Applied Artificial Intelligence Research

INNOVIAIR

BCL-YOLO: A YOLO-Based Model for Medical Image Detection

Bohan Zhang[a,*], Tingting Yang[b]

[a]School of Pharmacy, Jiangsu Ocean University, Lianyungang, Jiangsu 222005, China

[b]School of Electronic Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu 222005, China

ARTICLEINFO

Keywords:
Medical Imaging Detections
BSA
BEA
LDConv

ABSTRACT

Medical image detection requires accurate localization of subtle lesions while maintaining efficient inference. Methods:We propose BCL-YOLO, a YOLOv8-based detector that integrates the BC2fB attention module,the C2fIRB inverted-residual module, LDSPPF with linear deformable convolution, and the Multi-InterpIoU loss. Results: On the Br35H brain-tumour MRI dataset, BCL-YOLO achieved 97.2% mAP@0.5 and 74.6% mAP@0.5:0.95, improving mAP@0.5 by 3.4 percentage points over YOLOv8. Ablation and comparative experiments confirmed the contribution of each component and showed competitive precision, recall, efficiency, and cross-dataset robustness on ChestX-Det10. Conclusions:BCL-YOLO improves medical lesion detection by combining attention-guided feature extraction, multi-scale fusion, and efficient bounding-box regression.

1.Introduction

With the rapid development of medical imaging technology, medical imaging data in many fields such as motor systems and biological tissues have shown explosive growth, which has put forward higher requirements for accurate detection, classification and segmentation of images[1]. In actual clinical practice, the traditional manual interpretation mode has obvious limitations: doctors need to face massive image data, the work intensity is extremely large, and the diagnosis may be biased due to differences in experience, visual fatigue and other factors. Especially in the key point detection task of medical images, manual recognition is often difficult to ensure the consistency and accuracy of the results[2]. Artificial intelligence technology can autonomously learn features from massive images, significantly improve the sensitivity and specificity of abnormality detection, and provide strong support for early disease screening. The detection of abnormal areas in medical images is a key part of disease diagnosis, but due to the complexity and diversity of abnormal features, artificial interpretation poses great challenges[3]. The morphology of some early lesions is blurred, the boundaries are unclear, and the grayscale difference with normal tissues is extremely small, which is difficult to accurately distinguish by the human eye alone. In some complex cases, the lesions may show atypical morphology or distribution, which is different from the characteristics of common diseases. When faced with such situations, traditional artificial methods are prone to miss detection or misjudgment due to thinking formulations or empirical limitations[4]. When there are multiple lesions in the same image and they overlap with each other, manual recognition requires analysis of the nature and scope of each lesion, which is not only time-consuming and labor-intensive, but also may be misjudged due to information interference. Artificial intelligence technology can autonomously learn features from massive images, and form accurate recognition capabilities for various abnormal patterns through the training of a large amount of labeled data. Even subtle and rare lesion features can be effectively captured, thereby significantly improving the sensitivity and specificity of abnormal detection and providing strong support for early disease screening[5].

The rise of deep learning technology has opened up entirely new possibilities for medical image detection. Compared with traditional manual detection methods, deep learning, with its powerful feature learning and pattern recognition capabilities, can automatically extract key features from massive medical image data to achieve accurate detection and classification of medical images. This data-driven method can not only effectively reduce the dependence on the subjective experience of doctors, but also significantly improve the efficiency and consistency of detection, and

* Corresponding author.
E-mail address: 3533506203@qq.com.
https://doi.org/10.65455/j2rdvd27
Received 30 August 2026; Received in revised form 10 Septembert 2026; Accepted 11 September 2026; Available 21 September 2026
© The Author(s) 2026. This article is published with open access in Applied Artificial Intelligence Research under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

provide strong technical support for the early screening and diagnosis of medical imaging lesions[6].

With the rapid progress of deep learning technology, its excellent automatic feature learning ability has brought new progress to the field of medical image detection. By building a multi-layer neural network, deep learning can independently learn abstract and high-level feature expressions from massive data, and then more accurately capture the core features of medical images. As a very representative model in deep learning, convolutional neural networks(CNNs)have achieved remarkable results in the field of image recognition and have also been widely used in medical image detection[7]. CNNs can automatically extract local and global features from images by means of convolutional layers, pooling layers, and fully connected layers[8], which effectively improves the accuracy of recognition.

This study aims to improve the accuracy and efficiency of medical imaging detection to meet the growing demand for practical applications. The specific goal is to name the improved model as the BCL-YOLO model, so that it can accurately identify lesion areas in complex medical images while maintaining a high inference speed. The contribution is a task-driven architectural adaptation rather than a claim of a new elementary algorithm: the selected attention, lightweight structure, deformable convolution, and regression-loss techniques are combined because they address different medical-image bottlenecks. The design rationale and component compatibility are explained in Section 3, and the contribution of each component is verified in Section 4.

In order to achieve the above objectives, this study mainly carries out the following aspects of work:

(1)Improve the C2f module of the YOLOv8 model. Combining the BSA bipolar self-attention mechanism with the BEA bipolar high-efficiency attention, so that the model pays more attention to the key areas of the medical image during the feature extraction process, and enhances the ability to capture the lesion areas of the fine medical image;

(2)Refer to the inverted residual module IRB to improve the structure of the C2f module, optimize the calculation efficiency of the model, and improve the feature extraction effect without adding too much calculation;

(3)Improve the SPPF module. Introduce linear deformable convolution LDConv to optimize the spatial feature fusion mode of the SPPF module, so that it can better adapt to the regional features of medical image lesions at different scales, and further improve the model's ability to extract complex medical image features;

(4)Optimize the loss function. Use the expanded loss function Multi-InterpIoU to replace the original loss function and speed up the reasonable regression of the loss function, thereby improving the identification accuracy;

By integrating the above four innovative modules, BCL-YOLO forms an end-to-end detection framework that achieves a balance between accuracy and efficiency in medical image detection tasks.

The rest of this article is structured as follows: Section II focuses on work related to medical imaging testing. Section III describes the model framework and implementation details. The fourth section verifies the effectiveness of the algorithm through experiments. Section V is a summary of the study.

2.Related work

Wang Shaolin et al.[9] carried out research on the clinical focus of preoperative vision prediction for cataracts, and explored the application value of small hole mirrors in this scene. Through relevant experiments or clinical data observation and analysis, the research results show that small hole mirrors can play a certain role in preoperative vision prediction for cataracts, which can provide a reference for doctors to evaluate patients' postoperative vision recovery, contribute to more accurate surgical planning and evaluation of surgical results, and provide valuable practical basis for clinical diagnosis and treatment of cataract medical images.

Based on the Chest X-ray dataset, Guo Shaoning et al.[10] pretreated the chest X-ray images of the lungs, and used the DenseNet-121 model to carry out pneumonia classification experiments to achieve rapid and accurate clinical analysis and diagnosis 1. The innovation lies in the application of DenseNet-121 to pneumonia classification, using its dense connection characteristics to improve performance, with a test accuracy rate of 97.4%, which is better than ResNet.

Huang Flute et al.[11] proposed two models and developed a system based on self-supervised learning for the detection of key points in orthopedic images. Innovative points: First,combining self-supervised pre-training with edge prior information, using multi-stage networks and relationship vector loss to mine key points;second, based on self-supervision and knowledge distillation, improving cross-domain detection through pseudo-label updating, adversarial learning and data consistency disturbance; third, developing an integrated algorithm detection system. Accuracy:Knee joint data set MRE 2.41 mm, detection rate within 5 mm 81.87%; skull data set MRE 1.66 mm, detection rate within 4 mm 94.3%.

In computer vision, object detection is a critical task that automatically identifies specific objects in an image or video and annotates the locations and categories of those objects. With the rapid development of deep learning technology, object detection algorithms have made significant progress, and many excellent algorithms have also emerged, such as YOLOv8[12], RT-DETR[13], SSD[14], Faster R-CNN[15], etc. These algorithms play an important role in different application scenarios, promoting the wide application of target detection in security monitoring,automatic driving, face recognition and other fields.

Kumar A et al.[16] proposed an improved tiny YOLOv7 model variant EYE-YOLO. The model integrates multi-spatial pyramid pooling in the feature extraction network of tiny YOLOv7, and introduces the Focal-EIOU loss function in the detection network. Multi-space pyramid pooling can help extract richer features, and the Focal-EIOU loss function can better deal with various problems in target detection and improve detection accuracy. The experimental results show that the accuracy of the EYE-YOLO model is 28% higher than that of the tiny YOLOv7 model, the recall rate is 18% higher, the F1 score is 24% higher, and the mAP is 30.81% higher.

Lokesh L P et al.[17] evaluated the existing YOLO model using 788 normal images and 920 cataract images. Their innovation lies in introducing GhostConv into the YOLO

framework and proposing a detection model based on GhostYOLO to ensure detection accuracy while reducing computation and parameter scale and improving real-time performance. The study achieved a high detection performance with an accuracy rate of 96.8% and an average accuracy(mAP)of 95.2%.

Wang Yuhan et al.[18] applied the YOLOv8 model to the detection of hand fracture images, built datasets and trained the model through preprocessing. The innovation lies in the precise identification and positioning of hand fractures with the strong feature extraction and rapid detection capabilities of YOLOv8. Experiments show that the average accuracy of the improved model is 0.57 percentage points higher than that of the original model, which improves the detection accuracy.

Bao Qiangqiang et al.[19] improved the YOLOv8 model for multi-type pulmonary nodule detection: proposed the RepViTCAA module to improve the backbone C2f module to improve the accuracy of tiny pulmonary nodule detection and lightweight the model; proposed the ECLA-HSFPN module to reconstruct the feature fusion part to improve the accuracy of scale-unfixed pulmonary nodule detection; integrated into the Kan network to enhance the accuracy of tiny pulmonary nodule detection and the model generalization ability; improved the CIoU loss function based on the Inner-IoU auxiliary box idea to further solve the problem of scale-unfixed. The innovation lies in the improvement of the above modules and networks to solve the problems of tiny and irregular scales in the detection of multiple types of pulmonary nodules. In the Luna 16 dataset test, the accuracy rate of the improved model was 91.8 ± 0.05%, the recall rate was 90.8 ± 0.05%,the mAP @ 0.50 was 93.7 ± 0.08%, and the mAP @ 0.50:0.95 was 51.7 ± 0.09%. All indicators were better than the original model and the mainstream models such as YOLOv9 and RT-DETR. The detection effect on the four types of special datasets was also better than the original model.

Anbalagan K et al.[20] integrated target detection and temporal modeling technology for the first time, used YOLO-v7 to accurately locate key areas in eye images(such as optic neural disks), and then captured potential temporal features or contextual associations in images through GRU to improve the accuracy and robustness of glaucoma detection, providing a new paradigm for multi-modal intelligent diagnosis of eye diseases.

