AI

MDPI

Article

A Real-Time Laryngeal Disease Diagnosis Algorithm on Edge-AI

Yarong Liu [1,2], Dong Leng [1], Xiaolan Xie [1,2,*] and Zhiyu Li [3]

[1] School of Computer Science and Engineering, Guilin University of Technology, Guilin 541006, China; hblyr@glut.edu.cn (Y.L.); lengdong@glut.edu.cn (D.L.)
[2] Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541006, China
[3] Guilin People's Hospital, Guilin 541002, China
* Correspondence: xxl@glut.edu.cn

Abstract

Background: Laryngeal lesions represent a significant clinical challenge due to the complexity of the laryngeal structure, making manual diagnosis time-consuming and prone to subjective errors. Therefore, developing an accurate and lightweight automatic detection method is essential for improving the efficiency of laryngeal disease screening and diagnosis. **Methods:** This study proposes MSBA-YOLO, a lightweight laryngeal disease detection algorithm based on an improved YOLOv5s architecture. The method integrates FasterNet as the backbone network to reduce computational redundancy through partial convolutions and incorporates a Single-Head Self-Attention mechanism to capture long-range dependencies in complex lesion features. In addition, an MSBA-FIoU loss function is introduced to enhance the localization accuracy of multi-scale targets. **Results:** Experimental results show that MSBA-YOLO achieves a mean Average Precision (mAP) of 96.1% with a model size of only 6.4 MB, representing a 54.6% reduction in parameters compared with the baseline model. When deployed on the Jetson Orin Nano edge platform, the proposed method achieves real-time inference with a speed exceeding 50 FPS while maintaining low power consumption of 5.82 W. **Conclusions:** The results demonstrate that MSBA-YOLO effectively balances detection accuracy and computational efficiency, providing a robust and practical solution for portable and real-time clinical screening of laryngeal diseases on edge devices.

Keywords: laryngeal lesions; lightweight; loss function; edge platform

1. Introduction

Vocal cords are the folds of tissues related to voice creation. Benign tumors, such as cysts, granulomas, leukoplakia, nodules, and polyps, generated in the vicinity of the vocal cord, can induce several clinical complications, including wheezing, stridor, dysphonia, cough, asthma, and vocal cord palsy, and progress to malignant tumors or even cancer [1]. The diagnosis of laryngeal disease primarily relies on endoscopic examination, a process traditionally evaluated manually by trained physicians. However, due to the complex anatomical structure of the larynx, this approach is not only labor-intensive and time-consuming but also highly dependent on the practitioner's expertise and subjective interpretation. The advent of deep learning technology has significantly revolutionized numerous facets of the medical domain, ushering in an unparalleled epoch of precision and efficiency [2]. Consequently, leveraging deep learning technology for the recognition

Check for updates

Academic Editor: Achintya K. Bhowmik

Received: 4 February 2026
Revised: 13 March 2026
Accepted: 14 March 2026
Published: 18 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

and classification of laryngoscope images presents new opportunities for the auxiliary diagnosis of laryngeal disease.

Deep learning-based medical image classification algorithms offer distinct advantages, including autonomous learning capabilities, high diagnostic accuracy, and robust generalization performance. Especially in the upper aerodigestive tract endoscopy, such technology has demonstrated unprecedented power for automating and standardizing the analysis of laryngoscopic images [3]. For example, Baldini et al. [4] proposed a deep learning technique for real-time classification of laryngoscope images to automatically identify diagnostically informative frames. They utilized the ResNet-50 convolutional neural network, pre-trained via a self-supervised rotation prediction task to enhance the model's feature extraction capacity. Xiong et al. [5] developed a method using a deep convolutional neural network (DCNN) for the automatic diagnosis of laryngeal cancer and pre-cancerous lesions from laryngoscopic images. The network was initialized with ImageNet [6]-pre-trained weights to incorporate external contextual information, thereby improving the detection model's performance. Wei et al. [7] introduced the Non-Local Du-al-Stream Fusion Network, where YOLO first identifies regions of interest in laryngoscope images. The system processes data through parallel CNN and Transformer branches, with feature fusion modules integrating dual-stream inputs to enhance representation learning. Shao et al. [8] developed a hierarchical dynamic CNN for laryngoscope image classification, using a shallow ResNet18 for simple samples and a deep ResNet34 for complex cases. A feature reuse module merges shallow-network features into the deep network, improving classification accuracy. Ren et al. [9] designed a ResNet-101-based computer-aided diagnosis system to distinguish vocal cord nodules, polyps, mucosal leukocytes, and malignant tumors, demonstrating that deep learning classifiers outperform clinical visual assessment for most laryngeal lesion categories. Yin et al. [10] released the Laryngoscope8 dataset, containing normal and seven abnormal laryngeal image categories, and proposed a DenseNet-121 model with attention mechanisms, achieving 73% classification accuracy. Xiong et al. [11] developed a Mask R-CNN system for automated classification and segmentation of laryngeal leukoplakia, using ResNet-50 as the backbone and MS-COCO pre-trained weights. Optimized multi-task loss functions enabled the model to achieve 92% average precision in differentiating high-grade dysplasia/cancer (surgery-required) from low-grade dysplasia cases. Tonay et al. [12] fine-tuned pre-trained CNN models for histological laryngeal cancer image classification, with ResNet50 achieving the best test-set performance. Xin et al. [13] enhanced YOLOv8 with a multi-scale convolution module and fully convolutional design to improve small-lesion detection in laryngeal endoscopic images. Baldini et al. [14] proposed SRE-YOLO, a YOLOv8-derived model with a super-resolution branch for multi-scale lesion detection in laryngoscopic images. The branch is removed post-training to maintain efficiency while boosting detection performance over the YOLOv8 baseline.

Currently, in the field of laryngeal disease identification, many deep learning-based diagnostic methods face challenges due to a lack of emphasis on lightweight model design. Traditional models in this area are complex and have a large number of parameters, relying on high-performance computing resources during training and inference. This re-source-intensive nature makes them less compatible with portable medical devices that have limited computing and storage capabilities, hindering the practical application of deep learning in settings like mobile healthcare and rapid on-site diagnosis. To effectively integrate deep learning into medical devices, models need to meet the dual criteria of accurate diagnosis and lightweight implementation. When dealing with complex laryngeal medical images, models must excel in feature extraction to accurately differentiate various lesion characteristics, ensuring reliable and precise diagnostic results. Simultaneously, optimizing the model architecture is crucial to reduce computational complexity and

memory usage, enabling seamless operation on resource-constrained portable devices. This optimization facilitates real-time, efficient diagnostic inference that meets the strict time constraints of clinical diagnosis.

To address the aforementioned issues, this study proposes MSBA-YOLO, an enhanced laryngeal lesion diagnostic model based on YOLOv5s, optimized across multiple dimensions including network architecture, loss function refinement, and lightweight design. Furthermore, the model was deployed on an embedded platform for diagnostic performance analysis. The primary research contributions of this paper are summarized as follows:

1. Construction of a Lightweight Backbone: To address the constraints of limited computational power on embedded devices, the FasterNet architecture is introduced to replace the original backbone. By leveraging Partial Convolution, the model effectively reduces redundant computations and memory access frequency during feature extraction, significantly lowering spatial complexity and computational overhead while maintaining high performance.
2. Integration of Global Feature Perception: To overcome the insufficient feature extraction capabilities of lightweight networks when dealing with complex laryngeal lesions, a Single-Head Self-Attention mechanism is integrated into the feature fusion layers. This mechanism captures long-range dependencies within images with minimal parameter overhead, enhancing the model's ability to discriminate features of irregular pathological targets.
3. Optimization of Multi-Scale Regression Loss: Considering the characteristics of large scale variations and blurred boundaries in laryngeal lesions, the MSBA-FIoU loss function is proposed and applied. By introducing an auxiliary scaling factor and a focusing weight mechanism, this function adaptively balances the regression gradients of multi-scale samples, further improving the localization precision for subtle lesions.
4. Deployment and Validation on Edge Computing Platform: To verify the practical clinical potential, MSBA-YOLO was deployed on the NVIDIA Jetson Orin Nano edge platform using the DeepStream development framework. The experiments not only evaluated the model's high-frame-rate diagnostic performance at the edge but also analyzed its operational stability under low power consumption, providing technical validation for the development of portable real-time auxiliary diagnostic devices for laryngeal diseases.

2. Materials and Methods

2.1. Laryngoscope Datasets

The laryngoscopic images utilized in this study were derived from two open-source datasets: the Laryngeal NBI-Visible Light Dataset [15] and the Laryngoscope8 Dataset [10]. The Laryngeal NBI-Visible Light Dataset, an open-access resource released by Xidian University via the National Public Science Data Center for Basic Disciplines, comprises 559 laryngoscopic images. These images were categorized by senior clinicians into four classes: normal vocal folds, vocal fold nodules, vocal fold polyps, and vocal fold masses.

The publicly available Laryngoscope8 Dataset contains 3057 laryngoscopic images, which are divided into eight categories based on normal laryngeal morphology and the presence of various pathological conditions, including glottic carcinoma, Reinke's edema, granuloma, vocal fold leukoplakia, vocal fold cysts, vocal fold nodules, and vocal fold polyps. All images in this dataset were collected intraoperatively during head and neck surgical procedures at the Department of Otorhinolaryngology, the Sixth Medical Center of the Chinese PLA General Hospital, Beijing, China.

To augment the dataset, laryngoscopic images were subjected to rotation and cropping operations, yielding a final set of 7156 valid laryngoscopic images for the present study. All images were annotated using the LabelImg (1.8.6 Version) tool.

Representative examples of different laryngeal lesion categories are illustrated in Figure 1. These examples include glottic cancer, Reinke's edema, granuloma, vocal cord cyst, vocal cord leukoplakia, vocal cord polyps, vocal cord nodules, and normal laryngeal structures. Glottic cancer typically presents as irregular masses or ulcerative lesions on the vocal cords. Reinke's edema is characterized by diffuse swelling of the vocal folds caused by fluid accumulation in the superficial lamina propria. Granulomas usually appear as localized inflammatory nodules at the posterior glottis. Vocal cord cysts appear as smooth submucosal bulges within the vocal fold tissue. Vocal cord leukoplakia is characterized by white plaque-like lesions with blurred boundaries. Vocal cord polyps typically present as unilateral pedunculated or sessile lesions on the vocal folds, while vocal cord nodules are often bilateral symmetric protrusions caused by chronic vocal overuse. In contrast, normal samples showed symmetrical vocal folds without visible pathological abnormalities.

