Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automatic diagnosis of laryngeal diseases and segmentation of laryngeal cancer based on deep learning technolog

Xiaohong Jiang
Chongqing Jiaotong University
Lili Chen
clili522@cqjtu.edu.cn
Chongqing Jiaotong University
Xiang Wang
Chongqing Jiaotong University
Xianwen Wu
The First Affiliated Hospital of Chongqing Medical University
Nannan Zhang
Chongqing Medical University

Research Article

Keywords: Diagnosis of laryngeal diseases, lightweight network, knowledge distillation, multi-scale feature fusion, Laryngeal Cancer Segmentation

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9551273/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Automatic diagnosis of laryngeal diseases and segmentation of laryngeal cancer based on deep learning technolog y

Xiaohong Jiang[1], Lili Chen[1*], Xiang Wang[1], Xianwen Wu[2], Nannan Zhang[3]

[1]Chongqing Jiaotong University, Chongqing, 400074, China

[2]Department of Otolaryngology, The First Affiliated Hospital of Chongqing Medical University, Chongqing,400000, China

[3]Chongqing Medical University, Chongqing, 400016, China

[*]Corresponding author(s): Lili Chen: clili522@cqjtu.edu.cn

Contributing authors: Xiaohong Jiang: jxh0014@mails.cqjtu.edu.cn, Xiang Wang: 622230040002@mails.cqjtu.edu.cn, Xianwen Wu: cnwxwok@126.com, Nannan Zhang: 836711690@qq.com

Abstract

Background: Laryngeal cancer is currently a common clinical malignancy, and its early diagnosis is of great significance for improving patient survival rates. With the widespread application of deep learning technology in the field of medical image analysis, this study aims to establish an intelligent diagnostic system for early screening and precise diagnosis of laryngeal cancer.

Methods: We design the ScSE and GC modules and develop a lightweight laryngeal disease classification network named LLLDNet, employing a combined enhancement strategy of transfer learning and knowledge distillation to improve the accuracy of laryngeal disease image classification. we also design the HA and MSFF modules to develop a laryngeal cancer segmentation network named HAM-Unet, which integrates hybrid attention and multi-scale feature fusion. Finally, we conduct experiments on the proposed LLLDNet classification network and HAM-Unet segmentation network using datasets of laryngeal diseases and laryngeal cancer, respectively.

Results: Experimental results demonstrate that LLLDNet achieves an accuracy of 93.0% and a recall of 92.4%. While improving various metrics including accuracy, LLLDNet reduces the number of model parameters to only 0.425M and FLOPs to 294M, achieving sufficient lightweight design, conserving computational resources, and accelerating inference. The HAM-Unet network

attains an IOU of 90.86% and an accuracy of 93.92%, significantly outperforming other segmentation models and enabling more precise delineation of laryngeal cancer lesion regions.

Conclusion: These research findings indicate that the intelligent diagnostic framework constructed in this study holds significant potential for improving the efficiency and accuracy of laryngeal cancer diagnosis, providing a feasible technical solution for clinical computer-aided diagnosis.

Keywords: Diagnosis of laryngeal diseases, lightweight network, knowledge distillation, multi-scale feature fusion, Laryngeal Cancer Segmentation

1 Introduction

Laryngeal cancer is a common head and neck malignancy worldwide, for which early diagnosis and precise treatment are crucial to preserve laryngeal function and improve patient survival[1]. In the medical field, deep learning has demonstrated significant potential and broad prospects for development. Particularly in the field of medical image analysis, artificial intelligence technologies based on convolutional neural networks (CNNs) have achieved major breakthroughs[2].

Accurately classifying laryngeal lesions as benign conditions or malignant laryngeal cancer is a critical step in screening and preliminary diagnosis. Laryngeal medical images exhibit complex and diverse tissue structures, posing several challenges in image processing. First, the laryngeal region contains multiple intricate and morphologically varied anatomical structures, which often have blurred boundaries and low contrast—particularly under pathological conditions—making accurate identification difficult. Second, inherent limitations in medical imaging equipment, such as noise, artifacts, and restricted resolution, further degrade image quality and complicate feature extraction. Traditional diagnostic methods for laryngeal lesions rely heavily on clinicians' expertise and endoscopic examinations, which are often subjective, time-consuming, and require specialized operators[3].

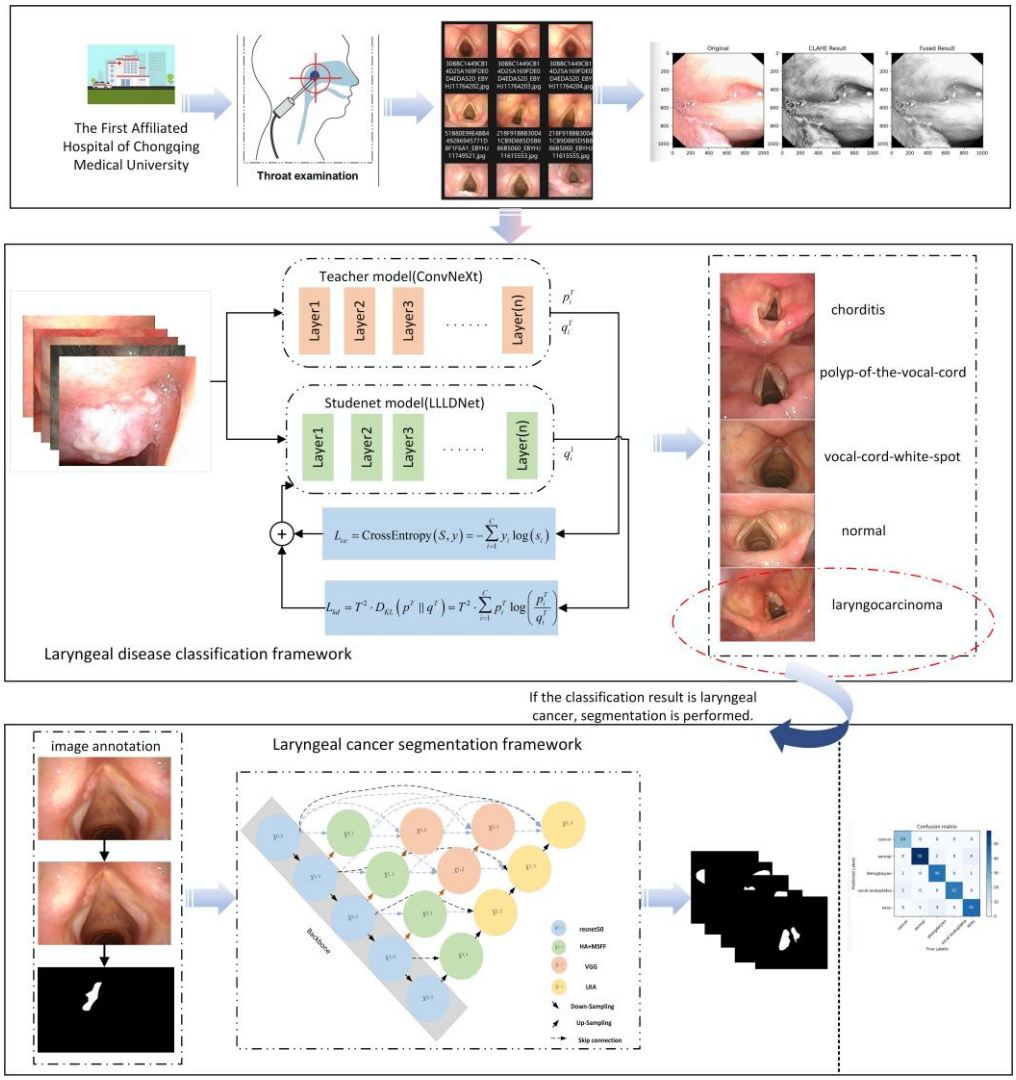


Fig.1 Overall framework for the intelligent diagnosis of laryngeal diseases. This framework integrates the lightweight classification network LLLDNet and the segmentation network HAM-Unet, realizing an automated analysis pipeline from multi-class laryngeal disease screening to fine-grained contour extraction of laryngeal cancer lesions.