Problems in the above literature:

1.Insufficient pertinence of feature extraction

Some models lack accuracy when capturing specific features of lesions in medical images. Key lesion features are easily interfered with by irrelevant background information, resulting in redundant feature extraction. For those early lesions with weak features, the feature extraction ability of the basic model is relatively limited, and it is easy to miss the detection, making it difficult to improve the recall rate.

2.Inadequate integration of multi-scale features

There are significant differences in the scale of the lesion in the medical image: for example, it may be a small local lesion or a large area of lesion, and the morphological changes of the lesions in different parts involve different scales. The feature fusion module of some models lacks the ability to integrate multi-scale features, which may lead to the omission of small-scale lesions or the loss of detailed features of large-scale lesions, which in turn affects the comprehensiveness of detection.

3.The difficulty of balancing calculation efficiency and accuracy

In order to improve the detection accuracy, some models have increased the scale of parameters and the amount of calculation, resulting in a decrease in real-time,which is difficult to meet the need for rapid response in clinical auxiliary diagnosis. Although the combination of traditional machine learning and deep learning methods improves the fine-grainedness to a certain extent, the multi-step fusion process may introduce redundant calculations and no weighted processing of key features, so there is still room for optimization in the balance between efficiency and accuracy.

In view of the above problems, this paper proposes a collaborative optimization detection framework BCL-YOLO, which achieves a breakthrough in performance through two core dimensions: attention fusion mechanism and multi-scale feature fusion capability. Specifically, the integration optimization is achieved by innovating the key modules of the model: designing the BC2fB module to integrate the BSA bipolar self-attention mechanism with the BEA bipolar high-efficiency attention, strengthening the focus on the key areas of medical images and the capture of subtle features; building the C2fIRB module to optimize the convolutional layer design based on the structure of the IRB inverted residual module to improve the efficiency and characterization ability of feature extraction; improving the LDSPPF module to enhance the cross-layer interaction and fusion accuracy of multi-scale features combined with LDConv. These integrated innovations enable BCL-YOLO to achieve high-performance detection in medical image detection scenarios, effectively balancing the contradiction between lightweighting and detection accuracy, while improving robustness in complex scenarios.

The five components were selected to address different failure modes in lesion detection rather than to assemble unrelated improvements. BSA supplies complementary spatial- and channel-attention cues for weak lesions; BEA re-weights the C2f output with efficient complementary-polarity context and reduces the cost of repeated attention; IRB constrains C2fIRB so richer feature extraction does not inflate parameters or GFLOPs; LDConv makes SPPF sampling adaptable to irregular lesion shapes and multi-scale contexts; and Multi-InterpIoU improves box regression when lesion boundaries are blurred. In BCL-YOLO, BSA and BEA operate sequentially on the input and output of the C2f branch, IRB changes the internal convolutional design, LDConv changes SPPF feature fusion, and Multi-InterpIoU acts only at the detection head. This division limits functional duplication and computational conflict. The ablation tests in Section 4 further isolate each contribution.

3.Method

3.1.YOLOv8 algorithm

As an advanced algorithm in the field of target detection, YOLOv8's network structure consists of three main parts: the backbone network, the neck and the detection head. Each part cooperates with each other to jointly achieve efficient detection of the target object. The YOLOv8 algorithm is shown in Figure 1.

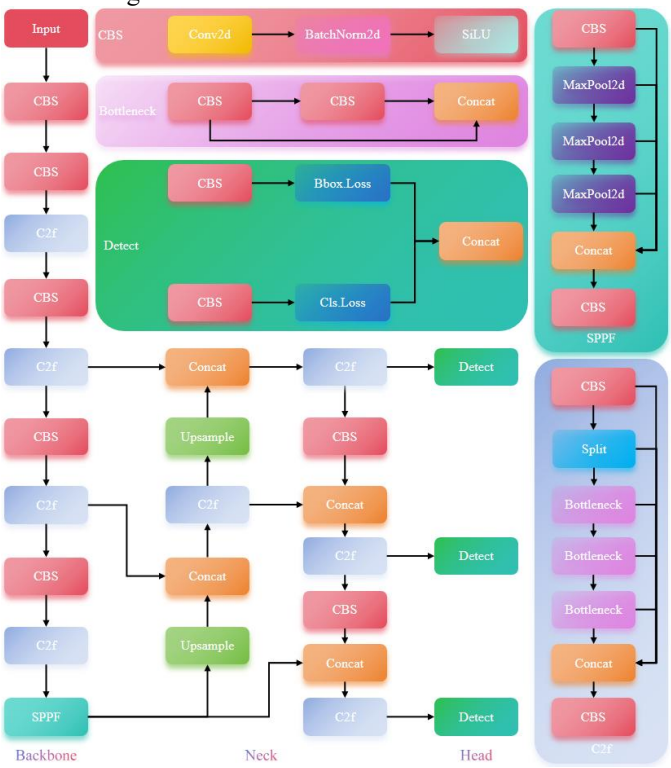


Fig 1. YOLOv8 algorithm

As the feature extraction part of YOLOv8, the backbone network undertakes the core task of extracting key information from the input image, providing high-quality feature support for the subsequent detection process. YOLOv8 uses an improved CSPDarknet53[21] as the backbone network, and the input features are downsampled five times to obtain features at five different scales. In addition, the Cross Stage Partial Connection design in the CSPDarknet53 structure also plays an important role in the backbone network. By dividing the feature map into two parts, one part is convolutionally operated, the other part is directly transmitted, and finally the two parts are fused. This design approach effectively reduces the amount of computation because it avoids double-counting some redundant information throughout the network. At the same time, by retaining the path directly transmitted, the original information in the feature map can be better retained, so that the network can maintain a high feature extraction capacity while reducing the amount of computation, which provides a solid basis for the subsequent detection process.

The neck structure connects the backbone network and the detection head, and is mainly responsible for fusing semantic information from different levels of the backbone network to enhance the detection ability of targets at different scales. Inspired by PANet, YOLOv8 was designed with a PAN-FPN structure in the neck[22]. With the help of top-down paths and horizontal connections, FPN will fuse the semantic information of the upper layer and the detail information of the lower layer to generate a multi-scale feature map with full semantic features and details. On this basis, pan was further optimized. Through the addition of bottom-up path aggregation, the information flow between different levels of features was strengthened, so that the fusion ability of the neck structure to multi-scale features was enhanced, and the detection performance of small targets was further improved. This design not only enhances the information transfer between different levels of features, but also enables features from different scales to be more fully integrated, thus improving the network's ability to detect multi-scale targets[23].

The detection head is the part of YOLOv8 that finally outputs the detection results. It predicts the bounding box coordinates of the target object, the category probability, and other related properties based on the multi-scale feature map of the neck output. YOLOv8 uses an anchorless mechanism to return directly to the parameters of the target box, that is,to predict the distance from each feature point to the four edges, thereby determining the position and size of the bounding box. This method avoids the need to preset the size and proportion of the anchor frame in the traditional anchor frame mechanism, simplifies the design of the model, and improves the flexibility and efficiency of detection. The detection head processes the feature map through a series of convolutional layers and then outputs the prediction results. For each predicted position, the binary classification loss function is used to generate the category, and the CIOU loss + DFL loss is used to generate the bounding box coordinates and the corresponding category probability[24], and finally the accurate detection results are obtained.

YOLOv8 contains a number of modules with a unique design and functionality, where C2f modules and SPPF modules play a key role in the feature extraction and fusion process.

The C2f module is an important part of the YOLOv8 backbone network, and it has an important impact on the model's feature extraction ability and computing efficiency. The design of the C2f module refers to the Elan structure in YOLOv7, combining C3 and Elan to form the C2f module[25]. The C2f module is designed to effectively reduce redundant computing and improve the training efficiency and feature representation ability of the model through the cross-phase partial connection mechanism. The module adopts a combination of residual connection and convolution operation in structure. By dividing the input feature map into two parts, one part directly passes through the convolution layer for feature extraction, and the other part is fused with the convolution-treated feature map through residual connection, as shown in Figure 2. This design enables the module to make full use of different levels of feature information without adding too much computation, enhancing the model's ability to extract complex features.

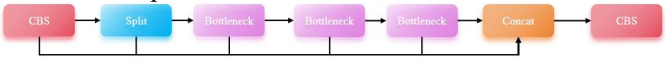


Fig 2. C2f module

The SPPF module is a key component in the YOLOv8 backbone network, focusing on the efficient aggregation of multi-scale features to provide more comprehensive semantic

information for subsequent detection tasks. As shown in Figure 3, the module adopts the innovative design of serial pooling + parallel fusion. First,the input feature map is serially processed by a plurality of maximum pooling layers of different sizes to quickly capture image features of different scales. Then, the original features are concatenated with the pooling results to enhance the richness of feature expression while retaining detailed information. Compared with the traditional spp structure, SPPF greatly reduces the computational complexity and reasoning time by optimizing the operation sequence of the pool while maintaining the same feature extraction ability, which not only improves the feature fusion efficiency, but also ensures the real-time performance of the model, providing an important support for the high-precision detection of YOLOv8.



Fig 3. SPPF module

3.2.BC2fB module

3.2.1.Bipolar self-attention

BSA(Bipolar Self-Attention) is a bipolar self-attention mechanism that captures information in different dimensions simultaneously by introducing spatial-focused attention and channel-focused attention to make better use of the characteristic relationships between pixels. So it's called a bipolar self-attention network. Each BSA module contains: Spatial Focus Attention (SFA) and Channel Focus Attention (CFA)[26], and the SFA and CFA modules are shown in Figure 4.

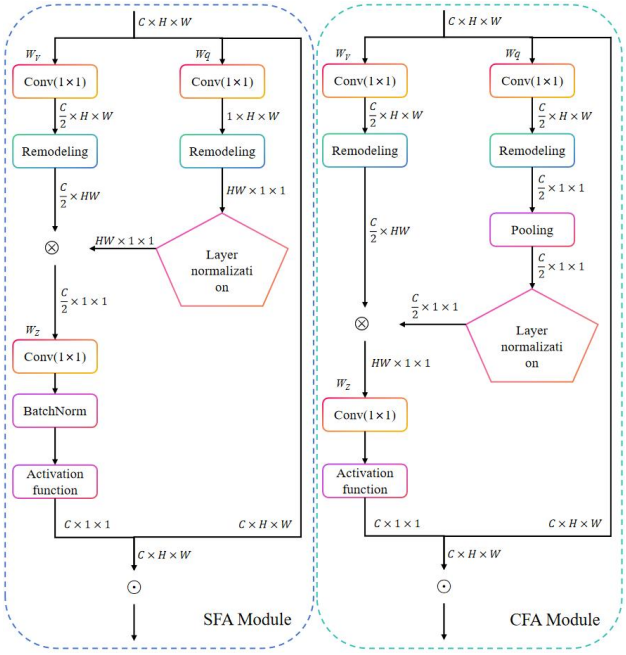


Fig 4. SFA and CFA modules

Each BSA module combines spatial and channel attention: spatial attention enhances the representation of local features by introducing broader contextual information; channel attention emphasizes the interdependence between feature maps,optimizing the expression of specific semantics.

