likelihood of finding the correct match within the top n retrieved results, thereby offering insight into the model's proficiency in effectively identifying relevant matches. A higher Rank-n value is indicative of superior retrieval performance. The formula for calculating mAP is presented in Equation 8.

$$mAP = \frac{\sum_{i=1}^{m} AP_i}{m} \quad (8)$$

Where m represents the total number of classes, and AP_i indicates the average precision for the i-th class.

C. Model Evaluation and Comparison Experiment

The principal performance metric used in this comparative analysis is mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5. This metric is computed by determining the Average Precision (AP) for each class, evaluating all images corresponding to that class, and then averaging these AP values across all classes. This approach provides a comprehensive evaluation of the model's detection accuracy, thereby serving as a critical performance indicator for the algorithm.

Beyond mAP, additional metrics were employed to ensure a well-rounded evaluation of the network models. These include the total number of trainable parameters, which is indicative of the model's capacity and inherent complexity, and the number of floating-point operations (FLOPs), which represents the model's computational cost and efficiency. Additionally, frames per second (FPS) was used to assess the model's real-time processing capability, which is crucial for many practical applications. Together, these metrics offer a thorough, multidimensional assessment of the detection algorithms' performance, efficiency, and computational feasibility. For a detailed comparison, Table 1 compiles the key performance indicators, enabling an analysis of the trade-offs between detection accuracy, computational overhead, and real-time processing abilities.

The Small Object Detection-YOLOv8 algorithm demonstrated significant improvements in terms of mAP and FPS compared to the original YOLOv8n, when tested on the LIDC-IDRI and LUNA16 datasets. Its mAP@0.5 scores reached 0.769 and 0.775, representing improvements of 0.055 and 0.051, respectively (Figure 11, LUNA16 dataset). To better understand the performance under varying levels of overlap, we also evaluated mAP@0.5:0.95, which computes precision by averaging across all IoU thresholds for all classes. This provides a more rigorous and nuanced measure of detection performance. Figure 12 highlights the enhanced performance of Small Object Detection-YOLOv8 compared to the baseline.

Despite a slight increase in computational complexity—FLOPs rising to 9.2 and 9.4, respectively—the Small Object Detection-YOLOv8 still maintains a lower parameter count than models like YOLOv5s, Faster R-CNN, and YOLOv7, all of which deliver inferior accuracy scores. Figure 13 illustrates the superior mAP@0.5 performance achieved by Small Object Detection-YOLOv8 on the LUNA16 dataset, highlighting its effectiveness for small object detection tasks.

Figure 14 further compares the training loss curves of YOLOv8n and Small Object Detection-YOLOv8, showing that the latter exhibits faster convergence along with a more stable and smoother training process. The increased stability demonstrates consistent optimization, minimizing fluctuations that often hinder convergence, thereby enhancing the training efficiency.

In summary, the improvements introduced in Small Object Detection-YOLOv8 make it more robust, effective, and adaptable compared to the original YOLOv8n model. It achieves a higher accuracy in detecting small targets, balances computational cost effectively, and maintains superior real-time processing capabilities, making it a strong candidate for practical deployment in small object detection scenarios.

TABLE I

COMPARISON WITH ADVANCED ALGORITHMS

Model	LIDC-IDRI				LUNA16			
	MAP@0.5	Params	Gflops	FPS	MAP@0.5	Params	Gflops	FPS
YOLOv7	0.357	36.8	104.7	61	0.278	37.1	105.1	50
Faster rcnn	0.339	142.1	283.4	37	0.285	137.099	370.21	45
Eff	0.546	3.924	5.2	63	0.261	3.874	5.234	59
YOLOv5s	0.523	7.215	16.5	156	0.685	7.013	15.8	62
SSD[24]	0.485	34.126	45.1	21	-	-	-	-
3D-CNN[25]	0.561	61.2	121.3	17	-	-	-	-
YOLOv8n	0.713	3.157	8.8	71	0.724	3.157	8.9	70
small object detection-YOLOv8	0.769	3.528	9.2	68	0.775	3.254	9.4	73

