DCLNet: Data Closed-Loop Network for Laryngoscopy Image Annotation and Classification

Xinxin Xu†, Xinyu Zhao†, Jian Cao*, Weichen Xu, Tianhao Fu, Ruilong Ren, Zicong Hu and Xing Zhang*

School of Software and Microelectronics, Peking University, Beijing, China
Email: caojian@ss.pku.edu.cn, zhx@pku.edu.cn
†These authors contributed equally to this work.

2024 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-5495-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/BioCAS61083.2024.10798144

***Abstract*—Laryngeal diseases diagnosis is a common challenge in medical image processing. Traditional methods usually use classification networks to process entire images or their cropped key area images. However, the former may not effectively focus on the crucial region within the image, while the latter is a two-stage task that cannot utilize global information. To address this problem, we propose a novel Data Closed-Loop Network (DCLNet) for laryngoscope image classification, fully utilizing the global features in the image and enabling end-to-end training. By introducing attention mechanisms, DCLNet can automatically expand the Laryngoscope8 classification dataset into a high-quality annotated object detection dataset. Moreover, to address some inherent issues of medical datasets, such as poor interpretability and a long-tailed distribution, we employ the methods of data closed-loop and curriculum learning. In this way, performance on the baseline dataset can be improved by optimizing the annotation of the training set without changing the model structure. Extensive experiments prove that our model achieves better performance compared to existing traditional methods.**

Keywords—medical image classification, data closed-loop, automatic annotation, curriculum learning

I. Introduction

According to statistics, the incidence rate of laryngeal diseases ranks third among all medical conditions [1], making laryngeal diseases a common health issue in our daily lives. Traditional laryngeal disease diagnosis relies on laryngoscope image scrutiny by experts, but this process is inefficient and time-consuming. Deep learning based AI methods for medical image classification [2]-[11] are in early stages, leaving room for improving diagnosis performance and efficacy.

Laryngeal disease diagnosis involves providing a laryngoscopic image to identify the specific disease type, constituting a conventional classification task. Previous approaches have predominantly relied on classical classification models or modified versions to diagnose laryngoscopic images [12]-[15]. Although these methods have demonstrated some effectiveness, they may not be optimal for the task at hand. In diagnostic workflows, doctors often focus on key regions in laryngoscopic images, rather than analyzing the entire image. This targeted analysis resembles an object detection task, wherein attention is selectively directed to key regions, enabling a more accurate focus and ultimately leading to more dependable diagnostic outcomes.

Currently, only one open-source laryngoscopic classification dataset, Laryngoscope8, with 3057 images, is available. This dataset comprises 8 classes, encompassing 7 disease categories and 1 normal category. However, when it comes to implementing object detection models for laryngeal disease diagnosis, the necessity of having a dedicated object detection dataset becomes paramount. Unfortunately, individuals lacking medical expertise may find it impractical to manually extend a medical classification dataset to an object detection dataset due to a lack of relevant knowledge. Moreover, manual annotation is generally time-consuming, presenting challenges even for experienced physicians with limited time. To tackle this issue, we propose a Data Closed-Loop Network (DCLNet) for automatic annotation and classification of laryngoscopic images, As is shown in Fig. 1. Through an iterative loop that combines a foundation model [16] with a small model, DCLNet transforms the Laryngoscope8 dataset from low-quality annotated images to a high-quality annotated object detection dataset.

Furthermore, medical image datasets often exhibit characteristics such as poor interpretability and a long-tailed distribution. In our approach, the iterative loop of data continuously modifies annotated regions. By tracking these changes in annotations, interpretability can be effectively enhanced. Additionally, we adopt a curriculum learning approach, gradually progressing from easy to challenging instances, to enhance the model's ability to generalize across the entire dataset, significantly improving its performance on few-samples categories. Experimental results indicate that our model outperforms existing traditional models, achieving optimal performance. Specifically, the accuracy reaches 83.2%, with an average precision of 0.7492, an average recall of 0.6638, and an average f-score of 0.7039.

In summary, our contributions in this paper can be outlined as follows:

- We design a data closed-loop method DCLNet, which achieves automatic annotation and classification of medical images iteratively.
- We expand the open-source laryngoscopic image classification dataset Laryngoscope8 into a corresponding object detection dataset.
- Extensive experiments demonstrate that our method achieves superior results compared to traditional classification models.

II. Related Work

We discuss previous related work from the perspectives of datasets and methodologies.