Upon identification of a laryngeal lesion as cancerous, pixel-level precise segmentation is performed to delineate the tumor boundary, which serves as a key basis for T-stage classification. The segmentation results enable quantitative assessment of invasion depth and extent, providing clinicians with more objective medical evidence. The task of laryngeal cancer segmentation faces even greater challenges: in early-stage laryngeal cancer or precancerous lesions, the color and texture differences between pathological and surrounding healthy mucosa are extremely subtle, resulting in blurred boundaries between lesions and normal tissues

and often leading to incomplete segmentation. Moreover, due to the deep anatomical location of the larynx, factors such as illumination and angle during image acquisition can introduce shadows, which may obscure textural features. This compromises the completeness of information extracted during segmentation and leads to inaccuracies in the results.

Traditional machine learning approaches for laryngeal lesion recognition typically rely on handcrafted feature extraction, where identification is based on conventional image characteristics such as texture, morphology, and color. Techniques including Support Vector Machines (SVM)[4], Random Forests[5], threshold segmentation[6], edge detection[7], and region growing operate under this paradigm. These methods depend heavily on manually designed features, often requiring expert involvement, and are sensitive to data quality, necessitating extensive preprocessing. Moreover, the feature extraction and model training stages are decoupled, which may lead to a misalignment between feature design and model optimization objectives. When applied to limited datasets, such models are prone to overfitting, thereby limiting their clinical applicability. With the advancement of deep learning technologies and their successful application in medical image analysis, an increasing number of studies have shifted toward deep learning-based approaches to overcome the limitations of traditional machine learning and further improve the performance of laryngeal lesion image segmentation. This paper presents an intelligent diagnostic system for laryngeal lesions, as illustrated in Fig. 1. The system first classifies laryngeal lesions into categories such as normal, benign diseases (vocal cord leukoplakia, vocal corditis, vocal polyps), and malignant lesions (laryngeal cancer). It then performs lesion segmentation on laryngeal cancer images. We propose a lightweight classification model, LLLDNet, along with a laryngeal cancer segmentation model, HAM-Unet, which incorporates a hybrid attention mechanism and a multi-scale feature fusion module.

The main contributions of this study are summarized as follows:

1) We propose two lightweight feature extraction modules, namely the GC and the ScSE. The GC module effectively reduces model dimensionality through group convolution, while the ScSE module employs a parallel attention mechanism to enhance the model's ability to capture complex features of laryngeal diseases.

2) We designed a lightweight model, LLLDNet, and adopted a joint enhancement strategy integrating transfer learning and knowledge distillation to improve the accuracy of laryngeal disease image classification.

3) We propose two modules, HA and MSFF, to enhance feature representation capabilities. The HA module improves feature representation by integrating multiple attention mechanisms, while the MSFF module employs dilated convolutions with varying rates to aggregate feature maps at different scales.

4) Based on the UNet++ architecture, we designed a multi-scale feature extraction model, termed HAM-Unet, for precise segmentation of laryngeal cancer.

2 Related work

2.1 Medical image classification method

Classification methods based on convolutional neural networks (CNNs) have been extensively studied in the field of medical imaging. To address the issue of information loss in multi-scale medical image classification tasks using traditional deep learning models, Liu et al. [8] proposed an adaptive deep learning module tailored to the characteristics of multi-scale medical image datasets. An et al. [9] introduced a visual attention mechanism into a deep learning model, constructing a novel multi-scale convolutional neural network that achieved higher accuracy than traditional machine learning algorithms on breast cancer image datasets. Alrowais et al. [10] developed a deep learning framework named LCDC-AOADL, which is enhanced by the Aquila optimization algorithm. Within this framework, the LCDC3 architecture is employed for feature extraction, and a Deep Belief Network (DBN) is utilized to perform classification tasks related to laryngeal cancer. Liu et al. [11] proposed a novel method utilizing semi-supervised adversarial learning and pseudo-labeling, which achieved an accuracy of 96.87% on a public dataset, outperforming most deep learning models. Mukhtorov et al. [12] introduced an interpretable deep learning approach based on ResNet152, integrated with Grad-CAM for endoscopic image classification, attaining high accuracy on the open-source KVASIR dataset. Liu et al. [13] developed a method based on the ConvNeXt pure convolutional neural network, incorporating an attention mechanism and adjusting the loss function, which demonstrated superior performance on COVID-19 data. Nuzaiha Mohamed et al. [14] proposed an

approach named ALCAD-DMODL, which employs a deep learning-based Dwarf Mongoose Optimization Algorithm for laryngeal cancer recognition. Wei et al. [15] proposed a deep learning-based automatic identifier by integrating output features from the deep neural network ResNet and Transformer to accurately recognize eight types of laryngeal lesions. Kwon et al. [16] constructed a convolutional neural network model based on vocal data and laryngeal images for the diagnosis of glottic cancer. Yin et al. [17] introduced an approach that utilizes an attention mechanism to classify laryngeal diseases using image tags. Given an input image, their method employs a pre-trained object localization model to identify key regions. While these improved convolutional neural networks have demonstrated strong performance in medical image classification tasks and achieved considerable accuracy, they often suffer from slow computational speed, large model size, and high hardware requirements. To address these limitations, this paper proposes a lightweight deep learning model specifically designed for laryngeal lesion diagnosis, named LLLDNet. The network aims to enhance computational efficiency and reduce model complexity, making it particularly suitable for clinical settings with limited medical resources and constrained computing equipment.

2.2 Medical image segmentation method

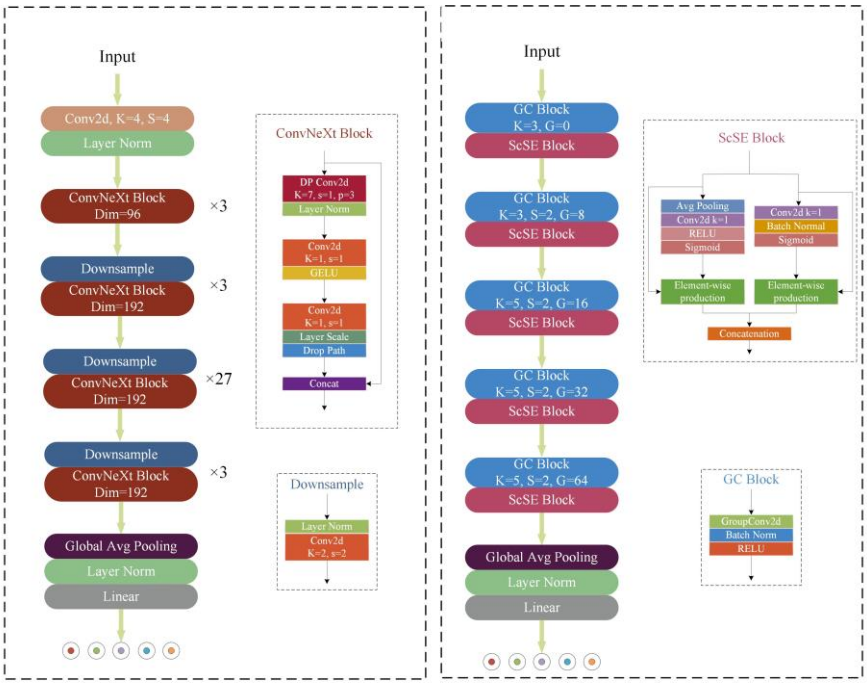
Convolutional neural networks have achieved notable research progress in medical image segmentation tasks. Shen et al. [18] proposed a dual-encoder image segmentation network incorporating HarDNet68 and a Transformer branch. This architecture enables the extraction of both local and global features from input images, allowing the segmentation network to learn richer image representations and thereby improving the effectiveness and accuracy of medical segmentation. Qian et al. [19] proposed an innovative dual-encoder image segmentation network architecture that integrates HarDNet68 and Transformer branches to comprehensively capture both local and global features of input images, significantly enhancing the effectiveness and accuracy of medical image segmentation. Yuan et al. [20] introduced a U-shaped network for colon cancer segmentation by combining depthwise separable convolutions with morphological operations, reducing model parameters while effectively improving segmentation precision. Zeng et al. [21] developed a quantitative image segmentation method for GICS based on deep reinforcement learning (DRL), which accurately distinguishes between test lines and control lines in GICS images. Yang et al. [22] proposed a U-shaped

image segmentation model named GPDENet, which employs a progressive enhancement strategy to refine decoded features and incorporates lightweight full-scale skip connections, outperforming most existing methods. Pham et al. [23] introduced a novel network called SegTransVAE, which utilizes a variational autoencoder branch coupled with a transformer to perform joint segmentation and reconstruction of input images. Ding et al. [24] developed a color normalization-based deep attention U-Net (CN-DA-UNet) to achieve end-to-end segmentation. Montalbo et al. [25] designed a lightweight U-Net incorporating separable depth-wise convolutions to reduce computational costs. While previous improvements in convolutional neural networks have enhanced the accuracy of medical image segmentation to some extent, the segmentation of specific medical images such as laryngeal cancer requires extremely high precision in delineating lesion areas. To address this need, this paper proposes a laryngeal cancer segmentation method based on UNet++, which integrates hybrid attention and multi-scale feature fusion to achieve precise delineation of laryngeal cancer lesions.

