Am J Transl Res 2025;17(5):3293-3306
www.ajtr.org /ISSN:1943-8141/AJTR0162602

Original Article

Improved YOLOv8-seg for laryngeal structure recognition in medical images

Haipo Cui[2*], Jinjing Wu[2*], Tianying Li[3*], Zui Zou[3], Wenhui Guo[3], Long Liu[2], Qianwen Zhang[4], Xiaoping Huang[1]

[1]Department of Anesthesiology, The First Hospital of Putian City, Putian 351100, Fujian, China; [2]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 201210, China; [3]Department of Anesthesiology, Naval Medical University, Shanghai 200433, China; [4]Department of Radiology, Changhai Hospital, Naval Medical University, Shanghai 200433, China. [*]Equal contributors and co-first authors.

Received December 11, 2024; Accepted April 14, 2025; Epub May 15, 2025; Published May 30, 2025

Abstract: Objectives: Tracheal intubation is a routine procedure in clinical surgeries and emergency situations, essential for maintaining respiration and ensuring airway patency. Due to the complexity of laryngeal structures and the need for rapid airway management in critically ill patients, real-time, accurate identification of key laryngeal structures is crucial for successful intubation. This study presents a real-time laryngeal structure recognition method based on an improved YOLOv8-seg model. Methods: Laryngeal images from retrospective intubation procedures were used to assist clinicians in the rapid and precise identification of critical laryngeal structures, such as the epiglottis, glottis, and vocal cords. The proposed model, named SlimMSDA-YOLO, integrates a lightweight neck structure, Slimneck, into the original YOLOv8n-seg model by combining GSConv and standard convolutions. This modification effectively reduces the floating-point operations and computational resource requirements. Additionally, a multi-scale dilation attention module was incorporated between the neck and head sections to enhance the network's ability to capture features across various receptive fields, thereby improving its focus on critical regions. Results: The SlimMSDA-YOLO model achieved a precision of 90.4%, recall of 84.2%, and mAP50 of 90.1%. The model's Giga Floating Point Operations Per Second was 11.4, and the number of parameters was 3,139,819. These results demonstrate the effectiveness of the proposed method in enhancing both model efficiency and performance. Conclusions: The SlimMSDA-YOLO model is lightweight and efficient, making it ideal for real-time laryngeal structure recognition during intubation. Comparative experiments with other lightweight segmentation networks highlight the effectiveness and superiority of the proposed approach.

Keywords: Tracheal intubation, YOLOv8, laryngeal structures, segmentation

Introduction

Tracheal intubation is a critical, life-saving procedure commonly used in surgeries, emergency medicine, and intensive care units (ICUs) to maintain airway patency and facilitate mechanical ventilation or anesthesia delivery [1]. This procedure is vital for preserving life and reducing mortality, particularly in cases of airway obstruction, respiratory failure, or during general anesthesia [2]. However, improper intubation can lead to complications such as misplacement into the esophagus, laryngeal trauma, or vocal cord injury, all of which increase morbidity and mortality risks [3]. These risks are further heightened in specific clinical situations. For example, difficult airway management-encountered in patients with anatomical abnormalities (e.g., obesity, cervical spine injuries, or congenital malformations) presents challenges such as limited visualization and increased procedural complexity [4]. Similarly, pediatric intubation is complicated by smaller anatomical structures, dynamic airway changes during growth, and a higher risk of tissue damage [5]. The success of endotracheal intubation largely depends on the operator's skill, especially in cases involving challenging airways. To address these difficulties, tools such as fiberoptic bronchoscopes and video laryngoscopes have been developed, providing operators with clear visualization of critical structures on a display screen [6]. Although these tools are effective, they require extensive training, and profi-

https://doi.org/10.62347/BIHI3707

ciency levels can vary significantly, particularly in emergency situations involving critically ill patients.

With rapid advancements in computer vision, image analysis has become increasingly important in medical applications, including the study of laryngeal structures. Ding et al. [7] proposed the CN-DA-Unet network based on U-Net for end-to-end segmentation of the glottis, incorporating color normalization of images before feature extraction and fusion, which improved segmentation results. Ren et al. [8] utilized a transfer learning approach using a pre-trained ResNet-101 model to classify normal laryngeal structures, vocal nodules, polyps, leukoplakia, and malignant tumors. Degala et al. [9] applied a basic U-Net model for glottis segmentation. Xiong et al. [10] used a deep convolutional neural network (CNN) to differentiate between laryngeal cancer, precancerous lesions, benign laryngeal tumors, and normal tissues.

Despite these advancements, no study has yet addressed laryngeal structure recognition in the context of endotracheal intubation, nor has a comprehensive recognition system been developed for various laryngeal structures during intubation. Existing methods primarily focus on static or pathological laryngeal analysis, which lacks the adaptability required for dynamic intubation scenarios. Unlike conventional static laryngeal recognition, dealing with still images or pathological conditions, the recognition of laryngeal structures during endotracheal intubation presents unique challenges. These include real-time detection of rapidly changing and often occluded anatomical features, as well as the need for high accuracy in critical clinical environments. The dynamic nature of intubation requires recognition systems that can not only detect and analyze laryngeal structures but also adapt to real-time changes in the airway, making it more complex than traditional static analyses.

To address these challenges and minimize errors during intubation, we integrated artificial intelligence (AI) into the procedure. To enhance the safety and efficiency of intubation, we developed a deep learning algorithm based on YOLO (You Only Look Once) [11], which enables real-time segmentation of laryngeal structures, including the tongue, palate, uvula, pharyngeal wall, cartilago epiglottica, supraglottic region, glottic fissure, vocal cords, and endotracheal tube. The results are displayed on a screen for real-time observation by the operator. YOLO is a single-stage object detection algorithm known for its superior detection speed compared to traditional two-stage algorithms. Unlike two-stage methods, which first generate region proposals and then classify them, YOLO directly predicts bounding boxes and class probabilities in a single pass, streamlining the detection process and significantly improving efficiency. The latest version of YOLO, the YOLOv8 model, supports both object detection [12] and instance segmentation [13], with the segmentation variant, YOLOv8-seg, providing precise object boundaries by segmenting objects at the pixel level in combination with detection boxes.

Given the complexity and variability of laryngeal anatomy, along with subtle color differences at the boundaries of physiological tissue structures, we made three primary contributions to the YOLOv8-seg model to improve recognition accuracy and real-time performance.

1. We replaced certain convolutions and the C2f module in the Neck structure of YOLOv8n-seg with the Group Shuffle Convolution (GSConv) and VoVGSCSP modules from the Slimneck structure. This modification reduced model parameters and improved processing speed. Experimental results demonstrated that the lightweight neck structure performed effectively in laryngeal dataset segmentation.

2. To enhance recognition accuracy for laryngeal structures, we introduced a Multi-Scale Dilation Attention (MSDA) mechanism at the Neck-Head junction. This mechanism ensures effective channel information transmission and refined processing, which is crucial for segmenting intricate and minute structures within laryngeal images. The MSDA mechanism strengthens feature representation while suppressing irrelevant or noisy information.

3. The proposed algorithm improves the accuracy of laryngeal structure recognition while reducing computational complexity and the number of parameters.