A. Laryngoscope Dataset

Currently, many researchers have studied the task of laryngoscope classification based on deep learning. However, there are very few open-source datasets available. To our knowledge, Li et al. [15] were the first to publicly release an 8-class laryngoscope dataset named Laryngoscope8. David et al. [13] trained and validated on their collected laryngeal videos, but did not make their collected dataset public. Ren et al. [17] obtained 24,667 laryngoscope images from the West China Hospital database of Sichuan University, but this

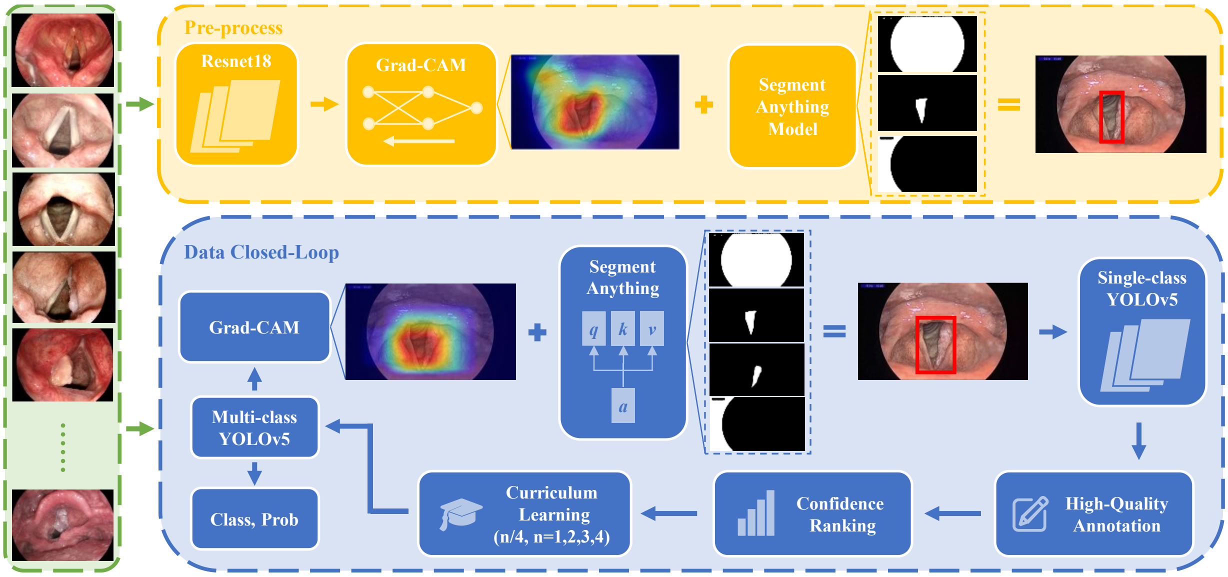

Fig. 1. The architecture of the DCLNet model. The model consists of two main parts: the pre-process and the data closed-loop. It can be observed that with the iterations of the data closed-loop, the activation maps generated by Grad-CAM become more accurate, and the granularity of the segmentation masks generated by SAM increases. This leads to more precise annotation box results and higher classification accuracy.

dataset has not been made open-source either. Thus, Laryngoscope8 stands as the first and only publicly available laryngoscope classification dataset. In this paper, we utilize Laryngoscope8 and extend it by employing an automated labeling approach, it converts the dataset from its initial classification format to the one for object detection, thus supporting the training of object detection models.

B. Methodology

With the widespread application of computer vision technology across various industries, Du et al. [18] employed traditional visual methods, manually delineating regions to extract features from laryngoscope images for diagnosing laryngeal diseases. As deep learning advanced, Gen et al. [19] adopted a deep learning approach, handling laryngoscope images end-to-end for the diagnosis of laryngeal diseases. Li et al. [15] initially used DenseNet121 to locate key regions within laryngoscope images. Wang et al. [14] introduced a novel network that adapts the handling of input images by automatically selecting different branches according to their categorization complexity. Yan et al. [20] employed the transformer architecture and proposed the dual-transformer model based on Vision Transformer. All of these methodologies entail the utilization of classification models for the diagnosis of laryngoscope images. In contrast to the existing approaches, our study capitalizes on the classification capability of a YOLOv5-based object detection model. We integrate attention mechanisms into the end-to-end model, thereby introducing an effective paradigm for laryngoscope image classification.

III. Method

Our method employs an object detection model YOLOv5 integrated with the attention mechanism through Grad-CAM to diagnose laryngeal diseases. Moreover, we utilize data closed-loop and curriculum learning in our proposed model to address inherent issues within the Laryngoscope8 dataset, including poor interpretability and a long-tailed distribution.

