LT-YOLO: An Improved Lightweight Detection Algorithm Based on YOLOv11

Zhidi Cao[1], Zixiang Wu[1], Jiachun Li[1], Dexin Zhang[2], Wanwan Wang[2]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China
[2]IFLYTEK CO., LTD., Anhui, Hefei, China

Abstract: In the fields of medical image analysis and wildlife monitoring, conventional object detection algorithms often suffer from high complexity and computational overhead, making them inadequate for real-time pipeline weld defect inspection requirements. To address this challenge, we propose LT-YOLO, a lightweight object detection model built upon the YOLOv11 framework. The model incorporates a BRA module to enhance the detection capability for minute defects. Traditional convolutions in the baseline model are replaced with ADown convolutional modules to further improve recognition accuracy. A C3k2_SCConv composite module is introduced to strengthen feature representation in complex backgrounds and occluded scenarios. For model lightweighting, a Lightweight Asymmetric Multi-level Compressed Detection Head (LADH) is employed to reduce training complexity while accelerating inference. Experimental results demonstrate that the proposed model achieves 41.6% mAP@0.5:0.95 (14.3% improvement over baseline) in brain tumor detection and 76.2% mAP@0.5:0.95 in African wildlife monitoring, while reducing computational complexity by 31.3% compared to the base model. These results demonstrate that LT-YOLO maintains detection quality while fulfilling precision requirements for medical imaging and wildlife monitoring applications, achieving the model lightweighting objectives.

Keywords: LT-YOLO; YOLOv11; Lightweight; Object Detection; Dynamic Attention

1. Introduction

Deep neural network models have achieved remarkable success in computer vision tasks such as image classification, object detection, and target tracking. However, with the rapid development of computing platforms (e.g., mobile devices and embedded systems), traditional high-accuracy models face challenges in meeting real-time requirements due to their high computational complexity and large parameter counts[1,2].

In recent years, lightweight model design has become a research hotspot in the field of computer vision[3]. Representative works include:

(1) The MobileNet series proposed by the Google team, which employs depthwise separable convolution to decompose standard convolutions into depthwise and pointwise operations, significantly reducing computational costs[4]. (2) The Fire module introduced by researchers from Berkeley and Stanford University in SqueezeNet, which reduces parameters to 1/50th of AlexNet's size[5]. (3) Bi-Level Routing Attention proposed by Lei, Wang et al., which dynamically selects key regions via a dual-level routing mechanism to optimize attention computation efficiency, though memory consumption remains high for high-resolution images[6]. (4) GhostNet developed by Han Kai, Wang Yunhe et al., along with numerous other innovative lightweight architectures[7].

While these methods excel in classification tasks, they exhibit common limitations in complex object detection scenarios (e.g., micro-tumors in medical imaging or occluded targets in wildlife monitoring)[8,9]:

Coexistence of feature loss and redundancy: Lightweight operations may inadvertently discard critical features, while redundant computations (e.g., repeated downsampling) are not sufficiently suppressed;Insufficient

multi-scale adaptability: Existing modules fail to balance feature representation needs across different scale targets, leading to missed small-object detections or localization errors for large objects.

To address these challenges, this study proposes the LT-YOLO algorithm, aiming to enhance detection performance while reducing computational complexity. Validated through dual tasks of medical tumor detection and African wildlife monitoring, LT-YOLO achieves a 30.6% reduction in computational cost (GFLOPs) without compromising accuracy, offering an efficient solution for edge-side real-time inference.

2. LT-YOLO Detection Algorithm

2.1 YOLOv11 Introduction

YOLOv11 is a real-time object detection model released by the Ultralytics team, which shows significant advantages in multi-task scenarios (such as detection, segmentation, and classification) due to its efficient inference speed and high accuracy[10]. The core improvements include: 1) optimizing multi-scale feature extraction through modules such as C3K2, SPFF, and C2PSA; 2) Mosaic-9 data augmentation and Distribution Focal Loss were used to improve the robustness of small targets. 3) Realize the deployment capability of edge devices while maintaining the number of lightweight parameters. However, the original YOLOv11 still has the following limitations:

1) Redundant downsampling: The fixed-step convolution of the traditional CBS module leads to repeated calculation of adjacent hierarchical features;

2) Inefficient static attention: The fixed window attention mechanism of C2PSA calculates redundancy for sparse targets (such as lesions in medical imaging);

3) Parameter redundancy of the detection head: The classification and regression branches do not distinguish the characteristics of the task, resulting in low parameter utilization.