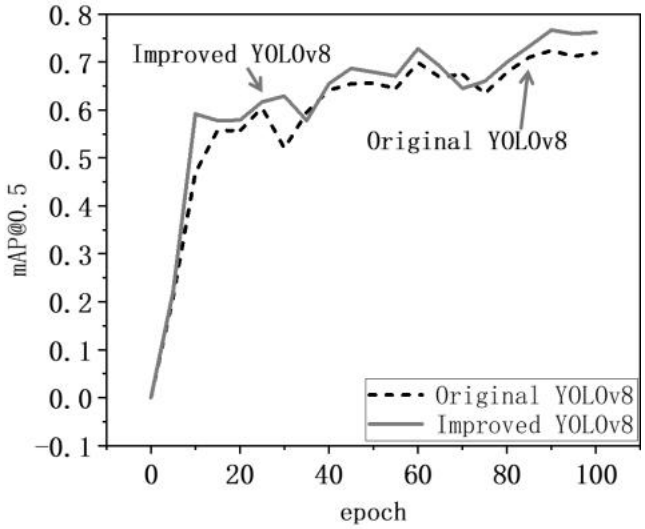


Fig. 11. Comparison chart of mAP@0.5 (LUNA 16)

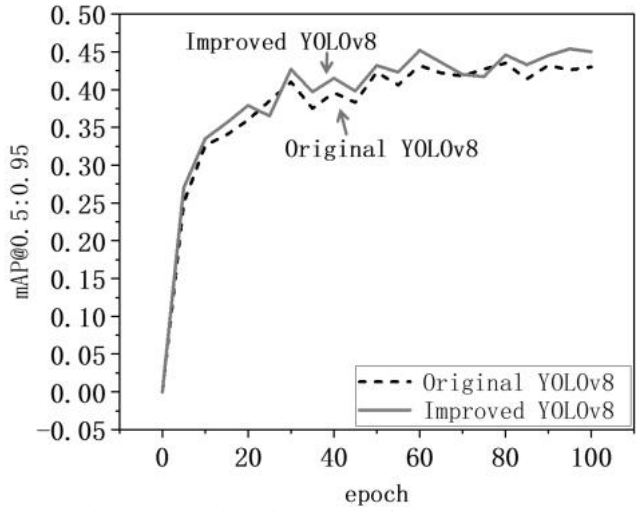


Fig. 12. Comparison chart of mAP@0.5:0.95 (LUNA 16)

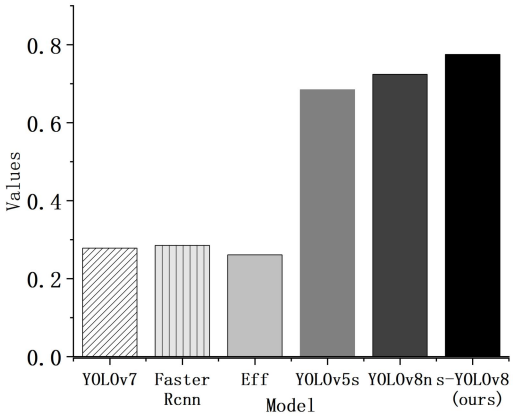


Fig. 13. Map@0.5 comparison chart (LUNA 16)

D. Ablation Experiment

To evaluate the impact of individual components in the enhanced algorithm, a comprehensive set of ablation experiments was performed to compare several configurations. The first configuration involved modifying the original YOLOv8n model by incorporating a Small Object Detection (SOD) head, which aims to improve the model's ability to identify small-scale objects that are often challenging for standard detection networks. In the second scenario, the original YOLOv8 model was augmented by integrating a Multi-scale Context Aggregation (MCA) attention mechanism. This attention module is designed to capture contextual information across different scales, thus enhancing the model's robustness in detecting objects of varying sizes and handling complex environments. Lastly, the third experiment focused on improving the loss function of the original YOLOv8 model by incorporating the MPDIoU (Mean Performance Distance Intersection over Union) metric. This improvement aims to optimize the bounding box regression by providing a more effective assessment of the overlap between predicted and actual bounding boxes, thereby leading to more precise localization of objects.

To ensure consistency in evaluating these modifications, all experiments were conducted under identical experimental conditions using the LUNA16 dataset. These conditions included controlling image resolutions, Non-Maximum Suppression (NMS) thresholds, confidence thresholds, and model weights (in megabytes, MB). By holding these parameters constant, it was possible to isolate the effects of the individual algorithmic enhancements and provide a fair comparison between different configurations. The ablation studies were particularly focused on comparing the detection accuracy across these models, with an emphasis on the specific contributions made by each enhancement.

The experimental results obtained under these controlled settings are thoroughly summarized in Table II. These results include insights into how the integration of the SOD head improved the model's ability to detect small objects, how the MCA attention mechanism enhanced feature extraction by aggregating contextual information, and how the MPDIoU loss function contributed to a more stable and accurate regression process. In addition to detection accuracy, the experiments also examined the influence of image resolution and the impact of various NMS and confidence thresholds on overall model performance.