3 Method

3.1 Classification model

We adopt a two-stage knowledge transfer strategy. In the first stage, transfer learning is employed using a teacher model based on the ConvNeXt architecture[26], which has been pre-trained on the large-scale natural image dataset ImageNet. The pre-trained weights of ConvNeXt are initially loaded for the laryngeal lesion diagnosis task, and the number of neurons in the final fully connected layer is adjusted to match the number of laryngeal lesion categories. Subsequently, the model is fine-tuned on a self-constructed laryngeal lesion dataset. Due to the difficulty in acquiring medical imaging data and the limited size of training samples, using models pre-trained on ImageNet as a starting point offers significant advantages. On one hand, these pre-trained models have already learned rich general visual features, substantially reducing the need for the model to learn basic features from scratch. On the other hand, this transfer learning strategy enables the model to focus more rapidly on learning medical imaging features specific to laryngeal lesions, while significantly reducing training time and computational resource consumption.



(a) Teacher Net (b) Student Net

Fig. 2 Lightweight convolutional neural network architecture

Building upon transfer learning, this study further introduces knowledge distillation to optimize the lightweight student network LLLDNet. The core idea of knowledge distillation is to guide the learning process of a compact student model (LLLDNet) using a fully trained large teacher model (ConvNeXt). Specifically, the student model learns not only from the true class labels (hard labels) but also from the class probability distributions (soft labels) generated by the teacher model. These soft labels, smoothed by a temperature parameter, incorporate the teacher's confidence information in its predictions, thereby revealing latent relationships among different laryngeal lesion categories. For instance, a sample classified as "vocal cord polyp" by the teacher model may also exhibit a certain probability of possessing features resembling "vocal cord leukoplakia." This probability distribution reflects the inherent similarities between different types of lesions. By simultaneously optimizing the loss functions for both hard and soft labels, the student model can inherit the rich knowledge acquired by the teacher model, achieving diagnostic performance comparable to that of large-scale models while maintaining its lightweight architecture.

3.1.1 Teacher network

ConvNeXt is a pure convolutional neural network architecture that replaces self-attention layers with depthwise convolutions, while pointwise convolutions substitute the linear layers in Transformers. As illustrated in Fig. 2(a), the teacher network is designed based on the advanced ConvNeXt architecture and adopts a hierarchical feature extraction structure. The network first performs initial feature extraction through a 4×4 convolutional layer with a stride of 4, followed by layer normalization to stabilize the data distribution. The core structure consists of four stages, each comprising a downsampling module and multiple ConvNeXt blocks. The downsampling module reduces the feature map size by half through convolutional operations while increasing the channel count from 96 to 192. The network ultimately compresses spatial information via global average pooling, followed by layer normalization, before the fully connected layer produces the final prediction. This design integrates the efficient local feature extraction capability of modern convolutional architectures with Transformer-style normalization and activation strategies, resulting in powerful feature representation capacity. The ConvNeXt block employs depthwise separable convolutions to capture spatial features with large receptive fields. This is followed by a sequence of operations including layer normalization, 1×1 convolution for channel expansion, GELU activation, and 1×1 convolution for channel reduction to enable cross-channel interaction. Additionally, layer scaling and stochastic depth (drop path) techniques are incorporated to enhance training stability and generalization capability.

3.1.2 Student network

Fig. 2(b) illustrates the student network architecture of the lightweight convolutional neural network LLLDNet, which employs a hierarchical structure composed of alternating GC blocks and ScSE modules. The network first extracts initial features through a standard convolutional block, followed by multiple downsampling stages. Each stage includes a GC block with a stride of 2 for feature extraction and spatial dimensionality reduction, where the number of groups increases progressively as 6, 16, 32, and 64 to enhance model capacity. Each GC block is followed by an ScSE attention module, which adaptively calibrates feature responses through parallel spatial and channel attention mechanisms, thereby enhancing critical features and suppressing noise. The network

ultimately compresses spatial information through global average pooling, stabilizes the feature distribution via layer normaliza-
tion, and finally outputs the classification results through a fully connected layer. This design achieves efficient feature extraction
and selective enhancement, improving the model's expressive power while maintaining computational efficiency. The GC module
consists of a group convolution, Batch Normalization, and a ReLU activation function. The ScSE module is a parallel attention
mechanism that enhances feature representation simultaneously from both spatial and channel dimensions. Upon receiving the
input feature map, the module processes it through two parallel branches: the spatial attention branch first obtains spatial statistics
via global average pooling, then generates a spatial weight map through a 1×1 convolution and an activation function, and finally
performs element-wise multiplication with the input features to emphasize important spatial locations. The channel attention branch
similarly acquires channel statistics through global average pooling, generates a channel weight vector via a 1×1 convolution,
batch normalization, and an activation function, and then performs channel-wise multiplication with the input features to reinforce
important feature channels. The outputs of the two branches are finally fused through concatenation, producing enhanced features
that emphasize both critical spatial regions and significant channels. This effectively improves the model's ability to perceive salient
features.

3.1.3 Distillation loss

To enhance the effectiveness of knowledge distillation, this paper introduces a temperature parameter T to soften the probability
distributions. Based on the cross-entropy loss (ce_loss) between the student model's predictions and the true labels, the distillation
loss (kd_loss) is calculated using KL divergence between the outputs of the student and teacher models. The final training loss is
obtained as a weighted combination of these two losses. This combined approach preserves the deterministic information provided
by the true labels while fully leveraging the inter-class correlation knowledge from the teacher model's outputs, thereby achieving
more effective knowledge transfer. Experiments demonstrate that by appropriately setting the temperature parameter T and the
weighting coefficient α, the student model can better learn the knowledge representation capability of the teacher model, attaining
superior diagnostic performance while maintaining its lightweight design.

The total loss function of the entire classification network is a weighted combination of two parts: the hard loss from the student model and the distillation loss.

$$L_{\text{total}} = \alpha \cdot L_{ce} + (1-\alpha) \cdot L_{kd}$$

The standard cross entropy loss between the student model output and the real hard label:

$$L_{ce} = \text{CrossEntropy}(S, y) = -\sum_{i=1}^{C} y_i \log(s_i)$$

The softened output of the teacher model:

$$p_i^T = \text{softmax}\left(z_i^T / T\right) = \frac{\exp\left(z_i^T / T\right)}{\sum_{j=1}^{C} \exp\left(z_j^T / T\right)}$$

Softened output of the student model:

$$q_i^T = \text{softmax}\left(z_i^S / T\right) = \frac{\exp\left(z_i^S / T\right)}{\sum_{j=1}^{C} \exp\left(z_j^S / T\right)}$$

KL divergence is used to measure the difference between the softening distribution q^{T} of the student model and the softening distribution p^{T} of the teacher model.

$$L_{kd} = T^2 \cdot D_{KL}\left(p^T \square q^T\right) = T^2 \cdot \sum_{i=1}^{C} p_i^T \log\left(\frac{p_i^T}{q_i^T}\right)$$

The coefficient α balances the contribution of the two loss terms. C denotes the total number of classes. The true label is represented as a one-hot vector y, where y_i is the label for the i-th class. The student model's raw logits are denoted by S, and the corresponding predicted probability for the i-th class is s_i, s_i=softmax(S)$_i$. The teacher's and student's raw outputs are z^T and z^S, respectively. The temperature parameter T softens the output distributions: T=1 yields the standard softmax, while T>1 produces a softer distribution with smaller inter-class probability differences. The distillation loss is measured by the KL divergence D_{KL} The loss is scaled by T^2 to counteract the $1/T^2$ reduction in gradient magnitude caused by the softening operation, thus making the effective weight of the distillation loss independent of the chosen T.