Methods

The YOLOv8-seg model

YOLO (You Only Look Once) is a highly efficient single-stage object detection algorithm that

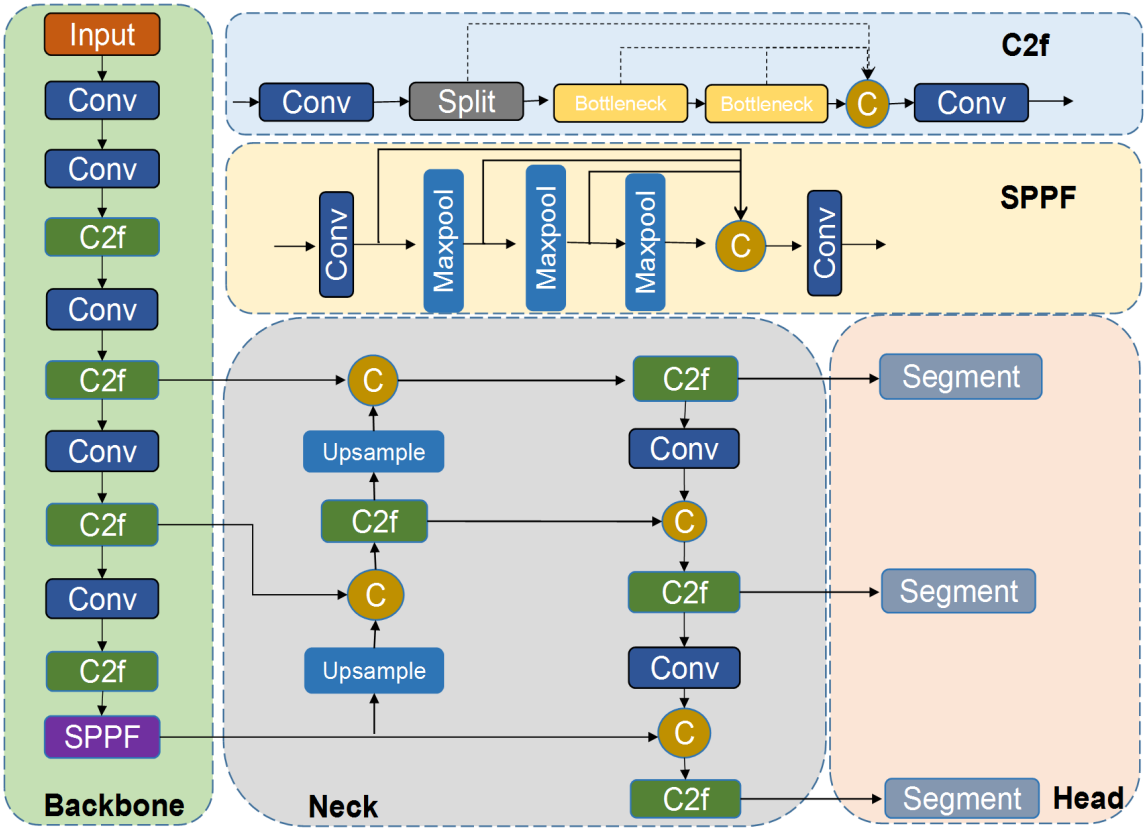

Figure 1. YOLOv8-seg network structure. C2f: Cross Feature Fusion Module; SPPF: Spatial Pyramid Pooling Fast.

performs both localization and classification in a single forward pass, significantly improving detection speed compared to traditional two-stage algorithms like Faster R-CNN [14]. YOLOv8 introduces several improvements in the YOLO series, as shown in **Figure 1**, including a lightweight convolutional network architecture that incorporates a C2f module to enhance feature reuse and computational efficiency. The backbone of YOLOv8 extracts multi-level features from the input image, while the SPPF module extends the receptive field using multi-scale pooling. The model's Neck integrates Feature Pyramid Network [15] and Path Aggregation Network [16] to fuse multi-level feature maps, thereby improving detection across various object scales. The detection head uses dynamic anchor-free bounding boxes, allowing the model to directly learn bounding box parameters, thus optimizing the detection process and improving segmentation accuracy.

YOLOv8 offers five model variants: YOLOv8n (nano), YOLOv8s (small), YOLOv8m (medium), YOLOv8L (large), and YOLOv8x (extra-large), primarily differing in the number of layers and parameters. YOLOv8-seg extends YOLOv8 to handle both object detection and semantic segmentation tasks. In addition to providing precise bounding boxes, YOLOv8-seg also delivers detailed classification information for each pixel. Compared to the standard YOLOv8 model, YOLOv8-seg introduces a segmentation branch that generates high-resolution segmentation masks through upsampling and convolution, ultimately producing segmented mask images. YOLOv8-seg incorporates several loss functions, including segmentation loss (seg_loss), bounding box loss (box_loss), classification loss (cls_loss), and distribution focal loss (DFL_loss) to enhance segmentation accuracy.

The model improvements in this study are based on the YOLOv8n-seg architecture. Since our goal is to integrate the algorithm into a device, we focused on designing a lightweight model structure.

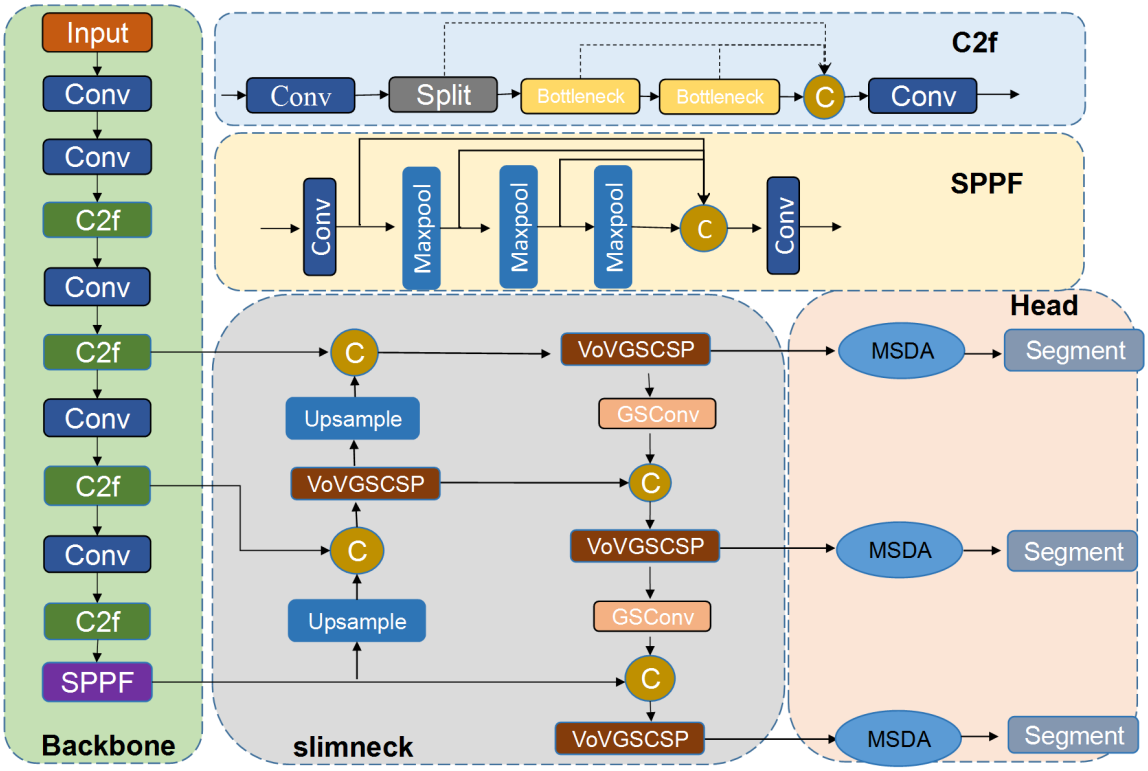

Figure 2. SlimMSDA-YOLO network structure. C2f: Cross Feature Fusion Module; SPPF: Spatial Pyramid Pooling Fast.

