A Cataract Detection Method Based on Improved YOLO11n

Yixiang Shang[1], Luping Qian[1], Haoran Li[1], Mengqi Liu[1], Mingyue Xue[2], Mingxu Li[3]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China
[2]School of Economics and Trade, Henan University of Technology, Zhengzhou, Henan, China
[3]IFLYTEK Co., Ltd., Hefei, Anhui, China

Abstract: Cataract is renowned as the world's leading cause of blindness. Current YOLO series models suffer from deficiencies such as severe feature interference, low detection rate for small lesions, and significant impact of artifacts. To address these issues, this paper proposes the ECL-YOLO11 model based on an improved YOLO11 approach. YOLO11 is divided into two layers: feature extraction and feature fusion. In the feature extraction stage, the backbone network is replaced with EfficientViT to enhance global feature modeling capability, and C3k2 is substituted with the C3k2_ContextGuided module to strengthen multi-scale context acquisition. In the feature fusion stage, SPPF is replaced with SPPF_LSKA to optimize micro-lesion response. In cataract detection experiments, the ECL-YOLO11n model achieves an accuracy of 95.52%, recall of 99.79%, F1-Score of 97.60%, $mAP_{0.5}$ of 99.35%, and $mAP_{0.5-0.95}$ of 77.46%. Compared with the YOLO11n baseline model, these metrics are improved by 2.34 to 3.7 percentage points. Compared with mainstream models such as Faster R-CNN and YOLOv8n, it shows varying degrees of improvement in classification metrics and an increase of 1.15 to 1.16 percentage points in target detection metrics. The $mAP_{0.5}$ values for "no cataract" and "with cataract" samples are 99.46% and 99.20% respectively.

Keywords: Cataract Screening; EfficientViT; C3k2_ContextGuided; SPPF_LSKA; YOLO11

1. Introduction

With the further acceleration of population aging, the issue of eye health among the elderly has become increasingly prominent. According to reports, cataract is a typical, progressive lens disease and currently the world's leading cause of blindness [1]. Approximately 94 million people worldwide suffer from moderate to severe visual impairment caused by cataracts, and more than two-thirds of them are elderly over 65 years old living at home or in collective elderly care institutions. Traditional cataract screening requires relevant equipment and professional medical staff, which cannot meet the needs of the large base of high-risk groups; at the same time, the initial symptoms of the disease are not easy to detect, which can easily delay treatment and cause irreversible visual impairment. Therefore, it is urgent to develop non-invasive early screening technology suitable for home-based health care and primary medical care.

Cataract detection technologies are divided into traditional image processing and deep learning-based methods. Traditional image processing methods manually design feature extraction operators, but they are easily affected by imaging quality and artifacts, have poor robustness and generalization ability, and cannot meet complex ocular imaging scenarios. The data-driven approach based on deep learning has become the main research method currently. A typical representative of Two-Stage is the Faster R-CNN [2] algorithm, which completes the target through two stages: region proposal and classification detection. Although it has high detection accuracy, it has high computational complexity and slow inference speed, making it unsuitable for edge devices. Classic representatives of One-Stage are SSD [3] and YOLO series, which have small model parameters and short inference delay, meeting the real-time requirements of embedded low-computing-power hardware and real-time analysis of fundus imaging. There have been many related research works on cataract detection based on deep learning. Sun et al [4]. proposed a dual-branch feature fusion classification network and an Edge-Enhanced

Dual Attention Segmentation Model (EE-DANet). Aiming at the difficulties such as similar image features of multiple ophthalmic diseases in fundus images and surgical images, mirror flipping and scale transformation of surgical images, they used self-attention and CNN dual-branch methods, multi-scale feature extraction modules, strip coordinate attention and other means to achieve accurate classification of multiple ophthalmic diseases and efficient segmentation of key regions in surgical images. Zhang et al [5]. constructed a global feature CNN (AlexNet), a local feature CNN and a global-local hybrid feature integration model. Through the visualization of the overall convolution inverse operation, they found the information loss of microvessels that cannot be well reflected by global features. They improved the model by combining the local detail features extracted from the variant dataset with ensemble learning to form the final hybrid model. Tests show that the classification accuracy of the hybrid model reaches 86.24%, which is significantly better than the single-feature model. Aiming at the problems of high noise interference in ultrasound images, limited equipment conditions in primary medical institutions, and lack of datasets, Wang et al [6]. implemented a YOLO-DS model for ultrasound images, which combines two-stage YOLOv3 region selection and DenseNet-161 classification. First, the first-stage model is used to accurately detect the eyeball and lens regions to remove redundant parts, and then the dense connection network in the second-stage model is used to enhance the advanced feature extraction ability. The average detection accuracy on the self-made ultrasound dataset reaches 90%, which is at the expert level.