The CFA weight is expressed as $A^{ch}(X)\epsilon R^{C\times1\times1}$, while the SFA weight is expressed as $A^{sp}(X)\epsilon R^{1\times H\times W}$. The output of the spatial branch is $Z_{sp}=A^{sp}(X)\odot^{sp}X\epsilon R^{C\times H\times W}$, where $\odot^{sp}$ denotes element-wise multiplication along the spatial dimensions and "+" denotes element-wise addition.

The design of BSA achieves a double advantage: it can accurately preserve the spatial details of the data and efficiently extract high-resolution features. This feature becomes a key support for analyzing complex data patterns, enabling BSA to deeply explore the value of high-resolution information when dealing with complex data structures of eye lesions, and thus more accurately capture the subtle features and intrinsic correlations in medical images.

Incorporating the BSA bipolar self-attention mechanism into the C2f module can bring multiple advantages to the model. The spatial and channel-focused attention of the BSA enables the model to more intelligently parse complex data as it processes image features. In medical imaging testing, the BSA is able to select branches to accurately capture this critical information for some subtle lesion areas, thereby improving the accuracy of medical imaging testing.

3.2.2.Bipolar efficient attention

BEA(Bipolar Efficient Attention) is a new type of attention mechanism designed to improve the efficiency and performance of the model's attention, and its design concept is based on the in-depth analysis and improvement of the traditional attention mechanism. When dealing with large-scale data and complex tasks, traditional attention mechanisms often face problems such as high computational complexity, large memory consumption, and insufficient ability to capture key information. The BEA attention mechanism effectively solves these problems through unique structural design and calculation, providing the model with more efficient attention allocation capabilities[27].

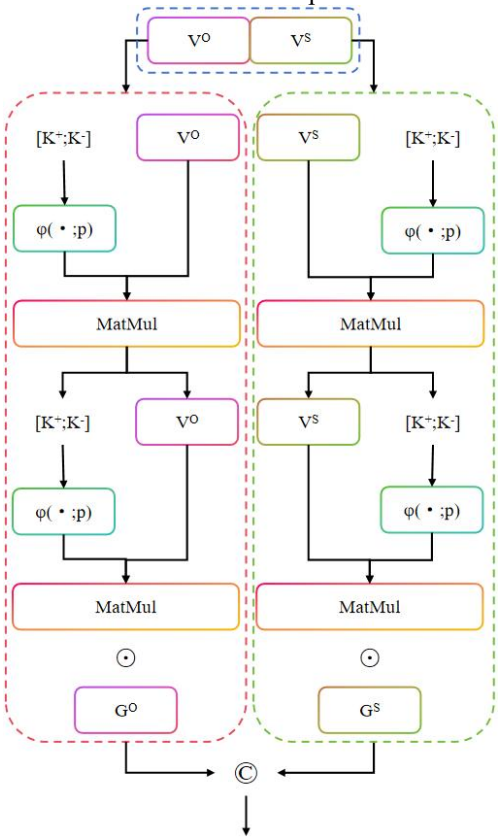


Fig 5. BEA attention mechanism

The core architecture of the BEA attention mechanism consists of a bipolar feature extraction module and an adaptive weight allocation module(as shown in Figure 5). Among them, the bipolar feature extraction module adopts bidirectional processing logic to synchronously extract the input features from the forward and reverse dimensions. This bidirectional mechanism can capture the feature association and context information more comprehensively. Taking the image data as an example, the forward extraction focuses on the spatial distribution from left to right and from top to bottom. The reverse extraction supplements the information dimension from right to left and from bottom to top, effectively avoiding the information loss problem that may occur in unidirectional extraction. Through such bipolar feature extraction methods, the model can obtain richer feature representations and significantly improve the ability to analyze complex patterns.

As another core component of BEA, the adaptive weight assignment module dynamically assigns weights based on the importance of the input features. Through a series of calculations and transformations, the module is able to learn key features autonomously and give them a higher weight. In medical imaging testing, for normal images and lesion areas, the module will automatically increase the weight to enhance the attention of the model, thereby improving the recognition accuracy; at the same time, it assigns low weights to irrelevant background information or noise, effectively reducing its interference with decision-making.

The BEA attention mechanism is calculated as follows:

$$\widetilde{Q}=[\mathscr{R}(Q/s)^p,\mathscr{R}(-Q/s)^p] \quad (1)$$

$$\widetilde{K}=[\mathscr{R}(K/s)^p,\mathscr{R}(-K/s)^p] \quad (2)$$

Here, Q and K denote the query and key matrices, respectively; s is a learnable scale parameter; p is a learnable polarity-power parameter; $\mathscr{R}$ denotes the ReLU activation function, and $[\cdot,\cdot]$ denotes feature concatenation. This positive-negative decomposition allows the network to capture complementary feature representations and enhance the representation of lesion details.

BEA is calculated in an efficient linear form:

$$Z_{sim}=\frac{\widehat{Q}_{sim}\cdot(\widehat{K}^T\cdot V_{1/\sqrt{n}})}{\widehat{Q}_{sim}\cdot\overline{K}^T+\varepsilon} \quad (3)$$

$$Z_{opp}=\frac{\widehat{Q}_{opp}\cdot(\widehat{K}^T\cdot V_{2/\sqrt{n}})}{\widehat{Q}_{opp}\cdot\overline{K}^T+\varepsilon} \quad (4)$$

Here, $\widehat{Q}_{sim}$,respectively,denote query representations with similar and opposite polarity; $\widehat{Q}_{opp}$ and V_1 are the two parts of the value matrix; V_2 denotes the mean key representation; n is the number of keys; and $\widehat{K}$ is a stability factor. The final attention output is modulated through gating and positional enhancement:

$$X_{out}=G\odot[Z_{sim}Z_{opp}]+\mathscr{F}_{DW}(V) \quad (5)$$

Here, G is the gating factor, $\odot$ denotes element-wise multiplication, and $\mathscr{F}_{DW}$ denotes depthwise separable convolution.

Incorporating the BEA attention mechanism into the C2f module can significantly improve the performance of the C2f module. The bipolar feature extraction method of BEA enables the C2f module to obtain more comprehensive information during the feature extraction process, and enhances the understanding and expression of the characteristics of the medical imaging lesion region. When processing medical images, the lesion area of the medical image can be captured more accurately, thereby improving the quality of feature extraction. The adaptive weight assignment module can help the C2f module to better focus on the key features of medical images and suppress the interference of irrelevant information. In the detection of medical images under complex backgrounds, the C2f module can highlight the lesion areas of medical images through the weight distribution mechanism of BEA, reduce the impact of background noise on the identification results, and improve the robustness and accuracy of the model. In addition, the efficiency of the BEA attention mechanism can also reduce the computational complexity of the C2f module to a certain extent, improve the operation efficiency of the model, and make it more advantageous in practical applications.

3.2.3.Attention fusion

The construction of the BC2fB module combines the BSA bipolar self-attention mechanism, the C2f module, and the BEA bipolar high-efficiency attention. This unique combination is designed to give full play to the advantages of each part and improve the performance of the model in target detection and feature extraction, as shown in Figure 6.

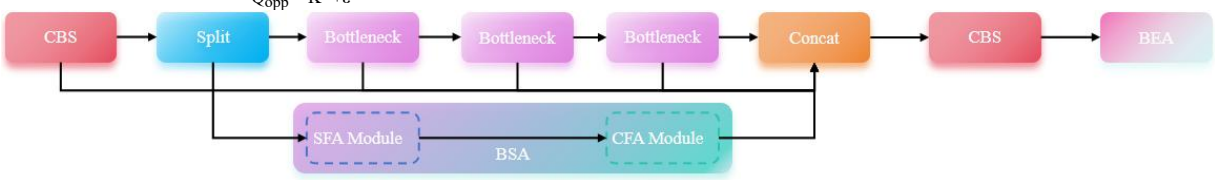


Fig 6. BC2fB module

As the Split single branch part of the BC2fB module, the BSA bipolar self-attention mechanism first inputs data into the SFA spatial focus attention module, and then the CFA channel focus attention module. After the two modules work together, the BSA bipolar self-attention mechanism can provide filtered and optimized feature representations for subsequent C2f modules, enabling the C2f module to more efficiently utilize computing resources while obtaining richer global and local information when processing data.

The C2f module is at the core of the BC2fB module and receives feature data from the BSA bipolar self-attention mechanism after processing. The C2f module is based on a cross-phase partial connection(CSP)structure design. By dividing the input feature map into two parts, one part is directly delivered to the output through a short connection, and the other part is processed by a plurality of stacked BottleNeck structures, and then the processed features are cross-fused with the original features. This design enhances the ability to express multi-scale features and effectively extract different levels of feature information. In the BC2fB module, the C2f module utilizes the optimized features provided by the BSA to further mine the key features and

patterns in the data, and improves the accuracy and robustness of feature extraction through technologies such as multi-branch fusion and residual connection.

The BEA Bipolar High Efficiency Attention, as the back-end part of the BC2fB module, reprocesses the characteristics of the C2f module output. Through the bipolar feature extraction module, BEA conducts in-depth mining of the features output by the C2f module from both forward and reverse directions, comprehensively capturing the relationship and context information between the features. Then,the adaptive weight assignment module dynamically assigns weights to different features according to their importance, highlighting key features and suppressing irrelevant information. After BEA processing, the features can more accurately reflect the essential characteristics of the data, providing a more reliable basis for subsequent testing tasks.

In the target detection process, the BC2fB module is mainly applied to the three key positions before detection. In the early stages of the backbone network, the BC2fB module can help the model quickly capture global information and key local features in the image, laying the foundation for subsequent feature extraction and targeting. In the middle stage of the backbone network, the BC2fB module is able to further explore the relationships between features and enhance the ability to understand complex targets and scenarios. At the connection between the backbone network and the neck structure, the BC2fB module can optimize and filter the characteristics of the backbone network output, making it more suitable for the multi-scale feature fusion requirements of the neck structure, so as to improve the detection ability of the entire model for different scales. By applying the BC2fB module in these three locations, the model can make full use of its advantages at different stages to improve the accuracy and efficiency of target detection.