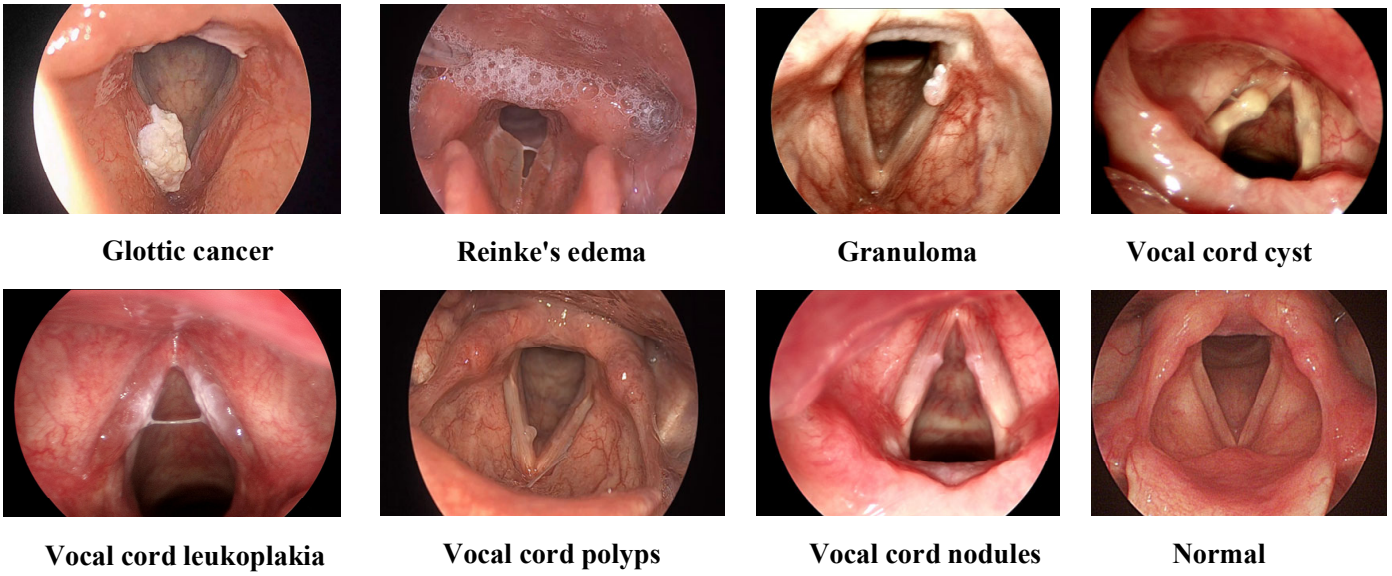


Figure 1. Representative examples of different laryngeal lesion categories in the dataset.

Among the 7156 images, only images containing pathological lesions were annotated with bounding boxes, while images with normal vocal fold structures were retained as negative samples and therefore do not appear in the bounding box annotation statistics.

The statistical distribution of lesion categories and bounding box annotations is presented in Figure 2. Statistical analysis shows that the center points of the lesions are highly concentrated near the center of the field of view, which is consistent with the acquisition characteristics of clinical laryngoscopy examinations, where physicians typically align the suspected lesion region within the center of the endoscopic view to facilitate observation and diagnosis. However, the lesion sizes are generally small, with widths and heights primarily distributed within the range of [0.05, 0.25], and they exhibit diverse shapes. These annotation characteristics—characterized by high central aggregation, extremely small target scales, and highly irregular morphologies—constitute the primary technical challenges for automated laryngeal lesion recognition and provide the experimental motivation for the lightweight multi-scale perception architecture proposed in this study.

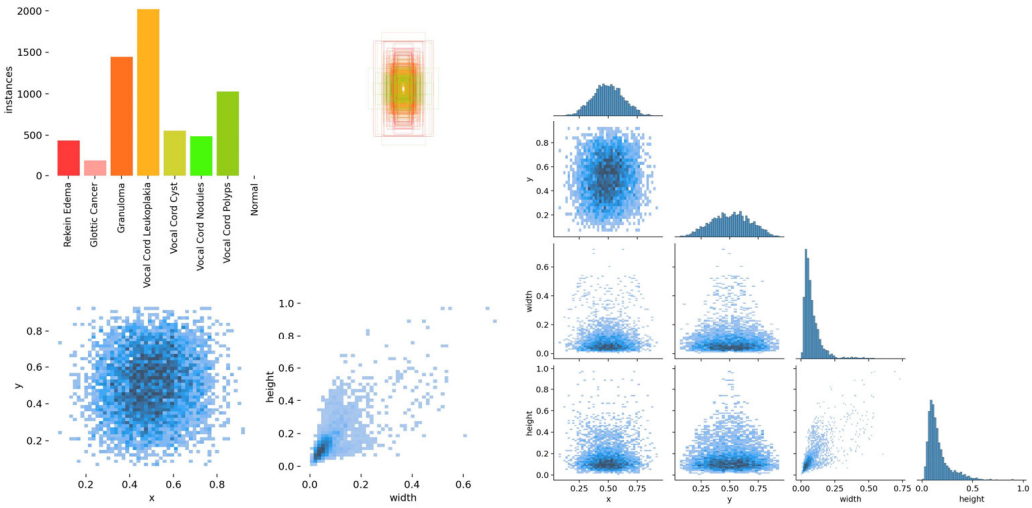


Figure 2. Schematic diagram of the label distribution for the laryngoscopy dataset.

To mitigate potential spatial bias during model training, several data augmentation strategies, including random rotation, cropping, and scaling, were applied to increase the spatial diversity of lesion locations. These augmentations help the model learn more robust spatial representations and improve its ability to detect lesions appearing near the boundaries of the image.

2.2. Baseline Network: YOLOv5

Although YOLOv8 has shown strong performance in many object detection tasks, YOLOv5s was selected as the baseline in this study for several practical considerations. YOLOv5 provides a highly stable and mature architecture that has been widely validated in lightweight deployment scenarios. In addition, its well-established engineering ecosystem and strong compatibility with TensorRT (10.3.0 Version)-based inference pipelines make it particularly suitable for real-time deployment on resource-constrained edge devices such as the Jetson platform. Therefore, YOLOv5s was adopted as the baseline architecture for further optimization in this work.

YOLOv5 is a single-stage object detection system based on Convolutional Neural Networks. It frames object detection as a regression problem by predicting spatial bounding boxes and class probabilities simultaneously. The overall architecture comprises three main components: the backbone network, the neck network, and the detection head. The backbone network adopts the CSPNet architecture to extract features such as object positions and sizes within an image. The neck integrates Spatial Pyramid Pooling and Path Aggregation Network for enhanced multi-scale feature map processing after feature extraction. The detection head serves as a connected mechanism that outputs the final object locations and corresponding label information [16]. The architecture of the YOLOv5 network is illustrated in Figure 3:

According to the scaling coefficients, YOLOv5 provides models in various sizes, including N/S/M/L/X scales, to meet the requirements of diverse scenarios [17]. The YOLOv5s model demonstrates moderate accuracy and has relatively fewer features compared to other YOLOv5 variants [18]. Given its compatibility with resource-constrained medical devices, we chose the YOLOv5s variant for improvement.

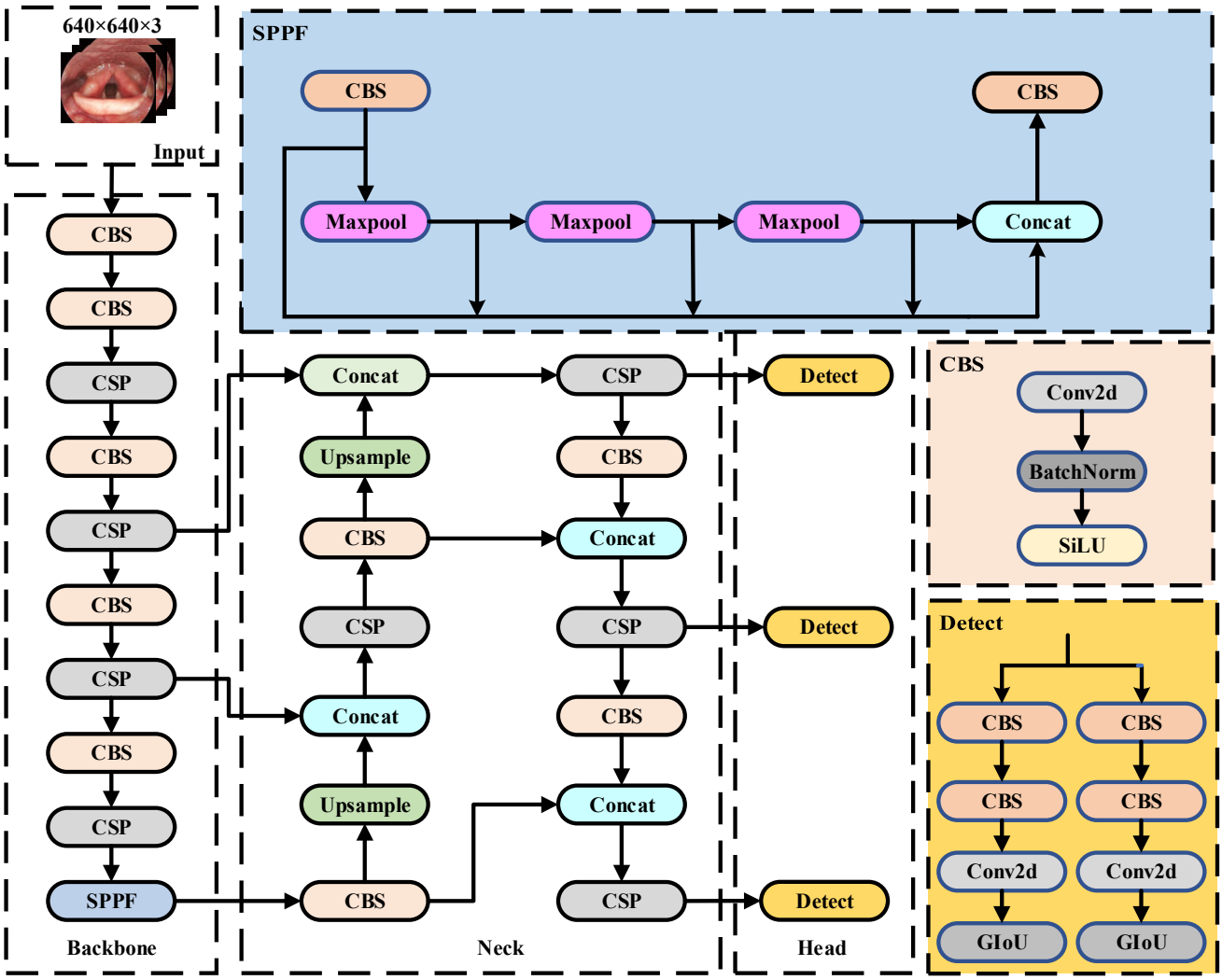


Figure 3. The architecture of the YOLOv5 network.

2.3. Overview of Proposed MSBA-YOLO

The architecture of MSBA-YOLO is illustrated in Figure 4.