Even though there are many deep learning methods available for cataract detection, there are still problems such as difficulty in distinguishing the differences between features, difficulty in effectively detecting early small lesions, and susceptibility to artifacts. In this case, hierarchical optimization is carried out based on the YOLO11n architecture, which is designed according to the technical logic of "feature extraction - feature fusion".

2. YOLOv11 Detection Algorithm

The main framework of the YOLO11 target detection model consists of three parts: backbone feature extraction network, neck feature fusion network, and detection head. The network structure is shown in Figure 1. The Conv convolution module extracts features of the input image, C3k2 is used for feature fusion and feature extraction, SPPF is used for feature information at different scales [7], C2PSA enhances feature extraction ability and target detection accuracy, Upsample is an upsampling module used to generate high-resolution feature maps. Concat is a concatenation module that concatenates different feature maps on the channel dimension [8].

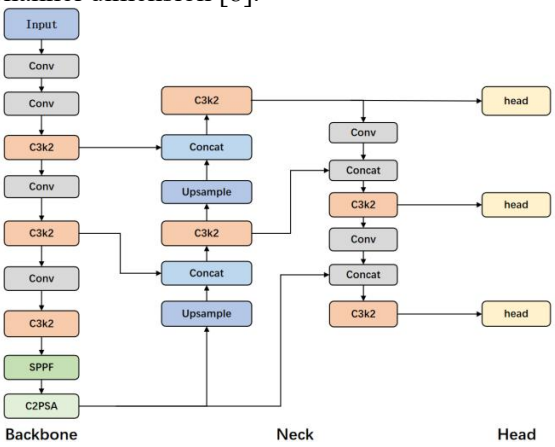

Figure 1. YOLO11 Architecture Diagram

3. Model Construction

Aiming at the problems of severe feature interference and low recognition rate of early small lesions in the YOLO11n baseline model in cataract detection, the model is improved around the goal of "strengthening feature extraction ability and optimizing feature fusion effect", and the improved model ECL-YOLO11n is formed through the improved scheme shown in Figure 2. The main improvement methods are: weakening the feature interference of cataract images themselves during input; improving the accuracy of the target frame output by the model.

Backbone network replacement: The original backbone feature extraction network of YOLO11n is completely replaced with the EfficientViT model, which can be used for global feature modeling of complex textures and boundary information of lens opacification regions, which is conducive to reducing the missed detection of early mild opacification cataracts.

Feature fusion module optimization: The Large Kernel Separable Convolution Attention mechanism (LSKA) is introduced to replace the SPPF module in the original model, so as to replace the original SPPF pooling structure,

enabling it to better discover tiny lesion features while avoiding being affected by artifacts.

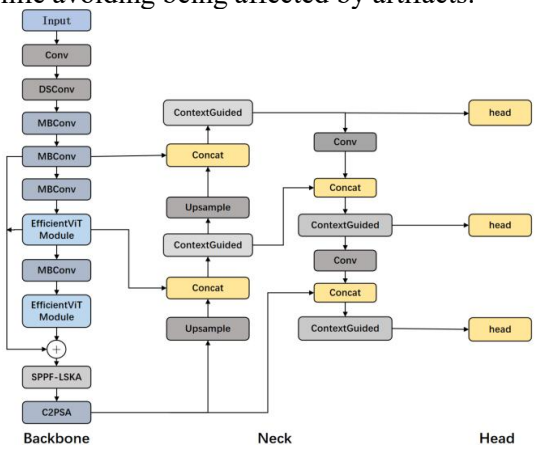

Figure 2. ECL-YOLO11Architecture Diagram

Local feature module upgrade: The module in the original C3k2 is replaced with the C3k2_ContextGuided module, which integrates local and global context information to improve the original problems. On this basis, the problems of insufficient global correlation and poor multi-scale adaptability in the original model are solved, and the detection robustness under complex backgrounds is improved.