3.3.C2fIRB module

3.3.1.Inverted residual block

Inverted Residual Block(IRB)is a structure widely used in lightweight convolutional neural networks. Its design concept is derived from the optimization and improvement of the traditional convolutional neural network structure. It aims to improve the efficiency and accuracy of feature extraction while keeping the model lightweight.

The structure of the inverted residual module IRB is mainly composed of three parts: expansion layer, depth separable convolution layer and compression layer, as shown in Figure 7.

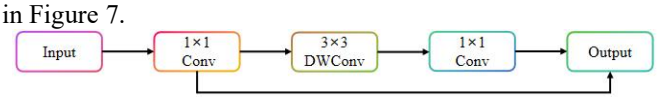


Fig 7. Inverted residual module

In the expansion layer, the number of channels of the feature map is increased by 1×1 convolution, which is similar to the expansion of the feature, allowing the model to capture more information in the low-dimensional feature space. This dilatation operation increases the nonlinear expression ability of the model and provides a broader feature basis for subsequent feature extraction. Taking the processing of image data as an example, when the number of feature map channels of the input image is small, the number of channels can be increased by expanding the layer, so that the model can learn more feature information of different dimensions.

The deep separable convolution layer is the core part of the IRB, which decomposes the traditional convolution operation into Depthwise Convolution and Pointwise Convolution[28]. The depth separable convolution is followed by two stages: first, the depth convolution, which convolves each input channel independently, significantly reducing the amount of computation; second, the point convolution, which restores the number of channels and introduces nonlinearity, enhances the expression ability of the model[29]. Deep convolution is an independent convolution operation for each channel, mainly used to extract the spatial characteristics of the image, which can effectively capture the detailed information of local areas in the image; point-by-point convolution is a convolution operation on the channel dimension on the basis of deep convolution, which is used to fuse the characteristic information of different channels. This decomposition method greatly reduces the amount of calculation and improves the operation efficiency of the model. Compared with the traditional convolution operation, the depth separable convolution can significantly reduce the computational complexity while maintaining the feature extraction ability, so that the model can operate efficiently in a resource-constrained environment.

The compression layer reduces the number of channels of the feature map by 1×1 convolution, "compresses" the feature map after the depth separable convolution process, and restores the number of channels similar to the input feature map to reduce the number of parameters and calculation amount of the model. While retaining the key feature information, this operation avoids the overfitting problem caused by the excessive number of channels, and improves the generalization ability of the model. Through the compression layer, the model is able to use computing resources more efficiently, focusing on extracting the features that are most valuable to the task.

In the feature extraction process, IRB uses its unique structure to effectively capture local and global features in images. The depth convolution in the depth separable convolution layer can capture the fine textures, edges, and other features of local areas in the image, while the point-by-point convolution can fuse these local features in the channel dimension to form a more representative feature representation. The cooperation between the expansion layer and the compression layer allows the model to flexibly adjust the number of channels of the feature map at different stages to meet the needs of feature extraction at different levels. When processing complex images, IRB can gradually extract low-level to advanced features from the image through multiple expansion, depth separable convolution and compression operations, providing strong support for subsequent classification,detection and other tasks.

3.3.2.Fusion inverted residual module

The C2f module is improved by referring to the IRB, aiming to combine the advantages of the two and improve the feature extraction ability and calculation efficiency of the model. This improvement is mainly reflected in the

adjustment and optimization of the internal structure of the C2f module.

In terms of improvement ideas, the convolutional layer in the C2f module was redesigned based on the expansion and compression ideas of the IRB. In the input part of the C2f module, add the CBS module to operate like the IRB expansion layer, appropriately increase the number of channels of the feature map by 1×1 convolution, and replace the activation function with SwiGLU to generate a new module CBSG, as shown in Figure 8. This operation allows the C2f module to obtain richer feature information when it begins to process features, providing a broader basis for subsequent feature fusion and extraction. When processing medical images, increasing the number of channels allows the model to learn more about the detailed features of medical images, such as pupils, eyes, etc., thereby improving the ability to capture medical image features. In the output part of the C2f module, refer to the design of the IRB compression layer to reduce the number of channels of the feature map by 1×1 convolution. This operation can avoid the overfitting problem caused by the excessive number of channels while retaining the key feature information, and improve the generalization ability of the model. After the feature extraction and fusion of the middle part, the compression operation of the output part can make the model focus more on the most valuable features of the task, reduce the interference of redundant information, and thus improve the accuracy of the model in the medical image detection task.



Fig 8. CBSG module

In order to further improve the feature extraction performance of the module, deep separable convolution was introduced into the CBS module to replace some traditional convolution operations, and the activation function was replaced with SwiGLU to generate a new module DBSG, as shown in Figure 9.



Fig 9. DBSG module

The deep separable convolution decomposes the convolution operation into deep convolution and point-by-point convolution, which can effectively extract spatial features and fusion channel information while reducing the computational load. The new Bottleneck-DBSG module is generated using the DBSG module instead of Bottleneck's CBS module, as shown in Figure 10. When processing large-scale image datasets, the use of deep separable convolution can significantly reduce the consumption of computing resources, so that the model can complete the feature extraction task in a shorter time, and meet the application scenarios with high real-time requirements.



Fig 10. Bottleneck-DBSG module

The improved C2fIRB module is more compact and efficient in structure. As shown in Figure 11, by reasonably adjusting the type and number of channels of the convolution layer, the balance between feature extraction ability and computational efficiency is achieved.

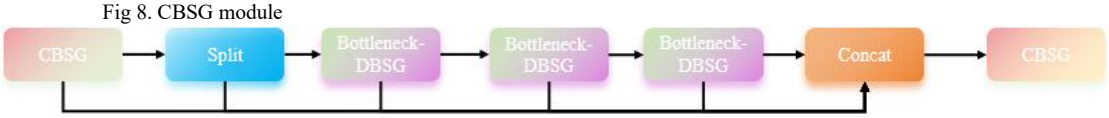


Fig 11. C2fIRB module

In practical applications, the C2fIRB module can replace the C2f module at the rest of the YOLOv8 model. In the different layers of the backbone network, the C2fIRB module can flexibly extract and fuse features according to the characteristics of each level. In the shallow network, it can quickly capture low-level features of the image, such as edges, textures, etc., by increasing the number of channels and adopting deep separable convolution; in the deep network, it can extract higher-level semantic features, such as the overall pattern of medical images, by compressing the number of channels and further feature fusion. Through this replacement, the model can improve the ability to extract and identify medical image features without significantly increasing the amount of computation, and improve the performance of the model in complex scenarios.

3.4.LDSPPF module

3.4.1.Linear deformable convolution

The main function of the linear deformable convolution(LDConv[30])module is to solve the performance limitations of standard convolution and traditional deformable convolution. Due to the fixed shape and size of the local sampling window, the standard convolution is difficult to adapt to the morphological differences of different objects. Although the traditional deformable convolution can flexibly adjust the sampling position, the number of parameters increases squarely with the size of the convolution kernel, resulting in low computational efficiency. LDConv uses a new coordinate generation algorithm to generate initial sampling positions for convolutional kernels of various sizes: first, establish a regular sampling mesh, then build an irregular mesh for the remaining sampling points and stitch it into a complete mesh. At the same time, the upper left corner is set as the sampling origin(0,0)to adapt to convolutional kernels of different sizes, which not only changes the growth trend of the number of parameters to linear, but also allows the convolutional kernel to have any number of parameters and sampling shapes, providing more possibilities for balancing network overhead and performance. The workflow is as follows: First, according to the number of convolution kernel parameters N, the initial shape and the original sampling coordinates are obtained by the coordinate generation algorithm; then, the offset with a dimension of (B, 2N, H, W) is obtained by the convolution operation, and the new sampling coordinates are obtained by adding the offset with the original coordinates to provide a variety of sample shapes for the convolution at each position of the feature map; finally,

the features at the new coordinates are interpolated and resampled, and then the feature extraction is completed by the convolution operation. The LDConv process is shown in Figure 12.

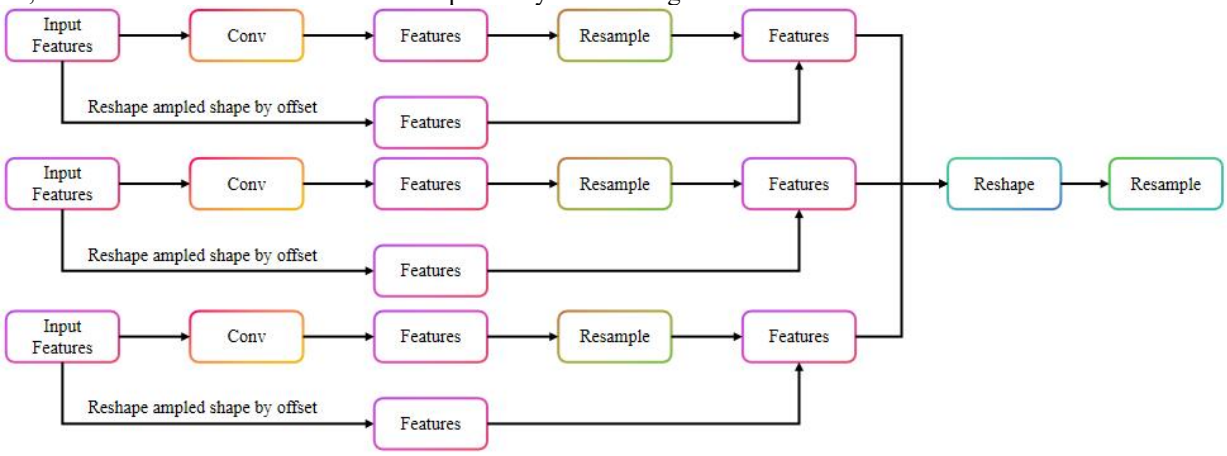


Fig 12. LDConv module

LDConv can replace the standard convolution in the YOLO model and effectively improve its performance. Related studies used different initial convolution kernel sizes and shapes, demonstrating the advantages of LDConv when dealing with diverse convolutional needs[31].

3.4.2.Optimizing the module structure

The SPPF module is mainly used to achieve multi-scale feature fusion in YOLO, and the feature maps of different scales are stitched by the maximum pooling for multiple consecutive times to improve the detection accuracy. In order to further simplify the parameters and calculation amount of the model, the SPPF module introduces the LDConv design concept and replaces the traditional convolution operation twice to optimize the performance. The improved LDSPPF module is more compact and efficient in structure. As shown in Figure 13, by reasonably adjusting the type and number of channels of the convolution layer, the balance between feature extraction ability and calculation efficiency is achieved.

