RESEARCH ARTICLE | MARCH 04 2026

An efficient YOLOv12n model is used for multi-scale brain tumor detection

Lan Liu ; Ronghuan Wu ; Chengyang Zhang; Jianhui Xu

AIP Advances 16, 035309 (2026)
https://doi.org/10.1063/5.0321006

An efficient YOLOv12n model is used for multi-scale brain tumor detection

Cite as: AIP Advances 16, 035309 (2026); doi: 10.1063/5.0321006
Submitted: 5 January 2026 • Accepted: 8 February 2026 •
Published Online: 4 March 2026

Lan Liu,[1] Ronghuan Wu,[2] Chengyang Zhang,[3] and Jianhui Xu[4,a)]

AFFILIATIONS

[1] Department of Basic Medicine, Changde Vocational Technical College, Hunan, Changde, 415000, China
[2] China Mobile Guizhou Co., Ltd. Guiyang Branch, Guiyang 550001, China
[3] Amazon, 410 Terry Avenue North, Seattle, Washington 98109, USA
[4] School of Medical Information Engineering, Shenyang Medical College, Shenyang, Liaoning 110043, China

[a)] **Author to whom correspondence should be addressed:** 329954097@qq.com

ABSTRACT

Accurate detection of brain tumors is critical for early diagnosis and treatment planning, yet it remains challenging due to the irregular shapes, blurred boundaries, and low contrast of lesions, particularly in Gliomas. To address the limitations of existing methods in balancing accuracy with computational efficiency, this paper proposes a lightweight brain tumor detection framework based on an improved YOLOv12n. We introduce three core innovations to enhance feature representation: (1) the A2C2f-CGLU-DYT module, which integrates dynamic activation functions to strengthen local nonlinear modeling and adaptability to input variations; (2) the C2TSSA module, which utilizes token statistics self-attention to efficiently capture global long-range dependencies without the heavy computational cost of traditional transformers; and (3) the CGAFusion module, which employs content-guided attention to effectively fuse shallow geometric details with deep semantic features. Experimental results on the Brain Tumor dataset demonstrate that the proposed method achieves a precision of 92.7%, a recall of 87.4%, and an mAP@0.5 of 93.8%, outperforming the baseline YOLOv12n by 0.7%, 0.9%, and 1.2%, respectively. Notably, in the challenging Glioma category, the model attains an accuracy of 84.6%, significantly surpassing YOLOv10n by a margin of 8.6%. Furthermore, validation on the Blood Cell Count dataset yields an mAP@0.5 of 94.1%, confirming the model's robust cross-dataset generalization ability.

I. INTRODUCTION

Brain tumors are a serious threat to human life and health, being a type of central nervous system disease with high disability and mortality rates.[1] Clinical studies have shown that early diagnosis and timely treatment are crucial for improving patient survival rates and prognosis.[2] Medical imaging technology, particularly Magnetic Resonance Imaging (MRI), with its advantages of non-invasiveness, high resolution, and multi-parameter imaging,[3] has become a core tool for clinical diagnosis, classification, and treatment planning for brain tumors. However, traditional manual image reading methods not only rely on the professional experience of doctors but also suffer from problems such as high subjectivity, poor reproducibility, and heavy workload, facing challenges in efficiency and accuracy when processing large volumes of medical images.

In recent years, deep learning (Deep Learning) techniques, especially Convolutional Neural Networks (CNNs),[4–6] have made significant breakthroughs in the fields of computer vision and medical image analysis. With their powerful feature extraction and representation capabilities, deep learning has shown outstanding performance in tasks such as medical image segmentation, classification, and object detection. U-Net[7] and its variants are widely used in medical image segmentation tasks, effectively retaining spatial detail information by relying on an encoder–decoder structure and skip connection mechanisms. However, in object detection, additional framework support is still required. Faster R-CNN,[8] as a typical two-stage detector, achieves high detection accuracy by utilizing the Region Proposal Network (RPN), but its inference speed is relatively slow, making it unsuitable for real-time diagnostic scenarios. In contrast, the YOLO[9–19] series, as a single-stage detector, combines the advantages of speed and accuracy, making it outstanding in medical image object detection, and thus has become a research hotspot in recent years.

Despite the numerous improvements made to the YOLO series, many shortcomings still exist in practical applications. For example, Kaya *et al.*[20] proposed a deep transfer learning method based on multi-model fusion (Fusion-Brain-Net), which combines VGG16, ResNet50, and MobileNetV2 for feature extraction and fusion. By leveraging fine-tuning and data augmentation techniques, the method significantly improved the accuracy of brain tumor classification; however, the multi-model fusion also resulted in higher computational costs and longer training times. Kang *et al.*[21] introduced an improved model named BGF-YOLO, which effectively enhances the performance of YOLOv8 by integrating a bidirectional routing attention mechanism (BRA), a generalized feature pyramid network (GFPN), and a fourth detection head. However, the model's high computational complexity increases resource consumption during training and inference. Bai *et al.*[22] proposed an improved model named SCC-YOLO, which integrates the SCConv (Spatial and Channel Reconstruction Convolution) module into YOLOv9 to reduce feature redundancy and improve brain tumor detection accuracy. However, the model has not been fully validated for generalization across other medical image datasets. Rahimi *et al.*[23] proposed a brain tumor detection method based on YOLOv4-tiny using Gaussian blur and Sobel filters for MRI image preprocessing to enhance features and fine-tune the model with transfer learning techniques. However, when detecting tumors with complex shapes, the inclusion of background areas in the bounding box affects detection accuracy. Sahoo *et al.*[24] proposed a hybrid model based on a GAN ensemble and modulated CNN for brain tumor detection and optimized classification performance through data augmentation and a soft voting method. However, due to the limited size of the dataset, the model's generalization ability remains insufficient. Mithun and Joseph Jawhar[25] proposed a method combining the YOLO NAS model, En-DeNet segmentation network, and mixed anisotropic diffusion filtering (MADF) for brain tumor detection and classification, achieving an accuracy of up to 99.7%. However, the model relies heavily on training data; its generalization ability has not been fully validated, and it has high computational costs. Chen *et al.*[26] proposed a model called YOLO-NeuroBoost, which significantly improves brain tumor detection accuracy by combining the dynamic convolution module KernelWarehouse, the attention mechanism CBAM, and the improved Inner-GIoU loss function, achieving a mAP50 of 99.48%. However, the model shows limited performance when handling low-contrast or high-noise images, and its generalization ability is affected by insufficient dataset diversity. Hariharan and Surya[27] proposed a brain tumor detection and classification method combining the YOLOv11 algorithm with CNN models using image preprocessing (such as resizing and Gaussian blur for denoising) and data augmentation techniques to achieve high detection accuracy. However, the model still suffers from misclassification in certain categories (such as background and non-tumor) and is limited by the use of a single-center dataset, which restricts its generalization ability.

