Applying Knowledge Distillation for Vocal Fold Identification in Laryngoscopy

Thao T.P Dao[1,2,3,4], Bich Anh Tran[5], Minh-Khoi Pham[1,2], Mai-Khiem Tran[1,2,3], Boi Ngoc Van[6], Chanh Cong Ha[7], Minh-Triet Tran[1,2,3]
[1]University of Science, Ho Chi Minh City, [2]Viet Nam National University, Ho Chi Minh City, [3]John von Neumann Institute,
[4]Thong Nhat Hospital, [5]Cho Ray Hospital, [6]Vinmec Central Park International Hospital, [7]7A Military Hospital

We are here for you
AAO-HNSF 2023
ANNUAL MEETING & OTO EXPERIENCE
SEPT. 30 – OCT. 4 | NASHVILLE, TN

Abstract

The increased need for more laryngoscopy datasets aims to bridge the gap between computer vision and laryngology research as well as support clinical practice. In this research, We present a novel vocal fold image dataset for image classification task and a lightweight deep learning network with reduced model weights but still high accuracy for mobile devices. The proposal involves AI assistance on smartphones for a portable laryngoscopy system aiding laryngoscopists in focusing on vocal fold areas.

Introduction

In the field of laryngoscopy, connecting scopes to smartphones offers a convenient and portable way for otolaryngologists to diagnose and plan treatments based on larynx images [1-2]. However, there haven't been many studies exploring the use of deep learning models for these smartphone-connected laryngoscopy devices to analyze vocal fold images. Therefore, it's essential to develop a lightweight deep learning network specifically designed for these handheld laryngoscopy systems.

Deep learning employs a technique called Knowledge Distillation (KD), which involves transferring expertise from a powerful but complex network (as a teacher model) to a simpler model (the student model) [3]. This process enhances the performance of the student model without affecting its efficiency. By passing on knowledge from teacher model, student model can achieve superior results compared to another student model that hasn't undergone this knowledge transfer process.

Consequently, we create a significantly smaller and more specialized model using KD for use in portable laryngoscopy devices.

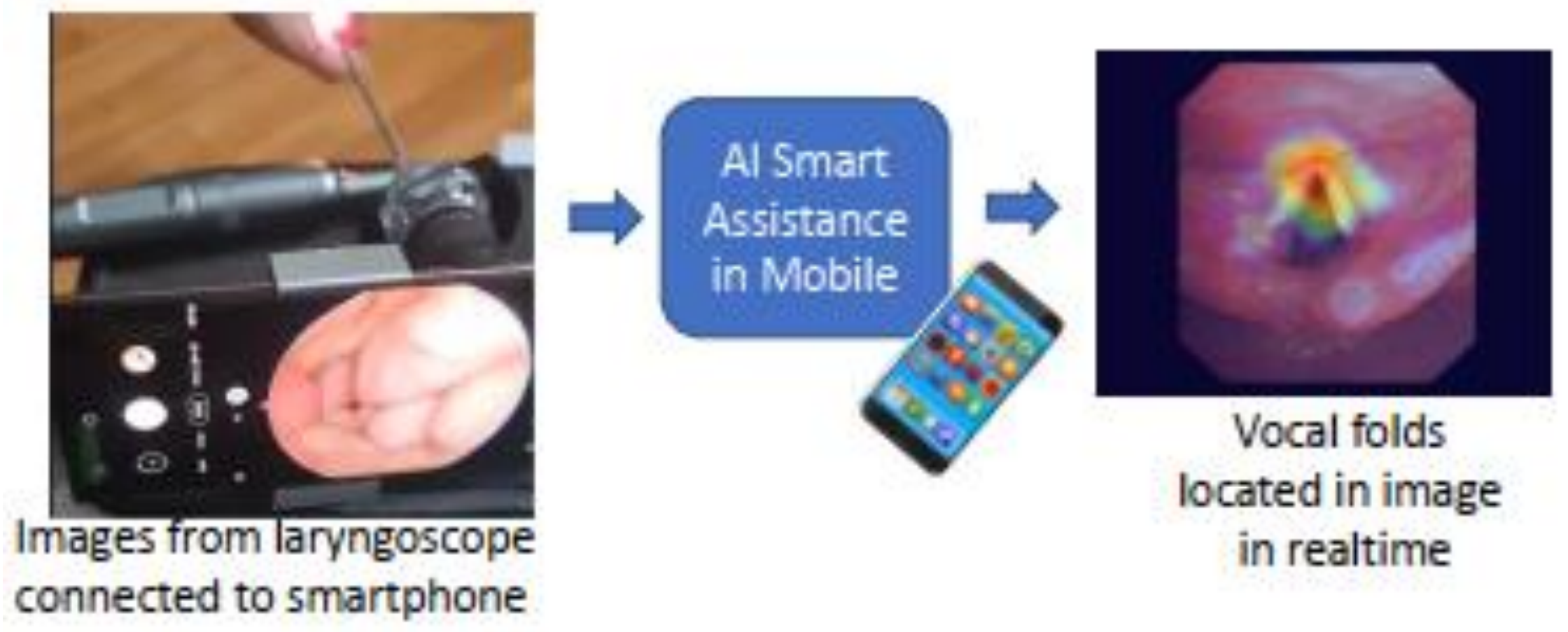
Images from laryngoscope connected to smartphone
AI Smart Assistance in Mobile
Vocal folds located in image in realtime