First, this paper proposes a lightweight backbone network based on an improved FasterNet. The core of this architecture employs PConv technology, which performs convolution operations only on a subset of the input channels. This approach significantly reduces redundant spatial computations and memory access frequency, thereby effectively optimizing the computational efficiency during the feature extraction stage. Second, a Single-Head Self-Attention mechanism is introduced into the feature extraction layers. By modeling long-range dependencies of global contextual information, it achieves the adaptive weight optimization of lesion features in the spatial dimension, markedly enhancing the network's feature discriminative power for irregular laryngeal pathological targets and its robustness against image noise interference. Finally, an improved MSBA-FIoU loss function is proposed. Building upon the original bounding box regression loss, it introduces an auxiliary scaling factor and a focusing weight mechanism. By dynamically learning the geometric relationship between the predicted and ground-truth boxes, it adaptively balances the regression gradients of multi-scale lesion samples, effectively improving the localization precision of subtle lesions in complex laryngoscopy environments.

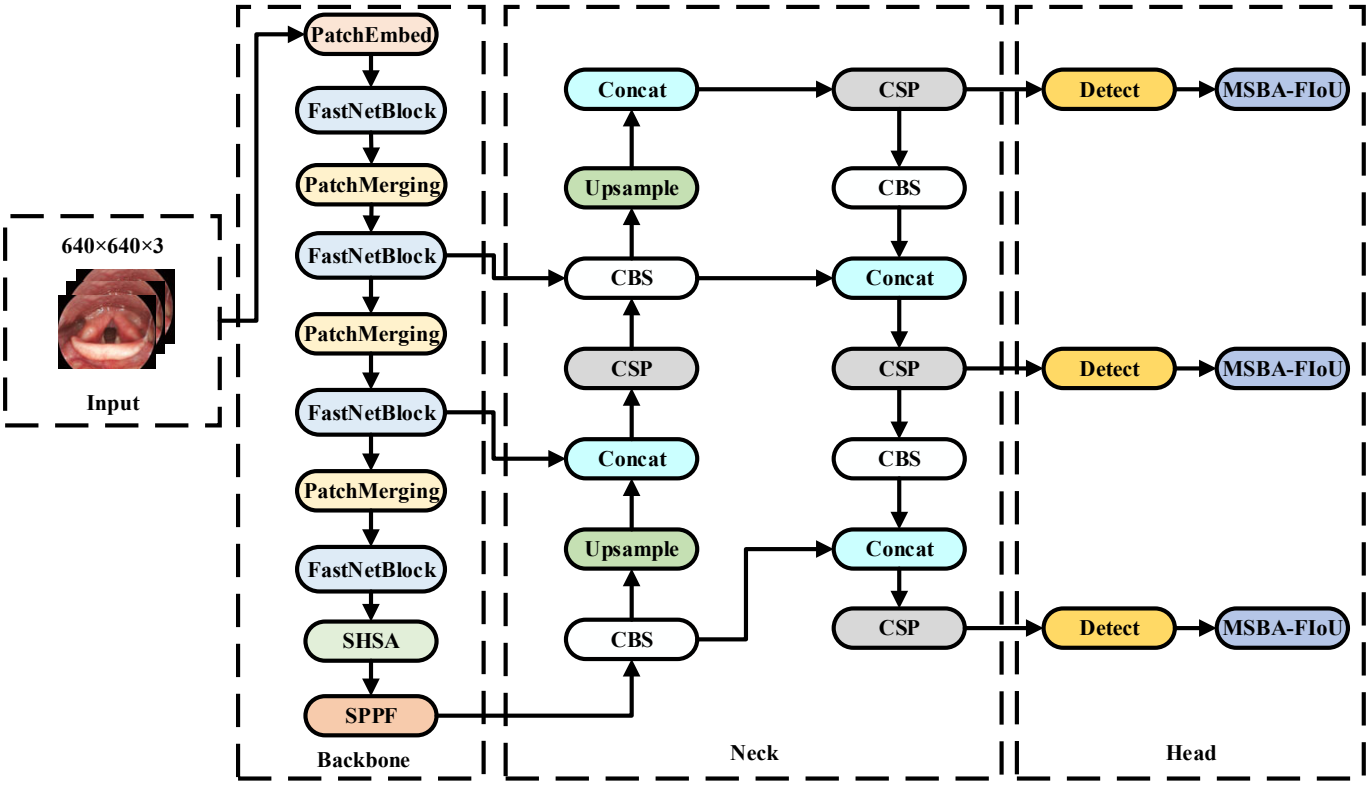


Figure 4. The architecture of the MSBA-YOLO.

2.4. Lightweight Backbone: FasterNet

As an innovative lightweight neural network architecture, FasterNet significantly enhances inference speed on edge devices while maintaining high accuracy, characterized by low latency and high computational efficiency. Tests on the ImageNet-1k dataset demonstrate that FasterNet's inference speed is 3.1 times faster than MobileViT-XXS on GPU/CPU and 2.5 times faster on ARM processors, with a 2.9% improvement in accuracy [19]. Considering the real-time requirements of laryngeal disease diagnosis and the computational constraints of Jetson devices, we selected FasterNet as the backbone of YOLOv5 to optimize detection performance at the edge. Currently, mainstream architectures such as ShuffleNet, MobileNet, and GhostNet are often adopted in lightweight model improvements. These structures utilize techniques like Depthwise Convolution (DWConv) [20] or Grouped Convolution [21] to reduce computational load and achieve lightweight detection. However, these methods increase memory access, which is disadvantageous for terminal deployment. Furthermore, redundant data operations incur additional inference time, making it difficult to meet the low-latency demands of real-time laryngoscopy diagnosis. The PConv mechanism of FasterNet selectively performs convolution on a subset of channels, achieving a dual reduction in both computational volume and memory consumption without compromising the quality of feature extraction. The specific structure is illustrated in Figure 5.

As illustrated in Figure 5, PConv utilizes the first or last few channels as feature maps for computation, where "$*$" denotes the convolution operator. This approach simplifies memory access, enhances computational efficiency, and reduces computational complexity, thereby improving detection efficiency. The computational overhead and memory access of PConv are defined by the following formulas:

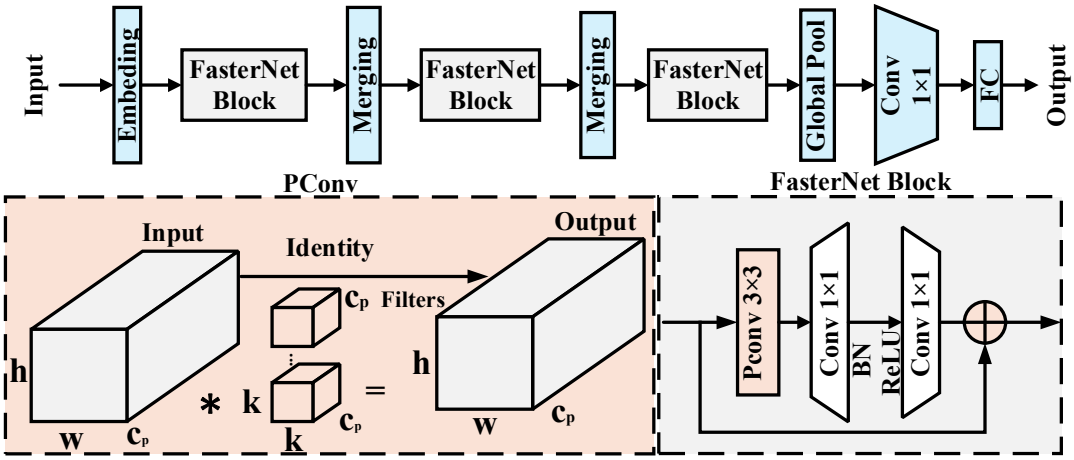


Figure 5. The architecture of FasterNet and the working principle of PConv.

$$\begin{cases} Conv = h \times w \times k^2 \times c^2 \\ Pconv = h \times w \times k^2 \times {c_p}^2 \end{cases} \tag{1}$$

$$\begin{cases} Conv = h \times w \times 2c + k^2 \times c^2 \approx h \times w \times 2c \\ Pconv = h \times w \times 2c_p + k^2 \times {c_p}^2 \approx h \times w \times 2c_p \end{cases} \tag{2}$$

In Equations (1) and (2), h and w denote the height and width of the feature map, respectively; k represents the convolution kernel size; c denotes the number of channels in a standard convolution, while c_p represents the number of channels processed by PConv. If we define the ratio as $r = c/c_p$, it can be derived from Equation (12) that after implementing PConv, the computational complexity of the convolution operation is reduced to $1/r^2$ of the original. This significantly alleviates the computational load and makes the model structure more lightweight. Furthermore, according to Equation (13), by employing the class inheritance mechanism to integrate PConv into the C3 modules for feature extraction, the memory access required by the convolution operation is reduced to $1/r$ of the original. This effectively lowers the memory bandwidth required during model execution and accelerates the inference speed.

2.5. Attention Mechanism: Single-Head Self-Attention Mechanism

In the task of laryngeal lesion recognition, the basic convolutional modules of the YOLOv5 model exhibit limitations in representation capability when extracting pathological features. They struggle to effectively capture the complexity of lesion morphologies and the fine structures of lesion boundaries, which compromises the discrimination accuracy for various types of laryngeal pathologies. To enhance feature representation, a common strategy is to introduce Multi-Head Self-Attention (MHSA) [22] modules to strengthen the modeling of key regions. Although MHSA can improve global contextual perception, its multi-branch attention structure significantly increases computational complexity and parameter overhead. In addition, the parallel attention heads introduce channel redundancy and require frequent tensor reconstruction operations (e.g., reshape and transpose), which further increase memory access cost and inference latency. These limitations make MHSA less suitable for deployment in resource-constrained embedded environments.

To address these issues, this study introduced a lightweight Single-Head Self-Attention (SHSA) module and integrated it into the feature fusion layers of the network. Unlike traditional MHSA, SHSA performs attention computation on only a subset of channels using a single attention head, while the remaining channels are directly propagated through a residual path. This design effectively reduces channel redundancy and significantly lowers

both the computational complexity and parameter overhead compared with conventional multi-head attention mechanisms. As a result, the SHSA module introduces only a minimal increase in model parameters while still enabling effective global context modeling. Consequently, the proposed module maintains strong feature representation capability while achieving higher computational efficiency and better compatibility with real-time edge deployment.