In summary, existing methods still struggle to achieve an ideal balance between detection accuracy, generalization ability, and computational efficiency and face multiple challenges. On the one hand, brain tumors exhibit high diversity in terms of shape, size, and location, with boundaries often presenting blurred characteristics. On the other hand, MRI data suffers from issues such as uneven sample distribution, cross-device imaging differences, and signal variations between different patients, which limit the model's generalization ability and robustness. To address these issues, this paper proposes a brain tumor detection method based on an improved deep learning framework. Through multi-scale feature fusion and efficient attention mechanisms, the method effectively enhances the model's performance in detecting tumors of various types, sizes, and with blurred boundaries. Extensive experimental validation is conducted on multiple publicly available MRI datasets. The main contributions of this paper are as follows:

- **A2C2f-CGLU-DYT Module:** We propose the A2C2f-CGLU-DYT module, which synergizes the Convolutional Gated Linear Unit (CGLU) with a dynamic Tanh activation function. This integration fortifies local feature extraction and nonlinear modeling capabilities, explicitly enhancing the model's robustness against input signal variations and magnitude heterogeneity.
- **C2TSSA Module:** We design the C2TSSA module, a lightweight global modeling mechanism grounded in Token Statistics Self-Attention. It effectively circumvents the high computational overhead of traditional transformers while preserving the capacity to capture long-range dependencies, thereby significantly refining the delineation of complex and irregular tumor morphologies.
- **CGAFusion Module:** We introduce the CGAFusion module to bridge the semantic gap between multi-scale features. By orchestrating content-guided channel and spatial attention, this module efficiently amalgamates shallow geometric details with deep semantic representations, critically improving the detection sensitivity for minute lesions and tumors with indistinct boundaries.

II. RELATED WORK

A. Bidirectional feature pyramid network module

BiFPN (Bidirectional Feature Pyramid Network)[28] is an efficient feature fusion module originally introduced in EfficientDet, designed to enhance the representation of multi-scale features. Unlike traditional FPN, BiFPN introduces bidirectional feature fusion paths, both top-down and bottom-up, allowing high-level semantic information and low-level detail information to complement each other. Additionally, BiFPN employs a learnable weighting mechanism, enabling the network to automatically adjust the importance of different feature layers during fusion, thereby enhancing the flexibility of feature representation. Furthermore, BiFPN streamlines redundant connections, retaining only effective fusion nodes with multiple inputs, and adds cross-scale connections, which improves its ability to perceive different object sizes while maintaining lightweight computation. Thanks to these designs, BiFPN has become a commonly used and efficient feature fusion module in multi-scale visual tasks, such as object detection and semantic segmentation.

B. Attention mechanism

The attention mechanism was first proposed by Bahdanau *et al.*[29] in the neural machine translation task to address the information bottleneck problem in traditional encoder–decoder structures

when processing long texts. By introducing a learnable weight mechanism, this method allows the model to selectively focus on different parts of the input sequence at each time step, significantly improving translation quality. Subsequently, Luong *et al.*[30] introduced more efficient local and global attention mechanisms, further advancing the application of attention mechanisms in natural language processing tasks. Since then, attention mechanisms have been widely applied in various fields, such as image recognition, speech recognition, and recommendation systems. An important milestone for the attention mechanism was the Transformer architecture proposed by Vaswani *et al.*[31] This architecture is entirely based on self-attention, discarding traditional convolutional and recurrent structures, which greatly enhances the model's parallel computation capability and long-range dependency modeling ability. Transformers and their variants have become the foundation of current mainstream natural language processing models, such as BERT,[32] GPT,[33] and T5.[34] In recent years, attention mechanisms have also been extended to graph neural networks (graph attention networks)[35] and vision tasks, such as Vision Transformer (ViT).[36] These works further validate the universality and effectiveness of attention mechanisms in modeling different modalities of data. To improve efficiency and scalability, researchers have also proposed various improved attention mechanisms, including sparse attention,[37] low-rank attention (Linformer),[38] and factorized attention (Performer),[39] which effectively alleviate the computational and storage bottlenecks of standard self-attention in large-scale data processing. In summary, the attention mechanism has evolved from an auxiliary component to a core element of modern deep learning architectures, playing a crucial role in various tasks. Future research can further focus on its applications and optimizations in cross-modal modeling, low-resource learning, and model compression.

C. YOLOv12 algorithm

YOLOv12,[40] released on February 18, 2025, is the latest attention-centered real-time object detection model in the YOLO series, significantly improving accuracy while maintaining high inference speed. Its core innovations include the regional attention mechanism (dividing the feature map into L equal-sized regions, with the default being 4, to achieve a large receptive field with low computational cost), the Residual Efficient Layer Aggregation Network (R-ELAN) (an improvement over ELAN's feature aggregation method, introducing scaled residual connections and bottleneck structures to optimize the training of large-scale attention models), and an optimized attention architecture (combining FlashAttention to reduce memory overhead, removing positional encoding, adjusting the MLP ratio to 1.2 or 2, reducing stacking depth, and introducing 7 × 7 separable convolutions for implicit positional encoding, among other improvements).