3.2 HAM-Unet

Based on UNet++ [27], we propose a new model named HAM-Unet for laryngeal cancer segmentation, as illustrated in Fig. 3. The model employs transfer learning by replacing the original VGG encoder with a ResNet50 encoder and utilizing pre-trained weights to accelerate convergence, thereby enhancing segmentation performance on small-scale laryngeal cancer datasets. Furthermore, we design an HA module in the encoder for attention-based feature enhancement and introduce an MSFF module between shallow and deep encoder layers to achieve multi-scale feature fusion. This enables the model to capture finer details and richer contextual information across scales, ultimately improving the accuracy of laryngeal cancer image segmentation. To further improve the model's performance in laryngeal cancer segmentation, a Large Kernel Attention (LKA) mechanism is introduced in the decoder. By leveraging large-sized convolutional kernels to extract features, it effectively integrates local detail information and global semantic context, thereby enhancing the segmentation precision along the edges of laryngeal cancer images.

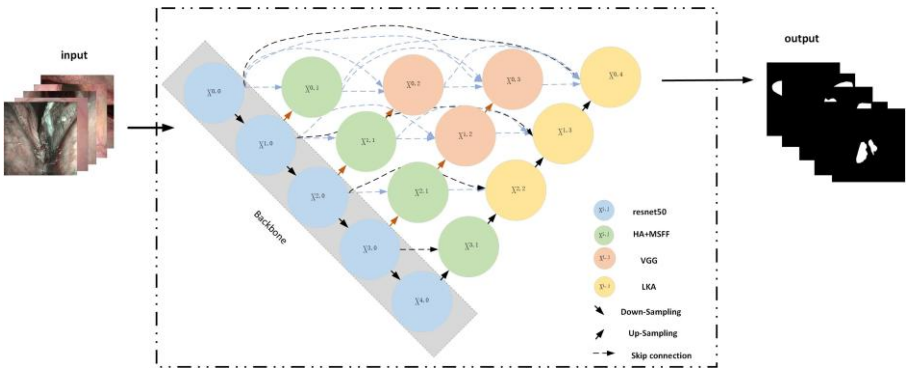


Fig. 3 HAM-Unet architecture

3.2.1 HA module

In laryngeal cancer images, lesion regions exhibit complex and diverse morphological characteristics with significant variations, and there is a lack of explicit modeling capability for global contextual information and multi-scale features. To further enhance the feature extraction ability of the encoder, this paper designs a HA module, whose structure is illustrated in Fig. 4. This module can more comprehensively extract and utilize feature information, thereby improving the feature representation capability of the model.

First, the input feature X is processed by a multi-head attention layer to capture global contextual information, generating a primary feature representation X_1. Through a permutation operation, the order of feature dimensions is adjusted, and a spatial attention mechanism is employed to compute a two-dimensional spatial weight matrix. After compression, this matrix is multiplied element-wise with the original features to achieve feature enhancement in spatial regions. Then, via a reverse permutation operation, the priority of the channel dimension is restored. A channel attention mechanism is adopted to generate a one-dimensional channel weight vector, which is applied to the feature map through a broadcasting mechanism for channel-wise weighting. Finally, the outputs of the dual branches are linearly fused to produce the optimized feature representation. The HA module achieves adaptive recalibration of feature maps in both spatial and channel domains through cascaded multi-stage dimension transformations and a learnable parametric weight generation mechanism. This approach effectively enhances the model's discriminative capability for key features while maintaining computational efficiency.

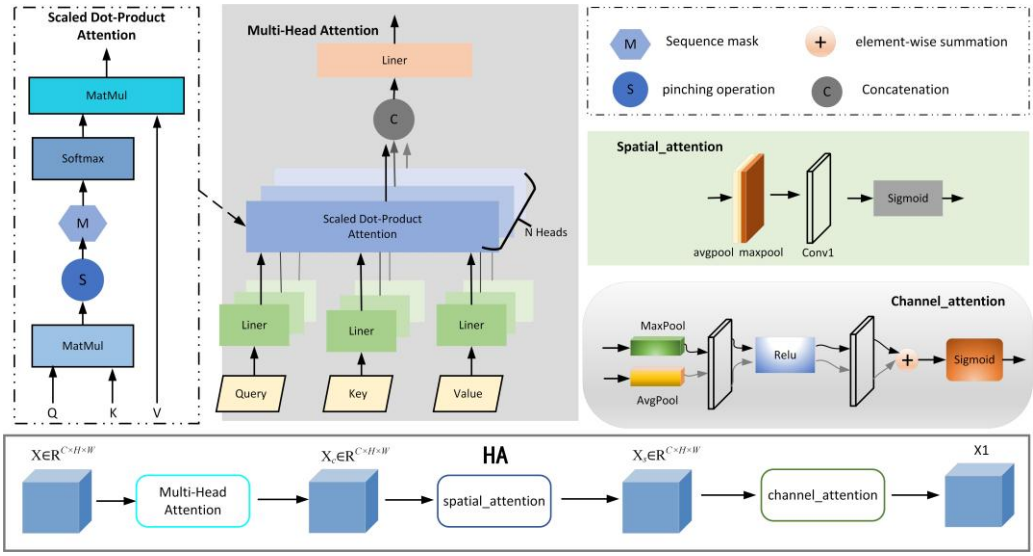


Fig. 4 HA architecture

3.2.2 MSFF module

In semantic segmentation tasks, it is often necessary to accurately segment targets of varying sizes. Traditional convolutional neural networks typically rely on single-scale feature representations and extract features progressively by stacking multiple convolutional layers. However, due to the limited receptive field of each convolutional layer, these networks struggle to capture both

local details and global contextual information simultaneously. To address this issue, this paper designs an MSFF module, whose architecture is illustrated in Fig. 5.

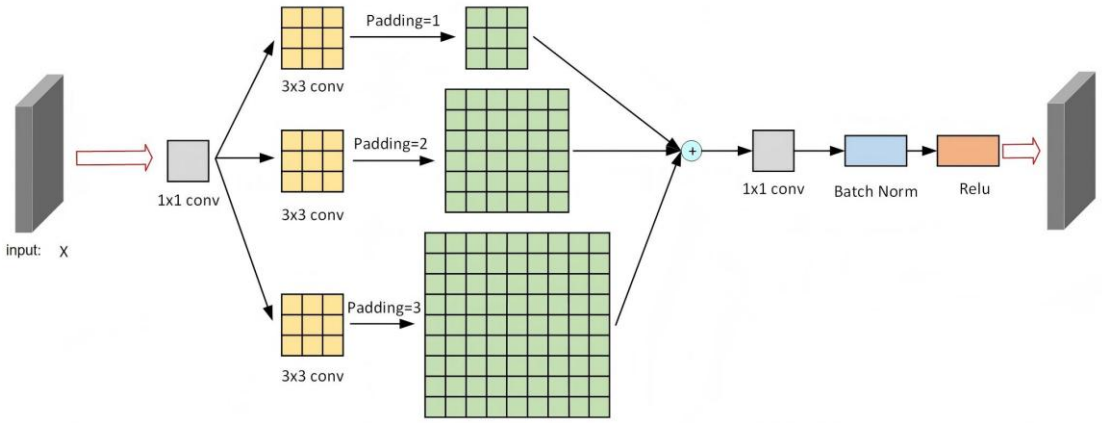


Fig. 5 MSFF structure

The MSFF module enhances the model's ability to comprehend and process multi-scale information in complex scenes by integrating feature maps at different scales. This capability is crucial for improving the segmentation accuracy of laryngeal cancer images, particularly when dealing with targets with irregular boundaries or complex shapes. Low-level features retain fine-grained spatial details such as textures and edges through local receptive fields, while high-level features, after multiple abstractions, develop strong representational capacity for high-level semantics including object categories and structural relationships. Multi-scale feature fusion thereby preserves both the precise local structures and the overall morphology of targets. The MSFF module first employs a 1×1 convolution to reduce the channel dimensionality of the input feature map. Subsequently, three parallel 3×3 convolutions with dilation rates of 1, 2, and 3 are applied to the reduced-dimension features, generating three sets of feature maps at different scales. These dilated convolutions enable the model to capture multi-scale spatial contextual information at the same hierarchical level. The outputs from the three convolutional branches are then concatenated and fused via a 1×1 convolution, which integrates the multi-scale features and enhances information utilization across different levels, thereby producing a more comprehensive feature representation. Finally, batch normalization and a ReLU activation function are applied to accelerate convergence, facilitate the learning of complex patterns, and mitigate the gradient vanishing problem.