The proposed SlimMSDA-YOLO model

The structure of the improved SlimMSDA-YOLO model is shown in **Figure 2**. In the original network, we incorporated a Slimneck structure and introduced a MSDA [17] mechanism between the neck and head sections. This improvement enhances model accuracy while maintaining a lightweight architecture. The following sections provide a detailed introduction to these improved modules.

The slimneck module

In this study, we optimized the YOLOv8n-seg model for laryngeal structure recognition by improving the Neck module. The original YOLOv8n-seg network was modified by introducing the Slimneck structure in the neck section, which is based on GSConv [18] and the GSBottleneck module. To create a lightweight network, we replaced the standard convolution in the neck with GSConv. The structure of GSConv is shown in **Figure 3**. GSConv is a lightweight convolution operation that combines Depthwise Separable Convolution [19], SC (Standard Convolution), and a shuffle operation. This reduces computational complexity while enhancing the model's feature extraction capability. The calculation formulas are as follows:

$$T_{SC} = O(K^2 \cdot C_{in} \cdot C_{out} \cdot H_{out} \cdot W_{out}) \quad (1)$$

$$T_{DSC} = O(K^2 \cdot C_{in} \cdot C_{out} \cdot H_{out} \cdot W_{out}) + O(C_{in} \cdot C_{out} \cdot H_{out} \cdot W_{out}) \quad (2)$$

$$T_{GSConv} = O(K^2 \cdot \frac{C_{in}}{G} \cdot \frac{C_{out}}{G} \cdot H_{out} \cdot W_{out}) \quad (3)$$

In the formula, K^2 represents the size of the convolution kernel, while C_{in} and C_{out} denote the number of channels in the input and output feature maps, respectively. H_{out} and W_{out} represent the height and width of the output feature map, and G denotes the number of groups, which divides the input channels into groups. As shown in the formula (1) (2) (3), the computational complexity of GSConv is minimal.

GSBottleneck is an enhanced module based on GSConv, as shown in **Figure 4**. By chaining two lightweight GSConv operations, GSBott-

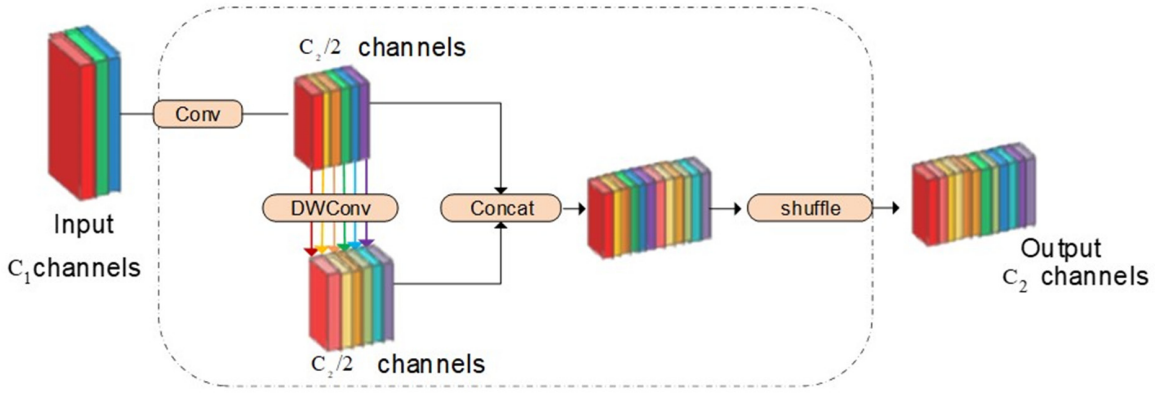

Figure 3. The GSConv module structure diagram. GSConv: Group Shuffle Convolution.

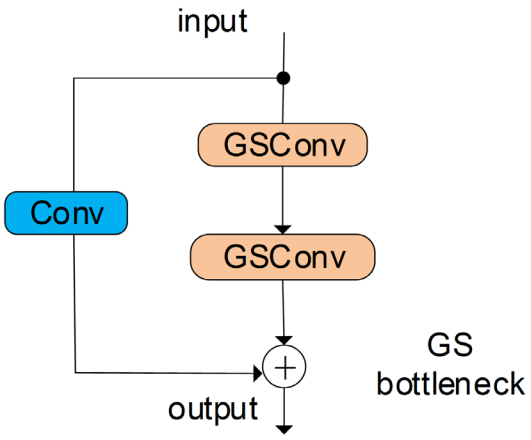

Figure 4. GSbottleneck module structure diagram.

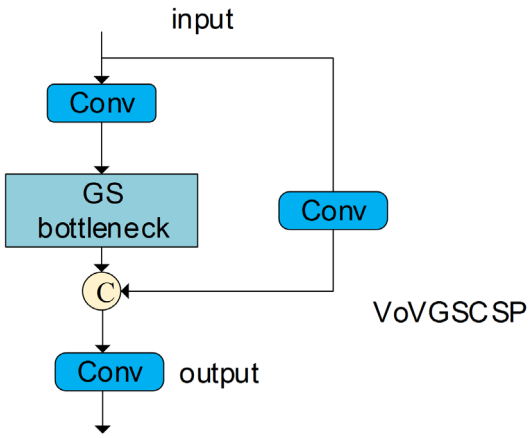

Figure 5. The VoVGSCSP structure diagram.

leneck creates an efficient feature extraction unit, significantly reducing the model's parameter count while maintaining strong feature representation.

Based on the GSBottleneck, the VoVGSCSCP module was further designed, as shown in **Figure 5**. VoVGSCSP is a multi-scale feature fusion module based on the GSBottleneck structure. This module relies on the GSBottleneck within its internal structure to perform feature extraction tasks. The lightweight convolution and optimized feature fusion approach enable the network to achieve strong performance with a reduced number of parameters, resulting in a stacked structure called "slim-neck". Specifically, in the original YOLOv8n-seg neck, we replaced the standard convolution with GSConv and substituted the original C2f module with the VoVGSCSP module.

The MSDA module

In the YOLOv8n-seg network, the Neck structure is responsible for further fusing and enhancing multi-scale features extracted from the backbone, while the Head performs final target classification, bounding box prediction, and segmentation mask generation. The connection between the Neck and Head is crucial for transferring feature information throughout the model, particularly for segmenting small and complex structures in medical images, where precise and effective information transfer is essential.

The anatomical complexity of laryngeal structures, with subtle and blurred feature details, presents challenges for traditional convolutional or feature fusion layers, which may struggle to capture multi-scale information. This limitation makes it difficult for the model to effective-

ly segment these complex features. To improve feature representation, we introduced a MSDA [18] mechanism after the VoVGSCSP module at the connection between the Neck and Head structure, as shown in **Figure 2**. This enhancement boosts the perception and segmentation accuracy of complex laryngeal features. The MSDA mechanism adaptively focuses on critical information at different scales, effectively highlighting important features while suppressing irrelevant or noisy information.