III. METHOD DESIGN

This paper presents a lightweight improved network structure for brain tumor detection, based on targeted optimizations of the YOLOv12n framework. The overall network consists of three parts: the improved backbone, the multi-scale feature fusion network (neck), and the detection head (head). The A2C2f-CGLU-DYT module is introduced in the backbone to enhance local feature extraction and nonlinear modeling capabilities. During the feature fusion stage, the C2TSSA module is added to capture global context information using an efficient self-attention mechanism. At the output end, the CGAFusion module is designed to fuse shallow detail features with deep semantic features, improving the detection accuracy for multi-scale tumors, especially small targets. The overall framework is shown in Fig. 1.

A. A2C2f-CGLU-DYT module

Brain tumors often exhibit varied shapes and blurred boundaries, requiring both detailed feature capture and global information acquisition. Ordinary convolutions are limited by their local receptive fields and lack the ability for global modeling, while conventional attention mechanisms are computationally expensive and not conducive to lightweight deployment. To address this, this paper introduces an improved A2C2f-CGLU-DYT module, which enhances feature extraction ability and computational efficiency while maintaining the structural advantages, as shown in Fig. 2.

First, the input feature map undergoes channel adjustment and preliminary feature extraction through convolution layers. It then passes into the backbone structure, composed of n stacked Ablock-CGLU-DYT units. In each Ablock-CGLU-DYT module, the Convolutional Gated Linear Unit (CGLU) restricts channel interactions within the local receptive field, thus reducing computational cost and improving parameter efficiency while ensuring feature expression capability. Building on this, the Dynamic Tanh (DYT)[41] mechanism is introduced at the CGLU[42] output. This mechanism adaptively adjusts the stretching range of Tanh based on the input magnitude, enhancing nonlinear modeling ability and improving robustness to changes in input magnitude. After multiple layers of stacking, the output features are fused through convolution layers and integrated with the initial input branch under the attention mechanism, ultimately generating enhanced feature maps that combine both global semantic information and local detail features. Compared to the original module, the A2C2f-CGLU-DYT significantly improves feature extraction capability and model robustness while retaining the advantages of the residual structure.

B. C2TSSA module

In brain tumor detection tasks, tumor regions often exhibit features with both blurred boundaries and significant global morphological differences. Relying solely on convolutions is insufficient for capturing long-range dependencies, while conventional self-attention mechanisms suffer from high computational complexity and slow inference speeds. To improve global information modeling ability while ensuring real-time performance, this paper introduces the C2TSSA module, as shown in Fig. 3.

The processing flow of the C2TSSA module is as follows: First, the input feature map is passed through convolution layers for channel adjustment and initial feature transformation for subsequent processing. Then, the convolved feature map is split along the channel dimension into two parts: one part is directly retained as a residual branch, which will be fused with the backbone features to enhance information flow and gradient propagation; the other part enters the backbone branch and sequentially passes through multiple stacked TSSA (Token Statistics Self-Attention) modules. The TSSA[43] module utilizes a variational maximum coding rate

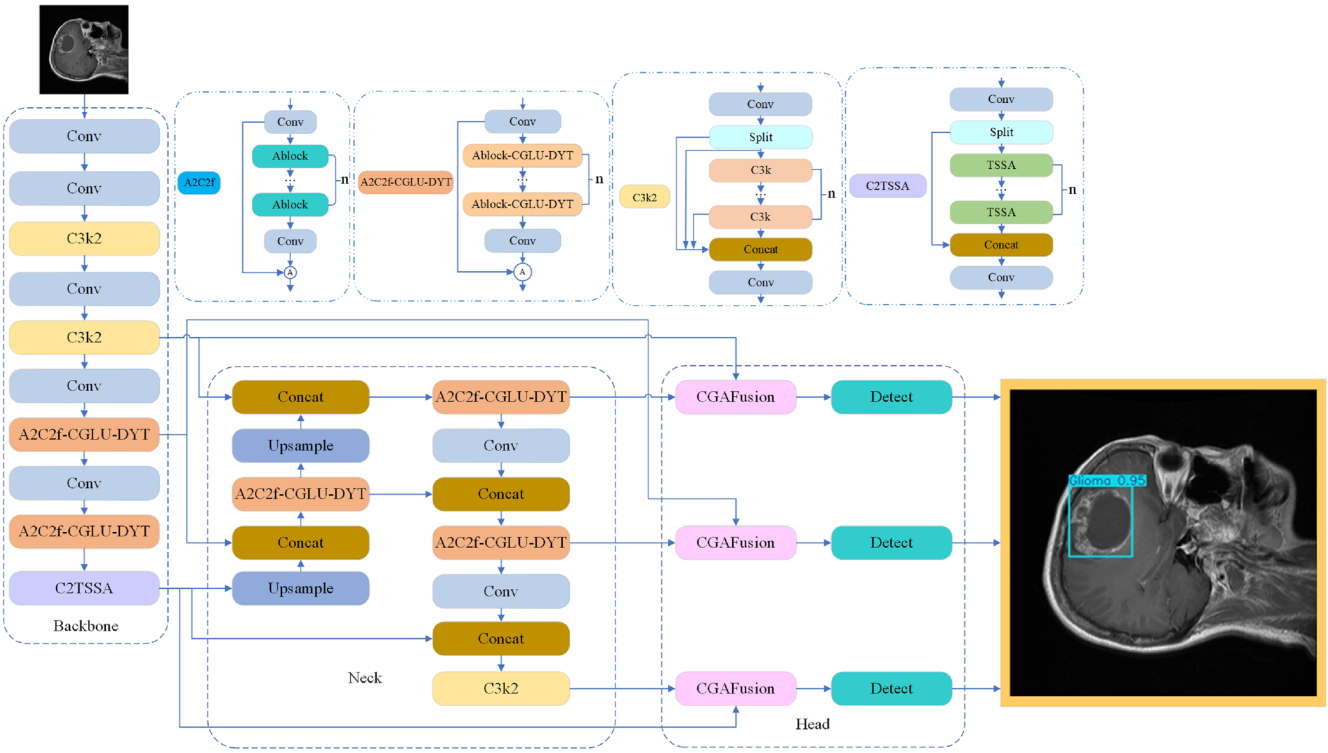


FIG. 1. Network structure diagram of the improved YOLOv12n.