Assume the input feature map is $X \in R^{C_{\text{in}} \times H \times W}$, where C_{in} is the number of input channels, H and W are the height and width respectively. The output feature map is $Y \in R^{C_{\text{fuse}} \times H \times W}$. Use a 1x1 convolution to reduce the number of channels of the input feature map from C_{in} to C_{red}. Set the convolution weight as $W_{\text{red}} \in R^{C_{\text{red}} \times C_{\text{in}} \times 1 \times 1}$ and the bias as $b_{\text{red}} \in R^{C_{\text{red}}}$. The convolution operation is expressed as:

$$X' = W_{\text{red}} * X + b_{\text{red}}$$

Apply three parallel 3x3 dilated convolutions with dilation rates of 1, 2, and 3 respectively to *X'*. The output channel number for each convolution is set to C_{out}. For each dilation rate d, the convolution weight is $W_{\text{d}} \in R^{C_{\text{out}} \times C_{\text{red}} \times 3 \times 3}$, and the bias is $b_{\text{d}} \in R^{C_{\text{out}}}$. The convolution operation maintains the spatial dimensions. The output feature maps are:

$$Y_d = W_d *_d X' + b_d, d \in \{1,2,3\}$$

The output feature maps of the three branches are spliced along the channel dimension:

$$Y_{\text{cat}} = [Y_1, Y_2, Y_3]$$

Through a 1×1 convolution to fuse the concatenated features, with the number of output channels set to C_{fuse}. Let the convolution weight be $W_{\text{fuse}} \in R^{C_{\text{fuse}} \times 3C_{\text{out}} \times 1 \times 1}$ and the bias be $b_{\text{fuse}} \in R^{C_{fuse}}$. The convolution operation is expressed as:

$$Z = W_{\text{fuse}} * Y_{\text{cat}} + b_{\text{fuse}}$$

Finally, apply the Batch Normalization and ReLU activation function. The batch statistics (mean μ and standard deviation σ) are calculated. The learnable parameters of BN are $\gamma \in R^{C_{fuse}}$ and $\beta \in R^{C_{fuse}}$. The operations are as follows:

$$\hat{Z} = \text{BN}(Z) = \gamma \odot \frac{Z - \mu}{\sigma} + \beta$$

$$Y = \text{ReLU}(Z) = \max(0, Z)$$

where $\odot$ denotes element-wise multiplication, and the output Y is the final feature map.

4 Experiment

4.1 dataset

This study collected images of laryngeal lesions using the Olympus electronic laryngoscope system OTV-S190 at the First Affiliated Hospital of Chongqing Medical University. The dataset consists of 280 laryngeal cancer images, 500 polyp of the vocal cord images, 500 chorditis images, 420 vocal cord white spot images, and 600 normal images. The dataset was divided into training, testing, and validation sets in an 8: 1: 1 ratio. Annotations of lesions in the 280 laryngeal cancer images were performed using ImageJ (version 1.54f). During annotation, the lesion areas in each image were delineated using curved edges and assigned categorical labels. The annotated images were saved in TIFF format. Subsequently, the color images were converted to grayscale in ImageJ, where the annotated lesion regions were set to white and all other areas to black, resulting in binary images that retained only the target regions. These binary images were saved in JPG format.

Laryngoscopic images often suffer from overexposure in central areas, underexposure in peripheral regions, and low contrast in vascular structures due to uneven illumination angles and mucosal reflections. To address these issues, we employ a CLAHE-based preprocessing pipeline. First, the image is converted from BGR to HSV color space, and CLAHE enhancement is applied exclusively to the V channel. Multi-scale detail integration is then achieved through Gaussian pyramid fusion. Subsequently, the image is converted to Lab color space, where the L channel undergoes a hybrid enhancement combining CLAHE and global histogram equalization, while the a and b channels are preserved to retain color information. Finally, the enhanced results from the HSV and Lab processing paths are fused with a 60%/40% weighting ratio, and a 3×3 Laplacian kernel is applied to further sharpen fine details.

4.2 Performance metrics

In the field of medical image classification and segmentation, the issue of class imbalance within datasets makes it difficult for a single evaluation metric to comprehensively reflect model performance. Therefore, this paper employs six metrics—Specificity, Precision, Recall, F1-score, Intersection over Union (IOU), and Dice coefficient—to conduct a comprehensive evaluation of the

proposed model's effectiveness. All metrics are computed based on a confusion matrix, which is defined using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

The formula for calculating specificity is presented as follows:

$$Specificity = \frac{TN}{TN + FP}$$

The formula for precision is given as follows:

$$precision = \frac{TP}{TP + FP}$$

The formula for recall is expressed as follows:

$$recall = \frac{TP}{TP + FN}$$

The formula for the F1 Score is depicted as follows:

$$F_1 Score = \frac{2 \times precision \times recall}{precision + recall}$$

The formula for Intersection over Union is presented as follows:

$$IOU = \frac{TP}{TP + FP + FN}$$

The formula for Dice Coefficient is presented as follows:

$$Dice = \frac{2TP}{2TP + FP + FN}$$

4.3 Implementation details

The experiments were conducted on a Windows 11-based deep learning platform. The hardware configuration consisted of an Intel® Xeon® Platinum 8481C processor and an NVIDIA RTX 4090D GPU with 24 GB of memory. The software environment included Python 3.10 and the PyTorch 2.1.2 deep learning framework, with the CUDA 11.8 parallel computing architecture enabled to accelerate model training. All experiments were performed on this unified platform to ensure the comparability and reproducibility of the results.

All images were uniformly resized to 256×256 pixels. During model training, we employed the SGD optimizer with an initial learning rate of 0.001, a momentum factor of 0.9, and a weight decay of 0.0001. The batch size and number of epochs were set to 20 and 200, respectively.

5 Results

5. 1 Classification experimental results

To validate the accuracy of the proposed classification network, this paper conducts comparative experiments pitting LLLDNet against other lightweight classification networks, including MobileNetV3[28], MobileVit[29], ShuffleNetV2[30], DenseNet121[31], EfficientNet_b0[32], RegNet[33], and ResNet34[34]. It can be seen from Table 1 and Fig. 6 that Precision, Recall, Specificity and F1 Score reached 93.96 %, 92.40 %, 98.20 % and 92.98 %, respectively. Notably, LLLDNet not only improves these performance metrics but also significantly reduces the model size to merely 0.425 million parameters, with a computational complexity of 294 million FLOPS. This design achieves substantial model lightweighting, effectively conserves computational resources, and accelerates inference speed. The confusion matrices of different classification models are shown in Fig. 7, indicating that our proposed LLLDNet model achieves the best performance.

Table 1 Comparison results of different models

Experiment	Precision(%)	Recall(%)	Specificity(%)	F1 Score(%)	Params(M)	Flops(M)
MobileNetV3	87.96	87.04	96.64	87.34	5.40	220
MobileVit	88.74	88.70	96.88	88.62	5.60	420
ShuffleNetV2	89.60	88.82	97.08	89.14	2.30	150
DenseNet121	90.62	90.02	97.58	90.28	20.00	2900
EfficientNet_b0	90.70	89.36	96.26	89.78	5.30	390
RegNet	85.02	85.00	95.86	84.76	20.60	400

ResNet34	91.22	89.80	97.40	90.28	21.80	3700
LLLDNet	93.96	92.40	98.20	92.98	0.425	294