MSDA captures multi-scale feature information by assigning different dilation rates to different attention heads. Specifically, the input feature map is linearly projected and divided into multiple heads, with each head performing the Sliding Window Dilated Attention (SWAD) operation using a distinct dilation rate. SWAD leverages the sparsity of the self-attention mechanism at different scales to aggregate semantic information within the receptive field at various scales. This approach effectively reduces redundancy in the self-attention mechanism without introducing additional computational overhead. The formulation of SWAD is described as follows:

$$X = SWDA(Q, K, V, r) \tag{4}$$

Here, Q, K, and V represent the query, key, and value matrices of the feature map X, respectively, while r denotes the dilation rate, which determines the number of keys and values participating in self-attention within the sliding window. In this mechanism, the dilation rate r controls the receptive field size within the sliding window, allowing the model to integrate features from different regions in a multi-scale manner. The formula for MSDA is as follows:

$$\begin{aligned} h_i &= SWDA(Q_t, K_t, V_t, r_t), 1 \le t \le n \\ X &= Linear\left(\text{Concat}\left[h_1, \ldots, h_n\right]\right) \end{aligned} \tag{5}$$

In this formula, the input feature maps Q_t, K_t, and V_t are slices of the current feature map's query, key, and value, respectively. Features are divided into n heads, with each head performing SWAD at a unique dilation rate r_t. The output of each head h_i is based on the attention results from these inputs and dilation rates. All heads h_1 to h_n are then concatenated and passed through a linear layer (fully connected layer) to aggregate features. This linear layer combines the multi-scale features from each head, yielding an overall feature representation X. **Figure 6** illustrates the working principle of MSDA, which, by default, uses a 3×3 convolution kernel with dilation rates of 1, 2, and 3, resulting in attention receptive field sizes of 3×3, 5×5, and 7×7 for different heads.

Experiments and results

Dataset and pre-processing

The dataset used in this study was collected during tracheal intubation procedures conducted in the Department of Anesthesiology at the First Hospital of Putian City from 2022 to 2023. It consists of 100 videos, encompassing a diverse range of patients in terms of age, gender, and clinical conditions. The images were captured using a video laryngoscope, ensuring the inclusion of eight key laryngeal structures and one device: the tongue, palate, uvula, pharyngeal wall, cartilago epiglottica, supraglottic region, glottic fissure, vocal cords, and endotracheal tube. The videos were split into frames, resulting in approximately 1,100 high-quality, non-redundant images selected for annotation.

Figure 7 illustrates the label distribution and bounding box characteristics of the laryngeal dataset used for YOLO model training. The visualization includes attributes such as class instance distribution, bounding box center coordinates, bounding box size, and spatial location distribution. The histogram in the top-left corner highlights a significant imbalance in the number of instances across different classes within the dataset. This class imbalance could potentially affect model performance, as some categories are underrepresented. To address this, YOLOv8's built-in data augmentation techniques were applied during data loading, helping to mitigate the imbalance and improving the model's robustness across all categories.

All images were manually annotated by experienced anesthesiologists using Labelme [20], an open-source image annotation tool supporting object detection and semantic segmentation. The annotation process employed polygon-based labeling, to ensure precise delineation of anatomical structures such as the cartilago epiglottica, glottic fissure, and vocal cords. These polygonal annotations were converted into segmentation masks, which are essential

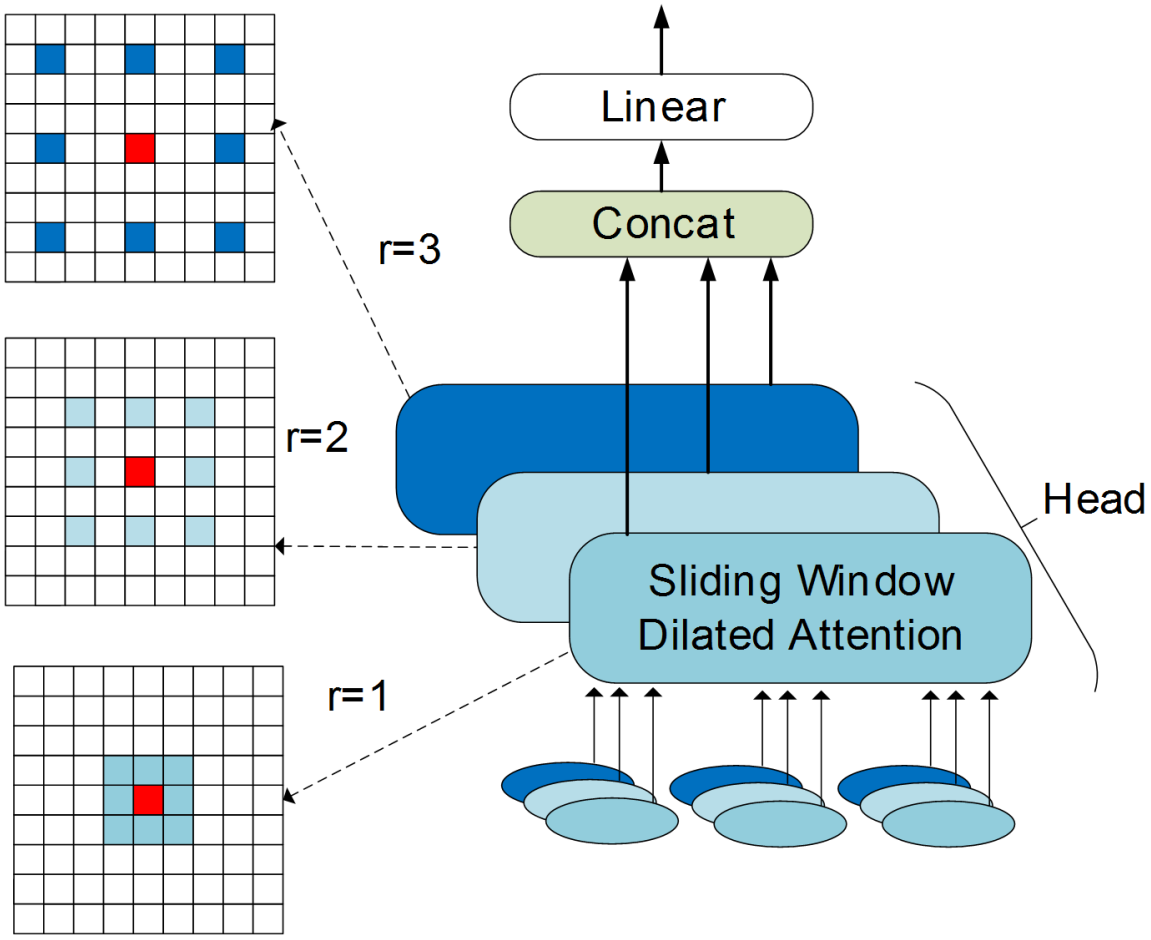

Figure 6. The MSDA structure diagram. MSDA: Multi-Scale Dilation Attention.

for training the YOLOv8n-seg model. To ensure accuracy and consistency, each image underwent multiple review rounds. The annotated dataset, referred to as the "VL dataset", consists of images with a resolution of 640×480. The dataset was split into an 8:2 ratio for training and validation, with an independent clinical dataset designated as the test set. This structured division ensures robust model evaluation and reduces overfitting. **Figure 8** presents sample images with their respective annotations.

Implementation details

The hardware configuration used for the experiments is as follows: an Intel(R) Xeon(R) Gold 6258R CPU @ 2.70 GHz, 256 GB of RAM, and an NVIDIA RTX A6000 GPU. The operating system is 64-bit Ubuntu 11.2.0. The software environment includes Visual Studio Code as the code editor, PyTorch 1.13.0 as the deep learning framework, and Python 3.9. The GPU driver version is 515.65.01, and the CUDA version is 11.7. Detailed parameter information is provided in **Table 1**.

Evaluation indicators

To evaluate the model's performance, we used metrics such as precision, recall, mAP50, Giga Floating Point Operations Per Second (GFLOPS), and the number of parameters. First, we define some variables used in the formulas: TP represents true positives, FP represents false positives, and FN represents false negatives. Additionally, weights, parameters, and GFLOPS were used to assess the model's complexity.