A. Design Overview

Fig. 1. illustrates the overall framework of our proposed DCLNet model. Our model consists of two main parts: pre-process and the data closed-loop. In pre-process, we first employ the ResNet18 model to train the Laryngoscope8 classification dataset. We extract the gradient-weighted class activation mapping values from correctly classifying images. Following this, based on the average activation values within several masks obtained through the Segment Anything Model applied to the corresponding original image, we select the mask with the maximum average activation value as the key mask. We then extend the key mask around its periphery to generate the bounding box as the critical areas. This process establishes the initial object detection dataset.

In the data closed-loop framework pipeline, we first train a single-class YOLOv5-based network, where the single class represents the key region in the input image. This model effectively producing high-quality annotations for crucial areas across all images. Following this model, we sort the output results by confidence, considering confidence to measure the sample's learning difficulty. Following the curriculum learning approach, we feed the samples into the multi-class YOLOv5 in ascending order of difficulty, from easy to challenging. Then the activation mapping derived from the multi-class model is utilized to further refine the annotated bounding boxes. After several iterations, a better annotated object detection dataset is obtained, resulting in the DCLNet model achieving excellent classification performance.

B. Automatic Annotation

Gradient-weighted Class Activation Mapping (Grad-CAM) [21] is a widely used technique in computer vision and deep learning. It serves as an interpretability tool [22]-[24]. Grad-CAM establishes a spatial correlation between the small model's predictions and the spatial locations in the input image, as illustrated in Fig. 2. The Segment Anything Model (SAM) [25] is an image segmentation model open-sourced by Meta. It is capable of segmenting any image based on a given prompt (including points, boxes, and text), and exhibits strong Zero-Shot capabilities for new samples. In our model, we choose to use a uniformly distributed grid as the prompt, which means SAM can search for potential objects on a global scale, as illustrated in Fig. 2. As a foundation model, SAM can perform

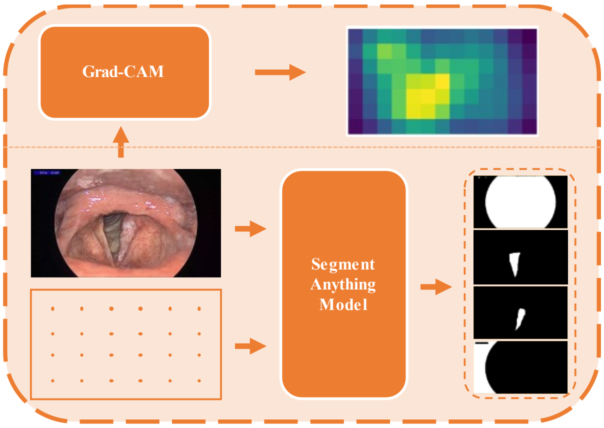

Fig. 2. The images go through the Grad-CAM module to generate region-level activation maps. Using SAM, multiple masks are obtained.

pixel-level segmentation on images, generating a series of masks with confidence scores. Due to differences in model size and parameter levels, the foundation model possesses stronger feature extraction capabilities but may be more prone to overfitting due to dataset-specific characteristics. Although the small model may not excel in granularity as the foundation model, it tends to produce more stable results in simpler tasks. Therefore, we combine the strengths of both foundation and small models, integrating the advantages of each, resulting in high-quality annotations for the Laryngoscope8 dataset.

C. *Data Closed-Loop*

In our DCLNet model, we utilize the "Grad-CAM & SAM Automatic Annotation" method on the trained model to refine the training set's annotations, enhancing the attention mechanism's focus. The updated set then further trains the model, enhancing performance and leading to more precise annotations. Subsequently, we use these new annotations to train the network for another iteration and continue this loop, as illustrated in Fig. 3. Furthermore, we increase the number of grid points used as prompts for the SAM model in each iteration. More points imply that the SAM model can generate finer-grained segmentation. Coordinating with more precise Grad-CAM activation maps in each round, our experiments show improved focus areas and classification accuracy.