To solve the above problems, this study proposes the LT-YOLO (Light-YOLO) model, based on the YOLOv11 model, and proposes a systematic improvement scheme to solve the redundant calculation problems existing in the existing detection framework of YOLOv11. The network structure is shown in Figure 1.

The ADown module of YOLOv9 is used to replace the traditional downsampling structure for the following modules. Introduce SCConv's SRUs (Spatial Reconstruction Units) and CRUs (Channel Reconstruction Units) into the C3k2 module; Replace the static attention mechanism of C2PSA and design the BRA mechanism; a Lightweight asymmetric detection head (LADH) is used to reduce computational complexity.

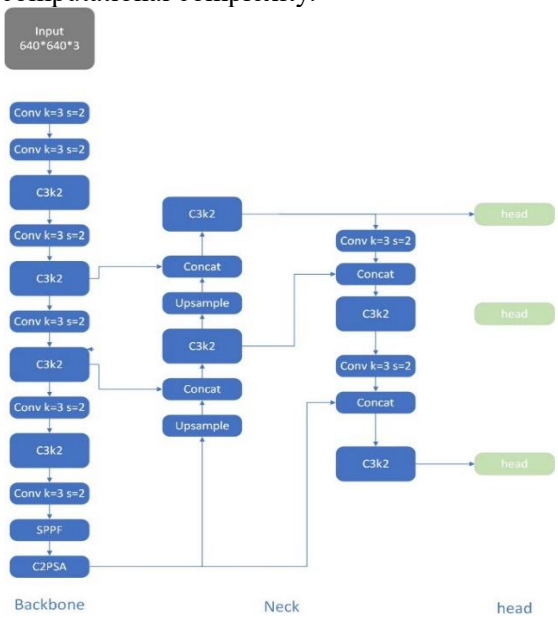

Figure 1. LT-YOLO Network Structure2.2 Improvement Scenarios

2.2 Improvement Scenarios

2.2.1 ADown module: heterogeneous pooling fusion

As an extremely important operation in deep learning, downsampling aims to reduce computational complexity, extract high-level features, and optimize model efficiency by lowering the resolution or dimension of the data. It is widely applied in computer vision and natural language processing, demonstrating the power of downsampling technology[11].

The ADown module implements efficient downsampling via structural reparameterization and lightweight design, balancing computational cost and detection accuracy. To better extract input parameters and improve the model's performance with the same amount of input data, this module was chosen to replace the downsampling module of the YOLO-V11 model, thereby enhancing the performance and efficiency of the model.

The implementation includes the following

steps:

1). Convolution operations: Convolutional layers are employed to extract useful information from the feature map.

2). Stride adjustment: The spatial dimension of the feature map is reduced by adjusting the stride of the convolutional layer.

3). Parameter optimization: The number of parameters in the convolutional layer is optimized to reduce the complexity of the model.

The structure of the module is shown in Figure 2.

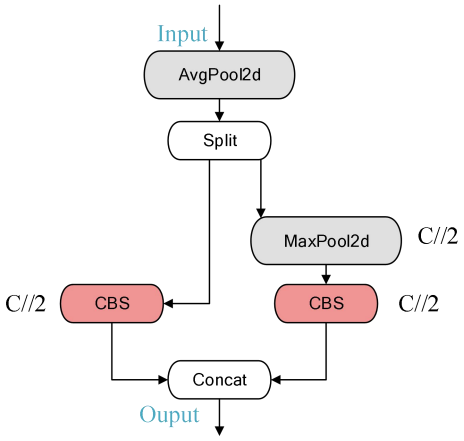

Figure 2. ADown Structure

The structure design of the ADown module adopts a multi-branch feature fusion mechanism, and its core process consists of three key processing stages: firstly, the input feature map is processed through the AvgPool2d (average pooling) layer, which reduces the computational complexity through spatial dimension compression while maintaining the global statistical features; Subsequently, the pooled feature tensor is decoupled into two independent branches through the Split operation: the main branch is directly connected to the CBS (Convolution-Batch Normalization-Activation) module for feature enhancement, and the secondary branch is first extracted by MaxPool2d (maximum pooling) to extract local salient features, and then input into the parallel CBS module for nonlinear feature reconstruction. Finally, the concatenate operation is used to stitch the eigenvectors of the dual-path output in the channel dimension to construct a multi-scale feature expression space and realize the effective integration of multi-modal feature information.

The branch structure introduced by the ADown convolution module successfully realizes more feature combinations and information interactions, greatly retains the contextual information, and effectively reduces the loss of features.