Figure 1. AI smart assistance on smartphone for vocal fold detection

Methods and Materials

We plan on performing several steps: collect samples, filter samples, establish the ground truth data, train and evaluate the different deep learning models, and finally compress the best model by knowledge distillation to apply on edge devices.

The Figure 2 show overview of our process to develop a smart vision-based assistance for vocal fold detection and localization.

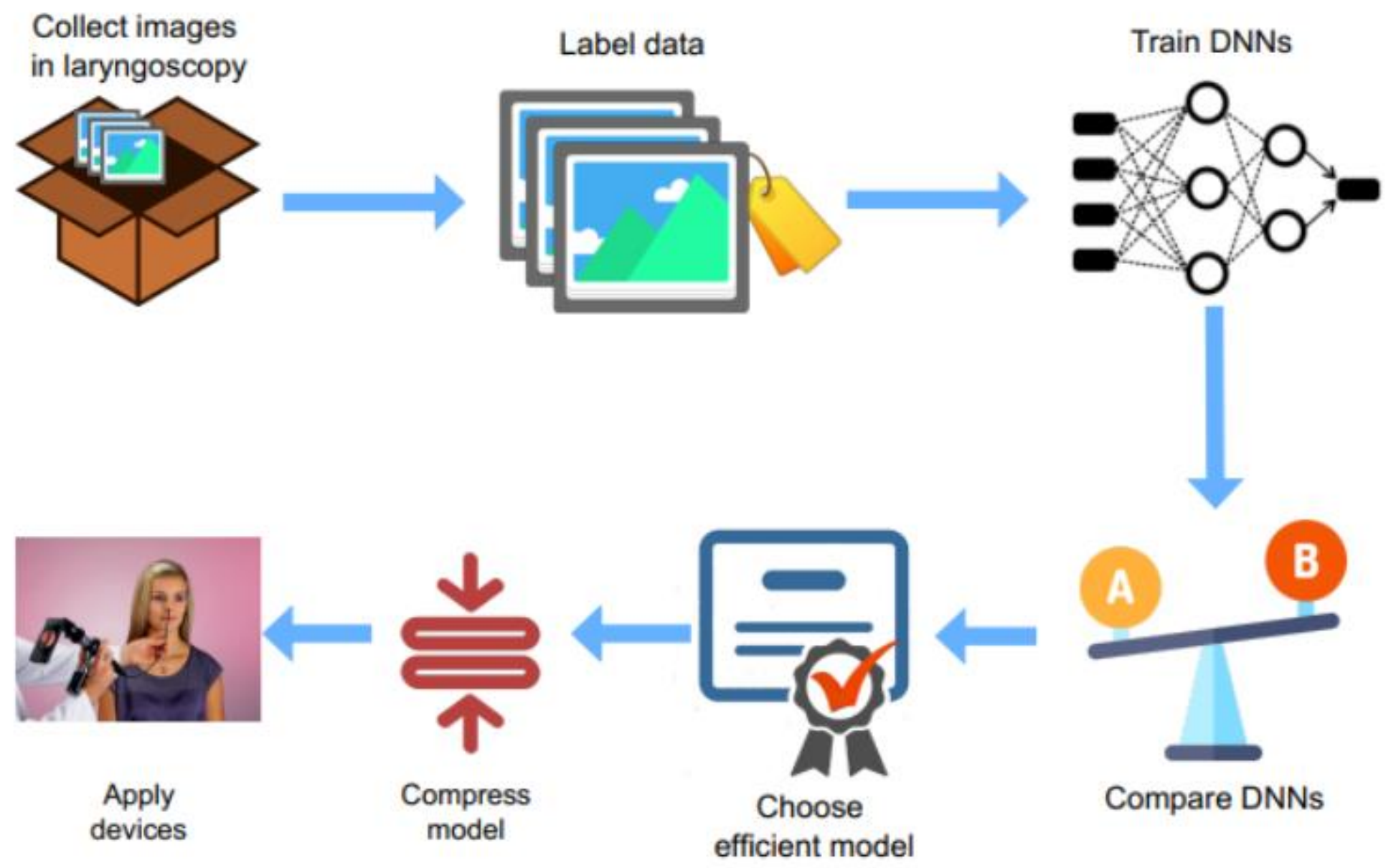

Figure 2. Overview of our process to develop a smart vision-based assistance

This below figure illustrates our developed KD architecture, which consists of a teacher network and a student network, to learn laryngoscopy images.