Specifically, the two standard convolution operations included in the original structure of the SPPF module have certain limitations in the feature extraction process. The fixed sampling window of the standard convolution is difficult to flexibly adapt to different shapes of the multi-scale feature map, and the number of parameters increases rapidly with the size of the convolution kernel, increasing the computational burden of the model. With its linear growth parameter characteristics and flexible sampling shape adjustment capabilities, LDConv is ideal for optimizing SPPF modules. In the LDSPPF module, the first replacement of LDConv is mainly used in the feature preprocessing stage before multi-scale feature fusion. The input feature map is targeted sampled through dynamically generated sampling coordinates, which can more accurately capture key information of different scale features and provide a more effective feature basis for subsequent maximum pooling operations. The second replacement of LDConv is applied to the feature integration stage after multi-scale feature stitching. The irregular convolution operation is used to deeply extract the mixed features after stitching, which enhances the correlation between the features of different scales and improves the quality of feature fusion.

In terms of structural design, the LDSPPF module adapted the number of channels of the convolution layer after introducing LDConv. Compared with the original SPPF module, under the premise of ensuring that the feature extraction ability does not decline, the number of channels in some convolutional layers is appropriately reduced. Combined with the advantages of linear growth of LDConv parameters, the overall parameter scale of the module is significantly reduced. At the same time, the module retains the original multi-scale maximum pooling structure of SPPF. After dimensional compression through 1×1 convolution, the obtained multi-scale features are stitched with the original features through the maximum pooling operation of different pooling core sizes in turn, and then feature integration and dimensional recovery are performed through the second LDConv, and finally the activation function is output. This structural design not only continues the advantages of the SPPF module in multi-scale feature fusion, but also improves the pertinence and efficiency of feature extraction with the flexibility of LDConv.

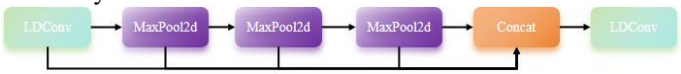


Fig 13. LDSPPF module

3.5.Optimizing the loss function

A range of extended IoU-based losses, such as GIoU[32], DIoU[33], CIoU[34], SIoU[35], and PIoU[36], are proposed to enhance IoU performance with additional geometric penalty factors (like center spacing, aspect ratio, or angle,etc.). Although some improvements have been made in practical applications, most of these geometric penalty factors are intended to alleviate the problem of gradient disappearance in non-overlapping scenarios, or to improve convergence efficiency by providing auxiliary geometric guidance. However, since they are not specifically designed to maximize IoU, these formulas may deviate from the optimization goals.

Liu H et al.[37] propose a new loss function InterpIoU, which eliminates handmade geometric penalties and supports pure IoU-driven formulas. Since the entire loss comes from IoU calculations only, InterpIoU retains the ideal optimization characteristics of the original IoU loss and is fully in line with the basic goal of BBR - maximizing IoU.

The interpolation box B_{int} is a weighted average of the ground-truth box B_{gt} and the predicted box B_{pred}, controlled by the coefficient α. The interpolation box B_{int} is defined in Equation(6).

$$B_{int}=(1-\alpha)B_{pred}+\alpha B_{gt} \quad (6)$$

where $0<\alpha<1$. This equation linearly interpolates their coordinates. A larger α places the interpolation box closer to the ground-truth box, whereas a smaller α places it closer to the predicted box.

The loss function definition is shown in Equation(7).

$$L_{InterpIoU}(B_{pred},B_{gt})=L_{IoU}(B_{pred},B_{gt})+L_{IoU}(B_{int},B_{gt}) \quad (7)$$

The InterpIoU loss function is illustrated in Figure 14. The interpolation box B_{int} is positioned at a fractional distance of$(1-\alpha)D$ from the center of the ground-truth box, where D is the distance between the centers of the predicted and ground-truth boxes. This ensures a meaningful, nonzero IoU between the interpolation box and the ground-truth box, providing a stable regression target.

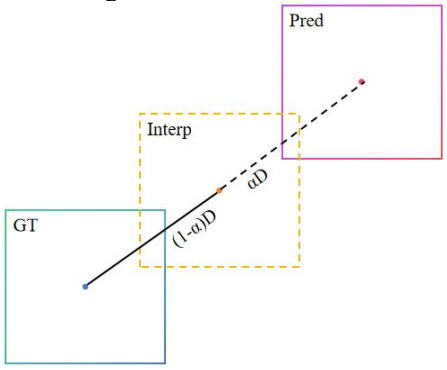


Fig 14. InterpIoU loss function

It can be seen from Equation(7) that the original InterpIoU ensures overlap with the real frame through a single-order interpolation box, which solves the problem of gradient disappearance in non-overlapping scenarios. However, when the prediction box is far away from the real box, the single-order interpolation box may not provide a sufficiently fine gradient signal, resulting in a slow or unstable optimization process. Therefore, consider adding a multi-order interpolation box on the basis of Equation(7), as shown in(8).

$$B_{int2}=(1-\beta)B_{pred}+\beta B_{gt} \quad (8)$$

where $\beta<\alpha$, which places B_{int2} closer to the ground-truth box B_{int} than B_{gt}. We name the resulting loss Multi-InterpIoU, as defined in Equation(9).

The Multi-InterpIoU loss function is illustrated in Figure 15. A new interpolation box, Interp2, is inserted between Interp and GT; its distance from the GT center is $\beta D(\beta<\alpha)$ and its distance from the Pred center is $(1-\beta)D$.

$$L_{Multi-InterpIoU}(B_{pred},B_{gt}) =L_{IoU}(B_{pred},B_{gt})+L_{IoU}(B_{int},B_{gt})+L_{IoU}(B_{int2},B_{gt}) \quad (9)$$

By constructing a multi-layer transition box between the prediction box and the real box, the Multi-InterpIoU forms a stepped gradient transfer path, which can more smoothly guide the prediction box to approach the real box, especially suitable for small targets, dense targets, and other scenarios that require fine adjustment. This improvement enhances gradient transfer through multi-layered interpolation, especially for scenarios where the prediction box is farther away from the real box. In addition, the multi-order interpolation box can further strengthen the consistency of the loss function with the IoU target. Since each interpolation box calculates the loss based on IoU, it avoids the interference of manual geometric penalties, and while improving the optimization efficiency, it can still maintain the direct optimization of IoU indicators, so that it performs more robustly in the dataset.

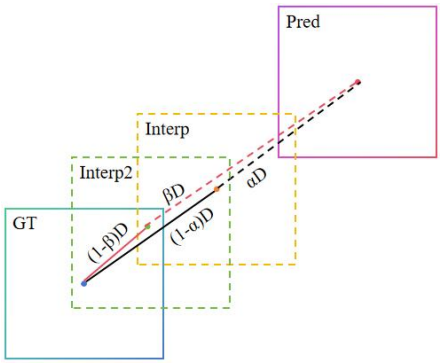


Fig 15. Multi-InterpIoU loss function

In summary, the experiment will optimize the loss function and replace the model loss function with Multi-InterpIoU.

3.6.Lesion-boundary refinement module

Medical images often show weak and blurred lesion boundaries. The newly designed Lesion-Boundary Refinement Module (LBRM) uses a lightweight gradient prior and channel gating to enhance boundary evidence while suppressing smooth background responses. The gradient gate is defined in Equation(10), and the refined feature is expressed in Equation(11).

$$G=\sigma(W_g[X;|\nabla X|]) \quad (10)$$

$$F_{LBR}=X+G\odot W_f[X;|\nabla X|] \quad (11)$$

where X is the input feature map, ∇X denotes the image-gradient prior, W_g and W_f are lightweight convolution layers, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication.

Figure 16 presents the internal structure of LBRM. The input feature is first converted into a gradient prior, which emphasizes weak and blurred boundary regions. The original features and the prior are concatenated and passed through a lightweight convolution branch. A sigmoid gate is then learned to suppress smooth background responses while retaining boundary evidence. Finally, the gated response is added back to the input through a residual connection, improving boundary localization without substantially increasing model complexity. This design is especially useful for MRI lesions with fuzzy edges and low contrast.

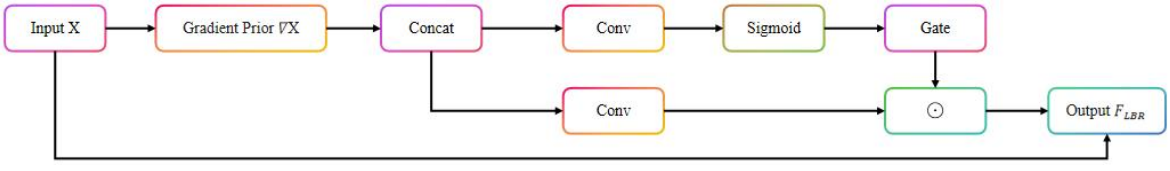


Fig 16. Lesion-boundary refinement module

3.7.Adaptive cross-scale fusion module

The newly designed Adaptive Cross-Scale Fusion Module (ACFM) replaces rigid multi-scale concatenation with a learnable normalized gate. It adaptively weights different pyramid levels and reduces interference between lesion scales. The scale weights are defined in Equation(12), and the fused output is expressed in Equation(13).

$$A_i=\text{softmax}(W_a\delta(W_b\text{GAP}(F_i))) \quad (12)$$

$$F_{\text{out}}=F_{\text{ACF}}+\lambda F_{\text{LBR}} \quad (13)$$

where F_i is a feature level, GAP is global average pooling, δ is ReLU, W_a and W_b are lightweight convolution layers, A_i is the normalized weight of each scale, F_{ACF} is the adaptively fused feature, F_{LBR} is the LBRM output, and λ is a learnable balance factor.

Figure 17 illustrates the workflow of ACFM. Each pyramid level is first compressed by global average pooling, and two lightweight convolution layers with ReLU and softmax generate scale-wise attention weights. These weights are used to recalibrate different pyramid features, reducing interference between lesion scales. The fused feature is then combined with the LBRM output through a learnable balance factor, enabling the detector to trade off semantic abstraction and lesion-boundary detail adaptively. This is important because lesions may appear at different scales in clinical images.

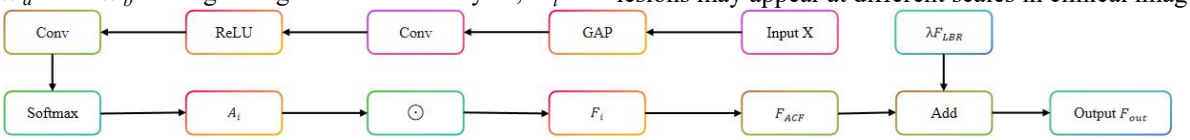


Fig 17. Adaptive cross-scale fusion module

3.8.BCL-YOLO algorithm

Based on the YOLOv8 algorithm, the BC2fB, C2fIRB, LDSPPF, ACFM, and LBRM modules are fused, and the loss function is replaced with Multi-InterpIoU. The improved architecture is shown in Figure 18, where the highlighted boxes mark the improved components used in BCL-YOLO.