3.1 EfficientViT Backbone Network

This section mainly focuses on the EfficientViT backbone network.

The YOLO11n backbone network is mainly to obtain a more lightweight model, which reduces the feature extraction ability, so it cannot obtain good lens opacification textures and edge details. In cataract detection, it is prone to missed detection of early mild lens opacification or false detection of normal aging textures of the lens, which is not conducive to the original intention of "suppressing feature interference and enhancing small lesion feature expression". In summary, EfficientViT is selected as the new backbone. It adapts to the computing power of edge devices through a hierarchical feature extraction method, and strengthens global and local feature modeling, which accurately makes up for the problems of the original module [9].

As can be seen from the EfficientViT Module [10] in Figure 3, the "multi-scale attention + feed-forward network" architecture reflects the corresponding process and helps determine the feature extraction path. Although compared with the original backbone structure, dimension reduction and pruning remove a large number of self-attention layers with high memory occupation, and EfficientViT adds some efficient feed-forward network layers, EfficientViT can still achieve the same level of lightweight as the original backbone structure. After increasing the speed parameter of the video, it can be found that the missed detection rate compared with the baseline has decreased; when increasing the convolution sliding window of the test dataset, it can be found that it still has the advantage of strong anti-opacification ability on the basis of reducing image resolution.

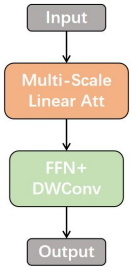

Figure 3. EfficientViT Module Architecture

3.2 SPPF_LSKA Module

First, the original SPPF module does not effectively select useful features, but simply aggregates features using multi-scale pooling. It cannot identify some micro-lesions such as early punctate opacification or mild edge opacification of the lens, and is prone to losing some important information of small lesions; moreover, it cannot distinguish real lesions from artifacts such as fundus reflection and eyelash occlusion, and often mistakenly judges noise as opacification regions. Therefore, this method needs to be improved to avoid its shortcomings.

Taking "first aggregate multi-scale features, then accurately screen effective information" as the basic idea, the basic components and sub-structures are used to optimize features collaboratively. First, the SPPF basic component is responsible for capturing multi-scale features, so as to achieve the effect of covering the feature set of "micro-lesion local details - lesion overall outline", and can completely solve the problem that a single scale is difficult to cover micro-lesions. LSKA [11] focuses on core optimization, splitting the original large-size 2D convolution kernel into a superimposed structure of "horizontal 1D convolution + vertical 1D convolution". On the basis of ensuring the amount of computation, it still has a large receptive field, can capture the local texture of the lesion, and uses dilated convolution to establish long-distance connections of the lesion to form a weight map, which weights the feature

map output by SPPF, that is, highlights effective information such as lens opacification edges and punctate lesions, and weakens invalid information such as artifacts and noise. The specific steps are shown in Figure 4. From the whole process, first, SPPF fully collects information of lesions of different scales, then obtains corresponding optimized features through effective feature enhancement and invalid feature suppression in LSKA, and finally outputs features for cataract detection.

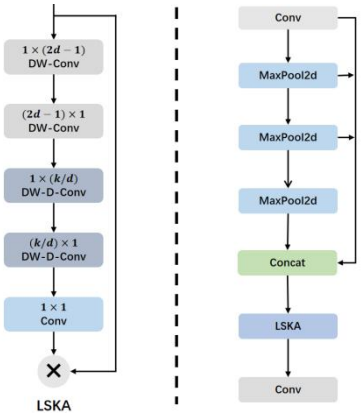

Figure 4. SPPF_LSKA Flow Chart

3.3 C3k2-ContextGuided Module

The C3k2 module in the original model relies heavily on local convolution, leading to insufficient global correlation, poor multi-scale adaptability, and weak robustness under complex backgrounds. To this end, the Context-Guided module is introduced.