The goal of these ablation studies was to rigorously determine the value added by each component in improving the detection capabilities of the model, especially in the context of small object detection, which is critical for medical imaging tasks such as identifying pulmonary nodules. By conducting these systematic experiments, the study was able to quantify the effects of each individual enhancement, thus providing deeper insight into how specific modifications affect the model's ability to generalize across various scenarios and image conditions. The detailed analysis, as presented in Table II, offers a comprehensive overview of the improvements achieved, ultimately establishing the efficacy of the proposed components in enhancing the overall performance of YOLOv8.

TABLE II

ABLATION EXPERIMENT

Model	Map	Imgsz	Iou	Conf	Wsz
YOLOv8n+SOD	76.6	640	0.7	0.01	5.99
YOLOv8n+MCA	72.6	640	0.7	0.01	5.96
YOLOv8n+MPDIoU	72.1	640	0.7	0.01	5.95

E. Visualization of Search Results

The dataset was carefully divided into five distinct subsets, where one subset was designated for testing purposes, while the remaining four were utilized for training. This partitioning scheme enabled a comprehensive evaluation of both the original YOLOv8n model and the YOLOv8 model enhanced with a small object detection mechanism. To ensure the robustness of the results, the experiment was conducted over five different iterations, as described in Table III, with each iteration representing an independent test cycle. This methodology allowed for a thorough assessment of the models' performance under varying conditions.

The average mean Average Precision at an Intersection over Union threshold of 0.5 (mAP@0.5) was recorded as 0.719 for the YOLOv8n model, while the small object detection-enhanced YOLOv8 achieved an average mAP@0.5 of 0.766 across the five iterations. These values highlight the superior detection capabilities, reliability, and stability of the enhanced YOLOv8 model in comparison to the original version. The enhanced model's consistent performance in achieving higher detection accuracy further supports its suitability for small object detection tasks, particularly in challenging environments.

Figure 15 provides a visual comparison of the detection results from both models, clearly illustrating the increased detection accuracy and improved localization of small objects by the enhanced YOLOv8. The results demonstrate that the small object detection-enhanced YOLOv8 more effectively identifies small-scale targets, reducing false negatives and improving overall detection performance.

The consistent success of the small object detection-enhanced YOLOv8 model across multiple experimental iterations reinforces its robustness and generalizability. This consistency indicates the model's ability to maintain stable performance across varying test conditions, establishing it as a reliable choice for applications involving small object detection, such as medical imaging or surveillance, where accuracy and dependability are critical. Overall, these findings underscore the advantages of incorporating small object detection mechanisms into YOLOv8, positioning it as a more capable and dependable model for specialized detection tasks.

TABLE III
CROSS VALIDATION YOLOV8N AND SMALL OBJECT DETECTION-YOLOV8

Model	YOLOv8n	small object detection-YOLOv8
Test1 *mAP*@0.5	0.718	0.763
Test2 *mAP*@0.5	0.723	0.767
Test3 *mAP*@0.5	0.716	0.771
Test4 *mAP*@0.5	0.725	0.761
Test5 *mAP*@0.5	0.712	0.766
Average Value	0.719	0.766