As illustrated in Figure 6, the SHSA module adopts a channel grouping and dynamic filtering mechanism. The input features undergo normalization and channel partitioning via the Split & GroupNorm module, resulting in feature groups X_1 and X_2. Specifically, the first C_p channels of X_1 are selected for attention computation, while the remaining $C - C_p$ channels in X_2 are kept in their original state and passed forward directly as a residual connection. This strategy effectively compresses the channel dimension and reduces both the computational and memory overhead. Within the attention computation path, X_1 generates Query (Q), Key (K), and Value (V) vectors through linear transformations, producing the enhanced feature representation X_1^* based on the computed attention weights. Subsequently, X_1^* and X_2 are concatenated along the channel dimension, followed by a linear projection to achieve cross-channel information fusion, thereby completing the feature enhancement process. The computation process of the SHSA layer can be formalized by Equations (3)–(6):

$$X_1, X_2 = Split(X, [C_p, C - C_p]) \tag{3}$$

$$Q = X_1 W^Q, K = X_1 W^K, V = X_1 W^V \tag{4}$$

$$X_1^* = Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{D_{qk}}})V \tag{5}$$

$$SHSA(X) = Concat(X_1^*, X_2)W^0 \tag{6}$$

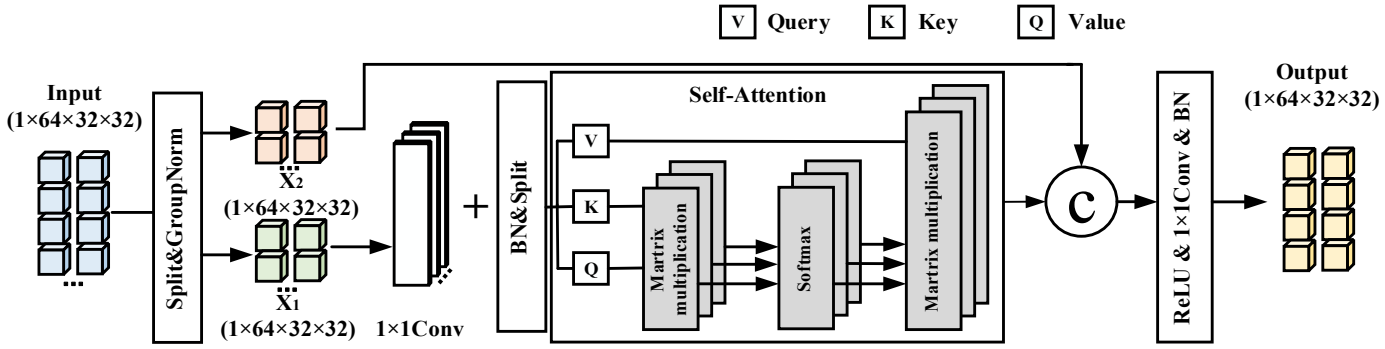


Figure 6. The architecture of SHSA.

In Equations (3)–(6), X denotes the input feature map. $Split(.)$ represents the channel partition operation that divides the input feature map into two groups, X_1 and X_2, where C_p denotes the number of channels used for attention computation and $C - C_p$ denotes the remaining channels. W^Q, W^K, and W^V denote the learnable projection matrices used to generate the Query, Key, and Value representations, respectively. D_{qk} represents the dimension of the Query–Key vectors used for scaling in the attention computation. $Softmax(.)$ denotes the normalization function applied to the attention weights. $Concat(.)$ represents the channel-wise concatenation operation, and W^0 denotes the learnable linear projection matrix used for feature fusion in the final output.

2.6. Optimization of the Bounding Box Regression Loss Function: MSBA-FIoU

In laryngeal disease detection tasks, the bounding box regression loss function directly determines the localization precision of the detection. The original YOLOv5 model primarily utilizes the Generalized Intersection over Union (GIoU) [23] as its loss function. Figure 7 provides a schematic diagram of GIoU, which was introduced to address the issue where predicted boxes and ground-truth boxes may not overlap. Compared to the original IoU, GIoU incorporates a minimum bounding rectangle C that completely encloses both the target box A and the predicted box B.

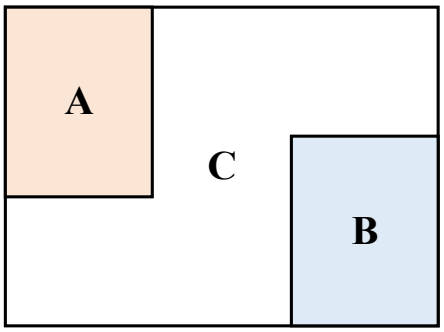


Figure 7. Schematic diagram of GIoU.

The calculation formulas for GIoU and the loss function L_{GIoU} are as follows:

$$L_{GIoU} = IoU - \frac{|S_c - (S_{A\cup B})|}{|S_c|} \; L_{GIoU} = 1 - GIoU \tag{7}$$

In the equations, S_C represents the area of the minimum bounding rectangle; $S_{A\cup B}$ denotes the union area of the predicted box and the ground-truth box; and IoU is a metric used in object detection to measure the overlap between the predicted and ground-truth boxes. The calculated GIoU serves as a distance metric to replace the standard IoU, while L_{GIoU} is the loss function; by minimizing this loss, the predicted box is driven closer to the ground-truth box. However, GIoU may degrade to standard IoU when two boxes are in an inclusive relationship or when they share the same height and width and are aligned on the same horizontal plane. Furthermore, when targets exhibit complex shapes or orientations, GIoU may fail to accurately reflect the overlap and positional relationship between the predicted and ground-truth boxes. In the context of laryngoscopy images, laryngeal lesions are typically minute; even a displacement of a few pixels can cause a drastic drop in the IoU value, leading to gradient fluctuations that make model convergence difficult. Additionally, the complex anatomical tissues surrounding the vocal cords generate a vast number of "easy-to-detect" negative samples. These samples dominate the training process, thereby suppressing the model's ability to learn "hard-to-detect" pathological features.

To address the aforementioned issues, inspired by Inner-IoU [24] and Focaler-IoU [25], we propose the Multi-Scale Boundary-Aware Focaler-IoU (MSBA-FIoU), which combines a dynamic auxiliary scaling mechanism with a linear redistribution strategy. The schematic diagram of Inner-IoU is illustrated as follows.

In Figure 8, the ground-truth bounding box and the auxiliary anchor box are denoted as B_{gt} and B, respectively. The center point of the ground-truth box is represented by (x_{cgt}, y_{cgt}), while (x_c, y_c) denotes the center of the auxiliary anchor box and the inner anchor box. The width and height of the ground-truth box are denoted as w_{gt} and h_{gt}, respectively, while the width and height of the anchor box are represented as w and h.

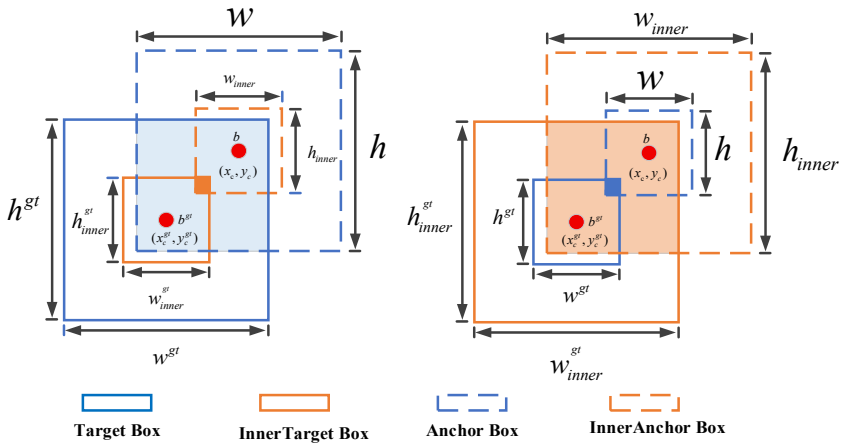


Figure 8. Schematic diagram of Inner-IoU.

In the original Inner-IoU, an auxiliary box is generated by introducing a fixed scale factor, namely Ratio, and the IoU is computed based on the generated inner box. The scale factor plays a dual role in the regression process. When a relatively small ratio is adopted, the auxiliary box shrinks, which provides stronger gradient increments for high-IoU samples and accelerates fine localization. In contrast, when a larger ratio is used, the effective regression region is enlarged, which is beneficial for improving regression performance in the low-IoU stage.

However, in the laryngeal lesion detection task, lesions exhibit substantial variations in scale and morphology. Therefore, a single fixed scale factor is insufficient to achieve a satisfactory balance for all lesion categories. For tiny lesions such as small polyps, the overlap between the predicted box and the ground-truth box is usually extremely limited during the early training stage. In this case, if a small fixed ratio is adopted, the auxiliary box may shrink excessively, resulting in weak gradient supervision and even regression failure. In contrast, for large lesions such as vocal cord masses or edema, although the target can be roughly localized, their boundaries are often ambiguous. If a large fixed ratio is used in this case, the regression constraint on the boundary becomes insufficient, leading to suboptimal localization accuracy.

To address this limitation, the fixed scale factor in Inner-IoU is replaced by a dynamic scale-aware factor λ, which is adaptively adjusted according to the scale of the lesion. Specifically, the area of the ground-truth box is first used to determine whether the target belongs to a small-scale or large-scale lesion, and then different values of λ are assigned to construct the auxiliary box. In this way, the regression process can be better matched to the scale characteristics of different lesion types. The auxiliary box is defined as follows:

$$b^{aux} = [x_c, y_c, w.\lambda, h.\lambda] \; \lambda = \begin{cases} 1.2, & S_{gt} < \delta \\ 0.8, & S_{gt} \geq \delta \end{cases} \tag{8}$$

where x_c and y_c denote the center coordinates of the predicted box, while w and h denote its width and height, respectively. S_{gt} represents the area of the ground-truth bounding box, and δ denotes the area threshold used to distinguish small lesions from large lesions. When $S_{gt} < \delta$, the target is regarded as a subtle lesion. In this case, a larger value of λ is assigned to enlarge the effective regression region of the auxiliary box, thereby providing stronger gradient signals during the early localization stage and improving the detection performance for small lesions. When $S_{gt} \geq \delta$, the target is considered relatively large. A smaller value of λ is adopted to impose a stricter localization constraint on the auxiliary box, which helps refine the boundary alignment and improve localization precision for lesions with blurred edges.

Based on the dynamically generated auxiliary boxes, the Inner-IoU is then calculated to measure the overlap between the predicted auxiliary box and the ground-truth auxiliary box. By replacing the fixed-ratio auxiliary box in the original Inner-IoU with a scale-adaptive one, the loss function becomes capable of perceiving lesion scale differences during regression. The resulting Inner-IoU is formulated as:

$$IoU_{inner} = \frac{Area(b_{aux}^{pred} \cap b_{aux}^{gt})}{Area(b_{aux}^{pred} \cup b_{aux}^{gt})} \tag{9}$$

where b_{aux}^{pred} and b_{aux}^{gt} denote the auxiliary boxes generated from the predicted box and the ground-truth box, respectively. Compared with the original Inner-IoU, this formulation enables the regression process to adaptively adjust its effective optimization region according to lesion size, thereby enhancing localization performance for multi-scale laryngeal lesions.