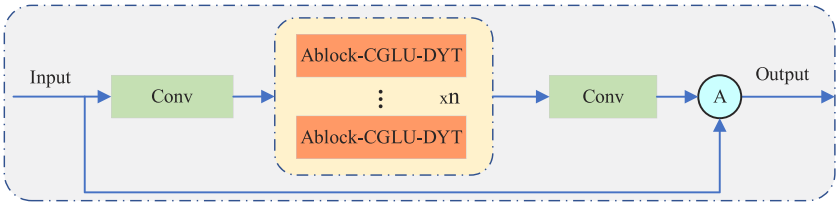


FIG. 2. A2C2f-CGLU-DYT module.

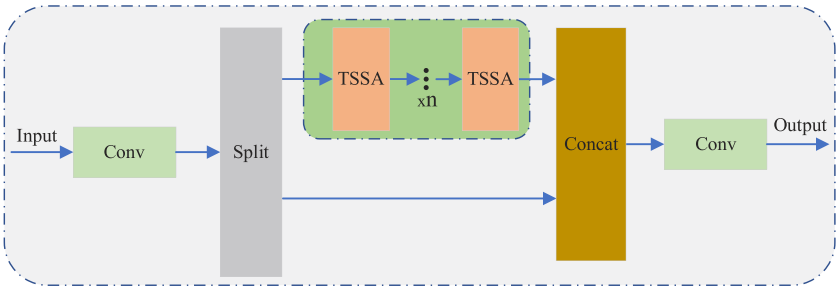


FIG. 3. C2TSSA module.

objective to achieve more efficient attention computation. The core idea is to divide the tokens into multiple semantic groups, making features within the same group more compact and features from different groups more dispersed, thereby reducing computation while enhancing the ability to distinguish features. With this mechanism, TSSA can effectively capture global information without the need to compute attention weights between tokens individually, and it combines local feature extraction ability to achieve multi-scale information fusion and global context modeling. After the TSSA stacking process, the features output from the backbone branch are concatenated with the features from the residual branch along the channel dimension, combining global attention features with the original local detail information. Finally, convolution layers are applied to fuse and re-encode the concatenated features, outputting feature

maps that retain details while enhancing global perception, providing more expressive feature representations for subsequent detection or recognition tasks.

C. CGAFusion (content-guided attention feature fusion) module

In brain tumor detection, accurately fusing features from different levels is key to improving model performance. Traditional feature fusion methods often focus on simple concatenation of shallow detail features and deep semantic features, but this approach tends to overlook the differences in importance between features from different levels when dealing with complex multi-scale objects. To address this issue, this paper introduces a content-guided feature fusion module—CGAFusion,[44] which uses a dynamic weighting mechanism to adaptively adjust the fusion of shallow and deep features based on the content at each position, thereby more effectively enhancing the accuracy of multi-scale tumor detection, as shown in Fig. 4.

This structure is designed to fully combine the advantages of low-level features F_{low} and high-level features F_{high} during the feature fusion stage, thereby obtaining more expressive fused features. The specific process is as follows: First, F_{low} and F_{high} are element-wise added to obtain an initial fused feature that contains both spatial details and semantic information, which is then input into the CGA module. The CGA generates a weight matrix W, where the values typically range from 0 to 1, used to dynamically measure the importance of low-level features at different positions. At the same time, 1−W is calculated as the weight for high-level features, achieving complementarity between the two features in spatial positions. Next, F_{low} is multiplied by W, and F_{high} is multiplied by 1−W to highlight their respective important parts. These two parts are then summed to achieve weighted fusion. Finally, a 1 × 1 convolution layer is applied to compress and mix the information along the channel dimension, generating the final fused feature F_{fuse}, which retains low-level detail information while introducing high-level global semantics, providing richer and more balanced feature representations for subsequent tasks.

IV. EXPERIMENT

A. Dataset

This study utilizes the publicly available Brain Tumor Detection dataset from Kaggle,[45] comprising anonymized MRI scans of meningioma, glioma, and pituitary tumors. The dataset was split into 2451 training and 613 testing images. As shown in Fig. 5, the distinct morphological characteristics, ranging from well-defined meningiomas to irregular, edema-surrounded gliomas and sellar region pituitary tumors, provide a diverse and challenging benchmark for evaluating the proposed method.

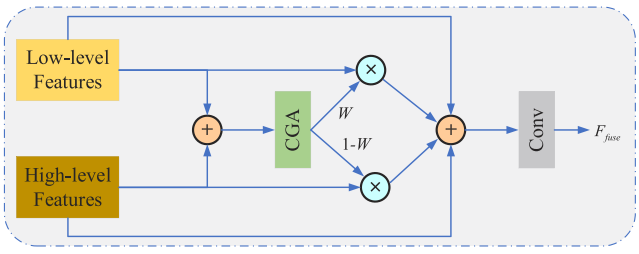


FIG. 4. CGAFusion module.

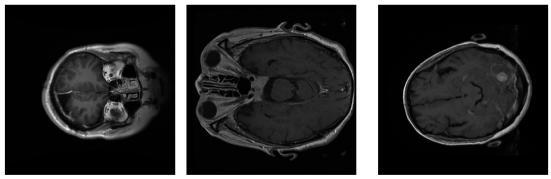


FIG. 5. Image samples of the brain tumor datasets.

B. Experimental platform and hyperparameter setting

Experiments were conducted on an Ubuntu 20.04 system using PyTorch 1.10.0, Python 3.8, and CUDA 11.3. The hardware setup included an NVIDIA RTX 4090 GPU (24 GB), an AMD EPYC 7T83 CPU, and 90 GB of RAM. Training employed an input resolution of 640 × 640, a batch size of 64, and 200 epochs. Optimization utilized an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005.