D. *Curriculum Learning*

Curriculum Learning [26]-[28] is a machine learning paradigm that draws inspiration from educational learning strategies. This approach trains models progressively, starting with simpler examples and gradually moving to more complex ones. To mitigate long-tail issues in Laryngoscopy8, we train our model with curriculum learning. To distinguish difficulty levels of various samples in the dataset, we sort the confidence scores of the single-class YOLOv5-based model's output results in each round of data closed-loop. The confidence score serves as a measure of sample difficulty. We then divide the training set according to the sorted scores into four subsets based on quartiles (1/4, 2/4, 3/4, 4/4) and train on each of these subsets during each training iteration, as illustrated in Fig. 4. The experimental results show that this approach significantly improves the classification performance of our DCLNet model on the categories with a small number of data volume.

IV. Experiments

We evaluate our model on the Laryngoscopy8 dataset and compare it with existing methods. The relevant results and experiments are presented below.

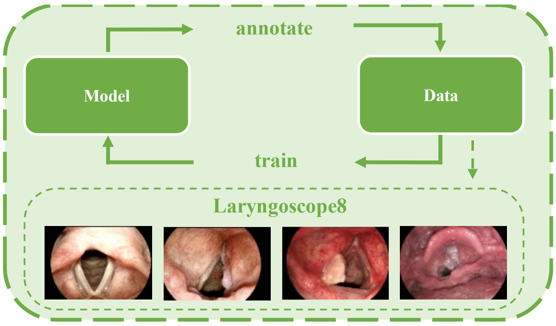

Fig. 3. Data closed-loop. A total of 4 iterations were conducted.

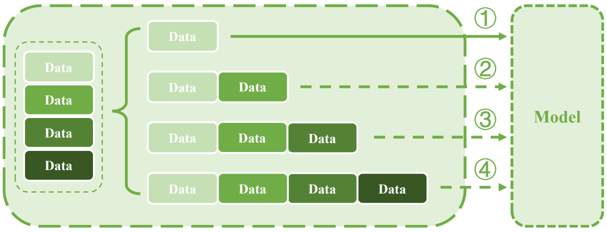

Fig. 4. Curriculum Learning. The training is conducted in four iterations progressively from easy to challenging.

A. *Experimental Setup*

The training is carried out on an NVIDIA 4090 GPU. The object detection models are trained with the SGD optimizer, beginning with a 0.01 learning rate. Each iteration is trained over 100 epochs, and the batch size is sequentially adjusted to 8, 16, or 32 as the volume of cyclic data increases. In the case of the SAM model, the points-per-side parameter is adjusted to 6, 8, or 10 with increasing cycles. These parameter configurations are carefully selected to refine the model to achieve the best possible results on the Laryngoscope8 dataset.

B. *Comparison With Other Methods*

We evaluate our method on the Laryngoscope8 dataset, with 70% for training and 30% for testing. We evaluate the performance of our proposed DCLNet model as shown in TABLE I. We used the same data augmentation on the Laryngoscope8 dataset to train classical classification models, with consistent hyperparameters. We compare these models with ours in terms of precision, recall, and f-score metrics. We observe that our model achieves optimal performance in 7 out of 8 categories. Moreover, we also compare the DCLNet model based on the accuracy metric as illustrated in TABLE II. The result indicates that our DCLNet surpasses the models proposed by previous researchers in terms of classification accuracy. Furthermore, in terms of visualization of results, Fig. 5. shows some examples of our DCLNet model's diagnostic results, indicating that our model can accurately focus on crucial areas in the image and correctly determine categories.

C. *Ablation Study*

We conduct a sequence of ablation experiments to validate the efficacy of the data closed-loop and curriculum learning approaches implemented in the DCLNet model.

1) Curriculum Learning: Compared to direct training on full data (Baseline), we incorporate curriculum learning throughout the training process. The outcomes of the comparative experiment are detailed in TABLE III. Initially, the experimental results of curriculum learning do not match the performance of Baseline in the initial rounds (as Baseline

TABLE I. PRECISION, RECALL, AND F-SCORE METRICS OF EACH CATEGORY FOR DIFFERENT METHODS.