Although the dynamic scale-aware factor improves the adaptability of Inner-IoU to lesions of different sizes, the imbalance between easy and hard samples still affects the optimization process. In particular, a large number of easy background samples may dominate training and weaken the model's focus on difficult lesion samples. Therefore, a Focaler operator is further introduced to remap the IoU_{inner} values into a specified difficulty interval $[u, v]$, enabling the regression process to focus more on informative samples. The Focaler-IoU is defined as:

$$FIoU = clamp(\frac{IoU_{inner} - u}{v - u}, 0, 1) \tag{10}$$

where u and v define the lower and upper bounds of the target difficulty interval, respectively, and clamp$(\cdot)$ restricts the mapped value to the range $[0, 1]$. Through this linear redistribution, samples with a higher optimization value are emphasized, while the influence of excessively easy samples is suppressed.

By combining the scale-aware auxiliary box generation mechanism with the Focaler-based redistribution strategy, the final MSBA-FIoU loss is formulated as:

$$L_{MSBA\text{-}FIoU} = 1 - FIoU \tag{11}$$

In summary, the proposed MSBA-FIoU extends the original Inner-IoU in two aspects. First, the fixed scale factor is replaced with a dynamic scale-aware factor to better handle lesions with large scale variations. Second, a Focaler-based remapping strategy is introduced to alleviate sample imbalance during training. As a result, the proposed loss function provides more stable regression gradients and improves localization accuracy in complex laryngoscopic environments.

3. Experiments and Results

To comprehensively evaluate the effectiveness of the proposed MSBA-YOLO algorithm, a series of experiments was conducted. First, comparative experiments with different lightweight backbone networks were performed to demonstrate the superiority of the proposed method. Second, several attention mechanisms were compared to validate the effectiveness of the SHSA module. Third, different bounding box regression loss functions were evaluated to analyze the convergence characteristics and localization performance of the proposed MSBA-FIoU loss. In addition, ablation studies were conducted to investigate the individual contributions of each component in the proposed algorithm. Finally, edge deployment experiments were carried out on the Jetson Orin Nano platform to verify the real-time inference capability and energy efficiency of the proposed method.

To reduce the influence of stochastic training factors, each model was trained multiple times with different random seeds, and the reported results correspond to the average performance across these runs.

3.1. Experimental Environment and Training Parameters

In this study, experiments were conducted using YOLOV5s as the baseline model. The input image size for the model was fixed at 640 × 640 pixels. The total number of training epochs was set to 250, incorporating an early stopping strategy: training would terminate prematurely if the model's performance on the validation set failed to improve for 100 consecutive epochs. The batch size was set to 64, with an initial learning rate of 0.0002 and a weight decay coefficient of 0.0005. Detailed configurations of the hardware platform and software environment utilized in this experiment are presented in Table 1.

Table 1. Experimental environment.

Parameter	Configuration
Operating System	Ubuntu 22.04.4 LTS
CPU	Intel Xeon Gold 6133
GPU	NVIDIA GeForce RTX 4090
Python Version	3.9.21
PyTorch Version	2.6.0
CUDA Version	12.4

3.2. Objective Evaluation Indicators

This study comprehensively evaluated model performance through three key detection metrics: Precision (P) for classification reliability, Recall (R) for lesion detection sensitivity, and mean Average Precision (mAP) for overall detection accuracy. Model complexity was quantified via: (1) parameter count, (2) computational load (GFLOPs), and (3) model size (MB). The mathematical formulations for P, R, and mAP are provided in Equations (11) and (12).

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN} \tag{12}$$

TP: True positives (correctly classified lesions);

FP: False positives (lesions misclassified as target type);

FN: False negatives (missed lesions).

Precision measures the model's classification reliability by quantifying the proportion of correct predictions among all positive classifications. Recall evaluates the detection sensitivity by measuring the fraction of actual positives correctly identified.

$$mAP = \frac{\sum \int_0^1 P(R)dR}{N} \tag{13}$$

mAP integrates precision–recall performance across all classes to assess overall diagnostic accuracy.

Parameter Count: The total number of trainable parameters, reflecting model scale and architectural complexity.

GFLOPs: Billions of floating-point operations required for a single inference pass, representing computational demand.

For models achieving comparable accuracy, lower values in both metrics indicate superior lightweight characteristics and operational efficiency.

3.3. Comparative Experiments of Different Lightweight Network Architectures

To verify the effectiveness of the improvement scheme proposed in this study, we conducted comparative tests among the original YOLOv5s model, the YOLOv5s model with EfficientViT [26] as the backbone, the YOLOv5s model with ShuffleNetV2 [27] as the backbone, the YOLOv5s model with GhostNet [28] as the backbone, and our improved MSBA-YOLO model. As illustrated in Figure 9, MSBA-YOLO significantly reduced the number of parameters and computational overhead while maintaining extremely high detection accuracy. The experimental results demonstrate that the proposed model achieved the optimal balance between precision and lightweight design, exhibiting superior overall performance.

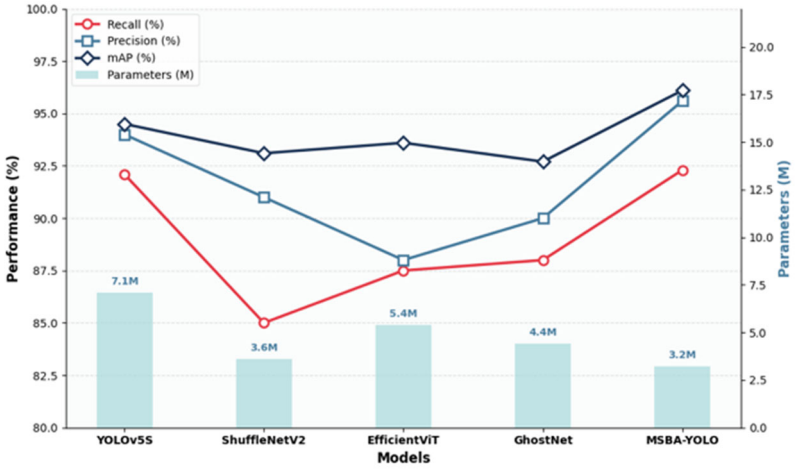


Figure 9. Comparison of lightweight network architectures.

As shown in Table 2, the detection results indicate that the lightweight network proposed in this study, MSBA-YOLO, significantly reduced the size of the original model by 54.6%, bringing it down to 6.4 MB. In comparison, YOLOv5s-ShuffleNetV2 achieved a 48.9% reduction, YOLOv5s-GhostNet achieved 34.8%, and YOLOv5s-EfficientViT achieved 23.4%. By employing FasterNet as the backbone, the MSBA-YOLO model effectively reduced model complexity while maintaining high detection efficiency. Compared to the original YOLOv5s model, MSBA-YOLO's GFLOPs decreased by 8.6 (a 54.4% reduction), and its total processing time on the desktop platform was only 9.7 ms, making it the fastest among all tested models.

Table 2. Detection results of different lightweight networks.

Model	Size/MB	GFLOPs	mAP%
YOLOv5s	14.1	15.8	94.5
YOLOv5s-ShuffleNetV2	7.2	7.5	93.1
YOLOv5s-EfficientViT	10.8	9.6	93.6
YOLOv5s-GhostNet	9.2	8.2	92.7
MSBA-YOLO	6.4	7.2	96.1

Furthermore, the ShuffleNetV2, GhostNet, and EfficientViT variants reduced GFLOPs by 8.3, 7.6, and 6.2, respectively, relative to the baseline model. This demonstrates that the MSBA-YOLO model possesses lower computational load and structural complexity. More importantly, the mean mAP of the MSBA-YOLO model reached 96.1%, which was 1.6% higher than the original model. Notably, while the other three lightweight networks also reduced model size, their mAP scores all fell below the baseline. This fully validates that

MSBA-YOLO effectively balances detection precision and speed while achieving a more lightweight architecture.

3.4. Comparison of Different Attention Mechanisms

To further verify the effectiveness of SHSA within the MSBA-YOLO model, this experiment compared it with the SE [29], CBAM [30], and SimAM [31] attention mechanisms. These four mechanisms were respectively integrated into the backbone network for experimental analysis, with the results summarized in Table 3.

Table 3. Performance comparison of different attention mechanisms.

Model	Parameters $\times 10^{-6}$	GFLOPs	Recall%	mAP@0.5/%
+SE	0.17	0.2	87.6	91.5
+CBAM	0.16	0.2	87.9	93.7
+SIMAM	0	0	83.4	89.4
+SHSA	0.14	0.1	92.3	96.1

Experimental data indicate that the model integrated with the SHSA mechanism outperformed the others across all key performance metrics, achieving a recall of 92.3% and an mAP@0.5 of 96.1%. This outcome strongly demonstrates the superiority of the SHSA mechanism in laryngeal lesion detection tasks, as it more effectively captures global contextual information. In contrast, although the SimAM mechanism achieved zero parameter increment, its detection accuracy was lower, with an mAP of only 89.4%. The parameter counts for the SE and CBAM mechanisms were 0.17 M and 0.16 M, respectively—slightly higher than SHSA's 0.14 M—yet their mAP performance (91.5% and 93.7%, respectively) fell short of SHSA. Therefore, considering both model complexity and detection performance, SHSA significantly enhances feature extraction capabilities with minimal parameter overhead.

3.5. Comparison of Different Loss Functions

To verify the effectiveness of the proposed MSBA-FIoU loss function, we conducted comparative training against the traditional CIoU [32] as well as other mainstream loss functions, including EIoU [33], SIoU [34], and Inner-IoU. As illustrated in Figure 10, we recorded the convergence of the model's Box Loss, Objectness Loss, and Classification Loss over 250 epochs.

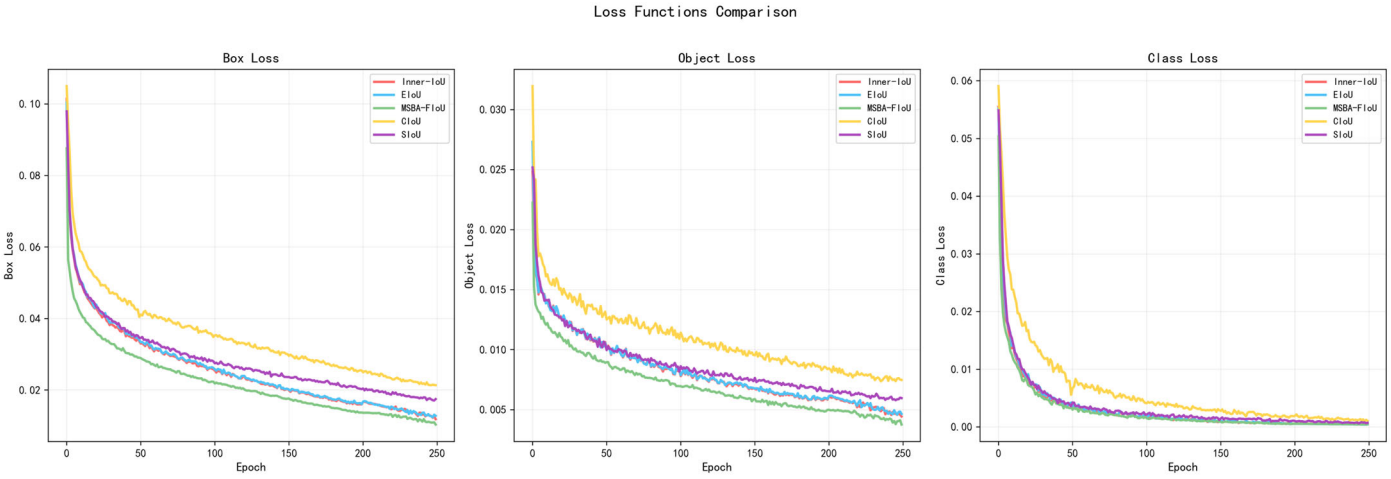


Figure 10. Performance comparison of different loss functions.