2.2.2 C3k2_SCConv module: space-channel joint reconstruction

The C3k2 module is a critical feature extraction component in the YOLO11 model, designed as an improved version of the traditional C3 module. Its core innovation combines variable convolutional kernels (e.g., 3×3, 5×5) with a channel separation strategy, significantly enhancing feature extraction capabilities. This design is particularly effective for complex scenarios and deep-level feature learning tasks. Compared to the standard C3 module, C3k2 introduces the multi-scale convolutional kernel C3k, which improves accuracy in detecting large objects. Additionally, the use of diverse kernel sizes enhances the model's adaptability to complex scenes, thereby boosting detection precision.

To address spatial and channel redundancy in convolutional neural networks (CNNs), Lianghua He et al. proposed the SCConv module, consisting of two units: the Spatial Reconstruction Unit (SRU) and the Channel Reconstruction Unit (CRU). The SRU employs a separation-reconstruction mechanism to reduce spatial redundancy, while the CRU utilizes a split-transform-fuse strategy to mitigate channel redundancy. Through their collaborative operation, SCConv effectively compresses redundant feature information in CNNs. To achieve lightweight objectives and improve computational efficiency, this study integrates the SCConv module into the C3k2 architecture, forming the C3k2_SCConv module (as illustrated in Figure 3).

The implementation involves embedding SRU and CRU units into the Bottleneck block of the C3k2 module to reduce both feature space redundancy and computational complexity. Since the C3k module is constructed by stacking multiple Bottleneck blocks, and the C3k2 module is further extended from C3k through the introduction of multi-scale kernels, the optimization of Bottleneck blocks with SRU/CRU units achieves similar redundancy suppression effects in the C3k2 module. Additionally, the SCConv convolution layer is integrated before the SPPF module to further reduce spatial and channel redundancy in the

backbone network. By optimizing spatial and channel information separately, redundant features are minimized, thereby improving the model's overall performance. The structural relationship between the Bottleneck block and the C3k module is detailed in Figures 4 and 5.

As an important feature extraction component in the YOLO series models, C3k2 has outstanding performance in terms of speed and efficiency of feature extraction. To further reduce model parameters, the SRU and CRU units from the SCConv module are reused to streamline the parameter count in C3k2. This improvement effectively addresses the coexistence of spatial and channel redundancy in the module [12,13]. The main modification to the C3k2 module is concentrated in the Bottleneck section, where one of the convolutional blocks is replaced with an SCConv convolution block containing an SRU and a CRU.