Method	Metrics	Edema	Cancer	Granuloma	Normal	Leukoplakia	Cyst	Nodules	Polyps	Average
Resnest50 [29]	Precision	0.8333	0.0000	0.5684	0.6101	0.5472	**1.0000**	1.0000	0.4483	0.6259
	Recall	0.1515	0.0000	0.2983	0.8889	0.5088	**0.1364**	0.1176	0.4643	0.3207
	F-score	0.2564	0.0000	0.3913	0.7235	0.5273	**0.2400**	0.2105	0.4561	0.4241
Densenet121 [30]	Precision	0.6667	0.0000	0.5965	0.6968	0.6279	0.4286	0.7000	0.4264	0.5179
	Recall	0.4848	0.0000	0.5635	0.7778	0.4737	**0.1364**	0.4118	0.5000	0.4185
	F-score	0.5614	0.0000	0.5795	0.7351	0.5400	0.2069	0.5185	0.4603	0.4629
ViT [20]	Precision	**1.0000**	0.5000	0.8878	0.5700	0.6636	0.5000	**1.0000**	0.6248	0.7183
	Recall	0.0606	0.3333	0.7099	0.8939	0.5105	**0.1364**	0.3196	0.5869	0.4439
	F-score	0.1143	0.4000	0.7889	0.6961	0.5771	0.2143	0.4844	0.6053	0.5487
DCLNet	Precision	0.8438	**1.0000**	**0.9181**	**0.8793**	**0.8246**	0.1538	0.6389	**0.7347**	**0.7492**
	Recall	**0.8182**	**0.5000**	**0.8674**	**0.9015**	**0.8246**	0.0909	**0.4510**	**0.8571**	**0.6638**
	F-score	**0.8308**	**0.6667**	**0.8920**	**0.8903**	**0.8246**	0.1143	**0.5288**	**0.7912**	**0.7039**

utilizes the entire training set). However, with an increasing number of rounds, the results of curriculum learning eventually surpass Baseline.

2) Data Closed-Loop: Based on curriculum learning, we introduce data closed-loop and conduct 4 iterations. The results in TABLE III demonstrate that incorporating data closed-loop consistently outperforms exclusive curriculum learning in classification accuracy. Moreover, the visualization results also verify the effectiveness of the data closed-loop, as shown in Fig. 6. As iterations increase, Grad-CAM's activation maps improve, enabling SAM to generate more precise masks, highlighting image's crucial areas with greater accuracy. Beyond average accuracy, we assess precision, recall, and f1-score for final iterations. The experiment results indicate that the data closed-loop and curriculum learning significantly improve the DCLNet model's overall classification performance, as illustrated in TABLE IV.

TABLE II. COMPARISON OF ACCURACY AMONG DIFFERENT MODELS.

Methods	Accuracy
CheXNet [31]	71.0%
AG-CNN [32]	71.0%
Inception_v3 [33]	71.0%
DenseNet-121 [15]	73.0%
HDCNet [14]	75.3%
ViT [20]	82.2%
Cross-ViT [20]	82.9%
Ours	**83.2%**

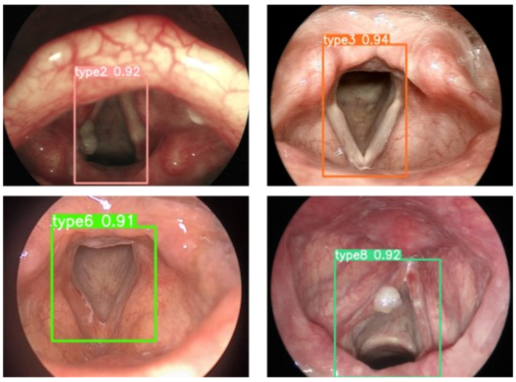

Fig. 5. Visualization of detection results.

TABLE III. COMPARISON OF ACCURACY AMONG DATA CLOSED-LOOP (DCL), CURRICULUM LEARNING (CL) AND BASELINE.

Round	Data Volume	DCL & CL	CL	Baseline
1	1/4	63.02%	62.04%	69.80%
2	2/4	75.16%	73.09%	80.31%
3	3/4	80.42%	78.77%	81.73%
4	1	**83.15%**	82.82%	82.60%

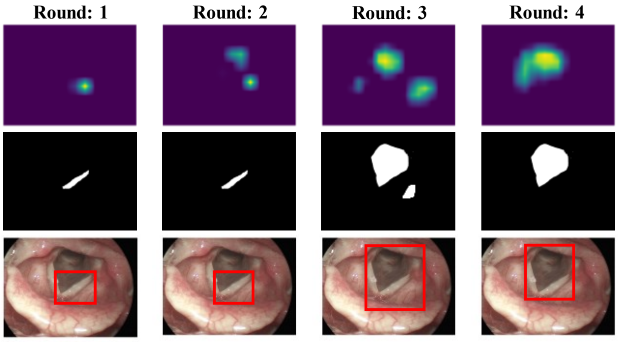

Fig. 6. Activation map, mask and annotation change with iteration.

TABLE IV. COMPARISON OF AVERAGE PRECISION, RECALL, AND F1-SCORE AMONG DIFFERENT METHODS.