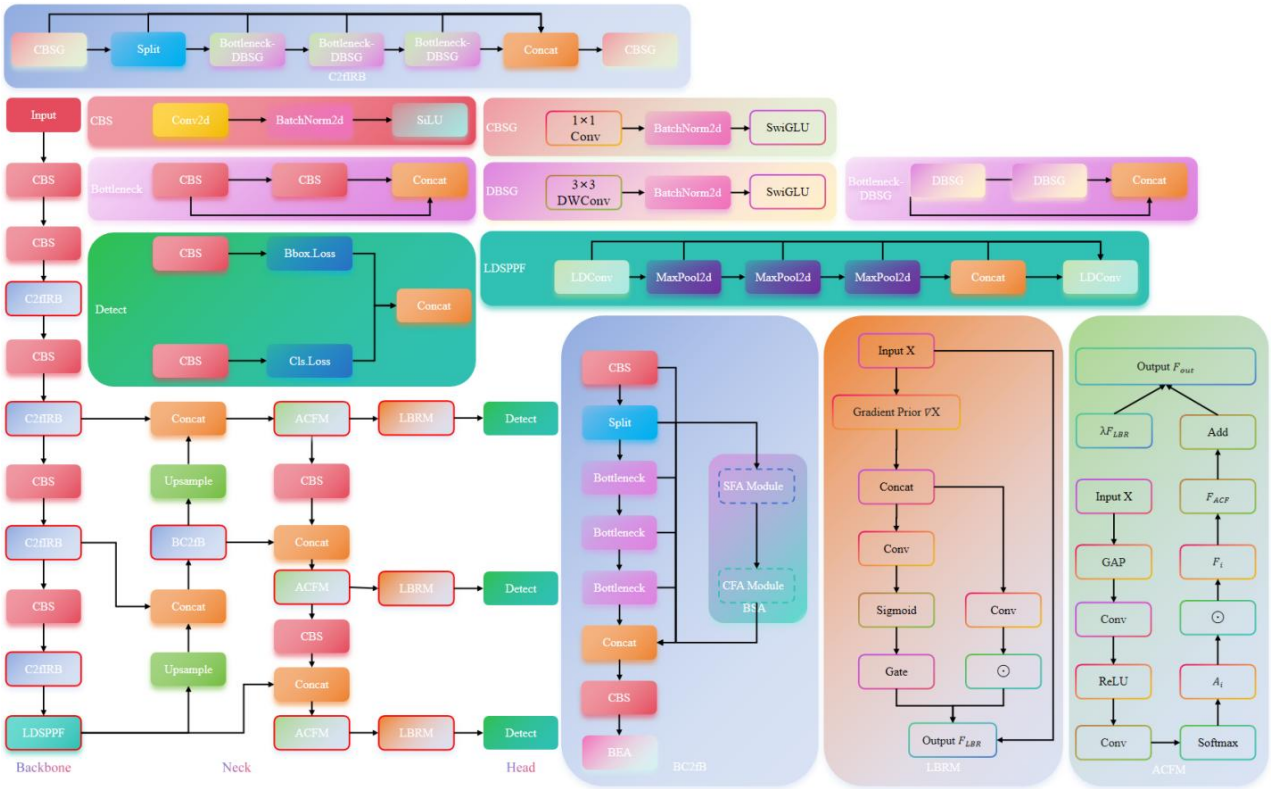


Fig 18. BCL-YOLO algorithm with ACFM and LBRM

Neck network: The backbone features are passed to the detection head through upsampling and splicing. BC2fB first strengthens attention to clinically relevant regions. ACFM then adaptively fuses the multi-scale features, and LBRM refines weak lesion-boundary evidence. This sequence suppresses background interference, improves boundary localization, and provides more reliable multi-scale features for the detection head.

4.Experiments and results analysis

4.1.Experimental environment and datasets

4.1.1.Hardware and software environmen

By rationally configuring the hardware environment and software environment, it provides a stable and efficient operating basis for the training and testing of the improved YOLOv8 medical image detection model to ensure the smooth operation of the experiment. The specific hardware and software environments are shown in Table 1.

Table 1. Environment configuration

Hardware environment	Software environment
NVIDIA RTX 3090 GPU	Python 3.8
Intel Core i9-12900K CPU	PyTorch 1.12.1
Windows 10 Professional	PyCharm Community Edition 2022.1.4

4.1.2.Br35H dataset

As an important public dataset in the field of neuroimaging and computer vision, Br35H[38] is specially created for relevant researchers[39]. The dataset contains 801 brain tumor MRI images, each with detailed and comprehensive labeling information, and is reasonably divided into three separate parts: training set, validation set, and test set, according to the study needs.

The training set contains 500 images, the validation set contains 201 images, and the test set contains 100 images. The test set is used for final performance evaluation. Representative images from this dataset are shown in Figure 19.

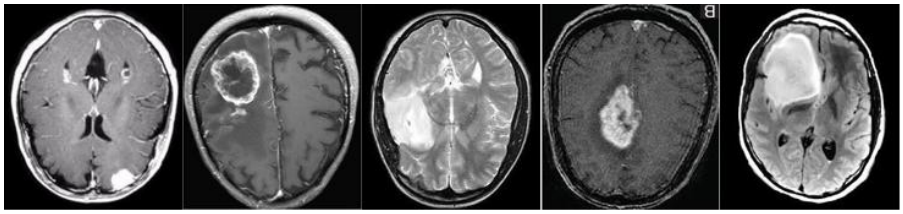


Fig 19. Dataset partial image

4.2.Model training and testing

4.2.1.Parameter setting and optimization strategies

Images were resized to 640×640 for both training and testing. Training used 100 epochs, batch size 32, SGD momentum 0.9, weight decay 0.0005, and initial learning rate 0.001. A three-epoch warmup phase preceded linear learning-rate decay to 1% of the initial value; no cosine schedule was used. Data augmentation used HSV perturbation (h=0.015, s=0.7, v=0.4), random translation 0.1, random scale 0.5, horizontal flip probability 0.5, and mosaic augmentation; mosaic was disabled during the final ten epochs. Rotation, shear, perspective distortion, vertical flipping, MixUp, and copy-paste augmentation were not used. All parameter settings are shown in Table 2.

Table 2. Experimental settings

Parameter	Value
Initial learning rate	0.001
Training epochs	100
Batch size	32
Weight decay	0.0005
Momentum	0.9
Optimizer	SGD
Input resolution	640×640
Data augmentation	HSV(0.015/0.7/0.4); translation 0.1; scale 0.5; horizontal flip 0.5; mosaic 1.0; disabled in last 10 epochs
Learning-rate schedule	3 warmup epochs; linear decay to 1% of initial LR

4.2.2.Test indicators and evaluation methods

After the model training, it is necessary to carry out a comprehensive and accurate evaluation of its performance to judge the actual performance of the model in the task of medical image detection. This experiment uses a variety of routine test indicators and scientific evaluation methods to ensure the reliability and effectiveness of the evaluation results.

(1). Accuracy

The accuracy rate is used to measure the prediction accuracy of the model as a whole, and the formula is shown in Equation(10)(14).

$$\text{Accuracy}=\frac{\text{Correct predictions}}{\text{Total samples}}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{14}$$

wherein:

TP(true positive):the number of samples from a given class that are correctly predicted as that class.

TN(true negative):the number of samples not belonging to a given class that are correctly predicted as not belonging to that class.

FP(false positive):the number of samples not belonging to a given class that are incorrectly predicted as that class.

FN(false negative):the number of samples from a given class that are incorrectly predicted as not belonging to that class.

(2). Recall

Recall measures the model's ability to identify samples from a given class(coverage), as defined in Equation(11)(15).

$$\text{Recall}=\frac{\text{Correctly identified samples}}{\text{Class samples}}=\frac{\text{TP}}{\text{TP+FN}} \tag{15}$$

(3). F1-score

The F1-score is the harmonic mean of precision and recall and provides a combined assessment of model performance.

Precision is calculated first, followed by the F1-score, as defined in Equations(12)(16) and(13), respectively.

$$\text{Precision}=\frac{\text{Correctly identified samples}}{\text{Predicted-as-class samples}}=\frac{\text{TP}}{\text{TP+FP}} \tag{16}$$

$$\text{F1}-\text{score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{17}$$

(4). Average Precision(AP)

AP measures the recognition accuracy for a single class and is the area under that class's precision-recall curve, as defined in Equation(14)(18).

$$\text{AP}=\int_0^1 \text{P(r)dr} \tag{18}$$

where P(r)denotes the precision at recall r.

(5). Mean Average Precision(mAP)

mAP measures overall performance across all classes and is the mean of the AP values for all classes, as defined in Equation(15)(19).

$$\text{mAP}=\frac{1}{\text{C}}\sum_{\text{i=1}}^{\text{C}}\text{AP}_\text{i} \tag{19}$$

where C is the number of classes and AP_i is the average precision for class i. We report mAP at two IoU thresholds: mAP@0.5 is the mean average precision at an IoU threshold of 0.5 and measures detection performance under a relatively

lenient localization criterion; mAP@0.5:0.95 is averaged over IoU thresholds from 0.5 to 0.95 at 0.05 intervals and better reflects bounding-box regression accuracy under strict localization criteria.

Together, these metrics and evaluation procedures provide a comprehensive and objective assessment of BCL-YOLO for medical image detection and support subsequent model optimization and application.

4.3.Experimental results and comparative analysis

4.3.1.Ablation study

To quantify the contribution of each modification, we conducted an ablation study by progressively removing the proposed components. All experiments were performed on the Br35H dataset. We report mAP@0.5,mAP@0.5:0.95, precision, recall, and F1-score, together with the number of parameters and GFLOPs to assess the efficiency impact of each component.