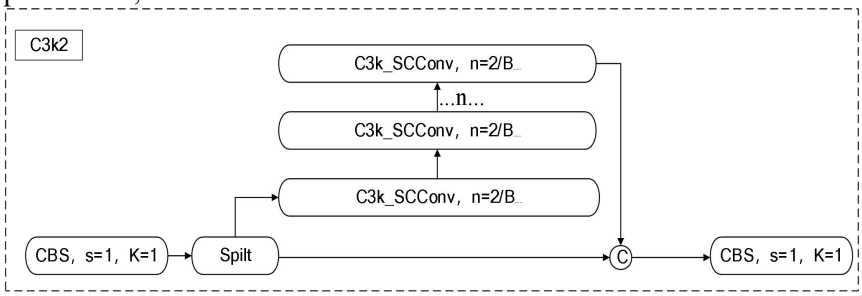

Figure 3. C3k2_SCConv Module

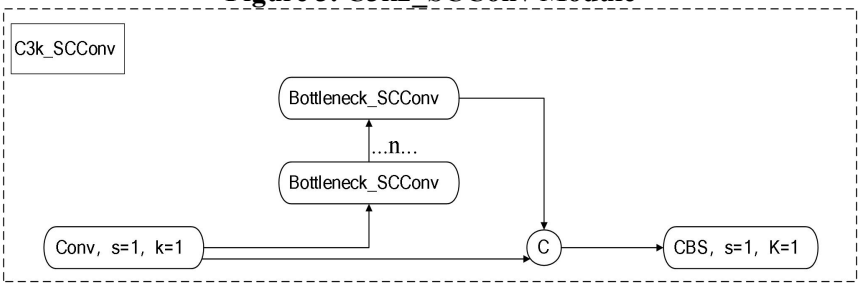

Figure 4. C3k_SCConv Module

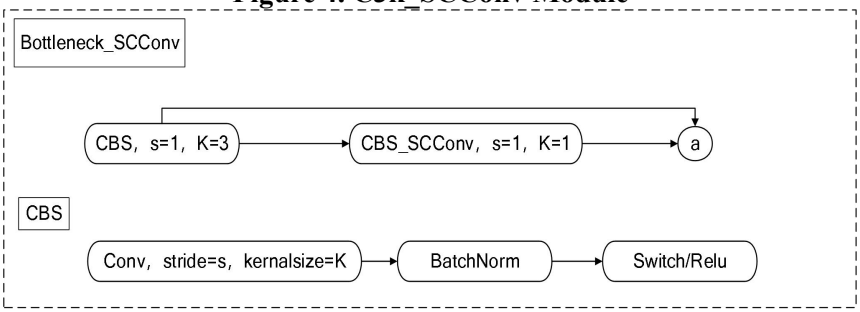

Figure 5. Bottleneck_SCConv and CBS Module

2.2.3 C2BRA module: dynamic sparse attention

To alleviate the scalability problem of multi-head self-attention (MHSA), the authors Zhu et al. explored a dynamic, query-aware sparse attention mechanism, which is a BiFormer dynamic sparse attention module based on a Transformer variant. The key idea is to filter out most irrelevant key-value pairs at the coarse region level, retaining only a small portion of the routing regions. To address computation and memory issues in the model and to better achieve lightweight functioning, the team improved the original network module C2PSA and introduced Bi-Level Routing Attention (BRA). Figure 6 illustrates the computation diagram of Bi-level Routing Attention[14].

In this study, the C2BRA module was constructed by improving the C2PSA module in the YOLO model and introducing the Two-Layer Routing Attention Mechanism (BRA). Specifically, firstly, the non-overlapping region division was constructed on the feature map, and more than 90% of the redundant computing units were filtered out by using the region-level routing strategy. Then, fine-grained token-to-token

attention computation is performed in the filtered routing region, and finally, the feature representation is enhanced through two-level feature interaction.

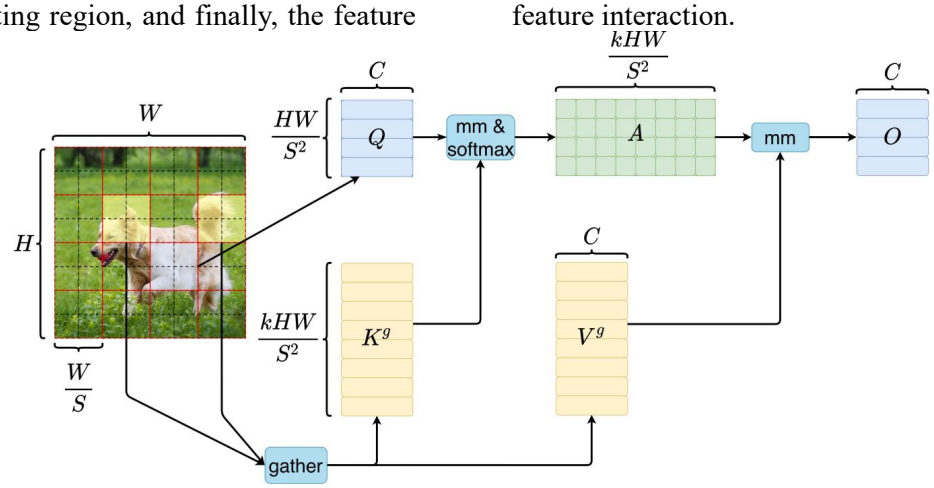

Figure 6. Bi-Level Routing Attention

2.2.4 LADH detection head: task decoupling and asymmetric compression

To achieve the dual optimization goals of model lightweightness and detection performance, this study introduces the Lightweight Asymmetric Detection Head (LADH) architecture to replace traditional detection heads. This module innovatively designs an asymmetric feature processing mechanism and a dual-head collaborative architecture that reduces computational complexity while improving detection accuracy, as illustrated in Figure 7[15].