Experimental results demonstrated that MSBA-FIoU exhibited significant superiority across all loss metrics. During the early stages of training, the MSBA-FIoU curves declined most sharply, enabling the model to reach a stable phase rapidly. In contrast, CIoU showed the slowest convergence speed and higher fluctuations, particularly in Box Loss, where it remained at a relatively high level even into the later stages of training. This proves that the proposed auxiliary scaling mechanism effectively amplifies the localization errors of minute lesions, thereby providing stronger gradient feedback. After 250 epochs of iteration, MSBA-FIoU achieved the lowest final loss values among all compared functions. Specifically, its convergence endpoints for Box Loss and Objectness Loss were markedly better than those of SIoU and EIoU. The Classification Loss curves indicate that while all loss functions performed similarly in classification recognition, MSBA-FIoU maintained the smoothest and lowest trajectory. This suggests that the loss function balances the weights of easy and hard samples through the Focaler mechanism, allowing the model to focus more on the regression of complex pathological regions and reducing oscillations during the training process.

In summary, the MSBA-FIoU loss function overcomes the difficulties posed by large size variations and blurred boundaries of laryngeal lesions by dynamically adjusting the focus of bounding box regression. Experimental evidence confirms that this improvement scheme not only substantially shortens the training cycle but also significantly enhances the model's localization precision for subtle lesions.

3.6. Ablation Study

To accurately evaluate the contribution of each module within the improved algorithm, we conducted systematic ablation experiments on our self-built dataset. The experimental results are summarized in Table 4.

Table 4. Results of the ablation study.

FasterNet	SHSA	MSBA-FIoU	Recall (%)	Precision (%)	mAP@0.5%	Params/M
-	-	-	92.1	94.0	94.5	7.1
-	√	-	92.7	94.3	94.9	7.3
√	-	-	87.3	90.2	92.9	3.1
√	√	-	90.7	93.3	94.3	3.2
√	√	√	92.3	95.6	96.1	3.2

Note: "√" indicates that the module is used, while "-" indicates that the module is not used.

Starting from the baseline YOLOv5s model, we first introduced the SHSA attention mechanism independently. As shown in Table 4, incorporating SHSA improved the mAP@0.5 from 94.5% to 94.9%, while recall increased from 92.1% to 92.7%. Although the parameter count increased slightly from 7.1 M to 7.3 M, the improvement in detection accuracy demonstrates that SHSA effectively enhances global contextual modeling with only minimal parameter overhead.

When only FasterNet was used to replace the original YOLOv5s backbone, the model size was significantly reduced from 7.1 M to 3.1 M, demonstrating the strong lightweight capability of the PConv-based backbone. However, the detection accuracy experienced a certain decline, with mAP@0.5 dropping from 94.5% to 92.9% and recall decreasing to 87.3%. This performance reduction can be attributed to the simplified convolution structure of FasterNet, which prioritizes computational efficiency but slightly weakens the feature extraction capability for complex laryngoscopic images.

After introducing the SHSA attention mechanism on top of the FasterNet backbone, the model performance was significantly restored. Recall increased from 87.3% to 90.7% and mAP@0.5 improved from 92.9% to 94.3%, while the parameter count remained almost

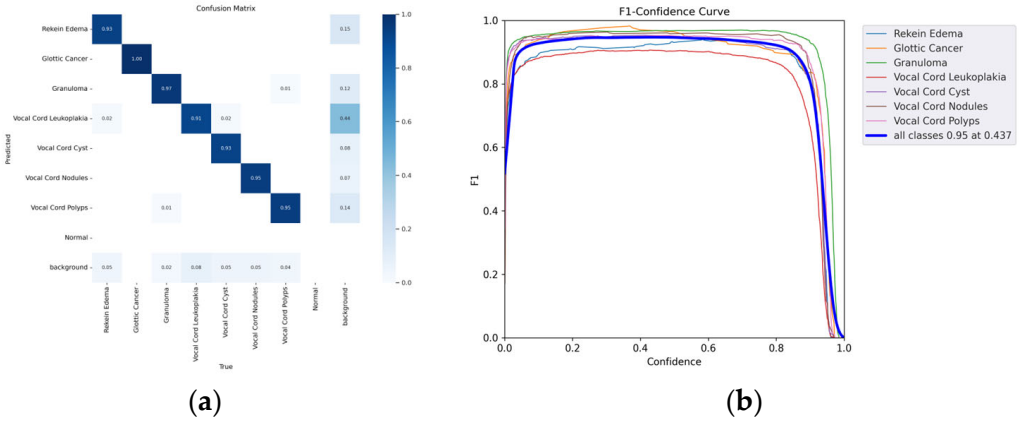
unchanged (3.2 M). This indicates that SHSA effectively compensates for the global feature modeling limitations of the lightweight backbone by capturing long-range contextual dependencies within lesion regions.

Finally, by incorporating the proposed MSBA-FIoU loss function, the model achieved its optimal performance. Compared with the previous configuration, recall further increased to 92.3%, precision reached 95.6%, and mAP@0.5 improved to 96.1%. These results demonstrate that MSBA-FIoU improves bounding box regression by adaptively focusing on difficult samples and handling large variations in lesion scale.

Overall, the ablation results demonstrate that although the introduction of a lightweight backbone slightly reduces the detection accuracy, the combination of SHSA and MSBA-FIoU effectively compensates for this limitation. The final MSBA-YOLO model achieved a 1.6% improvement in mAP compared with the original YOLOv5s while reducing the parameter count by more than 50%, demonstrating an effective balance between detection accuracy and model efficiency.

3.7. Visual Analysis of Detection Results

The confusion matrix scores in Figure 11 intuitively demonstrate the model's exceptional discriminative power for laryngeal diseases with typical anatomical features. Comparatively, the recognition accuracy for "Vocal Cord Leukoplakia" reached 91%. Statistics reveal that approximately 4% of background regions may be misreported as leukoplakia or related distractors, which aligns with clinical characteristics such as blurred margins and a coloration similar to the reflection of normal mucosa.



(**a**) (**b**)

Figure 11. (**a**) Confusion matrix and (**b**) F1-score curves of MSBA-YOLO.

The F1-score, which provides a comprehensive measure of precision and recall, reflects the model's robustness through its curve variation across different confidence thresholds. When the confidence threshold was set to 0.437, the mean F1-score for all classes reached a peak of 0.95. This indicates that at this specific threshold, the model maintains an extremely low missed detection rate while ensuring a high detection rate. The F1 curves for various lesion categories remained consistently high between confidence thresholds of 0.1 and 0.8, exhibiting a stable and broad trajectory. This suggests that MSBA-YOLO is not sensitive to threshold selection, demonstrating strong adaptability for practical clinical applications and the ability to stably output high-quality auxiliary diagnostic results.

The visualization results on the validation set are shown in Figure 12, demonstrating the effectiveness of the proposed method in accurately detecting different types of laryngeal lesions.

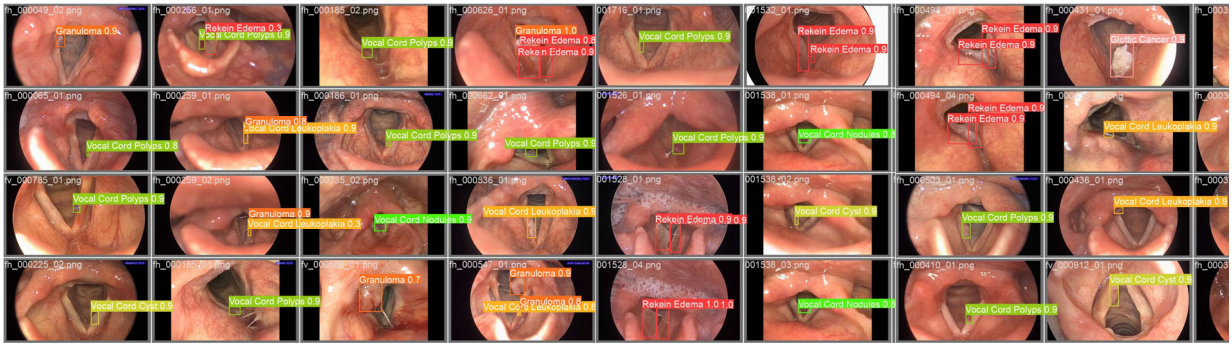


Figure 12. Schematic diagram of visualized detection results.

4. Edge Computing Platform Deployment

4.1. Deployment Platform

NVIDIA Jetson is currently an ideal embedded platform for deploying deep neural network models, offering advantages such as high computational efficiency, low power consumption, and ease of deployment for deep learning applications. To apply the proposed MSBA-YOLO model to practical clinical auxiliary diagnosis scenarios and achieve low-latency, high-performance real-time detection, we selected the Jetson Orin Nano as the edge computing deployment platform. The onboard resources of the Jetson Orin Nano are as Table 5:

Table 5. Hardware configuration of Jetson Orin Nano.

Component	Specification
CUDA	512
GPU Boost	1.5 GHz
RAM	4 GB
MAX Power	10 W
Camera	IMX219

4.2. Comparative Experiments

To verify the suitability of MSBA-YOLO in practical clinical edge computing environments, this study deployed the model on both a high-performance computing platform (RTX 4060) and an embedded edge computing platform (Jetson Orin Nano) for performance evaluation. The experimental results are summarized in Table 6.

Table 6. Comparative experiments on different platforms.

Processing Platform	Pre-Processing Time/ms	Inference Time/ms	NMS Time/ms	Total Time/ms	Power/w
RTX 4060 (8G)	0.6	6.7	2.4	9.7	17.8 w
JETSON Orin Nano (4G)	1.1	10.2	4.3	15.6	5.23 w

Experimental results show that the total processing time of MSBA-YOLO on the Jetson Orin Nano platform was only 15.6 ms, corresponding to a detection speed of approximately 64.1 FPS. Compared with the high-performance RTX 4060 platform, although there was a significant difference in hardware computing capability, the edge platform still achieved real-time inference performance due to the lightweight architecture of the proposed model. This response speed is sufficient to meet the real-time analysis requirements of electronic laryngoscopy video streams in clinical scenarios.