Precision measures the proportion of predicted positive pixels that are actually true positives. The formula is as follows:

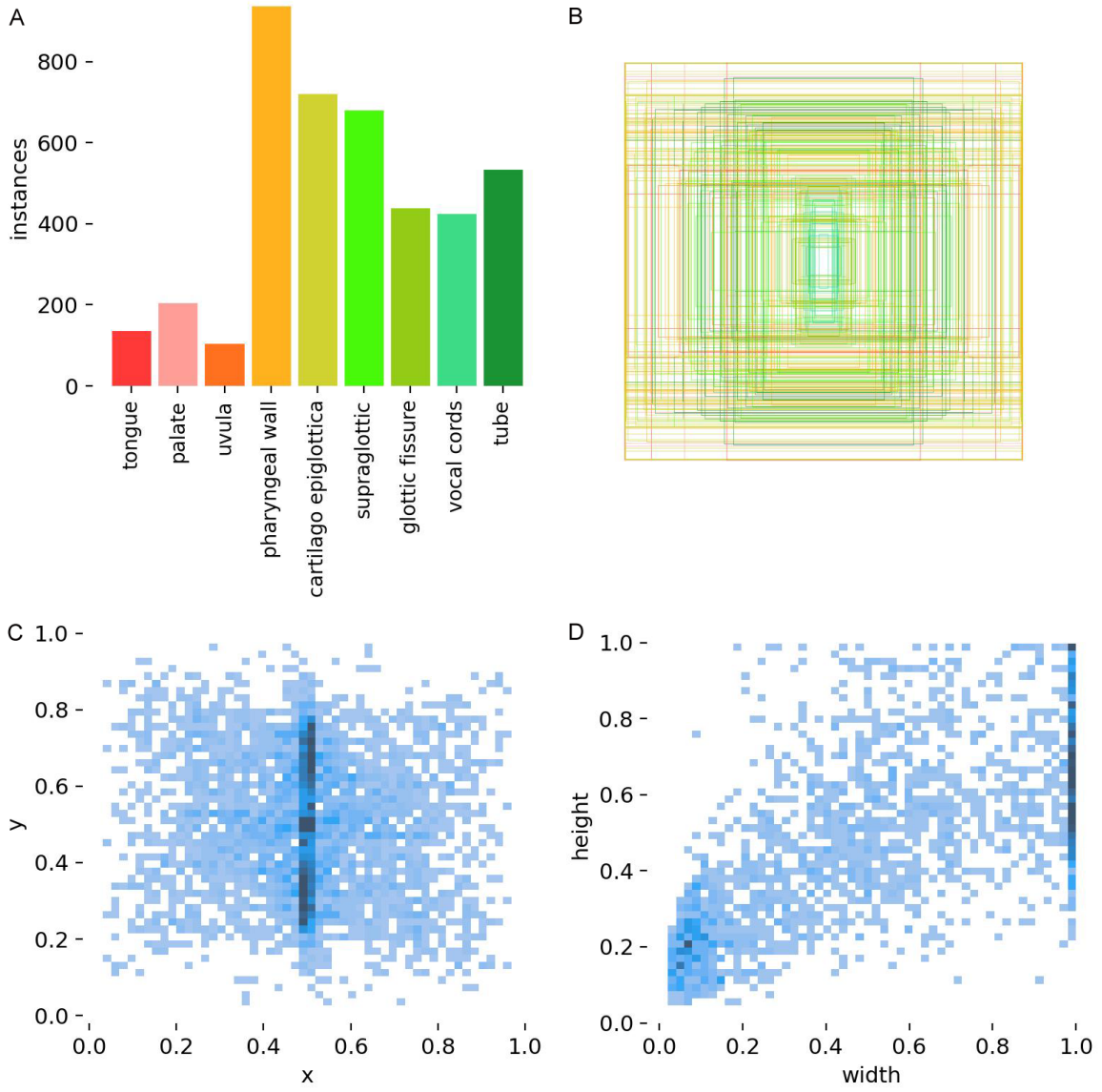

Figure 7. Visualization results of the dataset.

$$\text{precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall indicates the proportion of true positives correctly predicted by the model out of all actual positives, reflecting the model's ability to capture the target. The formula is as follows:

$$\text{recall} = \frac{TP}{TP + FN} \tag{7}$$

mAP50 is a key metric for evaluating overall performance in object detection and segmentation tasks and is widely used in YOLO models. mAP50 represents the mean precision across all classes at a 50% Intersection over Union (IoU) threshold. IoU measures the overlap between predicted and ground-truth boxes, with higher IoU values indicating better alignment. mAP50 is calculated by taking the weighted average precision across all classes, as shown in formulas (8) and (9). P(r) represents the precision value on the Precision-Recall curve.

$$\text{mAP50} = \frac{1}{N}\sum_{i=1}^{N} AP_{i,50} \tag{8}$$

$$\text{AP50} = \int_0^1 P(r)\,dr \tag{9}$$

GFLOPS, or Giga Floating Point Operations Per Second, represents the number of billions of

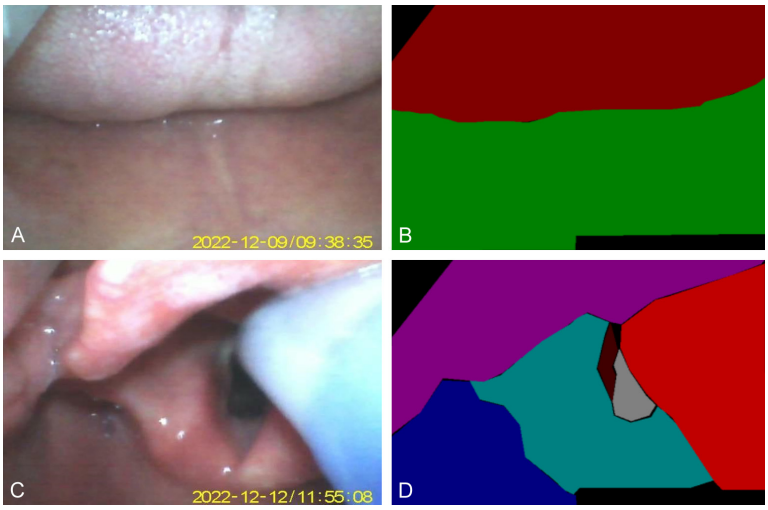

Figure 8. Throat images and corresponding labels. (A) and (C) images, (B) and (D) labels.

Table 1. Hardware specifications

Component	Specification
CPU	Intel(R) Xeon(R) Gold 6258R CPU @ 2.70 GHz
GPU	NVIDIA Corporation GA102GL [RTX A6000]
Memory	256
CUDA	CUDA 11.7
Python	Python 3.9
PyTorch	PyTorch1.13.0

Notes: CPU: Central Processing Unit; GPU: Graphics Processing Unit; CUDA: Compute Unified Device Architecture.

Table 2. Comparison results with baseline

Models	mPrecision	mRecall	mAP50	GFLOPS	Parameters
YOLOv8n-seg	85.4%	81.4%	89.1%	12.6	3398187
SlimMSDA-YOLO	90.4%	84.2%	90.1%	11.4	3139819

floating-point operations the computing device performs per second. Lower GFLOPS values can significantly improve runtime efficiency in practical applications.

The parameters refer to the total number of parameters that need to be trained and optimized in the model, typically measured in millions (M) or billions (B). A larger number of parameters generally indicates greater model capacity and representation power.