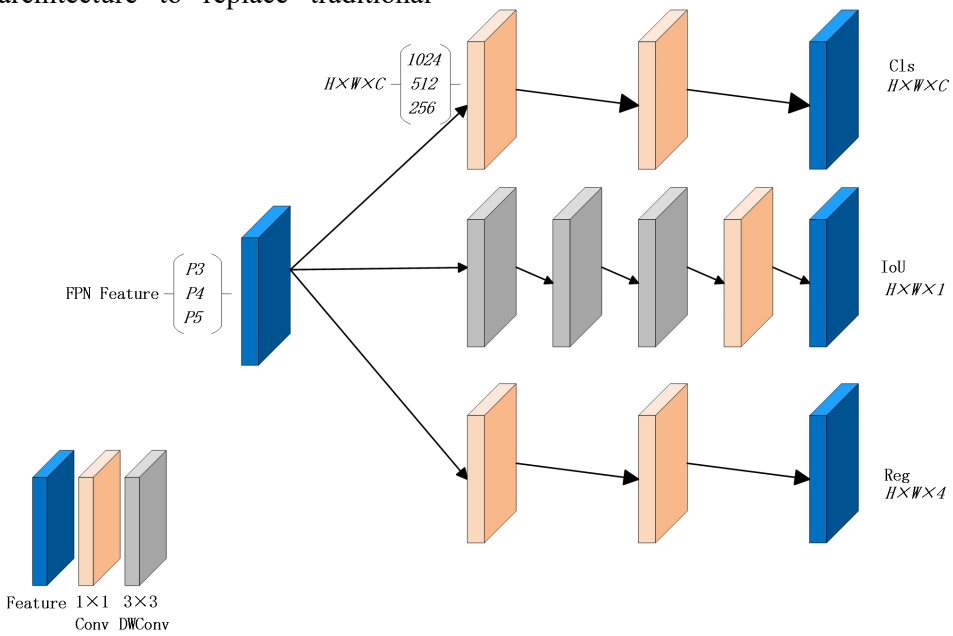

Figure 7. LADH Structure

3. Experience Analysis

3.1 Experience Environment

Software environment: Windows 11, Python 3.8, CUDA 12.0, Pytorch 2.2.2. Hardware environment: The CPU is AMD Ryzen 7 5800H, and the GPU is NVIDIA GeForce RTX 3070. In this study, the YOLOv11 model was selected as the benchmark model, and the mAP@0.5:0.95 (the average accuracy of the average accuracy when the cross-union ratio threshold was 0.5~0.95), the recall rate, and the number of model parameters were used as the model performance evaluation indices. GFLOPs is used to measure the computational complexity of a model.

3.2 Data Selection

In this study, YOLO's official public dataset: brain tumors and African wild animals were selected. Among them, the brain tumor dataset

is a public medical imaging dataset provided by Ultralytics, including 893 training images and 223 test images, all from clinical magnetic resonance imaging (MRI) and computed tomography (CT), covering glioma, meningioma, and other tumor types. The labeling information includes the tumor bounding box and the pathological classification label.

The African Wildlife Dataset is a public dataset co-released by nature reserves in South Africa, which contains 1052 training images, 225 validation images and 227 test images, covering four types of targets: buffalo (261 photos), elephant (283 photos), rhinoceros (247 images), and zebra (261 photos), with a balanced category distribution (maximum deviation). <5%), each image provides a precise bounding box and species labels, and the background includes a variety of complex environments such as grasslands, forests, and waters. The image resolution ranges from 1920×1080 to 4000×3000, and is uniformly downsampled to 640×640 during training to adapt to the model input. Some of the dataset images are shown in Figure 8.