C. Evaluation indicators

The experiments in this paper use F1 score, Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP) as evaluation metrics[46] and also consider the number of parameters (Parameters), with the calculation expressions as follows:

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \tag{1}$$

$$\text{Recall} = \frac{T_p}{T_p + F_N}, \tag{2}$$

$$\text{AP} = \int_0^1 P(R)\mathrm{d}R, \tag{3}$$

$$\text{mAP} = \frac{1}{n}\sum_{i=0}^{n} AP(i), \tag{4}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

where T_p represents the number of correctly detected objects, F_p represents the number of incorrectly detected objects, F_N represents the number of missed detections, n represents the number of classes, and $AP(i)$ represents the average precision of the i-th object class.

V. EXPERIMENTAL ANALYSIS

A. Comparison of results from multiple different models

To further validate the comprehensive performance advantages of the proposed method across different detection models, this section compares it with several mainstream object detection models on the Brain Tumor dataset. The comparison metrics include precision, recall, mAP@0.5, inference speed (FPS), computational complexity (GFLOPs), and the number of parameters (Params) to provide a thorough evaluation of the model's performance in detection accuracy, computational cost, and real-time capability.

Table I presents a comparison of the detection performance, inference speed, and computational complexity of YOLOv5s, YOLOv8n, YOLOv10n, YOLOv11n, YOLOv12n, RT-DETR-r18, and the proposed method on the Brain Tumor dataset. As shown, the proposed method achieves the best results in the three core evaluation metrics—precision, recall, and mAP@0.5, improving by 0.7%, 0.9%, and 1.2%, respectively, compared to the baseline model YOLOv12n. This indicates a clear advantage in overall detection accuracy and recall ability. Compared to other YOLO series models, the proposed method shows varying degrees of improvement in mAP@0.5, especially improving by 3.5% over YOLOv10n, demonstrating the effectiveness of the designed multi-scale attention and dynamic convolution structures in feature extraction and object recognition. In terms of inference speed, the proposed method achieves 476.1 FPS, which is lower than YOLOv12n and YOLOv10n but still significantly higher than RT-DETR-r18, fully meeting real-time detection requirements. Regarding computational complexity, the proposed method has 8.3 GFLOPs and 3.13M parameters, which is slightly higher than YOLOv12n but much lower than high-complexity models, such as RT-DETR-r18. Overall, the proposed method achieves comprehensive improvements in detection accuracy, recall, and mAP while maintaining a low computational cost, with inference speed still at a high level. This reflects a good balance between accuracy and speed and strong potential for application.

As shown in Table II, the proposed method achieves state-of-the-art performance across all three tumor categories. Most notably, in the challenging Glioma category, our model attains an accuracy of 84.6%, outperforming YOLOv10n by a significant margin of 8.6% and surpassing the latest YOLOv12n. Furthermore, our method maintains the highest precision in meningioma (98.6%) and pituitary tumor (98.2%) detection, demonstrating superior robustness and effectiveness in handling both complex and standard medical imaging tasks. This consistent superiority over the most recent YOLO iterations validates the efficacy of our architectural optimizations in capturing subtle pathological features that standard models often miss. Consequently, the proposed algorithm offers a highly reliable solution for computer-aided diagnosis, ensuring stable performance across cases with varying degrees of detection difficulty.

TABLE I. Comparison of results from multiple models.

Algorithms	Precision	Recall	mAP@0.5	FPS	GFLOPS	Params (M)
YOLOv5s	92.3	86.1	93.1	303.0	15.8	7.01
YOLOv8n	90.4	85.6	91.2	400.0	8.1	3.00
YOLOv10n	87.6	85.0	90.3	434.7	8.2	2.69
YOLOv11n	91.7	86.7	93.0	454.5	6.3	2.58
YOLOv12n	92.0	86.5	92.6	625.0	5.8	2.50
RT-DETR-r18	92.5	86.9	92.8	196.0	56.9	19.87
Ours	92.7	87.4	93.8	476.1	8.3	3.13

TABLE II. Comparison results of multiple algorithms in Average Precision (AP/%).

Category \ Algorithms	YOLOv8n	YOLOv10n	YOLOv11n	YOLOv12n	Ours
Meningioma	98.2	97.8	97.9	98.3	98.6
Glioma	79.3	76.0	84.0	82.5	84.6
Pituitary_tumor	96.1	97.3	97.2	97.1	98.2

To visually validate the effectiveness of the proposed method in complex clinical scenarios, we compared the visualization results of our improved model (ours) against mainstream YOLO series models (YOLOv8n-v12n) on three challenging cases. As shown in Fig. 6, our method demonstrated superior detection precision and confidence. Specifically, in the detection of a small Glioma sample with blurred boundaries (left panel), the latest YOLOv12n failed to detect the tumor entirely, and while YOLOv10n and YOLOv11n detected the target, they exhibited uncertainty with low confidence scores of 0.59 and 0.64, respectively; in contrast, our method accurately localized the lesion with a high confidence of 0.87. Furthermore, in the pituitary tumor classification task (middle panel), YOLOv8n critically misclassified the tumor as meningioma, and YOLOv12n showed a dangerously low confidence of 0.46, whereas our method correctly identified the tumor with a robust confidence of 0.86. Even in the simpler case (right panel), our method achieved the highest confidence score among all models. These qualitative results confirm that our improved algorithm effectively addresses the issues of missed detections, misclassifications, and low confidence present in existing models, significantly enhancing diagnostic reliability.

B. Analysis of the model's cross-dataset generalization ability

In the analysis of the model's cross-dataset generalization ability, Table III presents the performance of different object detection algorithms across several key metrics, including precision, recall, and mAP@0.5. By comparing the performance of different algorithms, a more comprehensive evaluation of their detection accuracy and computational efficiency on the Blood Cell Count dataset[47,48] can be achieved.