The Context-Guided Block is the main body of the lightweight semantic segmentation network CGNet [12]. It brings global and local context information into the model, making it easier for the model to grasp the structure of input data and the characteristics of features, and can obtain higher accuracy with fewer parameters. The ContextGuided Block [13] consists of four parts, as shown in Figure 5: Extractors use 3×3 ordinary convolution to extract local features at positions; Expansive Extractors use 3×3 dilated convolution to extract surrounding context features; Two extractors concatenate the outputs of +, then use BN to accelerate training convergence, use PRelu activation unit to make the network more robust and insensitive to network structure and hyperparameters, and finally obtain joint features; pool finally performs global average pooling and multi-layer perceptron weighting on the joint features, and multiplies the weighted result with the input to obtain global context features.

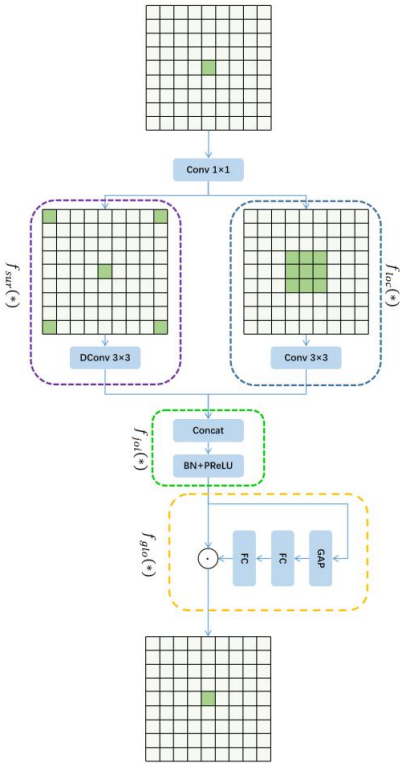

Figure 5. ContextGuided Architecture Diagram

This method can combine local details, neighborhood context and global semantic information. The ContextGuided Block is introduced into the C3k2 module to obtain the C3k2_ContextGuided module, and the specific structure is shown in Figure 6. Compared with C3k2, it reduces the overall size of the network without increasing the number of network parameters; and by introducing spatial dependence and context information, the learned joint features of local features and context features are improved by global context features, which can better capture useful information, remove the influence of redundant and invalid information, and improve the network learning ability.

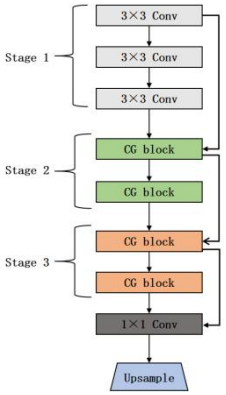

Figure 6. C3k2-ContextGuided Module Structure

4. Experimental Results and Analysis

4.1 Dataset

Based on the public dataset Eye Detection Dataset, among the 1976 samples in the dataset, 1715 samples have cataract features and 264 are normal. The annotation categories are divided into "cataract" (with cataract) and "normal" (without cataract), corresponding to labels 0 and 1; the annotation boxes are represented by normalized coordinates. The cataract detection model is trained by dividing the ocular images into training set, validation set and test set in the ratio of 7: 2: 1 (1385, 396 and 198 respectively).

4.2 Experimental Environment and Parameter Indicators

To ensure the fairness of all model experimental results, this paper uses NVIDIA GeForce RTX 3080Ti GPU, Windows 11, 64-bit operating system. PyTorch 2.2.1 is used as the deep learning framework, CUDA 12.1 as the parallel computing platform, and Python 3.8 as the programming language. The SGD optimizer is used, the training batch size is 32, the number of iterations is 100 epochs, and the initial learning rate is 0.01. The experiment uses accuracy (Precision, P), recall (Recall, R), F1-Score, mean average precision (mAP0.5) and mean average precision (mAP0.5-0.95) as the main evaluation indicators of model performance.

4.3 Comparison Experiments of Different Models

To verify the performance of the improved model, the selected models are trained using the cataract dataset under the same experimental environment and the test set data are collected.