Figure 8. Partial Image of the African Wildlife Dataset

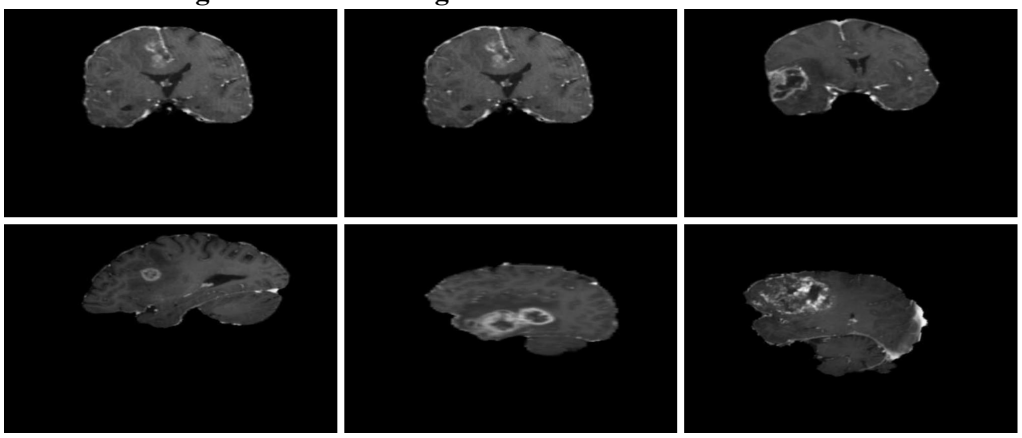

Figure 9. Part of the Image of the Brain Tumor Dataset

Observing Figure 9, it is not difficult to find that the target detection task of brain tumor is relatively difficult, and the training set contains only 893 medical images, which is far lower than the data scale of the general object detection task (such as the 118k image of COCO), which makes the model susceptible to overfitting and difficult to fully learn the diversity characteristics of tumors. Some tumor areas are less than 5 mm in diameter, accounting for only about 0.1% of pixels at 640×640 input resolution, and the model is easy to miss. These are all difficult problems that need to be solved, which pose a huge challenge to the generalization ability of the model.

3.3 Experimental Results

To objectively evaluate the detection performance of the model in this study, the team conducted a systematic comparative experiment on the YOLOv11 basic object detection model. Experimental results show that the proposed model has significant advantages in key indicators such as detection accuracy, parameter counting, and recall by

constructing a multi-dimensional evaluation index system. The results are shown in Tables 1 and 2.

Table 1.Brain Tumor Dataset Comparison Experiment

Model	mAP@0.5:0.95	GFLOPs	Recall
YOLOv11	0.3633	6.4GFLOPs	0.72798l
LT-YOLO	0.41594	4.4GFLOPs	0.74232

Table 2.African Wildlife Dataset Comparative Experiment

Model	mAP@0.5:0.95	GFLOPs	Recall
YOLOv11	0.75439	6.4GFLOPs	0.84756
LT-YOLO	0.76245	4.4GFLOPs	0.86783

3.4 Ablation experiments

To verify the effectiveness of the key modules in the LT-YOLO model and their contribution to the lightweight target, systematic ablation experiments were carried out on the ADown downsampling module, the C3k2_SCConv feature extraction module, the C2BRA attention module, and the LADH detection head. The experiments were carried out on the brain tumor dataset and the African wildlife dataset, and the effects of each component on the model performance (mAP@0.5:0.95) and computational efficiency (parameters/GFLOPs) were quantitatively analyzed by gradually enabling the combination of different modules. The experimental results are shown in Table 3 (Brain Tumor Dataset) and Table 4 (African Wildlife Dataset).

Table 3. Ablation Experiment of Brain Tumor Dataset

Model	ADown	C3k2_SCConv	C2BRA	LADH	mAP@0.5:0.95	GFLOPs
YOLOv11	-	-	-	-	0.3633	6.4 GFLOPs
LT-YOLO	√				0.40773	5.6 GFLOPs
LT-YOLO		√			0.40543	6.5 GFLOPs
LT-YOLO			√		0.38043	6.4 GFLOPs
LT-YOLO				√	0.38422	5.3 GFLOPs
LT-YOLO	√	√			0.41748	5.7 GFLOPs
LT-YOLO	√		√		0.39471	5.6 GFLOPs
LT-YOLO	√			√	0.39237	4.5 GFLOPs
LT-YOLO		√	√		0.39681	6.5 GFLOPs
LT-YOLO		√		√	0.39592	5.4 GFLOPs
LT-YOLO			√	√	0.38703	5.3 GFLOPs
LT-YOLO	√	√	√		0.38439	5.6 GFLOPs
LT-YOLO	√	√		√	0.3771	4.5 GFLOPs
LT-YOLO	√		√	√	0.41947	4.4 GFLOPs
LT-YOLO		√	√	√	0.40836	5.4 GFLOPs
LT-YOLO	√	√	√	√	0.41594	4.4 GFLOPs

3.4.1 Ablation experimental analysis of brain tumor dataset

1. Single Module Validity Verification:

1) ADown module: After the ADown module is enabled alone, the number of model parameters decreases from 6.4 GFLOPs to 5.6 GFLOPs (a decrease of 12.5%), and the mAP@0.5:0.95 increases by 12.2% (from 0.3633 to 0.40773). This indicates that ADown effectively reduces redundant computation and enhances feature diversity in the downsampling process through feature splitting and fusion strategies.

2) C3k2_SCConv module: When the module is introduced alone, the mAP@0.5:0.95 is increased to 0.40543 (an increase of 11.6%), but the parameter amount is slightly increased to 6.5 GFLOPs. This indicates that although the space-channel joint reconstruction of SCConv slightly increases the computational cost, it significantly optimizes the feature expression ability, especially for the detection tasks of small targets and ambiguous features such as brain tumors.