YOLOv8 and SlimMSDA-YOLO experimental results

To compare the performance of the improved model with the original YOLOv8 on our dataset, this section analyzes the experimental results before and after the improvement. **Table 2** presents the performance comparison in terms of precision, recall, mAP50, GFLOPS, and parameter count. From **Table 2**, we observed that in the original YOLOv8n-seg network, precision, recall, and mAP50 were 85.4%, 81.4%, and 89.1%, respectively, while in the improved model, these metrics increased to 90.4%, 84.2%, and 90.1%, respectively. Additionally, GFLOPS decreased from 12.6 to 11.4, and the number of parameters reduced from 3.4M to 3.1M. These improvements indicate that the model offers higher segmentation accuracy while reducing computational complexity, providing clinicians with a more real-time visual experience and laying a solid foundation for future clinical applications.

Table 3 presents the segmentation metrics for each laryngeal class in both the YOLOv8n-seg and SlimMSDA-YOLO models, comparing three representative precision indicators. The vocal cords are a critical structure to avoid during intubation, as a clearly visible vocal cord boundary enables doctors to maneuver the tube accurately through the glottis without injuring the vocal cords. High-precision segmentation of the vocal cords is thus essential for surgical safety. For vocal cord segmentation, the improved model increased precision from 70.3% to 93.7%, recall from 46.4% to 64.2%, and AP50 from 59% to 74.8%, significantly reducing the probability of false positives.

The cartilago epiglottica, which covers the tracheal opening, is another key structure for preventing aspiration. During intubation, it is essential for physicians to clearly identify and avoid the cartilago epiglottica to ensure smooth intubation. The improved model increased

Table 3. Comparison results table with baseline for each class

	YOLOv8n-seg			SlimMSDA-YOLO		
	Precision%	Recall%	AP50%	Precision%	Recall%	AP50%
Tongue	84.1	86	99.5	94	87	93.4
Palate	86.2	84.6	85.5	86.6	80.5	86.7
Uvula	80	75	94.5	90.5	92.9	97.6
Pharyngeal wall	86	90.1	91.5	84.3	85.7	89.2
Cartilago epiglottica	92.6	94	99.5	96.2	94.1	97.2
Supraglottic	91.8	81.5	86.1	88.5	87.9	92.7
Glottic fissure	80.1	77.5	87	81.2	68.5	79.6
Vocal cords	70.3	46.4	59	93.7	64.2	74.8
Tube	97.9	98	99.5	98.9	97.3	99.4

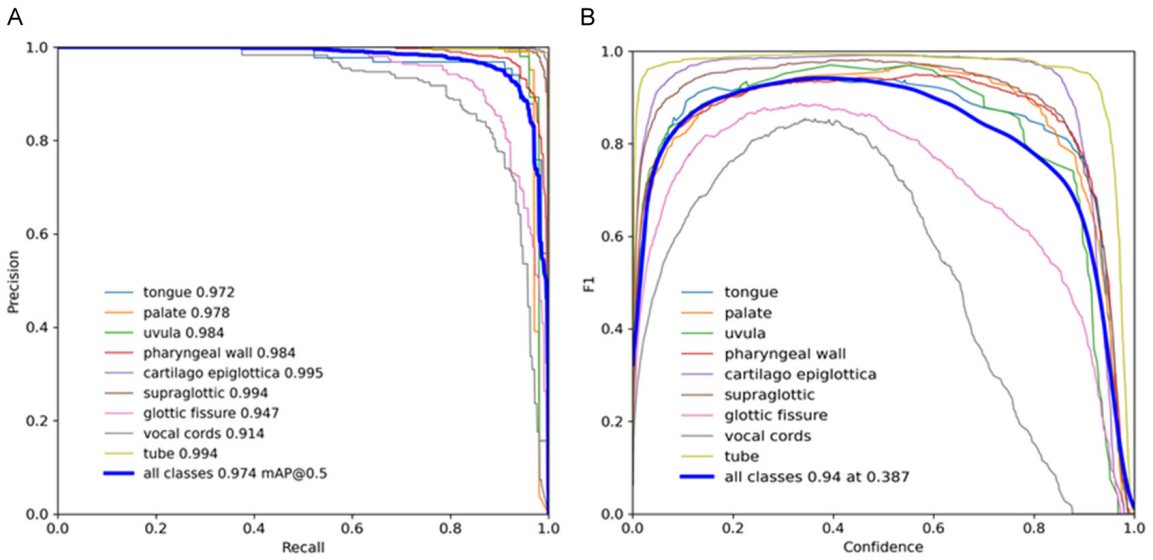

Figure 9. Precision-Recall curve (A) and F1-Confidence of curve (B).

precision for cartilago epiglottica segmentation from 92.6% to 96.2%, with AP50 remaining at 97.2%. This enhancement allows the model to more accurately segment the cartilago epiglottica in complex environments, aiding doctors in safely bypassing this structure.

For the tracheal tube, precision improved from 97.9% to 98.9%, ensuring that clinicians can reliably track the tube's position in real-time within the laryngoscopic view, reducing the risk of errors during intubation. Although the uvula is not a primary target in the intubation process, accurate segmentation can provide physicians with a better understanding of the overall laryngeal structure. In the uvula segmentation task, the improved model showed a significant increase, with precision rising from 80% to 90.5%, recall from 75% to 92.9%, and AP50 from 94.5% to 97.6%.

The glottic fissure, a key entry point for endotracheal intubation, saw an improvement in precision from 80.1% to 81.2% with the enhanced model, although recall slightly decreased, maintaining a stable overall performance. Accurate localization of the glottis is essential for minimizing patient discomfort.

In addition to improved segmentation accuracy, the SlimMSDA-YOLO model, with its lightweight Slimneck structure, reduced GFLOPS from 12.6 to 11.4, significantly lowering computational complexity. This optimization enables real-time operation on low-resource devices. In procedures like endotracheal intu-