It can be seen from Table 1 that the accuracy of ECL-YOLO11 is 95.52%, the recall rate is 99.79%, the F1-Score is 97.6%, the $mAP_{0.5}$ is 99.35%, and the $mAP_{0.5\text{-}0.95}$ is 77.46%, which are 2.34%, 3.7%, 2.93%, 1.15% and 1.16% higher than the baseline model YOLO11n respectively. Among them, the precision is improved to varying degrees compared with the baseline model YOLO11n, which is because the accuracy of the ECL-YOLO11 algorithm is improved while maintaining good robustness; and after a large number of experimental comparisons, it is concluded that all indicators are leading, which can verify the efficiency of the proposed model in classification and target detection.

In cataract detection, the pupils of cataract patients can present various conditions, and normal people also have differences in pupil size due to different reasons, which makes cataract detection more difficult and puts forward higher requirements for the detection accuracy of the model; avoiding the problem that the improvement of the accuracy of a certain disease affects the overall detection accuracy of the model. Table 2 compares the mAP0.5 values of different disease categories between YOLO11n and ECL-YOLO11 models. It can be seen that the precision of each category of ECL-YOLO11 has been improved, and there is no abnormal change in the indicators of any category. Therefore, it can be verified that the improved model has a good effect on various diseases.

Table 1. Comparison Results of ECL-YOLO11 with Other Models

Models	Precision P/%	Recall R/%	F1-Score/%	$mAP_{0.5}$/%	$mAP_{0.5\text{-}0.95}$/%
Faster R-CNN	82.19	78.28	79.98	65.02	32.46
YOLOv8n	94.27	96.40	95.29	97.95	76.21
YOLOv10n	92.74	90.86	91.68	96.12	74.26
YOLO11n	93.18	96.27	94.67	98.20	76.31
YOLOv12n	91.34	96.80	93.86	98.21	75.12
YOLOv13n	95.48	90.69	92.76	96.46	76.36
ECL-YOLO11n	95.52	99.79	97.60	99.35	77.46

Table 2. Comparison of $mAP_{0.5}$ Indicators for Detecting Cataract

Cataract Status	YOLO11n	ECL-YOLO11n
No Cataract	99.05	99.46
With Cataract	97.36	99.20
All	98.20	99.35

4.4 Ablation Experiments

Ablation experiment is a commonly used deep learning experimental method to judge whether different network branches are beneficial to the entire network model. To better illustrate the impact of the three modules C3k2_ContextGuided, SPPF_LSKA, and EfficientViT on YOLOv11 respectively, and confirm the role of each module in ECL-YOLOv11, the YOLOv11 model is first used as a basic model, and then

C3k2_ContextGuided, SPPF_LSKA, and EfficientViT are gradually introduced for ablation experiments to analyze the contribution of each module and verify its strengthening ability in the overall structure.

It can be seen from Table 3 that the proposed C3k2_ContextGuided, SPPF_LSKA, and EfficientViT can all improve the model detection effect. SPPF_LSKA can increase accuracy and detection precision, but the recall rate decreases slightly. Adding C3k2_ContextGuided enhances the recall rate and improves the average detection value. Then introducing all three modules to form ECL-YOLO11 achieves the best balance when the precision $mAP_{0.5}$=99.35%, $mAP_{0.5\text{-}0.95}$=77.46%, classification performance accuracy Precision P=95.52%, recall rate Recall R=99.79%, and F1-Score=97.60%, which indicates that each module can achieve the advantages of detection performance and resource saving.

Table 3. Ablation Experiment Results

C3k2_ContextGuided	SPPF_LSKA	EfficientViT	Precision *P*/%	Recall *R*/%	F1-Score/%	$mAP_{0.5}$/%	$mAP_{0.5\text{-}0.95}$/%
			93.18	96.27	94.67	98.20	76.31
√			89.76	97.98	93.69	98.05	77.12
	√		97.24	93.18	95.13	98.21	77.46
		√	91.14	97.65	94.27	98.11	76.50
√		√	96.32	89.16	92.54	96.85	75.75
	√	√	95.70	89.36	92.17	95.63	75.42
√	√		93.13	96.55	94.79	97.42	77.48
√	√	√	95.52	99.79	97.60	99.35	77.46

4.5 Visualization of Experimental Results

This chapter visualizes the detection results of the YOLO11n model and the ECL-YOLO11 model on the dataset.

It can be seen from Figure 7 that YOLO11n has missed detection and false detection when detecting cataract diseases. In contrast, the improved ECL-YOLO11 model has significantly improved detection accuracy and stability in determining whether there is a cataract target, showing better target positioning ability and higher confidence distribution.