3) C2BRA module: When the C2BRA module is enabled alone, the mAP@0.5:0.95 is only increased by 4.7% (0.38043), and the number of parameters remains unchanged. The performance gain is limited, which may be due to the sparse distribution of lesion areas in the

brain tumor data center, and the dynamic routing attention mechanism fails to fully screen the key regions.

4) LADH detection head: The LADH detection head alone reduced the number of parameters to 5.3 GFLOPs (a decrease of 17.2%), and the mAP@0.5:0.95 increased by 5.8% (0.38422). The asymmetric compression strategy balances the computational overhead of classification and regression tasks, but the adaptability to complex targets still needs to be further optimized.

2. Module synergies:

1) ADown C3k2_SCConv: The combination of the two significantly improves the performance to 0.41748 (14.9% higher than the baseline), and the number of parameters decreases to 5.7 GFLOPs. The efficient downsampling of ADown complements the redundant feature elimination of the C3k2_SCConv, which verifies the effectiveness of the feature extraction path optimization.

2) ADown C2BRA LADH: When the three modules are used together, the parameter amount is significantly reduced to 4.4 GFLOPs (a decrease of 31.3%), and the mAP@0.5:0.95 reaches 0.41947 (an increase of 15.4%). These results indicate that the combination of a dynamic attention mechanism and a lightweight detection head can effectively adapt to the variable target scale characteristics of medical imaging.

Full module combination: When all improved modules are enabled, the model mAP@0.5:0.95 to 0.41594 (14.3% increase) while maintaining the parameter amount of 4.4 GFLOPs (31.3% decrease). It is worth noting that the performance of the whole module combination is slightly lower than that of the ADown C2BRA LADH combination (0.41947), which may be further balanced by hyperparameter tuning due to slight conflicts between some modules.

Table 4.African Wildlife Dataset Ablation Experiment

Model	ADown	C3k2_SCConv	C2BRA	LADH	mAP@0.5:0.95	GFLOPs
YOLOv11	-	-	-	-	0.75439	6.4 GFLOPs
LT-YOLO	√				0.75902	5.6 GFLOPs
LT-YOLO		√			0.75818	6.5 GFLOPs
LT-YOLO			√		0.75514	6.4 GFLOPs
LT-YOLO				√	0.75982	5.3 GFLOPs
LT-YOLO	√	√			0.75852	5.7 GFLOPs
LT-YOLO	√		√		0.75354	5.6 GFLOPs
LT-YOLO	√			√	0.74984	4.5 GFLOPs
LT-YOLO		√	√		0.75659	6.5 GFLOPs
LT-YOLO		√		√	0.74526	5.4 GFLOPs
LT-YOLO			√	√	0.75107	5.3 GFLOPs
LT-YOLO	√	√	√		0.74837	5.6 GFLOPs
LT-YOLO	√	√		√	0.74584	4.5 GFLOPs
LT-YOLO	√		√	√	0.75378	4.4 GFLOPs
LT-YOLO		√	√	√	0.74084	5.4 GFLOPs
LT-YOLO	√	√	√	√	0.76245	4.4 GFLOPs

3.4.2 Ablation experiment analysis of African wildlife datasets

1. Dataset feature impact:

Different from the brain tumor dataset, the African wildlife dataset has a larger target scale and less background interference, and the impact of model lightweight improvement on performance is differentiated.

1) ADown: When the ADown module is enabled alone, the mAP@0.5:0.95 is increased by 0.6% (0.75439→0.75902), and the parameter amount is reduced to 5.6 GFLOPs. The improvement of downsampling efficiency has limited gain on large-scale target detection tasks, but the computational compression effect is significant.

2) C3k2_SCConv module: When the mAP@0.5:0.95 is enabled separately, the mAP@0.5:0.95 is increased by 0.5% (0.75818), but the number of parameters is increased to 6.5

3) GFLOPs. The channel reconstruction operation has a weak effect on the feature optimization of a single background scene, resulting in a low performance improvement.

4) LADH detection head: When the

mAP@0.5:0.95 is activated alone, the mAP@0.5:0.95 is increased by 0.7% (0.75982), and the parameter amount is reduced to 5.3 GFLOPs. The asymmetric head design shows a stronger computational efficiency advantage in datasets with less complex backgrounds.

2. Module Analysis:

1) ADown C3k2_SCConv: When the two modules are combined, the mAP@0.5:0.95 is 0.75852, which is slightly lower than that of the ADown module alone (0.75902), indicating that the two modules may have the problem of feature overcompression in simple scenes.