Table 3. Ablation study results

BC2fB	C2fIRB	LDSPPF	Multi-InterpIoU	mAP@0.5(%)	mAP@0.5:0.95(%)	Precision(%)	Recall(%)	F1	Params(M)	GFLOPs
×	×	×	×	93.8	71.2	94.5	89.3	0.918	3.16	8.9
√	×	×	×	96.2	73.5	96.8	91.5	0.941	3.42	9.3
×	√	×	×	94.8	72.1	95.2	90.1	0.926	2.89	8.5
×	×	√	×	95.1	72.8	95.6	90.8	0.931	3.25	9.1
×	×	×	√	95.1	73.2	95.4	90.5	0.929	3.16	8.9
√	√	√	√	97.2	74.6	97.5	92.8	0.951	2.88	8.6

Experiment 1 uses the unmodified YOLOv8 baseline, which achieves 93.8% mAP@0.5 and 71.2% mAP@0.5:0.95. With BC2fB alone(Experiment 2), mAP@0.5 increases to 96.2%, an improvement of 2.4 percentage points; mAP@0.5:0.95 increases by 2.3 points and recall increases by 2.2 points. With C2fIRB alone(Experiment 3), mAP@0.5 reaches 94.8%, while the parameter count decreases by 8.5% from 3.16 M to 2.89 M.LDSPPF alone(Experiment 4) yields 95.1% mAP@0.5, a 1.3-point gain. Replacing the loss with Multi-InterpIoU alone(Experiment 5) also yields 95.1% mAP@0.5 and improves mAP@0.5:0.95 by 2.0 points. When all three modules and Multi-InterpIoU are combined (Experiment 6), BCL-YOLO achieves the best performance: 97.2% mAP@0.5 and 74.6% mAP@0.5:0.95, gains of 3.4 points over the baseline for both metrics.

Overall, every proposed component improves detection performance when used alone, with BC2fB making the largest contribution to recall and reducing missed detections. The combined optimization maximizes performance. Notably, C2fIRB reduces the final parameter count to 2.88 M, 8.9% below the 3.16 M baseline, while the full model improves mAP@0.5 by 3.4 points, demonstrating an effective balance between accuracy and efficiency.

The ablation results show that the modules are complementary. BC2fB improves feature extraction and attention to clinically relevant regions, especially through boundary-aware attention; C2fIRB improves computational efficiency while maintaining accuracy; and LDSPPF improves the handling of multi-scale medical-image features. Their combination strengthens BCL-YOLO across the medical image detection task.

4.3.2.Comparative study

To compare the strengths and limitations of BCL-YOLO, we evaluated it against SSD, Faster R-CNN, RT-DETR, YOLOv9, and YOLOv10. All models were trained in the same experimental environment. The comparison includes mAP@0.5, mAP@0.5:0.95, precision, recall, parameters, GFLOPs, and inference speed(FPS), as summarized in Table 4.

Table 4. Comparative results on the Br35H dataset

Method	mAP@0.5(%)	mAP@0.5:0.95(%)	Precision(%)	Recall(%)	Params(M)	GFLOPs	FPS
Faster R-CNN	88.7	64.3	90.2	84.5	41.52	134.8	28
SSD	91.7	67.8	93.1	86.8	23.74	84.6	45
RT-DETR	91.1	66.5	92.5	86.2	20.18	60.2	72
YOLOv8n	93.8	71.2	94.5	89.3	3.16	8.9	156
YOLOv9	94.5	72.1	95.3	90.1	3.24	9.2	148
YOLOv10	94.2	71.8	94.8	89.8	2.76	8.4	162
BCL-YOLO	97.2	74.6	97.5	92.8	2.88	8.6	154

BCL-YOLO shows clear advantages on the medical image detection task. Its mAP@0.5 reaches 97.2%, improving on YOLOv8n by 3.4 percentage points, YOLOv9 by 2.7 points, and YOLOv10 by 3.0 points. Its mAP@0.5:0.95 reaches 74.6%, a 3.4-point improvement over YOLOv8n. BCL-YOLO uses 2.88 M parameters, 8.9% fewer than YOLOv8n(3.16 M), requires 8.6 GFLOPs versus 8.9 GFLOPs, and runs at 154 FPS, which is close to the 156 FPS baseline. These results indicate that the proposed combination improves accuracy without a substantial computational cost. BCL-YOLO also outperforms Faster R-CNN and SSD in both accuracy and speed.

4.3.3.Visualization of experimental results

Figure 20 shows detection results for brain-tumor MRI images in the Br35H dataset. The model accurately identifies well-defined solid tumors, irregular infiltrative tumors, and smaller tumors, with predicted boxes closely matching the annotated lesion extents.

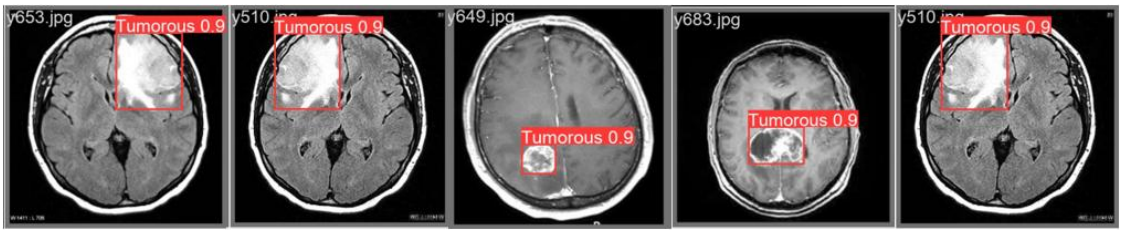


Fig 20. Model detection results

Further analysis indicates that the model distinguishes tumor margins from normal brain tissue effectively, reducing missed detections and misclassification caused by blurred boundaries. This behavior is supported by the bipolar attention integrated in BC2fB, which focuses on tumor cores and peripheral infiltration zones while suppressing background interference such as cerebrospinal fluid and vascular shadows. The model remains stable on MRI images with artifacts or sequence variation, reflecting the complementary benefits of C2fIRB feature extraction and LDSPPF multi-scale fusion.

4.3.4.Generalization experiment

To further evaluate the robustness of BCL-YOLO, we conducted experiments on another public medical image dataset.

ChestX-Det10[40] is an instance-level chest X-ray dataset for disease classification and localization. It contains ten common disease or abnormality categories: atelectasis, calcification, consolidation, effusion, emphysema, fibrosis, fracture, mass, nodule, and pneumothorax. Three certified radiologists annotated the images using bounding boxes: two radiologists annotated independently, and a third, more experienced radiologist adjudicated between their annotations or added new annotations. ChestX-Det10 is a subset of NIH ChestX-14 and contains 3,543 images, randomly divided into 3,001 training images and 542 test images[41]. The data include samples with varied viewpoints, brightness, and acquisition conditions.

Representative images from the dataset are shown in Figure 21.

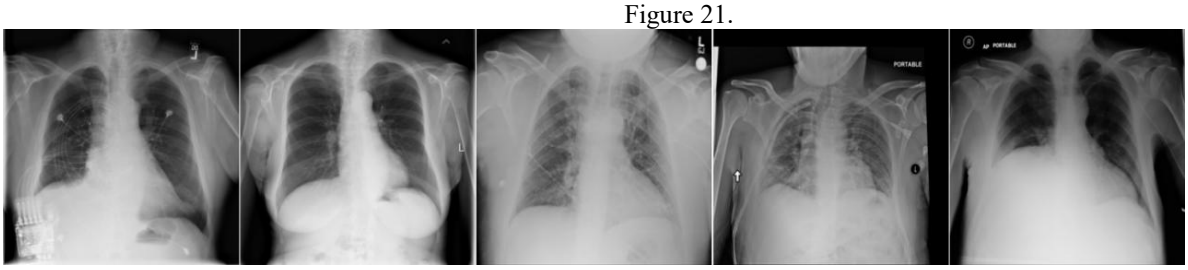


Fig 21. Representative images from the dataset

To further compare BCL-YOLO with alternative detectors on medical images, we evaluated BCL-YOLO, SSD, Faster R-CNN,RT-DETR, YOLOv9, and YOLOv10. All results are reported as mAP@0.5, as shown in Table 5.

Table 5. Comparative results on the ChestX-Det10 dataset(mAP@0.5)

Method	mAP@0.5(%)	Params(M)	GFLOPs	FPS
Faster R-CNN	28.3	41.52	134.8	28
SSD	35.6	23.74	84.6	45
RT-DETR	38.2	20.18	60.2	72
YOLOv8n	48.5	3.16	8.9	156
YOLOv9	50.1	3.24	9.2	148
YOLOv10	49.3	2.76	8.4	162
BCL-YOLO	53.8	2.88	8.6	154

ChestX-Det10 contains ten imbalanced thoracic-abnormality classes, with very few examples in some categories; therefore, its overall mAP@0.5 is lower than on Br35H. The reported values are consistent with public YOLOv8 benchmarks of approximately 48-56% mAP@0.5. BCL-YOLO obtains 53.8% mAP@0.5, improving on the YOLOv8n baseline by 5.3 percentage points and supporting cross-modality generalization.

4.3.5.Statistical analysis of object sizes

To objectively assess lesion-size distribution in Br35H, we counted all 690 annotated boxes. Under the COCO area convention, objects were categorized as small(area < 32×32 = 1,024 px²), medium(32×32 to 96×96 px²), and large(>96×96 px²). The results are given in Table 6.

Table 6. Object-size distribution in the Br35H dataset

Object-size category	Count	Proportion	Description
Small(<32×32 px)	3	0.4%	Almost no COCO-defined small objects
Medium(32×32 to 96×96 px)	42	6.1%	A small proportion of lesions are medium-sized
Large(>96×96 px)	645	93.5%	Most lesions are large
Total	690	100%	-

The results show that 93.5% of annotated boxes in Br35H are large, whereas only 0.4% meet the COCO definition of a small object. Accordingly, this work is framed as medical-image lesion detection rather than small-object detection. Although the boundary-aware attention in BC2fB remains useful for large lesions by strengthening boundary features and localization, the main contribution is the combined adaptation to blurred lesion boundaries and low contrast in medical images.

4.3.6.Repeated experiments and statistical reliability

Because Br35H is limited to 801 images, we performed five independent runs with random seeds 0, 42, 123, 456, and 789 under the same experimental configuration. We report the mean and sample standard deviation for each metric in Table 7.

Table 7. Results of five independent runs on the Br35H dataset

Metric	R1(seed=0)	R2(seed=42)	R3(seed=123)	R4(seed=456)	R5(seed=789)	Mean ± SD
mAP@0.5(%)	97.2	96.8	97.4	96.5	97.1	97.0 ± 0.35
mAP@0.5:0.95(%)	74.6	73.9	74.8	73.5	74.3	74.2 ± 0.52
Precision(%)	97.5	97.1	97.6	96.8	97.3	97.3 ± 0.31
Recall(%)	92.8	92.3	93.1	91.9	92.6	92.5 ± 0.45

Across the five independent runs, the sample standard deviation of mAP@0.5 is 0.35 percentage points. The sample standard deviations of all reported metrics are below 0.6 points, indicating stable results across the tested random seeds. These results support repeatability, although the limited dataset size remains a study limitation.

4.3.7.Failure case analysis

To present the model limitations objectively, we analyzed BCL-YOLO failure cases in the validation set. The failures fall into three main categories:

(1)False negatives: the model can miss lesions that are small or located deep in brain tissue, especially when the intensity difference between the tumor boundary and normal tissue is extremely small. Such weak boundaries are difficult for boundary-aware attention to distinguish reliably.