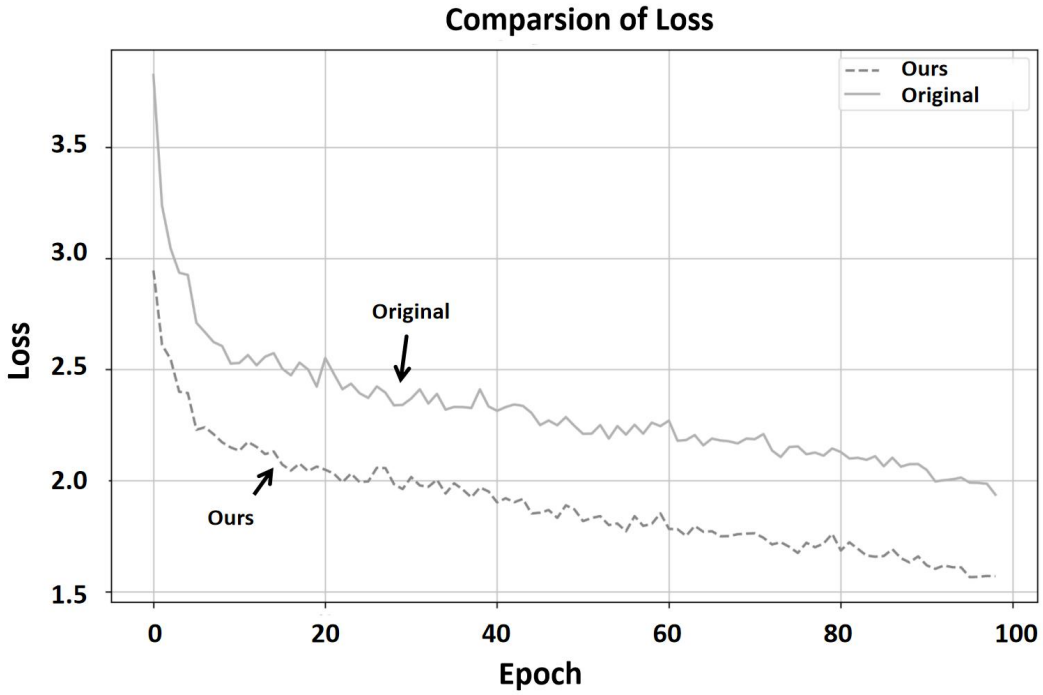
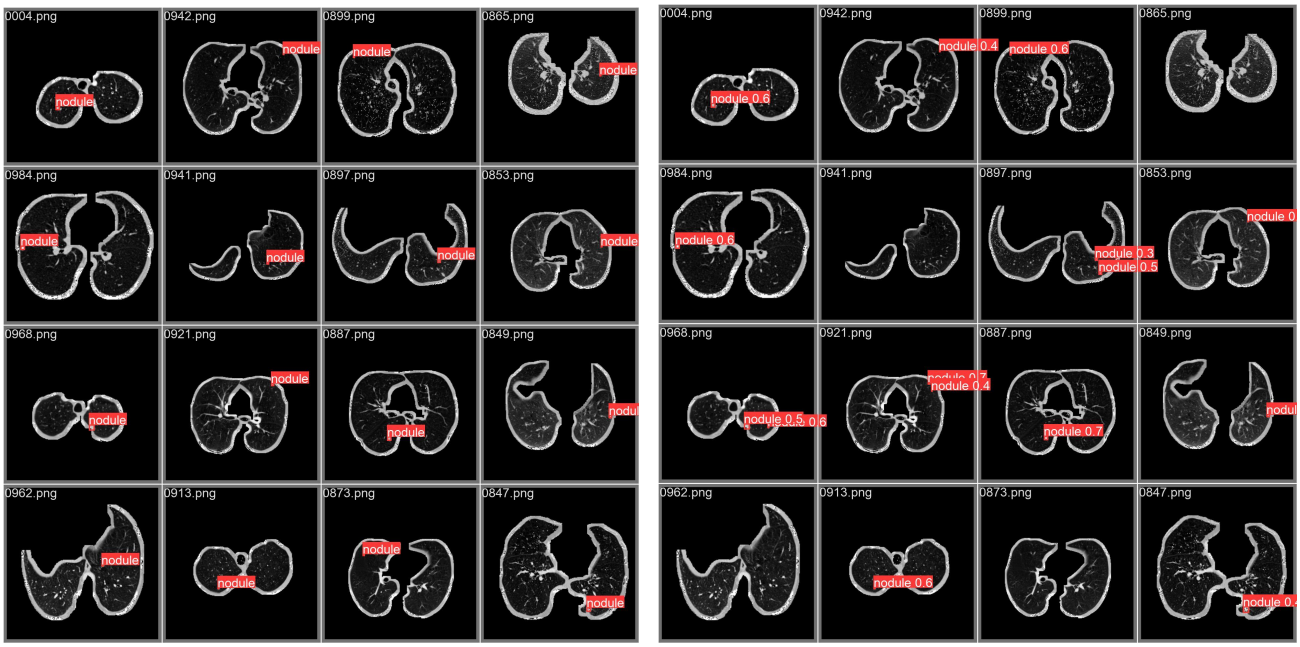


Fig. 14. Comparison of the loss functions (LUNA 16)



(a) YOLOv8n (b) small object detection-YOLOv8

Fig. 15. Detection Results of the YOLOv8n and small object detection-YOLOv8s (LUNA 16)

V. Conclusions

This research introduces Small Object Detection-YOLOv8, an advanced algorithm specifically designed to enhance the detection of pulmonary nodules in medical imaging. By incorporating the Multi-scale Context Aggregation (MCA) attention mechanism into the baseline YOLOv8n model, the proposed approach significantly improves its ability to capture global contextual information through bidirectional cross-attention. This enhanced attention mechanism enables the model to focus more effectively on relevant regions in medical images, thereby increasing the overall detection accuracy for small and subtle targets such as pulmonary nodules.