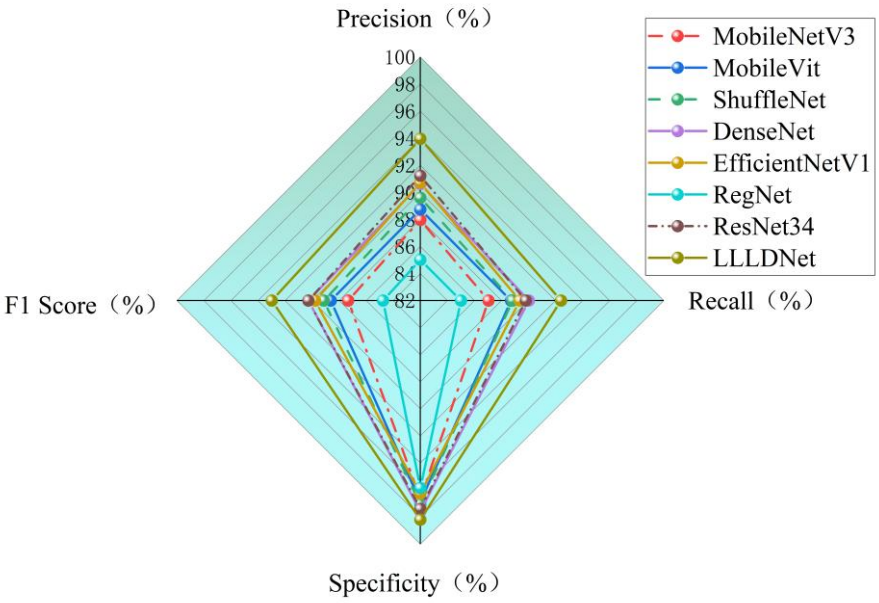


Fig. 6 Index radar chart of classification network

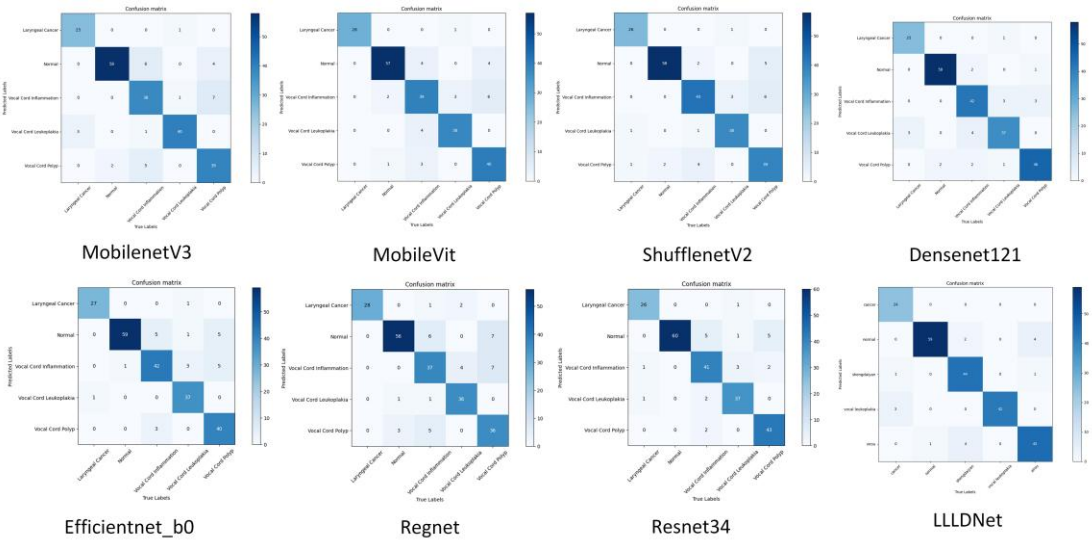


Fig. 7 Confusion matrix of different deep learning models

5.2 Segmentation experimental results

To validate the effectiveness of the HAM-Unet module, a series of detailed ablation studies were conducted to meticulously analyze the contribution of each individual component to the overall performance. These components include the pre-trained ResNet50 weights, the MSFF module, the HA module, and the LKA module. The specific network configurations for different ablation experiments are detailed in Table 2. Given the limited number of laryngeal cancer images, data augmentation techniques including rotation and cropping were applied to prevent model underfitting, expanding the dataset from 280 to 1,045 samples. In the experiments, UNet++ was first adopted as the baseline model and trained on the constructed laryngeal cancer dataset. The performance of each proposed module was evaluated using IOU, Precision, Recall, and Dice. Subsequently, different components were incrementally integrated into the baseline model. The corresponding experimental results are summarized in Table 3.

Table 2 Components of ablation experiment

Experiment	**Specific module combination**
Module1	Baseline
Module2	Baseline+resnet50
Module3	Baseline+resnet50+MSFF
Module4	Baseline+resnet50+MSFF+HA
HAM-Unet	Baseline+resnet50+MSFF+HA+LKA

Each key module in the HAM-Unet contributes to improving model performance to varying degrees, with the proposed MSFF module demonstrating particularly significant gains. When replacing the VGG backbone in the baseline UNet++ with pretrained ResNet50 weights using transfer learning, the IOU increased by only 0.02%. However, integrating the MSFF module resulted in a substantial improvement of 3.37% in IOU, attributable to its ability to capture multi-scale details and contextual information.

Furthermore, the incorporation of the HA module led to an increase of 3.46% in IOU and 2.65% in Precision over the baseline, as it enables more comprehensive feature extraction and utilization, thereby enhancing the model's representational capacity. The addition of the Large Kernel Attention (LKA) module yielded the highest scores across all metrics, achieving 90.86% IOU, 94.32% Precision, 96.52% Recall, and 94.30% Dice.

Table 3 Ablation experimental results of different network architectures

Experiment	IOU(%)	Precision(%)	Recall(%)	Dice(%)
Baseline	85.24	91.62	92.39	91.62
Module1	85.26	92.18	91.98	90.86
Module2	88.59	93.92	93.71	92.53
Module3	89.47	94.27	93.70	92.18
Module3	90.86	94.32	96.52	94.30

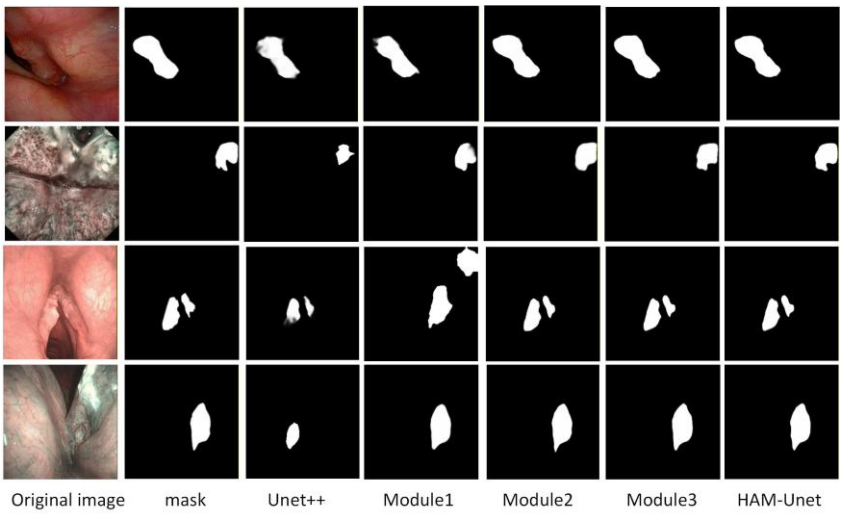


Fig. 8 Ablation module for laryngeal cancer lesion segmentation results comparison

A comparison of segmentation results from the ablation studies is illustrated in Fig. 8. For the same laryngeal cancer image, as the pre-trained ResNet50 weights, MSFF module, HA module, and Large Kernel Attention (LKA) mechanism are incrementally

incorporated, the segmented lesion regions progressively converge toward the ground truth. When using transfer learning with the baseline UNet++ model, a significant discrepancy exists between the segmented lesion areas and the ground truth. To address this, the MSFF module is introduced to integrate multi-scale feature maps, enabling the model to better comprehend and process multi-scale information in complex scenarios. Consequently, the overlap between the segmented lesion regions and the ground truth is progressively improved. Given the complex characteristics of laryngeal cancer images, which pose challenges for accurate boundary segmentation, the proposed HA module and the incorporated Large Kernel Attention (LKA) mechanism effectively address this issue. These components enable the HAM-Unet network to comprehensively extract and utilize feature information, with enhanced focus on lesion regions, thereby improving the model's feature representation capability. As observed in the segmentation results of the HAM-Unet model in Fig. 8, the predicted lesion areas exhibit near-complete overlap with the ground truth annotations. This demonstrates that the proposed key modules significantly enhance the accuracy of laryngeal cancer segmentation and play a crucial role in the task.