As shown in Table III, the proposed algorithm outperforms all comparative models across all evaluation metrics. Specifically, it achieves an mAP@0.5 of 94.1%, surpassing YOLOv11n by 1.7%, which demonstrates stable comprehensive performance. Most critically, our method attains a recall of 92.8%, significantly reducing the missed diagnosis rate; this is vital for early clinical screening. Simultaneously, we maintain the highest precision (86.2%), effectively avoiding the false positive issues observed in RT-DETR-r18 (80.6%). These quantitative results confirm that the improved

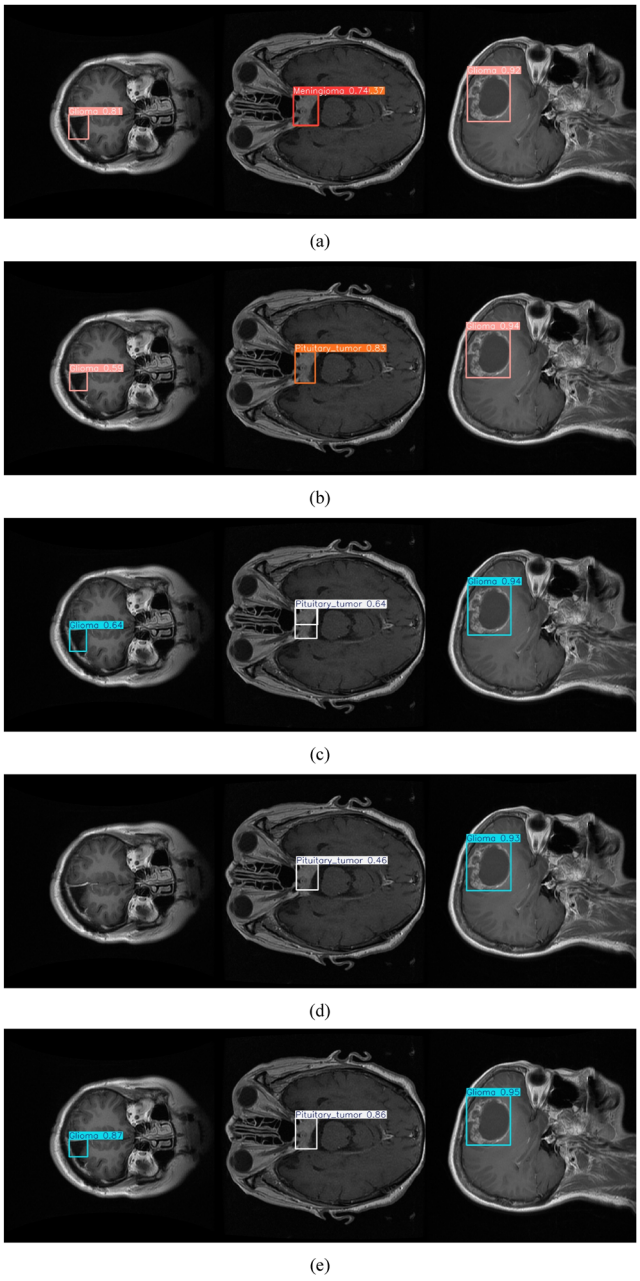


FIG. 6. Comparison of detection results of different models on multi-class brain tumor images: (a) YOLOv8n, (b) YOLOv10n, (c) YOLOv11n, (d) YOLOv12n, and (e) ours.

strategies not only significantly enhance feature extraction capabilities but also achieve an optimal balance between precision and recall, demonstrating superiority over existing YOLO variants and transformer-based algorithms.

Figure 7 shows the practical application results of YOLOv12n and ours in the cell detection task, particularly in the performance

TABLE III. Accuracy comparison of different models on the Blood Cell Count dataset.

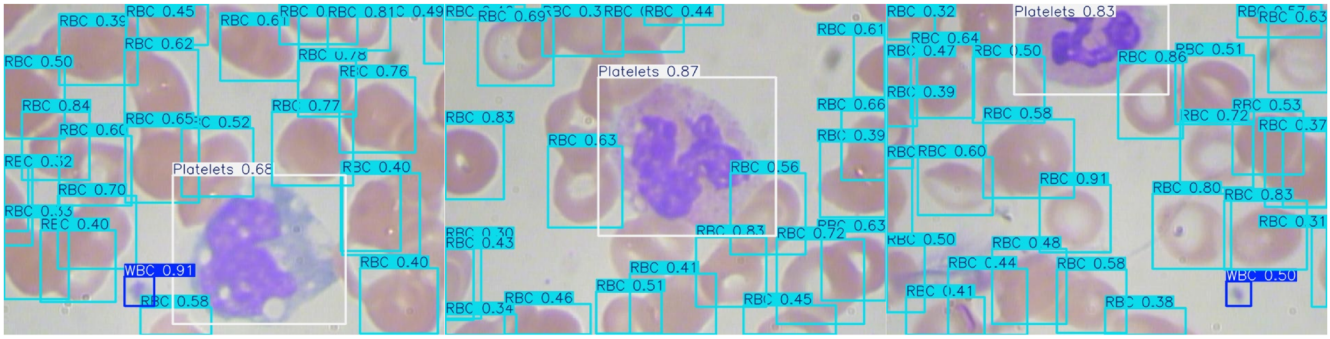
Algorithms	Precision	Recall	mAP@0.5
YOLOv8n	85.7	89.3	92.2
YOLOv10n	79.4	87.2	88.1
YOLOv11n	85.8	90.2	92.4
RT-DETR-r18	80.6	89.1	90.0
Ours	86.2	92.8	94.1

of object detection in cell images. From the detection boxes and corresponding confidence values, we can observe some significant differences. First, YOLOv12n [Fig. 7(a)] accurately detects most cell types in the image, such as red blood cells (RBCs) and white blood cells (WBCs), and assigns corresponding confidence values. However, some of the denser targets have lower confidence values. Additionally, YOLOv12n has some redundant detection boxes for certain cells, which may result in multiple boxes detecting the same target in some cases. In contrast, ours [Fig. 7(b)] demonstrates higher detection accuracy, particularly in the confidence values of the cells, where most detection boxes have confidence values significantly higher than YOLOv12n. For example, the confidence for platelets reaches 0.96 and 0.95, and RBCs also often receive high confidence values, indicating that ours is more accurate in detecting dense targets and can handle target recognition better. Moreover, the detection box distribution of ours is more precise, with fewer redundant boxes, showing better detection efficiency and accuracy. Overall, ours outperforms YOLOv12n in terms of precision, object detection accuracy, and robustness. In particular, in dense target detection and high-confidence predictions, ours demonstrates stronger capability, providing more reliable detection results in diverse cell images. These advantages make the custom model notably competitive in cross-dataset generalization and practical applications.