(2)False positives: artifacts, including metal artifacts and motion blur in MRI images, can be incorrectly identified as tumors.

(3)Localization errors: for highly irregular tumors, the IoU between the predicted and annotated boxes may fall between 0.3 and 0.5. Multi-InterpIoU offers partial improvement in such cases, but further refinement is needed.

These failure cases suggest that improved boundary-feature extraction, artifact-suppression preprocessing, and more refined bounding-box regression could further improve model performance.

4.3.8.Robustness test

To evaluate robustness under image degradation, we applied Gaussian noise, brightness changes, contrast changes, Gaussian blur, and JPEG compression to the Br35H test set and assessed detection performance at different degradation levels. The results are shown in Table 8.

Table 8. Robustness test results for BCL-YOLO under different image degradations(mAP@0.5)

Perturbation	Parameter	mAP@0.5(%)	Relative decrease
None(baseline)	-	97.2	-
Gaussian noise	σ=0.01	95.8	-1.4
Gaussian noise	σ=0.05	91.3	-5.9
Gaussian noise	σ=0.10	85.6	-11.6
Brightness change	+20%	96.5	-0.7
Brightness change	-20%	95.2	-2.0
Brightness change	-40%	88.7	-8.5
Contrast change	+20%	96.8	-0.4
Contrast change	-20%	94.1	-3.1
Gaussian blur	kernel=3	96.1	-1.1
Gaussian blur	kernel=5	93.5	-3.7
Gaussian blur	kernel=7	89.2	-8.0
JPEG compression	quality=50	96.3	-0.9
JPEG compression	quality=30	94.7	-2.5
JPEG compression	quality=10	90.8	-6.4

BCL-YOLO is robust to mild perturbations(noise σ=0.01, brightness ±20%, and Gaussian blur with kernel=3), with mAP@0.5 decreases below 2 percentage points. Under severe degradation(noise σ=0.10, brightness −40%, and Gaussian blur with kernel=7), performance decreases by 8–12 percentage points. The large brightness decrease(−40%)has the greatest impact, consistent with the limitation discussed in the Conclusions that performance declines under extreme illumination changes.

5.Conclusions and future work

This study investigates the use of YOLOv8 for medical image detection. By adapting core modules to the characteristics of medical images, BCL-YOLO improves detection capability in complex scenarios while retaining efficient inference.

The contribution is a task-driven engineering integration for medical image detection rather than a claim of new primitive algorithms. BSA, BEA, IRB, LDConv, and InterpIoU were selected because they target complementary bottlenecks: lesion-background contrast, efficient contextual reweighting, parameter cost, variable lesion morphology, and gradient quality in box regression. BC2fB combines BSA and BEA serially to strengthen clinically relevant regions without duplicating the same attention computation; C2fIRB adapts the inverted residual design to retain representational capacity with fewer parameters; LDSPPF uses LDConv to improve multi-scale fusion for irregular lesions; and Multi-InterpIoU improves localization at the detection head. The ablation and comparative results validate both the need for each adaptation and the advantage of their coordinated use in brain-tumor MRI detection.

On Br35H, BCL-YOLO reaches 97.2% mAP@0.5 and 74.6% mAP@0.5:0.95, gains of 3.4 percentage points over the YOLOv8n baseline for both metrics. The model uses 8.9% fewer parameters and runs at 154 FPS. Across five independent runs, every reported sample standard deviation is below 0.6 points. BCL-YOLO also outperforms the compared methods in the ChestX-Det10 cross-modality experiment and remains stable under mild image degradation.

This study has limitations. Performance declines under severe brightness reduction, where mAP@0.5 decreases by 8.5 points at −40% brightness. The failure-case analysis also shows that weak tumor boundaries and artifacts can cause missed detections or false positives. Future work could investigate stronger boundary representations, illumination-compensation methods, artifact suppression, more efficient lesion detectors, Transformer-based architectures, and multimodal data fusion.

Author contributions Bohan Zhang: data collection, figure preparation, writing — original draft, result interpretation, manuscript revision. Tingting Yang: conceptualization, methodology, model implementation, experimental

design and statistical analysis. All authors have read and approved the final version of the manuscript.

Funding No funding was received for this work.

Data availability The datasets used in this paper are publicly available open-source datasets. Br35H and ChestX-Det10 datasets can be obtained from Kaggle. All relevant references to these datasets are provided in the manuscript. No new datasets were generated in this work.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] GUO C, HE Y, LIU H. Research progress on applications of artificial intelligence in musculoskeletal medical imaging. Chinese Journal of Medical Ultrasound (Electronic Edition), 2025, 22(3): 266.

[2] CHEN H, CHEN S. Application and prospects of keypoint detection technologies in medical imaging. Software Engineering and Applications, 2025, 14: 38.

[3] WANG N, LIN S, QI F, et al. Anomaly detection in medical images based on self-supervised learning. Journal of Computer-Aided Design and Computer Graphics, 2025, 37(3): 474-483.

[4] ZHANG S, ZHAO T, QI F, et al. An intelligent biomedical image classification algorithm based on wavelet transform. Advances in Applied Mathematics, 2025, 14: 16.

[5] HUANG S, ZHANG S. A medical image segmentation method using hierarchical fusion and multi-scale attention. Modeling and Simulation, 2025, 14: 887.

[6] WU X, LIU L, CHEN J, et al. Application of an artificial-intelligence-assisted cataract diagnosis system in community screening. Chinese Journal of Ophthalmology, 2021.

[7] LIN C, CHEN M. A diversified evaluation of classroom teaching effectiveness based on computer vision. Intelligent Computer and Applications, 2025, 15(4): 125-131.

[8] HUANG Y, CHEN H, XU S. Causes and detection methods of appearance defects in fair-faced concrete. Guangdong Building Materials, 2025, 41(7): 96-100.

[9] WANG S, ZHANG H, LIU Q, et al. The role of a pinhole mirror in predicting visual acuity before cataract surgery. Ophthalmology, 2025, 14(1): 43-47.

[10] GUO S, ZHANG W, DONG Y. Intelligent detection of pulmonary medical images based on DenseNet. Journal of Harbin University of Commerce (Natural Sciences), 2023, 39(2): 164-168.

[11] HUANG D. Research on keypoint detection in orthopedic images based on self-supervised learning. Changchun: Jilin University, 2024.

[12] REIS D, KUPEC J, HONG J, et al. Real-time flying object detection with YOLOv8[Z/OL]. arXiv:2305.09972, 2023.

[13] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[Z/OL]. arXiv:2005.12872, 2020.

[14] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector//European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[15] CHEN X, GUPTA A. An implementation of faster RCNN with study for region sampling[Z/OL]. arXiv:1702.02138, 2017.

[16] KUMAR A, DHANALAKSHMI R. EYE-YOLO: a multi-spatial pyramid pooling and Focal-EIOU loss inspired tiny YOLOv7 for fundus eye disease detection. International Journal of Intelligent Computing and Cybernetics, 2024, 17(3): 503-522.

[17] LOKESH L P, POOLA R G, GORREPATI L P, et al. Real-Time Cataract Diagnosis with GhostYOLO: A GhostConv-enhanced YOLO Model. Engineering, Technology & Applied Science Research, 2025, 15(3): 22945-22952.

[18] WANG Y, YANG Y. A YOLOv8-based detection algorithm for hand-bone fracture images. Artificial Intelligence and Robotics Research, 2025, 14: 621.

[19] BAO Q, TANG S, LI Q, et al. Improved YOLOv8 for multi-type pulmonary nodule detection. Journal of Frontiers of Computer Science and Technology, 2025, 19(2): 429-442.

[20] ANBALAGAN K, RADHIKA K, RAMACHANDRAN V M, et al. Artificial Intelligence in Healthcare: YOLO-v7 and GRU Integration for Automated Glaucoma Detection//2025 International Conference on Computational Robotics, Testing and Engineering Evaluation. Piscataway, NJ: IEEE, 2025: 1-6.

[21] WANG G, CHEN Y, AN P, et al. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. Sensors, 2023, 23: 7190.

[22] ZHANG H, LIANG M, WANG Y. YOLO-BS: a traffic sign detection algorithm based on YOLOv8. Scientific Reports, 2025, 15(1): 7558.

[23] RASHEED A F, ZARKOOSH M. Optimized YOLOv8 for multi-scale object detection. Journal of Real-Time Image Processing, 2025, 22(1): 6.

[24] LOU H, DUAN X, GUO J, et al. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. Electronics, 2023, 12: 2323.

[25] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[26] WANG Y, XU D. A hyperspectral classification network integrating spectral-fusion GAT and bipolar self-attention. Optical Technique, 2025, 51(3): 367-376.

[27] WANG K, HE G, LI X. HSF-DETR: A Special Vehicle Detection Algorithm Based on Hypergraph Spatial Features and Bipolar Attention. Sensors, 2025, 25: 4381.

[28] WANG Y, ZUO Y, WU L, et al. Image super-resolution reconstruction based on an attention mechanism. Journal of Computer Applications, 2021, 41(3): 845-850.

[29] HAN Y. Research on object detection algorithms based on attention mechanisms and inverted residual blocks. Jinan: Shandong Normal University, 2024.

[30] ZHANG X, SONG Y, SONG T, et al. LDConv: linear deformable convolution for improving convolutional neural networks. Image and Vision Computing, 2024, 149: 105190.

[31] DENG L, SHA W. SCA-YOLO: Real-time object detection based on deformable convolution and context-aware attention. Journal of Guangdong University of Technology, 2025, 42(0).

[32] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 658-666.

[33] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2020: 12993-13000.

[34] ZHENG Z H, WANG P, REN D W, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Transactions on Cybernetics, 2021, 52(8): 8574-8586.

[35] GEVORGYAN Z. SIoU loss: More powerful learning for bounding box regression. 2022.

[36] LIU C, WANG K, LI Q, et al. Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism. Neural Networks, 2024, 170: 276-284.

[37] LIU H, WATANABE H. InterpIoU: Rethinking Bounding Box Regression with Interpolation-Based IoU Optimization[Z/OL]. arXiv:2507.12420, 2025.

[38] HAMADA A. Br35h: Brain tumor detection 2020[Z/OL]. Kaggle, 2021. https://www.kaggle.com/datasets/ahmedhamada0/braintumor-detection.

[39] YU Y. Research on medical imaging based on deep learning. Chengdu: University of Electronic Science and Technology of China, 2024.

[40] LIU J, LIAN J, YU Y. ChestX-det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities[Z/OL]. arXiv:2006.10550, 2020.

[41] LI C, XU D, JIN S, et al. Chest X-ray emphysema detection based on channel attention and dilated convolution. Journal of Lanzhou University of Technology, 2022, 48(2): 81-89.