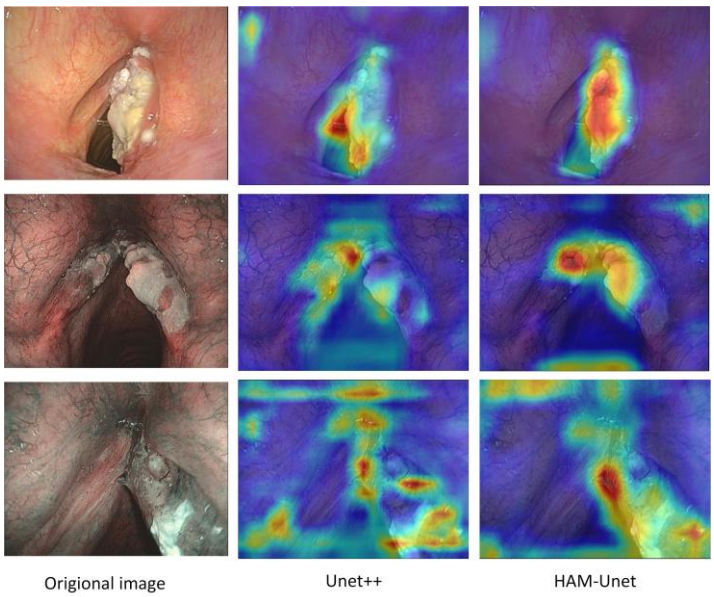
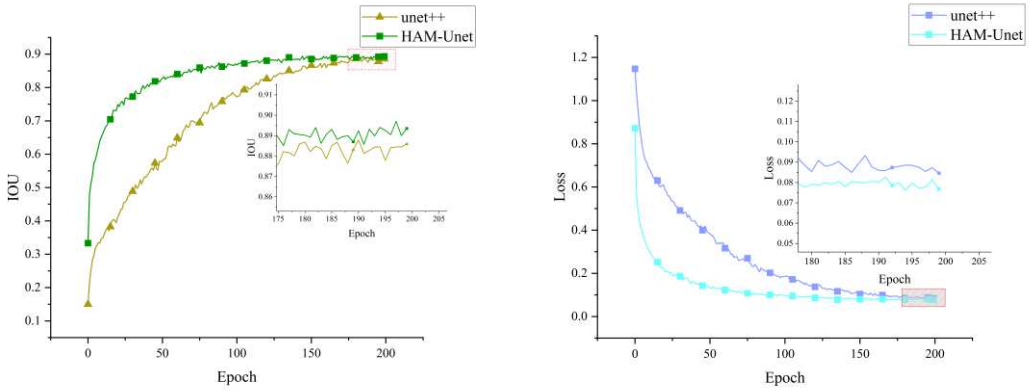


Fig. 9 Comparison of heat maps before and after improvement

To better understand the effectiveness of the proposed module, a visual analysis was conducted using Grad-CAM[35] to compare the model's behavior before and after the improvement. The visualization results are presented in Fig. 9. It is evident that the enhanced model exhibits a significantly stronger focus on the laryngeal cancer lesion regions. This indicates that the improved

model can more effectively suppress the influence of distracting features, thereby concentrating more precisely on identifying and analyzing the specific locations of lesions. Such an enhancement contributes to the model's capability for accurate lesion localization within complex medical imaging scenarios.

Fig. 10 presents a comparative analysis of the training curves for IOU, and Loss between the baseline and improved models. It can be observed that the enhanced HAM-Unet model achieves faster convergence speed and higher final values across all evaluation metrics.



(a) (b)

Fig. 10 Comparison of indicators before and after improvement; (a) IOU, (b) Loss

We compare the proposed HAM-Unet model against a range of mainstream segmentation models, including: Attention UNet[36], CE-Net[37], DeepLabv3[38], FCN[39], IS[40], UNet[41], and the baseline model UNet++, the comparison results of different segmentation models are shown in Table 4. As evidently demonstrated, the proposed HAM-Unet model exhibits significant advantages. Compared to the original UNet++ baseline, HAM-Unet achieves improvements of 5.62% in IOU, 2.3% in Precision, 4.13% in Recall, and 2.68% in Dice, indicating its overall superior capability over UNet++. When compared to other commonly used segmentation models—namely Attention UNet, CE-Net, DeepLabv3, FCN, IS, and UNet. HAM-Unet also demonstrates outstanding segmentation performance, with IOU increases of 22.41%, 24.09%, 25.14%, 20.76%, 19.53%, and 8.70%, respectively. These results confirm that HAM-Unet possesses stronger capability for laryngeal cancer segmentation.

Table 4 Comparison of experimental results of different models

Model	IOU(%)	Precision(%)	Recall(%)	Dice(%)
AttentionUNet	68.45	77.84	78.92	66.80
Cenet	66.77	77.96	75.53	64.70
Deeplabv3	65.72	83.06	70.95	64.70
Fcn	70.10	79.40	80.30	66.90
IS	71.33	78.75	83.83	68.70
UNet	82.15	90.37	90.05	89.35
UNet++	85.24	91.62	92.39	91.62
HAM-Unet	90.86	93.92	96.52	94.30

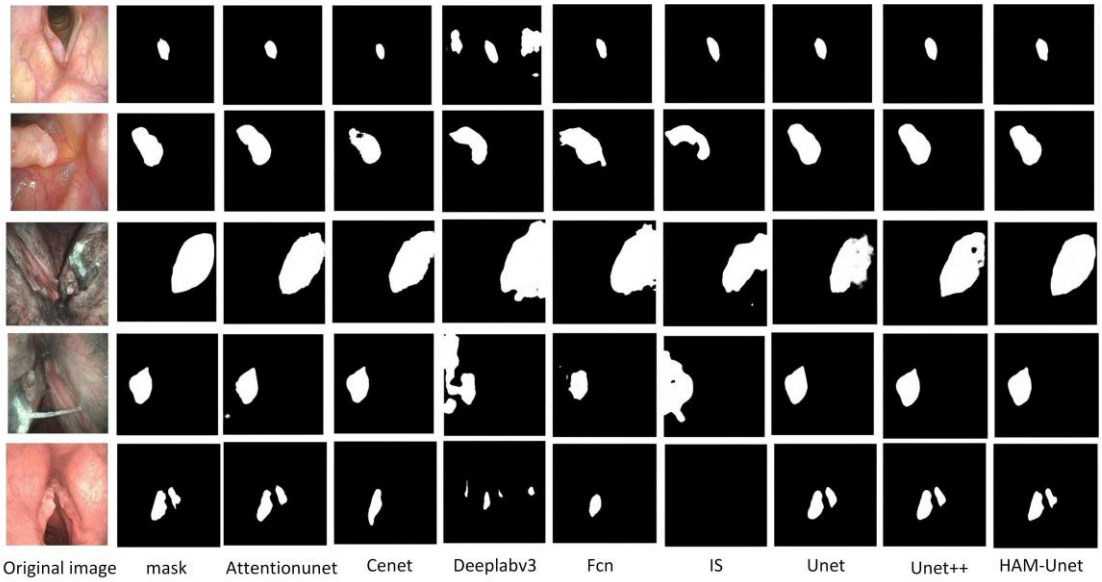


Fig. 11 Comparison of different segmentation model results

Fig. 11 provides visual examples of segmentation results produced by different networks. It can be observed that the proposed HAM-Unet model delivers superior performance in segmenting laryngeal cancer lesion regions while maintaining a lower false

positive rate. In comparison to other segmentation networks, HAM-Unet delineates lesion areas more accurately and completely, demonstrating its higher accuracy and reliability in the task of laryngeal cancer image segmentation.

6 Conclusion

As a common head and neck tumor, the clinical diagnosis of laryngeal cancer primarily relies on visual examination by physicians, which often leads to misdiagnosis and missed detection. With the advancement of artificial intelligence, deep learning has gradually gained traction in the medical field, yet certain limitations remain. To address issues such as low classification accuracy, low computational efficiency, and incomplete segmentation of laryngeal cancer images, this paper constructs an intelligent diagnostic system for laryngeal lesions aimed at facilitating early diagnosis of laryngeal cancer.

We design an LLLDNet for laryngeal lesion classification, which employs a joint enhancement strategy combining transfer learning and knowledge distillation to improve the accuracy of laryngeal disease classification. Furthermore, by incorporating the proposed HA and MSFF modules into UNet++, we propose a HAM-Unet network for laryngeal cancer segmentation. Experimental results on the laryngeal disease dataset demonstrate that LLLDNet achieves an accuracy of 93.0%, recall of 92.0%, specificity of 98.20%, and an F1-score of 92.98%. Notably, LLLDNet not only enhances these performance metrics but also significantly reduces the model size to only 0.425 million parameters, which is substantially lower than other classification models. With a computational complexity of 294 million FLOPS, it also remains more lightweight than many existing models, thereby conserving computational resources and increasing inference speed. Meanwhile, the HAM-Unet network achieves an IOU of 90.86%, precision of 93.92%, recall of 96.52%, and a Dice score of 94.30%, significantly outperforming other segmentation models and enabling more precise delineation of laryngeal cancer lesion regions.