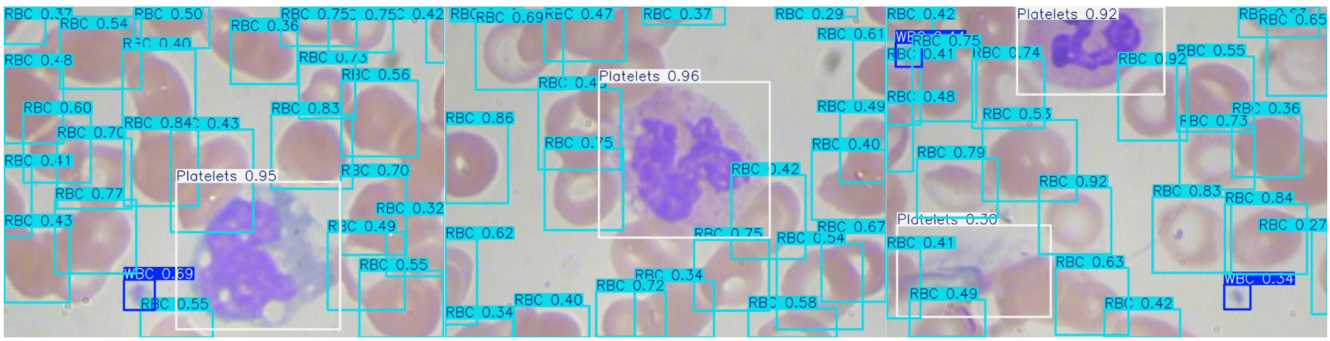
C. Ablation experiment

To evaluate the impact of each improvement module on model performance and computational overhead, ablation experiments were conducted based on YOLOv12n, testing the effects of CGAFusion, C2TSSA, A2C2f-CGLU-DYT, and their combinations. By comparing metrics such as precision, recall, and mAP@0.5, the role and trade-offs of different modules in improving detection accuracy, recall, and speed were analyzed.

The ablation experimental results in Table IV clearly demonstrate the advantages of our method over the original YOLOv12n. Based on the YOLOv12n baseline, the introduction of CGAFusion, C2TSSA, and A2C2f-CGLU-DYT modules leads to improvements across different performance metrics. Specifically, CGAFusion increases precision by 1.2%, while C2TSSA and A2C2f-CGLU-DYT raise recall by 2.4% and 4.4%, respectively. Ultimately, with the combined multi-module approach, mAP@0.5 reaches 93.8%, representing a 1.2% improvement over the baseline and achieving the best overall performance. These results indicate that our proposed strategy can significantly enhance recall while maintaining high precision, thereby comprehensively outperforming YOLOv12n.



(a)



(b)

FIG. 7. Comparison of detection results of different models on the Blood Cell Count dataset: (a) YOLOv12n and (b) ours.

TABLE IV. Results of the ablation experiment.

Number	Experiments	Precision	Recall	mAP@0.5
1	YOLOv12n	92.0	86.5	92.6
2	YOLOv12n+CGAFusion	93.2	87.1	93.5
3	YOLOv12n+C2TSSA	91.5	88.9	93.3
4	YOLOv12n+A2C2f-CGLU-DYT	91.7	90.9	93.1
5	YOLOv12n+A2C2f-CGLU-DYT+ C2TSSA+CGAFusion	92.7	87.4	93.8

VI. CONCLUSION

This paper addresses the challenges of low contrast, small targets, and blurred boundaries in brain tumor detection in medical images and proposes a lightweight improved detection method based on the YOLOv12n framework. By introducing the A2C2f-CGLU-DYT module to enhance local feature extraction and nonlinear modeling capabilities, utilizing the C2TSSA module to improve global context information modeling, and combining the CGAFusion module for efficient fusion of shallow detail and deep semantic features, the proposed method significantly improves the model's detection performance. Experimental results show that the method achieves 92.7% precision, 87.4% recall, and 93.8% mAP@0.5 on the Brain Tumor dataset and 94.1% mAP@0.5 on the Blood Cell Count dataset, significantly outperforming the baseline YOLOv12n model. Ablation experiments also validate the effectiveness of the individual improvement modules.

Overall, the proposed method achieves a balance between precision and speed while maintaining low computational overhead, with high potential for clinical application. However, the current method still has limitations, such as relatively high computational resource consumption in ultra-high-resolution medical images and room for further improvement in detecting very small lesions. Future work will focus on enhancing the detection of small tumors by exploring high-resolution feature retention and adaptive attention mechanisms. Additionally, we plan to validate the model's robustness across diverse medical imaging datasets and clinical settings to further strengthen its generalization capability and clinical applicability.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

L.L. and R.W. contributed equally to this paper.

Lan Liu: Conceptualization (equal); Methodology (equal); Writing – original draft (equal). **Ronghuan Wu**: Formal analysis (equal); Investigation (equal); Visualization (equal). **Chengyang Zhang**: Data curation (equal); Software (equal); Validation (equal). **Jianhui Xu**: Funding acquisition (equal); Project administration (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] P. Jyothi and A. R. Singh, "Deep learning models and traditional automated techniques for brain tumor segmentation in MRI: A review," Artif. Intell. Rev. **56**(4), 2923–2969 (2023).