Method	Precision	Recall	F-score
Baseline	0.6450	0.6083	0.6261
CL	0.6891	0.6246	0.6553
DCL & CL	**0.7491**	**0.6638**	**0.7039**

V. CONCLUSION AND FUTURE WORK

In this study, we introduce an innovative approach to annotating laryngeal images. It changes a dataset designed for classification into one intended for object detection. by leveraging the strengths of both the foundation model and the small model. Moreover, through the training approach of data closed-loop and curriculum learning, our proposed DCLNet model enhances the interpretability of medical datasets and alleviates the impacts of long-tail distribution, and exhibits competitive results on the Laryngoscopy8 dataset.

REFERENCES

[1] N.G.M.M. Agency, "Analysis of China's oral and throat disease market. " Admen 000(006) (2010). 28–28.

[2] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis." Annual review of biomedical engineering 19 (2017): 221-248.

[3] Litjens, Geert, et al. "A survey on deep learning in medical image analysis." Medical image analysis 42 (2017): 60-88.

[4] Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." IEEE transactions on medical imaging 35.5 (2016): 1285-1298.

[5] Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." nature 542.7639 (2017): 115-118.

[6] Rodrigues, et al. "A new approach for classification skin lesion based on transfer learning, deep learning, and IoT system." Pattern Recognition Letters 136 (2020): 8-15.

[7] Polsinelli, Matteo, Luigi Cinque, and Giuseppe Placidi. "A light CNN for detecting COVID-19 from CT scans of the chest." Pattern Recognition Letters 140 (2020): 95-100.

[8] Jiang, Huiyan, et al. "A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation." Computers in Biology and Medicine (2023): 106726.

[9] Alqahtani, Tariq Mohammed. "Big Data Analytics with Optimal Deep Learning Model for Medical Image Classification." Comput. Syst. Sci. Eng. 44.2 (2023): 1433-1449.

[10] Kim, Hee E., et al. "Transfer learning for medical image classification: a literature review." BMC medical imaging 22.1 (2022): 69.

[11] Rajaraman, Sivaramakrishnan, Prasanth Ganesan, and Sameer Antani. "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks." PloS one 17.1 (2022): e0262838.

[12] Esmaeili, Nazila, et al. "Contact Endoscopy–Narrow Band Imaging (CE-NBI) data set for laryngeal lesion assessment." Scientific Data 10.1 (2023): 733.

[13] Wellenstein, David J., et al. "Detection of laryngeal carcinoma during endoscopy using artificial intelligence." Head & Neck 45.9 (2023): 2217-2226.

[14] Wang, Shaoli, et al. "Hierarchical dynamic convolutional neural network for laryngeal disease classification." Scientific Reports 12.1 (2022): 13914.

[15] Yin, Li, et al. "Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism." Pattern Recognition Letters 150 (2021): 207-213.

[16] Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).

[17] Ren, Jianjun, et al. "Automatic recognition of laryngoscopic images using a deep‐learning technique." The Laryngoscope 130.11 (2020): E686-E693.

[18] Du, Chen, et al. "Validation of the laryngopharyngeal reflux color and texture recognition compared to pH‐probe monitoring." The Laryngoscope 127.3 (2017): 665-670.

[19] Ye, Gen, et al. "Deep learning for laryngopharyngeal reflux diagnosis." Applied Sciences 11.11 (2021): 4753.

[20] Yan, Fangyuan, Bin Yan, and Mingtao Pei. "Dual Transformer Encoder Model for Medical Image Classification." 2023 IEEE International Conference on Image Processing (ICIP). IEEE, 2023.

[21] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[22] Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.

[24] Omeiza, Daniel, et al. "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models." arXiv preprint arXiv:1908.01224 (2019).

[25] Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).

[26] Bengio, Yoshua, et al. "Curriculum learning." Proceedings of the 26th annual international conference on machine learning. 2009.

[27] Wang, Xin, Yudong Chen, and Wenwu Zhu. "A survey on curriculum learning." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.9 (2021): 4555-4576.

[28] Soviany, Petru, et al. "Curriculum learning: A survey." International Journal of Computer Vision 130.6 (2022): 1526-1565.

[29] Zhang, Hang, et al. "Resnest: Split-attention networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[30] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[31] Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225 (2017).

[32] Guan, Qingji, et al. "Thorax disease classification with attention guided convolutional neural network." Pattern Recognition Letters 131 (2020): 38-45.

[33] Xiong, Hao, et al. "Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images." EBioMedicine 48 (2019): 92-99.