In terms of power efficiency, the Jetson Orin Nano demonstrated clear advantages for edge deployment. The measured power consumption during inference was approximately 5.23 W, while the RTX 4060 platform consumed about 17.8 W under the same experimental configuration. This indicates that the edge platform achieves comparable real-time detection capability while requiring significantly lower energy consumption. These results further demonstrate that the proposed MSBA-YOLO algorithm effectively balances detection accuracy, computational efficiency, and energy consumption, making it well-suited for deployment in resource-constrained medical edge devices.

4.3. Real-Time Diagnostic Performance Analysis

To evaluate the real-time diagnostic performance of the proposed system in a clinical-like environment, a real-time laryngeal lesion detection system was implemented on the Jetson Orin Nano platform using the NVIDIA DeepStream framework. The overall workflow of the system is illustrated in Figure 13. During operation, electronic laryngoscopy images were continuously captured and processed through the DeepStream pipeline, while the optimized MSBA-YOLO model was accelerated using the TensorRT inference engine.

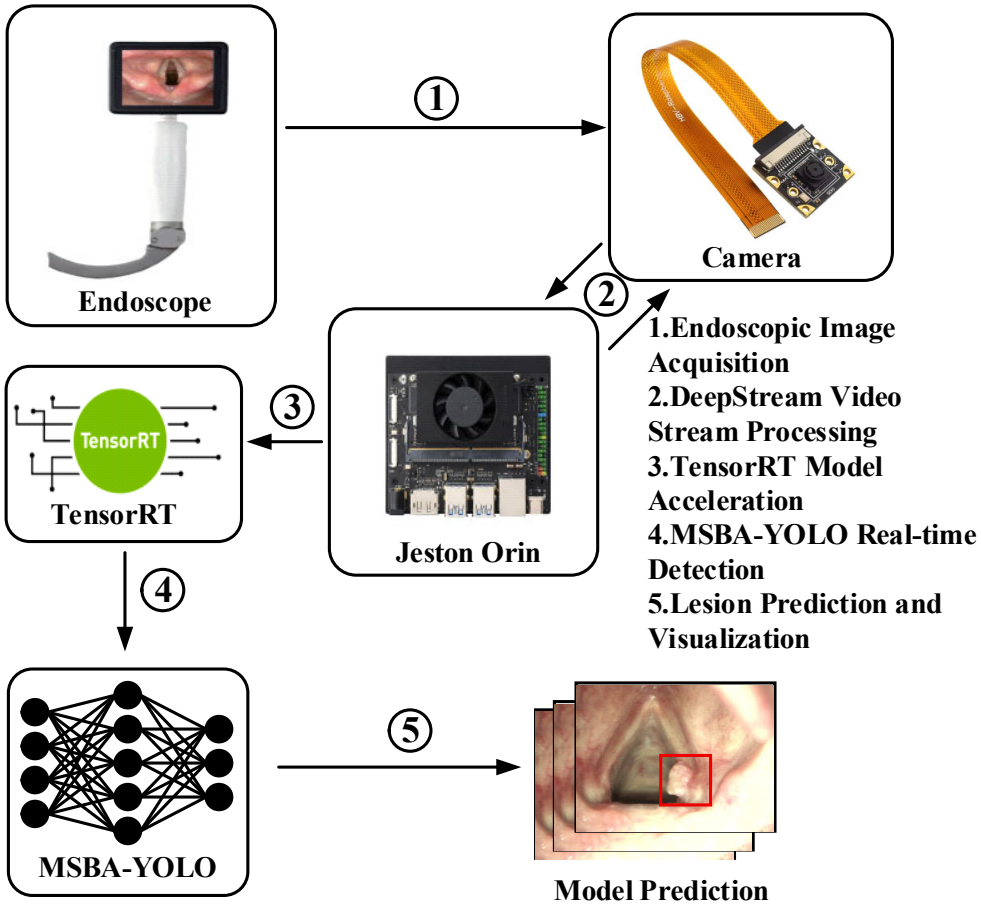


Figure 13. Real-time diagnostic architecture.

To further quantify the deployment advantages of the proposed system, we conducted a comparative benchmark between the Jetson Orin Nano and a desktop-grade NVIDIA RTX 4060 GPU. The evaluation focused on three critical dimensions: real-time frame rate, power consumption, and energy efficiency. The experimental results, captured over a continuous 300-s operation, are illustrated in Figure 14.

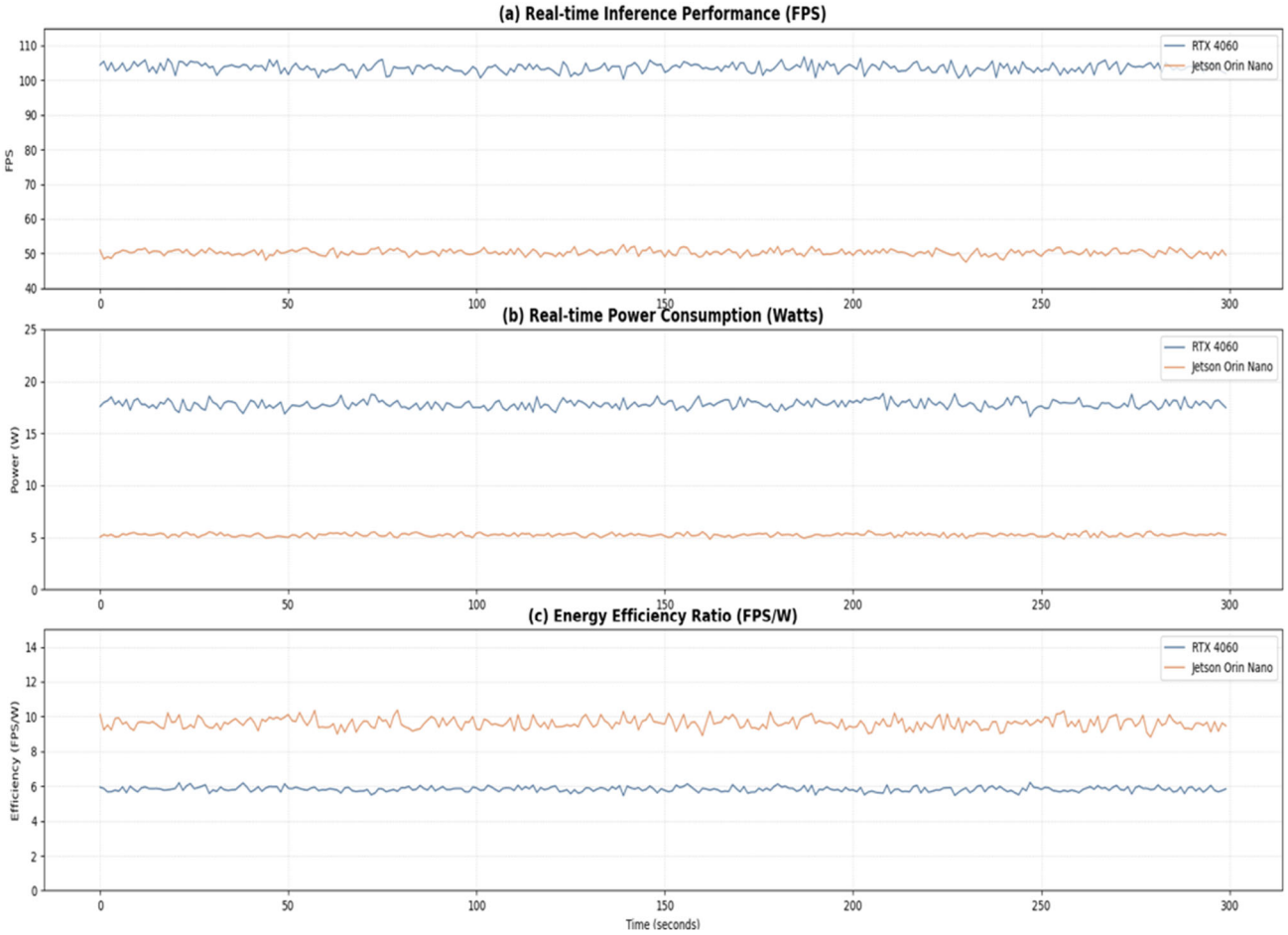


Figure 14. Real-time performance comparison.

As shown in Figure 14a, while the RTX 4060 achieved a high throughput of approximately 103.4 FPS, the Jetson Orin Nano maintained a stable and consistent frame rate of 50.27 FPS when processing the live video stream through the DeepStream pipeline. This confirms that the embedded system is fully capable of providing instantaneous diagnostic feedback during clinical procedures.

The power consumption trajectories in Figure 14b highlight the significant energy advantage of the embedded platform. The RTX 4060 operated at an average power of 17.77 W, whereas the Jetson Orin Nano required only 5.25 W—a reduction of approximately 70.5% compared to the desktop-grade GPU.

To evaluate the overall utility of the hardware in a mobile medical context, the energy efficiency ratio (EER) was calculated as the ratio of FPS to power consumption. As illustrated in Figure 14c:

The Jetson Orin Nano reached an average EER of 9.57 FPS/W.

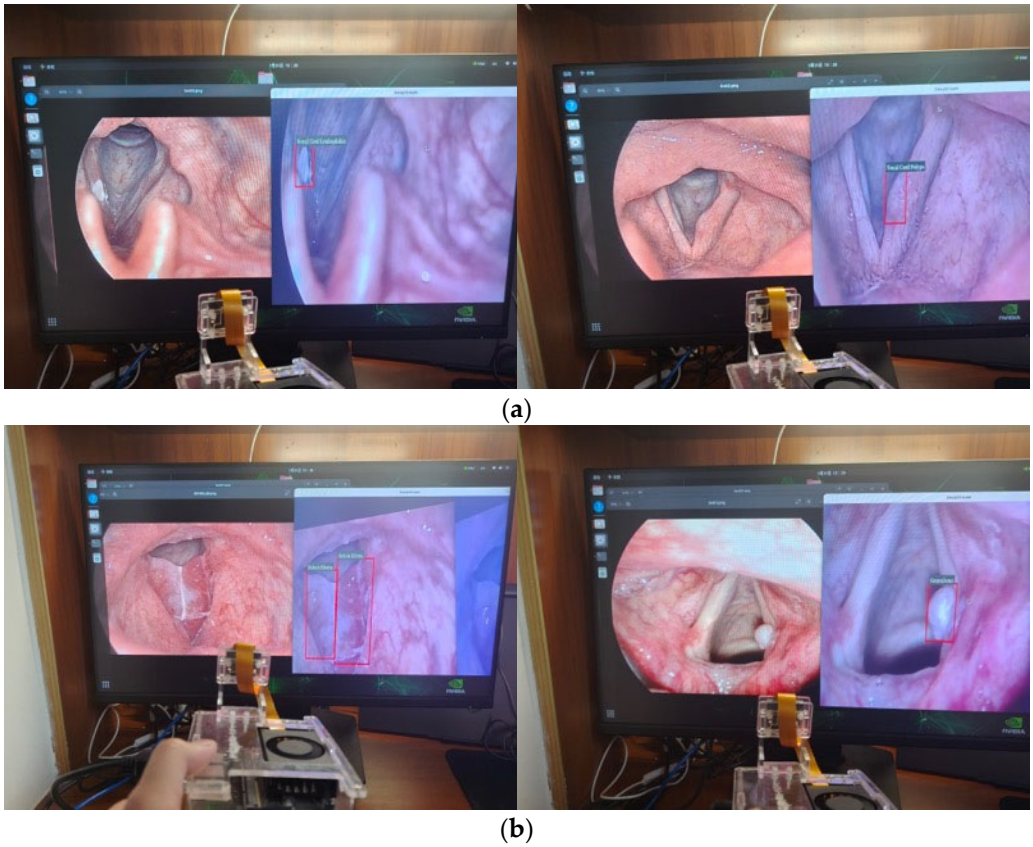
The RTX 4060 yielded an average EER of 5.82 FPS/W.