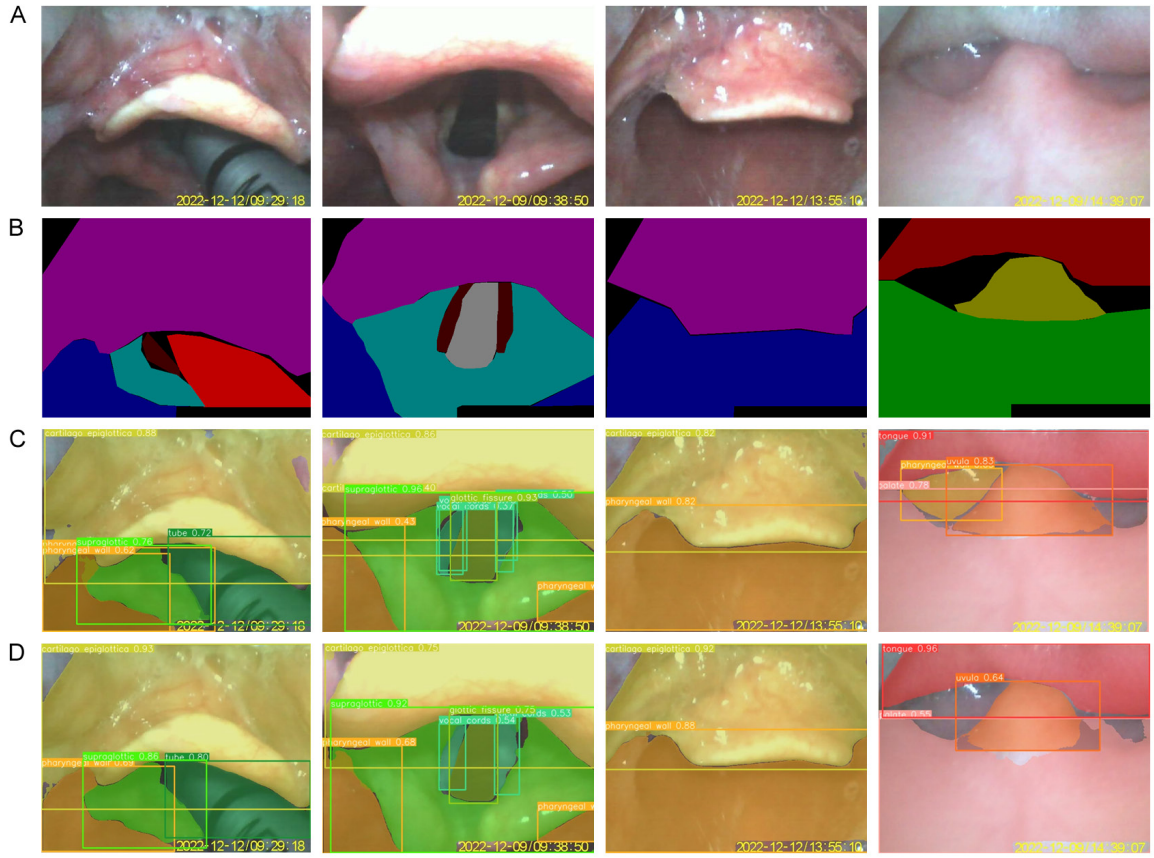

Figure 10. Result Visualization: (A) Original Image, (B) Ground Truth, (C) YOLOv8n-seg Segmentation Results, (D) SlimMSDA-YOLO Segmentation Results. MSDA: Multi-Scale Dilation Attention.

bation, where rapid response is critical, delays in segmentation can disrupt the clinician's workflow and even increase surgical risks. The improved model not only enhances accuracy but also maintains speed, making it suitable for deployment in laryngoscopic devices.

Figure 9A presents the PR curve of the SlimMSDA-YOLO algorithm on the test set, with the x-axis representing recall and the y-axis representing precision. The closer the PR curve is to the top right corner, the better the segmentation performance.

Figure 9B shows the F1 curve of the SlimMSDA-YOLO model on the test set. The F1 curve reflects the balance between the model's precision and recall, with values ranging from 0 to 1. The curve annotation indicates "all classes 0.94 at 0.387", meaning that when the confidence threshold is set to 0.387, the model achieves an overall F1 score of 94% across all classes. This suggests that selecting this confidence threshold allows the model to achieve a good balance between precision and recall for segmentation. The confidence threshold helps filter out pixels or regions with confidence levels above a certain threshold, ensuring stability and reliability in segmentation results. By adjusting the confidence threshold, an optimal balance between accuracy and efficiency can be found, leading to better segmentation performance. In summary, the improved SlimMSDA-YOLO model demonstrates excellent performance on the custom laryngeal dataset, accurately segmenting and identifying each target class.

To more intuitively observe the segmentation performance of the model before and after improvement, we selected representative segmentation results for display in **Figure 10**. The improved SlimMSDA-YOLO effectively detects and labels different laryngeal regions, achieving both segmentation mask labeling and bounding box detection. The detection box dis-

Table 4. Comparative analysis of ablation experiments

Models	mPrecision%	mRecall%	mAP50%	GFLOPS	Parameters
YOLOv8n-seg	85.4	81.4	89.1	12.6	3398187
YOLOv8n-seg + Slimneck	90	83.4	89.1	11.2	3052331
YOLOv8n-seg + MSDA	91.2	84	89.9	12.8	3405679
Ours	90.4	84.2	90.1	11.4	3139819

Notes: GFLOPS: Giga Floating Point Operations Per Second; MSDA: Multi-Scale Dilation Attention.

Table 5. Compare the results with other models

Models	mPrecision%	mRecall%	mAP50%	GFLOPS	Parameters
YOLOv5n-seg	85	81	85	15.7	4259845
YOLOv7n-seg	87	83	88	13.5	3969372
YOLOv10n-seg	87	84	87	14.2	3605675
Ours	90.4	84.2	90.1	11.4	3139819

Note: GFLOPS: Giga Floating Point Operations Per Second.

plays the confidence level, indicating the model's confidence in the detection result, with values ranging from 0 to 1; the closer the value is to 1, the higher the model's confidence in the detection. This model efficiently segments and labels various laryngeal regions. As shown in the figure, the model accurately segments target areas across all nine laryngeal structure categories, demonstrating the robustness and anti-interference capabilities of the SlimMSDA-YOLO model in the segmentation task. This is particularly evident in detailed areas like the glottis and vocal cords, where the model exhibits high accuracy and applicability.

Ablation experiment

We integrated two different optimization methods into the original YOLOv8n-seg network to construct ablation experiments and verify the effectiveness of each module. Four configurations were developed: YOLOv8n-seg, Slimneck + YOLOv8n-seg, MSDA + YOLOv8n-seg, and Slimneck + MSDA + YOLOv8n-seg. These networks were trained on our custom laryngeal structure dataset for 150 epochs under identical experimental configurations. The experimental results are shown in **Table 4**.

As shown in **Table 4**, the improved model demonstrated significant performance gains compared to the original YOLOv8n-seg model. First, with the introduction of the Slimneck structure, GFLOPS reduced to 11.2, and the parameter count decreased to 3,052,331. This reduction in computational complexity and parameter count was accompanied by an improvement in precision to 90%. This enhancement reduces computational load and contributes to smoother intubation procedures by enabling real-time identification of laryngeal structures. Second, when the MSDA attention mechanism was added, precision increased to 91.2%, recall rose to 84%, and mAP50 improved to 89.9%. Although GFLOPS and parameters slightly increased from 12.6 to 12.8, the MSDA mechanism significantly enhanced the model's ability to recognize fine details in laryngeal structures, aiding in the accurate localization of critical regions. Finally, after integrating both the Slimneck and MSDA modules, further improvements were observed in both precision and model speed. The combination of these two modules demonstrated superior performance in real-time segmentation of laryngeal structures.

The ablation experiments show that the combination of the Slimneck lightweight design and the MSDA attention mechanism for global feature awareness effectively improves the model's performance in recognizing laryngeal structures in complex environments. This has important implications for clinical applications in endotracheal intubation, enhancing procedural safety and ensuring real-time operation.

Comparison experiments

To validate the detection performance of the proposed network on the laryngeal dataset, this section presents comparative experiments

with other lightweight segmentation models, including YOLOv5n-seg [21], YOLOv7n-seg [22], and YOLOv10n-seg [23]. All experiments were conducted under the same experimental configurations and on the same dataset. The following section analyzes the experimental results.

Table 5 compares the improved SlimMSDA-YOLO algorithm with other mainstream lightweight segmentation algorithms. In terms of mprecision, the SlimMSDA-YOLO model achieved an accuracy of 90.4%, significantly outperforming YOLOv5n-seg at 85% and YOLOv7n-seg and YOLOv10n-seg at 87%, indicating that the improved model has stronger accuracy in segmentation tasks. For recall, SlimMSDA-YOLO achieved 84.2%, slightly higher than YOLOv7n-seg at 83% and YOLOv5n-seg at 81%, and comparable to YOLOv10n-seg at 84%. This demonstrates that the SlimMSDA-YOLO model maintains high precision while retaining good target detection capability. In terms of mAP50, SlimMSDA-YOLO performed well with a score of 90.1%, exceeding YOLOv5n-seg at 85%, YOLOv7n-seg at 88%, and YOLOv10n-seg at 87%, further confirming the segmentation effectiveness of the improved model. Additionally, the improved model excels in computational complexity (GFLOPS), with a GFLOPS of 11.4, lower than other models, which range from 13.5 to 15.7. This indicates that SlimMSDA-YOLO has lower computational overhead and higher operational efficiency. Finally, in terms of parameter count, SlimMSDA-YOLO has 3,139,819 parameters, far fewer than YOLOv5n-seg (4,259,845), YOLOv7n-seg (3,969,372), and YOLOv10n-seg (3,605,675). This demonstrates that the improved model achieves a more lightweight design while maintaining precision, making it more suitable for resource-limited environments. Overall, the improved SlimMSDA-YOLO outperforms other models in terms of precision, recall, mAP50, and GFLOPS, exhibiting a clear comprehensive advantage.