In summary, the intelligent diagnostic method for laryngeal diseases proposed in this study provides a theoretical foundation for clinical medical applications. However, further in-depth validation is required before it can be formally deployed in clinical practice. A primary limitation is that the training and validation datasets were obtained from a single medical center, and the amount

of laryngeal cancer data remains relatively limited. Future work will involve collecting larger datasets from multiple healthcare institutions to advance research on intelligent diagnosis of laryngeal diseases.

Declarations

- Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of the First Affiliated Hospital of Chongqing Medical University (Protocol No. (2024) Ethical Review (Clinical) No. 2024-095-01; Approval Date: January 18, 2024). This was a retrospective study and therefore did not require clinical trial registration. Informed consent was obtained from all participants involved in this study.

- Consent for publication

Not applicable

- Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

- Competing interests

The authors declare that they have no competing interests

- Funding

This study was supported by the National Natural Science Foundation of China (62376044).

- Acknowledgement

Not applicable

- Author contribution

XHJ: Coding implementation, Writing, experimental design. LLC: Conceptualization. XW: Coding implementation. XWW: data acquisition. NNZ: data annotation. All authors have read and agreed to the published version of the manuscript.

References

[1] Bao X, Wang S, Liu X, et al. Highly sensitive detection of CYFRA21-1 with a SERS sensing platform based on the MBs enrichment strategy and antibody-DNA-mediated CHA amplification[J]. Frontiers in Bjkjioengineering and Biotechnology, 2023, 11(1): 10. DOI:10.3389/fbioe.2023.1251595.

[2] Al-Haija Q A, Droos A. A comprehensive survey on deep learning-based intrusion detection systems in Internet of Things (IoT)[J]. Expert Systems, 2025, 42(2). DOI:10.1111/exsy.13726.

[3] Cho W K, Lee Y J, Joo H A, et al. Diagnostic accuracies of laryngeal diseases using a convolutional neural network-based image classification system[J]. The Laryngoscope, 2021, 131(11): 2558-2566.

[4] Chandra M A, Bedi S S. Survey on SVM and their application in image classification[J]. International Journal of Information Technology, 2021, 13(5): 1-11.

[5] Rigatti S J. Random forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[6] Xin Y, Martínez José-Fernán, Martina E, et al. An Improved Otsu Threshold Segmentation Method for Underwater Simultaneous Localization and Mapping-Based Navigation[J]. Sensors, 2016, 16(7): 1148-.DOI:10.3390/s16071148.

[7] Rong W, Li Z, Zhang W, et al. An improved Canny edge detection algorithm[C]//IEEE International Conference on Mechatronics & Automation. IEEE, 2014. DOI:10.1109/ICMA.2014.6885761.

[8] LiuXiangbin, HeJiesheng, SongLiping, et al. Medical Image Classification based on an Adaptive Size Deep Learning Model[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021.DOI:10.1145/3465220.

[9] An F, Li X, Ma X. Medical Image Classification Algorithm Based on Visual Attention Mechanism-MCNN[J]. Oxidative Medicine and, Cellular Longevity, 2021, 2021(2): 1-12. DOI:10.1155/2021/6280690.

[10] Alrowais F, Mahmood K, Alotaibi S S, et al. Laryngeal cancer detection and classification using aquila optimization algorithm with deep learning on throat region images[J]. Ieee Access, 2023, 11: 115306-115315.

[11] Liu K, Ning X, Liu S. Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling[J]. Sensors, 2022, 22(24): 12. DOI:10.3390/s22249967.

[12] Mukhtorov D, Rakhmonova M, Muksimova Y I. Endoscopic Image Classification Based on Explainable Deep Learning[J]. sensors, 2023, 23(6). DOI:10.3390/s23063176.

[13] Liu F, Zang C, Shi J, et al. An Improved COVID-19 Lung X-Ray Image Classification Algorithm Based on ConvNeXt Network[J]. International Journal of Image & Graphics, 2024, 24(3). DOI:10.1142/S0219467824500360.

[14] Mohamed N, Almutairi R L, Abdelrahim S. Automated laryngeal cancer detection and classification using dwarf mongoose optimization algorithm with deep learning. Cancers (Basel). 2023[EB/OL]. https://www.mdpi.com/2072-6694/16/1/181.

[15] Wei R, Liang Y, Geng L, et al. A non-local dual-stream fusion network for laryngoscope recognition[J]. American Journal of Otolaryngology, 2025, 46(1): 104565.

[16] Kwon I, Wang S-G, Shin S-C, et al. Diagnosis of Early Glottic Cancer Using Laryngeal Image and Voice Based on Ensemble Learning of Convolutional Neural Network Classifiers[J]. J Voice, 2022.

[17] Yin L, Liu Y, Pei M, et al. Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism[J]. Pattern Recognition Letters, 2021, 150: 207-213.

[18] Shen T, Xu H. Medical image segmentation based on Transformer and HarDNet structures[J]. IEEE Access, 2023, 11: 16621-16630

[19] Qian L, Huang H, **a X, et al. Automatic segmentation method using FCN with multi-scale dilated convolution for medical ultrasound image[J]. The Visual Computer, 2023, 39(11): 5953-5969

[20] Yuan H, Chen L, He X. MMUNet: Morphological feature enhancement network for colon cancer segmentation in pathological images[J]. Biomedical Signal Processing and Control, 2024, 91(000). DOI:10.1016/j.bspc.2023.105927.

[21] Zeng N, Li H, Wang Z, et al. Deep-Reinforcement-Learning-Based Images Segmentation for Quantitative Analysis of Gold Immunochromatographic Strip[J]. Neurocomputing, 2020. DOI:10.1016/j.neucom.2020.04.001.

[22] Yang Y, Li L, Wang L, et al.GPDENet: Combining gradient-path design and dynamic enhancement strategy for medical image segmentation[J].Biomedical Signal Processing and Control, 110[2025-09-12].

[23] Pham Q D, Nguyen-Truong H, Phuong N N, et al. Segtransvae: Hybrid cnn-transformer with regularization for medical image segmentation[C]//2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, 2022: 1-5.

[24] Ding H, Cen Q, Si X, et al. Automatic glottis segmentation for laryngeal endoscopic images based on U-Net[J]. Biomedical Signal Processing and Control, 2022, 71: 103116.

[25] Montalbo F J P. S3AR U-Net: A separable squeezed similarity attention-gated residual U-Net for glottis segmentation[J]. Biomedical Signal Processing and Control, 2024, 92: 106047.

[26] Liu Z, Mao H, Wu C Y, et al. A ConvNet for the 2020s[J]. arXiv e-prints, 2022. DOI:10.48550/arXiv.2201.03545.

[27] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation[J]. 2018. DOI:10.1007/978-3-030-00889-5_1.

[28] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[J]. IEEE, 2020.DOI:10.1109/ICCV.2019.00140.

[29] Mehta S, Rastegari M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer[J]. 2021. DOI:10.48550/arXiv.2110.02178.

[30] Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[J]. Springer, Cham, 2018. DOI:10.1007/978-3-030-01264-9_8.

[31] Huang G, Liu Z, Laurens V D M, et al.Densely Connected Convolutional Networks[J].IEEE Computer Society, 2016.DOI:10.1109/CVPR.2017.243.

[32] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[J]. 2019. DOI:10.48550/arXiv.1905.11946.

[33] Radosavovic I, Kosaraju R P, Girshick R, et al. Designing Network Design Spaces[C]//Computer Vision and Pattern Recognition. IEEE, 2020. DOI:10.1109/CVPR42600.2020.01044.

[34] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016. DOI:10.1109/CVPR.2016.90.

[35] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[36] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning Where to Look for the Pancreas[J]. 2018. DOI:10.48550/arXiv.1804.03999.

[37] Gu Z, Cheng J, Fu H, et al. CE-Net: Context Encoder Network for 2D Medical Image Segmentation[J]. IEEE Transactions on Medical Imaging, 2019: 1-1. DOI:10.1109/TMI.2019.2903562.

[38] Chen L C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. 2017. DOI:10.48550/arXiv.1706.05587.

[39] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651. DOI:10.1109/CVPR.2015.7298965.

[40] Salimans T, Goodfellow I, Zaremba W, et al. Improved Techniques for Training GANs[J]. 2016.DOI:10.48550/arXiv.1606.03498.

[41] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Springer, Cham, 2015. DOI:10.1007/978-3-662-54345-0_3.