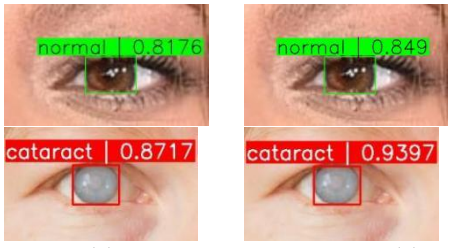

Figure 7. Detection Results of YOLO11n and ECL-YOLO11 Models

5. Conclusion

The main achievement of this research is the development of the ECL-YOLO11n model, which can meet the deployment requirements of cataract screening systems in resource-constrained scenarios such as mobile devices and embedded devices with a smaller model size. It also breaks through the previous problems of small coverage of traditional cataract screening and the defects of existing YOLO models in detection. Through iterating "measures - problems - effects", the iterative closed loop is improved. In "measures - problems - effects", EfficientViT is used to replace the original backbone, which solves the problem of weak global feature modeling of the original backbone and easy missed detection of small lesions in the early stage, and improves the recognition ability of early lens opacification; the SPPF_LSKA module is used to replace the original SPPF module, which enhances its perception ability of tiny details of micro-lesions and improves the suppression ability of artifacts; the C3k2-ContextGuided module is used to replace the original C3k2, which supplements context information, reduces background misjudgment, and improves the multi-scale and anti-interference ability of the model.

References

[1] Yang Bin, Chen Yi, Zong Yixin. Modern Light Environment, Visual Health Problems and Research Status. Science and Technology China, 2020, (02):15-17.

[2] Ren Shaoqing, He Kaiming, Girshick Ross & Sun Jian. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.https://doi.org/10.1109/TPAMI.2016.2577031.

[3] Li Zhenwei. Improvement and System Design of Strawberry Fruit Detection Method Based on Deep Learning. Nanjing University of Information Science and Technology, 2025. DOI:10.27248/d.cnki.gnjqc.2025.000895.

[4] Sun Bin. Research on Cataract Auxiliary Detection and Surgical Image Segmentation Algorithm Based on Deep Learning. Taiyuan University of Technology, 2024. DOI:10.27352/d.cnki.gylgu.2024.001586.

[5] Zhang Linlin. Research on Automatic Cataract Classification Based on Deep Learning. Beijing University of Technology, 2018.

[6] Wang Yong. Ultrasound Cataract Detection Algorithm Based on Deep Learning. Modern Computer, 2021, (11):97-101.

[7] Chen Sihan, Liu Yong, He Xiang. Factory Pedestrian Detection Algorithm Based on Improved YOLOv8. Modern Electronics Technique, 2024, 47(24):160-166. DOI: 10.16652/j.issn.1004-373x.2024.24.025.

[8] Zhang Xinran, Wang Xinzhong, Zhou Kuizhou, et al. Multi-Target Dairy Cow Daily Behavior Detection in Complex Environments Based on Improved YOLO11n. Transactions of the Chinese Society of Agricultural Engineering, 2025, 41(14):155-164.

[9] Xiang Houxue, Xu Guiyang, Zhang Yuhua, et al. Intelligent Recognition Algorithm for Typical Rail Defects Based on Improved YOLOv8. Science Technology and Engineering, 2025, 25(18):7785-7792.

[10] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks // International Conference on Machine Learning. Chambridge: PMLR, 2019: 6105-6114.

[11] Wu Shichen, Mao Yuming, Hu Huizhong, et al. Grape Leaf Disease Detection Method Based on Improved YOLO11n. Transactions of the Chinese Society of Agricultural Engineering, 2025, 41(14):140-147.

[12] Zhang Z, PANG Y. CGNet: cross-guidance network for semantic segmentation. Science China (Information Sciences), 2020, 63(02):49-64.

[13] Li Jun, Zhou Keyu, Zou Jun, et al. Detection Algorithm for Protective Equipment Wearing in Construction Scenarios Based on Improved YOLOv8n. Journal of Zhengzhou University (Engineering Science Edition), 2025, 46(03):19-25+104. DOI:10.13705/j.issn.1671- 6833.2025.03.002.