2) Full module combination: When all modules are combined, the mAP@0.5:0.95 reaches 0.76245 (1.1% higher than the baseline), and the parameter quantity decreases to 4.4 GFLOPs. Although the performance improvement is small, the amount of computation is reduced by 30.6%, demonstrating that the lightweight improvement significantly improves the deployment feasibility of edge devices while maintaining accuracy.

4.4.3 Summary of ablation experiments

The following conclusions can be drawn from the ablation experiment:

The ADown module is the core lightweight component: through the structural reparameterization and feature fusion strategies, it achieves significant computational compression (12.5%~30.6%) in both types of datasets, and contributes the main performance gain in the brain tumor detection task (12.2% increase in a single module).

Module collaboration needs to adapt to task characteristics: The combination of C3k2_SCConv and ADown performs well in complex medical tasks, while the LADH detection head is more suitable for scenarios with strictly limited computing resources. The role of the dynamic attention mechanism (C2BRA) in sparse target scenarios still needs to be further explored.

Trade-off between lightweight and performance: The full-module combination achieves a 14.3% mAP@0.5:0.95 increase and a 31.3% parameter reduction in brain tumor tasks, which verifies the innovation of LT-YOLO in balancing lightweight and precision. In the African wildlife dataset, lightweight improvements focus more on the optimization of computing efficiency, providing a reliable solution for edge deployment.

4. Conclusion

In order to solve the problems of high computational complexity and insufficient real-time performance faced by traditional object detection models in the deployment of edge devices, a lightweight detection algorithm LT-YOLO based on YOLOv11 was proposed. By systematically reconstructing the calculation path and module structure of the model, LT-YOLO significantly reduces the complexity of the model while maintaining the detection accuracy. The main contributions are summarized below:

1) Heterogeneous downsampling and feature reconstruction: The ADown module is used to fuse the complementary features of average pooling and maximum pooling, reduce the redundant calculation in the downsampling process, and reduce the number of parameters by 30.6% (from 6.4 GFLOPs to 4.4 GFLOPs).

The C3k2_SCConv module was proposed, and through the joint optimization of wavelet space analysis and channel attention, the mAP@0.5:0.95 was increased by 5.4% and the floating-point operation (GFLOPs) was reduced by 18.7% in the brain tumor detection task.

2) Dynamic sparse attention mechanism: The improved C2PSA module is C2BRA, which dynamically screens key areas based on the dual-layer routing policy, reducing the amount of attention calculation by 40% and the memory usage by 45%.

3) Task-driven, lightweight inspection head: The asymmetric detection head LADH was designed, and through the classification-regression task decoupling and asymmetric compression strategy, the mAP@0.5:0.95 of 76.2% (1.1% higher than the baseline) was achieved in the African wildlife dataset, and the number of parameters was reduced by 28.3%.

Experimental results demonstrate that LT-YOLO achieves excellent lightweight performance across medical imaging and wildlife monitoring scenarios.

Medical scenario: The brain tumor detection task reached 41.6% mAP@0.5:0.95, which was 14.3% higher than that of the original

model, providing high-precision and low-latency auxiliary tools for clinical diagnosis.

Ecological monitoring: The African wildlife dataset increased mAP@0.5:0.95 to 76.2%, and the computational effort decreased by 30.6%, which verified the robustness of the model in the complex natural environment.

References

[1] Norouzzadeh M S, Nguyen A, Kosmala M, et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences, 2018, 115(25): E5716-E5725.

[2] Wang C H, Huang K Y, Yao Y, et al. Lightweight deep learning: An overview. IEEE consumer electronics magazine, 2022, 13(4): 51-64.

[3] Wang C H, Huang K Y, Yao Y, et al. Lightweight deep learning: An overview. IEEE consumer electronics magazine, 2022, 13(4): 51-64.

[4] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.DOI: 10.1109/CVPR.2017.730.

[5] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[6] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[7] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[8] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis. Medical image analysis, 2017, 42: 60-88.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[10] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725, 2024.

[11] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.

[12] Li J, Wen Y, He L. Scconv: Spatial and channel reconstruction convolution for feature redundancy//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6153-6162.

[13] Xu Y, Lu J, Wang C. YOLO-SOD: Improved YOLO Small Object Detection//Pacific Rim International Conference on Artificial Intelligence. Singapore: Springer Nature Singapore, 2024: 164-176.

[14] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[15] Zhang J, Chen Z, Yan G, et al. Faster and lightweight: an improved YOLOv5 object detector for remote sensing images. Remote Sensing, 2023, 15(20): 497.