Additionally, an extra small-object detection head has been integrated alongside the original three prediction heads in the model. The introduction of this specialized prediction head contributes to reducing overfitting, particularly in datasets with limited samples, which is often the case in medical image analysis. The extra head specifically targets small object detection, resulting in improved precision for identifying pulmonary nodules amidst challenging visual contexts.

The model's loss function was also optimized by substituting the Complete Intersection over Union (CIoU) loss with the Mean Performance Distance Intersection over Union (MPDIoU) loss. The MPDIoU loss function was chosen for its streamlined computation and enhanced capability for multi-scale object localization. This substitution leads to more stable and accurate bounding box regression, which is crucial for correctly identifying the boundaries of small nodules, thereby improving the reliability of predictions.

When evaluated on benchmark medical imaging datasets, the enhanced Small Object Detection-YOLOv8 demonstrated significant performance improvements compared to the original YOLOv8n. Specifically, the model achieved an increase in mean Average Precision (mAP) at an Intersection over Union threshold of 0.5 (mAP@0.5) by 0.055 and 0.051 on the LIDC-IDRI and LUNA16 datasets, respectively. These gains highlight the superior detection capabilities of the improved model. Despite a moderate increase in floating-point operations (FLOPs) by 0.4 and 0.5 for each respective dataset, this computational trade-off is well-justified by the enhanced detection accuracy, particularly for small object instances.

Moving forward, future research will focus on further refining and optimizing the Small Object Detection-YOLOv8 algorithm to address the remaining challenges inherent to small object detection tasks, especially those related to pulmonary nodules. These efforts will include exploring more sophisticated feature extraction techniques, enhancing the robustness of the model against noisy data, and potentially integrating semi-supervised learning approaches to leverage unlabeled medical images. Such advancements aim to push the boundaries of detection accuracy and ensure the model's practical applicability in real-world clinical settings, where early detection of pulmonary nodules is critical for improving patient outcomes in lung cancer treatment.

References

[1] M. Hussain, "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, 2023.

[2] 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, 2015.

[3] 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, June 2017, doi: 10.1109/TPAMI.2016.2577031.

[4] R. B. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2013.

[5] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2015.
[6] A. Bochkovskiy, C. Wang and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," ArXiv, abs/2004.10934, 2020.
[7] C. Y. Wang, A. Bochkovskiy and H. Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464-7475, 2022.
[8] S. Liu, L. Qi, H. F. Qin, J. P. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," CoRR, abs/1803.01534, 2018.
[9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936-944, 2016.
[10] H. Kim and J. Ko, "Fast Monte-Carlo Approximation of the Attention Mechanism," in AAAI Conference on Artificial Intelligence, 2022.
[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR, abs/1409.1556, 2014.
[12] K. He, X. Zhang, S. Ren, J. Sun and Microsoft Research, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2015.
[13] G. Huang, Z. Liu, L. van der Maaten, Tsinghua University and Facebook AI Research, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2016.
[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," ArXiv, abs/1704.04861, 2017.
[15] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," ArXiv, abs/1905.11946, 2019.
[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992-10002, 2021.
[17] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 2020.
[18] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," ArXiv, abs/1804.02767, 2018.
[19] H. Y. Zhang, Y. Wang, F. Dayoub and N. Sünderhauf, "VarifocalNet: An IoU-aware Dense Object Detector," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8510-8519, 2020.
[20] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 12993-13000, 2020.
[21] S. Ma and Y. Xu, "MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression," ArXiv, abs/2307.07662, 2023.
[22] I. Naseer, S. Akram, T. Masood, A. Jaffar, M. A. Khan and A. Mosavi, "Performance Analysis of State-of-the-Art CNN Architectures for LUNA16," Sensors, vol. 22, p. 4426, 2022.
[23] M. B. Lin, R. R. Ji, Y. Wang, Y. C. Zhang, B. C. Zhang, Y. H. Tian and L. Shao, "HRank: Filter Pruning Using High-Rank Feature Map," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1526-1535, 2020.
[24] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European Conference on Computer Vision (pp. 21-37). Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2
[25] Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., & Heng, P. A. (2017). Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Transactions on Medical Imaging, 35(5), 1182-1195. https://doi.org/10.1109/TMI.2016.2528129