The real-time detection results on the Jetson Orin Nano platform are shown in Figure 15.

The data indicate that the Jetson Orin Nano is approximately 64.4% more energy-efficient than the RTX 4060. This high EER demonstrates that the MSBA-YOLO model, when combined with TensorRT acceleration on the Jetson platform, achieves an optimal balance between performance and energy economy.

In conclusion, the experimental results demonstrate that while desktop GPUs offer higher raw performance, the Jetson Orin Nano platform provides a more viable and optimized solution for the deployment of real-time laryngeal lesion detection. It successfully

satisfies the stringent requirements of clinical real-time imaging while maintaining a minimal power footprint, paving the way for the development of high-efficiency, portable AI-assisted diagnostic terminals.



(**a**)

(**b**)

Figure 15. Real-time laryngeal disease diagnosis on Jetson Orin Nano. (**a**) Vocal cord leukoplakia and vocal cord cyst. (**b**) Reinke's edema and laryngeal Granuloma.

5. Conclusions

This study proposes MSBA-YOLO, a lightweight laryngeal lesion detection framework designed for real-time deployment on edge devices. The proposed method improves the baseline YOLOv5s architecture through three key components: a FasterNet-based lightweight backbone, a Single-Head Self-Attention (SHSA) module for enhanced global feature perception, and the MSBA-FIoU loss function for improved localization of multi-scale lesions.

Experimental results demonstrate that the proposed model effectively balances detection accuracy and computational efficiency. Compared with the original YOLOv5s model, MSBA-YOLO reduced the parameter size from 14.1 MB to 6.4 MB, representing a 54.6% reduction, while improving the detection accuracy from 94.5% to 96.1% mAP@0.5. The ablation study further verifies that the SHSA module effectively enhances global contextual modeling with minimal parameter overhead, while the proposed MSBA-FIoU loss improves bounding box regression for lesions with large scale variations and blurred boundaries.

In addition, the proposed model was successfully deployed on the Jetson Orin Nano edge platform. Benchmark results show that the system achieved a single-frame inference latency of 15.6 ms, including preprocessing, model inference, and non-maximum suppression. When processing live video streams through the DeepStream pipeline with a camera input, the system maintained a stable real-time detection speed of approximately 50 FPS. Compared with the desktop GPU platform (RTX 4060), the embedded system

achieves competitive real-time performance while significantly reducing energy consumption, demonstrating strong suitability for portable medical diagnostic devices.

Overall, the proposed MSBA-YOLO framework provides an effective solution for real-time laryngeal lesion detection in resource-constrained environments.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L. and D.L.; software, D.L.; validation, Y.L., D.L. and Z.L.; formal analysis, Y.L., D.L. and Z.L.; investigation, Y.L., D.L. and Z.L.; resources, X.X. and Z.L.; data curation, D.L. and Z.L.; writing—original draft preparation, Y.L. and D.L.; writing—review and editing, Y.L. and D.L.; visualization, Y.L. and D.L.; supervision, Y.L., X.X. and Z.L.; project administration, X.X.; funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China, grant number 62262011; Guangxi Key Technologies R&D Program, grant number GuikeAB23026004; and Guangxi Key Technologies R&D Program, grant number GuikeAB23026036.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Medical Ethics Committee of Guilin People's Hospital (Protocol Code: 2025-002KY, approved on 10 January 2025) for the secondary analysis of laryngeal image data.

Informed Consent Statement: This study utilized an open-source, de-identified dataset (Laryngoscope8). Informed consent was obtained by the original investigators at the time of data collection.

Data Availability Statement: The Laryngoscope8 dataset used in this study is publicly available at https://github.com/greenyin/Laryngoscope8 (accessed on 12 December 2025). The NBI Visible Light Dataset of Pharyngeal and Laryngeal Images can be accessed by application through the National Basic Science Data Center.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kim, G.H.; Hwang, Y.J.; Lee, H.; Sung, E.-S.; Nam, K.W. Convolutional neural network-based vocal cord tumor classification technique for home-based self-prescreening purpose. *BioMed. Eng. OnLine* **2023**, *22*, 81. [CrossRef]
2. Haug, C.J.; Drazen, J.M. Artificial intelligence and machine learning in clinical medicine, 2023. *N. Engl. J. Med.* **2023**, *388*, 1201–1208. [CrossRef] [PubMed]
3. Sampieri, C.; Baldini, C.; Azam, M.A.; Moccia, S.; Mattos, L.S.; Vilaseca, I.; Peretti, G.; Ioppi, A. Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: A guide for physicians and state-of-the-art review. *Otolaryngol.–Head Neck Surg.* **2023**, *169*, 811–829. [CrossRef] [PubMed]
4. Baldini, C.; Migliorelli, L.; Berardini, D.; Azam, M.A.; Sampieri, C.; Ioppi, A.; Srivastava, R.; Peretti, G.; Mattos, L.S. Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO. *Comput. Methods Programs Biomed.* **2025**, *260*, 108539. [CrossRef]
5. Xiong, H.; Lin, P.; Yu, J.-G.; Ye, J.; Xiao, L.; Tao, Y.; Jiang, Z.; Lin, W.; Liu, M.; Xu, J. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine* **2019**, *48*, 92–99. [CrossRef] [PubMed]
6. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
7. Wei, R.; Liang, Y.; Geng, L.; Wang, W.; Wei, M. A non-local dual-stream fusion network for laryngoscope recognition. *Am. J. Otolaryngol.* **2025**, *46*, 104565. [CrossRef]
8. Wang, S.; Chen, Y.; Chen, S.; Zhong, Q.; Zhang, K. Hierarchical dynamic convolutional neural network for laryngeal disease classification. *Sci. Rep.* **2022**, *12*, 13914. [CrossRef]
9. Ren, J.; Jing, X.; Wang, J.; Ren, X.; Xu, Y.; Yang, Q.; Ma, L.; Sun, Y.; Xu, W.; Yang, N. Automatic recognition of laryngoscopic images using a deep-learning technique. *Laryngoscope* **2020**, *130*, E686–E693. [CrossRef]
10. Yin, L.; Liu, Y.; Pei, M.; Li, J.; Wu, M.; Jia, Y. Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism. *Pattern Recognit. Lett.* **2021**, *150*, 207–213. [CrossRef]
11. Xiong, M.; Luo, J.-W.; Ren, J.; Hu, J.-J.; Lan, L.; Zhang, Y.; Lv, D.; Zhou, X.-B.; Yang, H. Applying Deep Learning with Convolutional Neural Networks to Laryngoscopic Imaging for Automated Segmentation and Classification of Vocal Cord Leukoplakia. *Ear Nose Throat J.* **2024**. [CrossRef]

12. Tonay, M.A.K.; Dihan, S.S.; Zarif, F.I.; Adnan, O.R.; Ali, M.N.Y. Laryngeal Cancer Detection and Classification Using Deep Learning on Histopathological Images. In Proceedings of the 2025 29th International Conference on Information Technology (IT), Zabljak, Montenegro, 19–22 February 2025; pp. 1–5.
13. Nie, X.; Zhang, X.; Wang, D.; Liu, Y.; Xing, L.; Liu, W. Laryngeal cancer diagnosis based on improved YOLOv8 algorithm. *Mach. Learn. Sci. Technol.* **2025**, *6*, 015011. [CrossRef]
14. Baldini, C.; Azam, M.A.; Sampieri, C.; Ioppi, A.; Ruiz-Sevilla, L.; Vilaseca, I.; Alegre, B.; Tirrito, A.; Pennacchi, A.; Peretti, G. An automated approach for real-time informative frames classification in laryngeal endoscopy using deep learning. *Eur. Arch. Oto-Rhino-Laryngol.* **2024**, *281*, 4255–4264. [CrossRef] [PubMed]
15. Shen, P. Throat NBI Visible Light Dataset (V1). Xi'an University of Electronic Science and Technology, National Basic Science Data Center. 4 April 2023. Available online: https://cstr.cn/16666.11.nbsdc.rWxpsDHj (accessed on 12 December 2025).
16. Wang, M.; Yang, W.; Wang, L.; Chen, D.; Wei, F.; KeZiErBieKe, H.; Liao, Y. FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103752. [CrossRef]
17. Wu, T.-H.; Wang, T.-W.; Liu, Y.-Q. Real-time vehicle and distance detection based on improved yolo v5 network. In Proceedings of the 2021 3rd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 18–20 June 2021; pp. 24–28.
18. Ni, Y.; Qing, Z.; Ruan, H.; Xiong, X.; Hu, X.; Chang, J. UAV Image Target Detection Algorithm Based on Improved YOLOv5. *J. Jianghan Univ. (Nat. Sci. Ed.)* **2025**, *53*, 87–91.
19. Lu, Y.; Kang, S.; Wu, S.; He, C. Nonwoven defect detection by fusing selection kernel attention. *Comput. Eng. Appl.* **2024**, *60*, 331–339.
20. Poudel, R.P.; Bonde, U.; Liwicki, S.; Zach, C. Contextnet: Exploring context and detail for semantic segmentation in real-time. *arXiv* **2018**, arXiv:1805.04554. [CrossRef]
21. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
22. Li, P.; Zheng, J.; Li, P.; Long, H.; Li, M.; Gao, L. Tomato maturity detection and counting model based on MHSA-YOLOv8. *Sensors* **2023**, *23*, 6701. [CrossRef]
23. Tan, H.; Liu, X.; Yin, B.; Li, X. MHSA-Net: Multihead self-attention network for occluded person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 8210–8224. [CrossRef]
24. Zhang, H.; Xu, C.; Zhang, S. Inner-IoU: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.
25. Zhang, H.; Zhang, S. Focaler-iou: More focused intersection over union loss. *arXiv* **2024**, arXiv:2401.10525.
26. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. EfficientViT: Memory efficient vision transformer with cascaded group attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14420–14430.
27. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
28. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
30. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
31. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11863–11874.
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
33. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
34. Loss, G.Z.S. More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.