Discussion

In this study, we propose an improved lightweight YOLO model, SlimMSDA-YOLO, designed to enhance segmentation and detection accuracy in laryngeal images, with a particular focus on identifying key laryngeal structures and detecting pathological regions. Experimental results show significant improvements in precision, recall, and mAP50, while effectively reducing the parameter count and computational complexity. These advancements make SlimMSDA-YOLO well-suited for deployment in resource-limited embedded devices and real-time applications. The main innovations include the introduction of the Slimneck structure and the MSDA attention mechanism. The Slimneck structure optimizes both the lightweight convolution module and feature fusion module, reducing the model's size while improving its adaptability to laryngeal anatomical structures and maintaining high detection accuracy. The MSDA attention mechanism captures multi-scale spatial information, enhancing the model's focus on target regions and significantly improving segmentation and detection performance in the complex laryngeal anatomical environment. Additionally, experiments on public laryngeal datasets confirm the generalization and robustness of SlimMSDA-YOLO.

While SlimMSDA-YOLO demonstrates notable improvements in both accuracy and efficiency, its real-time performance in highly constrained environments (such as edge devices or embedded systems) could still face limitations due to hardware restrictions. Future work could explore the integration of advanced attention mechanisms, such as self-attention or transformer-based methods, to improve feature focusing, particularly for complex laryngeal structures where distinguishing subtle details is crucial. Moreover, multimodal data fusion, incorporating complementary information such as acoustic data or clinical text, could provide more comprehensive insights and improve the model's robustness across diverse clinical scenarios. These advancements could enhance both the accuracy and interpretability of the model, enabling it to better adapt to a variety of clinical needs. Despite these limitations, SlimMSDA-YOLO shows substantial promise in clinical applications, supporting anesthesiologists in managing complex airways and offering valuable assistance in training junior doctors. As technology advances, this model could become a core component of future intelligent medical devices, driving the integration of AI with medical image analysis.

Acknowledgements

This study was supported by the Putian Science and Technology Plan Project (2024SY002).

Disclosure of conflict of interest

None.

Address correspondence to: Xiaoping Huang, The First Hospital of Putian City, No. 449, Nanmen West Road, Chengxiang District, Putian 351100, Fujian, China. E-mail: 18950719090@189.cn; Qianwen Zhang, Department of Radiology, Changhai Hospital, Naval Medical University, Changhai Road 168, Shanghai 200433, China. E-mail: zhangqianwen_smmu@126.com

References

[1] Lapinsky SE. Endotracheal intubation in the ICU. Crit Care 2015; 19: 258.

[2] Higgs A, McGrath BA, Goddard C, Rangasami J, Suntharalingam G, Gale R and Cook TM. Difficult Airway Society; Intensive Care Society; Faculty of Intensive Care Medicine; Royal College of Anaesthetists. Guidelines for the management of tracheal intubation in critically ill adults. Br J Anaesth 2018; 120: 323-352.

[3] Blanc VF and Tremblay NA. The complications of tracheal intubation: a new classification with a review of the literature. Anesth Analg 1974; 53: 202-213.

[4] Apfelbaum JL, Hagberg CA, Connis RT, Abdelmalak BB, Agarkar M, Dutton RP, Fiadjoe JE, Greif R, Klock PA, Mercier D, Myatra SN, O'Sullivan EP, Rosenblatt WH, Sorbello M and Tung A. 2022 American Society of Anesthesiologists practice guidelines for management of the difficult airway. Anesthesiology 2022; 136: 31-81.

[5] Weiss M and Engelhardt T. Proposal for the management of the unexpected difficult pediatric airway. Paediatr Anaesth 2010; 20: 454-464.

[6] Kaplan MB, Ward DS and Berci G. A new video laryngoscope-an aid to intubation and teaching. J Clin Anesth 2002; 14: 620-626.

[7] Ding H, Cen Q, Si X, Pan Z and Chen X. Automatic glottis segmentation for laryngeal endoscopic images based on U-Net. Biomed Signal Process Control 2022; 71: 103116.

[8] Ren J, Jing X, Wang J, Ren X, Xu Y, Yang Q, Ma L, Sun Y, Xu W, Yang N, Zou J, Zheng Y, Chen M, Gan W, Xiang T, An J, Liu R, Lv C, Lin K, Zheng X, Lou F, Rao Y, Yang H, Liu K, Liu G, Lu T, Zheng X and Zhao Y. Automatic recognition of laryngoscopic images using a deep-learning technique. Laryngoscope 2020; 130: E686-E693.

[9] Degala D, Rao MA, Krishnamurthy R, Gopikishore P, Priyadharshini V, Prakash T and Ghosh PK. Automatic glottis detection and segmentation in stroboscopic videos using convolutional networks. Interspeech 2020; 4801-4805.

[10] Xiong H, Lin P, Yu JG, Ye J, Xiao L, Tao Y, Jiang Z, Lin W, Liu M, Xu J, Hu W, Lu Y, Liu H, Li Y, Zheng Y and Yang H. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. EBioMedicine 2019; 48: 92-99.

[11] Han X, Chang J and Wang KJPCS. You only look once: unified, real-time object detection. 2021; 183: 61-72.

[12] Reis D, Kupec J, Hong J and Daoudi A. Real-time flying object detection with YOLOv8. ArXiv 2023; [Epub ahead of print].

[13] Yue X, Qi K, Na X, Zhang Y, Liu Y and Liu C. Improved YOLOv8-Seg network for instance segmentation of healthy and diseased tomato plants in the growth stage. Agriculture 2023; 13: 1643.

[14] Ren S, He K, Girshick R and Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017; 39: 1137-1149.

[15] Lin TY, Dollár P, Girshick R, He K, Hariharan B and Belongie S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017. pp. 2117-2125.

[16] Liu S, Qi L, Qin H, Shi J and Jia J. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2018. pp. 8759-8768.

[17] Jiao J, Tang YM, Lin KY, Gao Y, Ma AJ, Wang Y and Zheng WS. Dilateformer: multi-scale dilated transformer for visual recognition. IEEE Transactions on Multimedia; 2023. pp. 8906-8919.

[18] Li H, Li J, Wei H, Liu Z, Zhan Z and Ren Q. Slim-neck by GSConv: a better design paradigm of detector architectures for autonomous vehicles. ArXiv 2022; [Epub ahead of print].

[19] Chollet F. Xception: deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017. pp. 1251-1258.

[20] Russell BC, Torralba A, Murphy KP and Freeman WT. LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 2008; 77: 157-173.

[21] Jiang S, Ao J, Yang H, Xie F, Liu Z, Yang S, Wei Y and Deng X. Fine-grained recognition of bitter gourd maturity based on improved YOLOv5-seg model. Sci Rep 2024; 14: 10856.

[22] Cao L, Zheng X and Fang L. The semantic segmentation of standing tree images based on the Yolo V7 deep learning algorithm. Electronics 2023; 12: 929.

[23] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J and Ding G. Yolov10: real-time end-to-end object detection. ArXiv 2024; [Epub ahead of print].