[2] E. Q. Lee *et al.*, "Barriers to accrual and enrollment in brain tumor trials," Neuro. Oncol. **21**(9), 1100–1117 (2019).

[3] T. A. G. M. Huisman, "Tumor-like lesions of the brain," Cancer Imaging **9**(Special issue A), S10 (2009).

[4] T. Hossain *et al.*, "Brain tumor detection using convolutional neural network," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (IEEE, 2019).

[5] G. Kumar, P. Kumar, and D. Kumar, "Brain tumor detection using convolutional neural network," in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)* (IEEE, 2021).

[6] M. Arabahmadi, R. Farahbakhsh, and J. Rezazadeh, "Deep learning for smart healthcare—A survey on brain tumor detection from medical imaging," Sensors **22**(5), 1960 (2022).

[7] H. Dong *et al.*, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Annual Conference on Medical Image Understanding and Analysis* (Springer International Publishing, Cham, 2017).

[8] R. Ezhilarasi and P. Varalakshmi, "Tumor detection in the brain using faster R-CNN," in *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on* (IEEE, 2018).

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017).

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767 (2018).

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934 (2020).

[13] T. Wang, Y. Zhai, Y. Li *et al.*, "Insulator defect detection based on ML-YOLOv5 algorithm," Sensors **24**(1), 204 (2023).

[14] Y. Wang, P. Zhang, and S. Tian, "Tomato leaf disease detection based on attention mechanism and multi-scale feature fusion," Front. Plant Sci. **15**, 1382802 (2024).

[15] 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)* (IEEE, Vancouver, BC, 2023), pp. 7464–7475.

[16] M. Sohan, T. Sai Ram, and C. Venkata Rami Reddy, "A review on YOLOv8 and its advancements," in *Algorithms for Intelligent Systems* (Springer, 2024), pp. 529–545.

[17] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," arXiv:2402.13616 (2024).

[18] A. Wang *et al.*, "YOLOv10: Real-time end-to-end object detection," arXiv:2405.14458 (2024).

[19] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," arXiv:2410.17725 (2024).

[20] Y. Kaya, E. Akat, and S. Yıldırım, "Fusion-brain-net: A novel deep fusion model for brain tumor classification," Brain Behav. **15**(5), e70520 (2025).

[21] M. Kang *et al.*, "BGF-YOLO: Enhanced YOLOv8 with multiscale attentional feature fusion for brain tumor detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer Nature Switzerland, Cham, 2024).

[22] R. Bai, G. Xu, and Y. Shi, "SCC-YOLO: An improved object detector for assisting in brain tumor diagnosis," in *Proceedings of the 2025 International Conference on Health Big Data* (ACM, 2025).

[23] M. Rahimi, M. Mostafavi, and A. Arabameri, "Automatic detection of brain tumor on MRI images using a YOLO-based algorithm," in *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)* (IEEE, 2024).

[24] S. Sahoo *et al.*, "An augmented modulated deep learning based intelligent predictive model for brain tumor detection using GAN ensemble," Sensors **23**(15), 6930 (2023).

[25] M. S. Mithun and S. Joseph Jawhar, "Detection and classification on MRI images of brain tumor using YOLO NAS deep learning model," J. Radiat. Res. Appl. Sci. **17**(4), 101113 (2024).

[26] A. Chen, D. Lin, and Q. Gao, "Enhancing brain tumor detection in MRI images using YOLO-NeuroBoost model," Front. Neurol. **15**, 1445882 (2024).

[27] P. Hariharan and N. Surya, "Automated brain tumor detection and localization using YOLO-based deep learning algorithm," Am. J. Psychiatr. Rehabil. **28**(5), 198–203 (2025).

[28] L. Zhu *et al.*, "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer, 2018).

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations* (ICLR, San Juan, PR, 2016).

[30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv:1508.04025 (2015).

[31] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems* (MIT Press, 2017), p. 30.

[32] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)* (Association for Computational Linguistics, 2019).

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei, "Language models are few-shot learners," Adv. Neur. Inform. Proc. Syst. **33**, 1877–1901 (2020).

[34] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res. **21**(140), 1–67 (2020).

[35] P. Velickovic *et al.*, "Graph attention networks," arXiv:1710.10903 (2017).

[36] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," arXiv:2010.11929 (2020).

[37] R. Child *et al.*, "Generating long sequences with sparse transformers," arXiv:1904.10509 (2019).

[38] S. Wang *et al.*, "Linformer: Self-attention with linear complexity," arXiv:2006.04768 (2020).

[39] K. Choromanski *et al.*, "Rethinking attention with performers," arXiv:2009.14794 (2020).

[40] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," arXiv:2502.12524 (2025).

[41] J. Zhu *et al.*, "Transformers without normalization," in *Proceedings of the Computer Vision and Pattern Recognition Conference* (IEEE, 2025).

[42] D. Shi, "Transnext: Robust foveal visual perception for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024).

[43]Z. Wu *et al.*, "Token statistics transformer: Linear-time attention via variational rate reduction," arXiv:2412.17810 (2024).
[44]Y. Gu *et al.*, "A conditionally parameterized feature fusion U-Net for building change detection," Sustainability **16**(21), 9232 (2024).
[45]P. K. Darabi, Medical Image Dataset: Brain Tumor Detection, Kaggle, 2023, www.kaggle.com/datasets/pkdarabi/medical-image-dataset-brain-tumor-detection.
[46]M. Chen, Y. Xu, W. Qin, Y. Li, and J. Yu, "Tomato ripeness detection method based on FasterNet block and attention mechanism," AIP Adv. **15**(6), 065117 (2025).
[47]B. Liu, "Blood cell count and detection method based on YOLO," Highl. Sci. Eng. Technol. **27**, 594–599 (2022).
[48]H. Sazak and M. Kotan, "Automated blood cell detection and classification in microscopic images using YOLOv11 and optimized weights," Diagnostics **15**(1), 22 (2024).