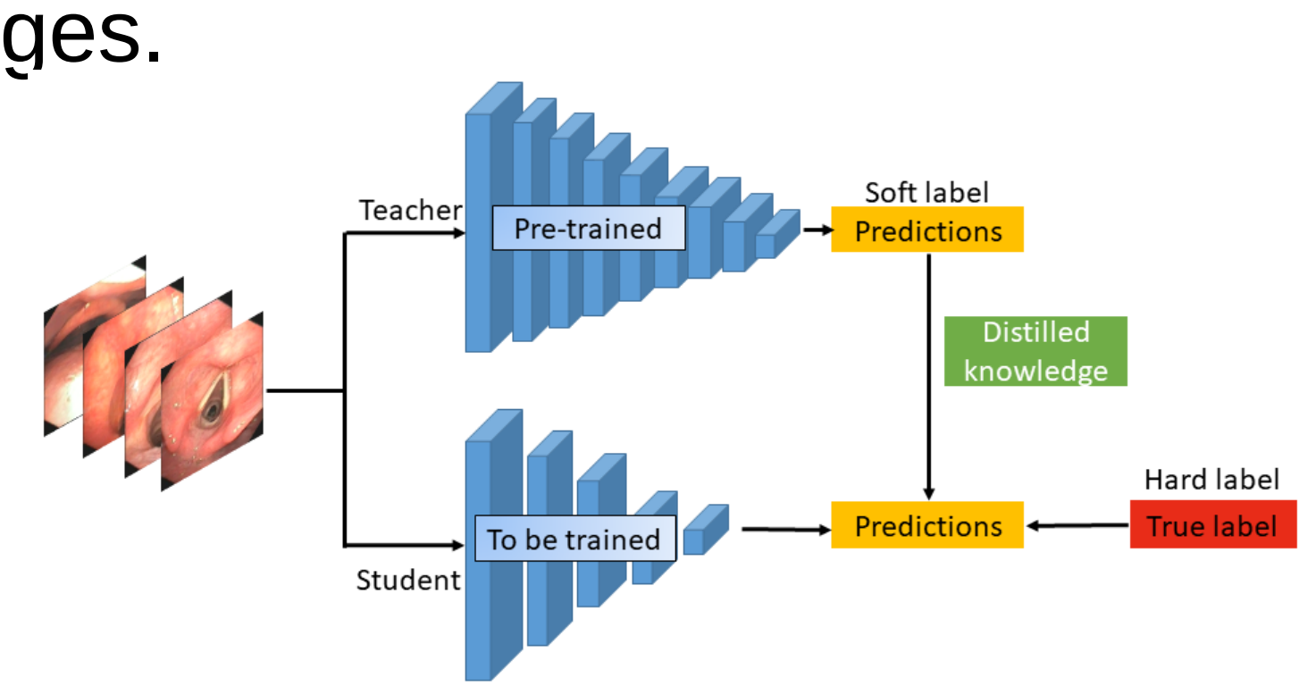

Figure 3. Overview of our KD architecture

Results and Discussion

As illustrated in Table 1, EfficientNetB1 shows extremely high accuracy, recall, and precision in assessing images of vocal fold appearance.

Table 1. Results of state-of-the-art backbones.

	VGG19	ResNet50V2	MobileNetV2	InceptionV3	DenseNet201	Xception	EfficientNetB1
Accuracy (%)	91.8	98.5	96.1	98.3	98.2	98.2	**98.7**
Recall (%)							
Non vocal fold	94.2	98.8	96.7	98.3	98.3	98.5	**99.2**
Vocal folds	88.7	98.0	95.3	**98.3**	98.0	97.8	98.0
Precision(%)							
Non vocal fold	91.4	98.5	96.4	**98.6**	98.5	98.3	98.5
Vocal folds	92.3	98.5	95.8	97.8	97.8	98.0	**99.0**

The Table 2 illustrates quantitative results. Distillation helps the student to approximately match the teacher's performance and also helps faster convergence and require minimal computing resource.

Table 2. Comparison between student and teacher performance on our validation set

Methods	Accuracy	No. parameters
EfficientNet-B1	98.7%	6.7M
MobileNetV2	96.1%	2.4M
Simple-ResNet-Scratch	96.7%	0.8M
Simple-ResNet-Distilled	98.4%	0.8M

In Figure 4, EfficientNetB1 exhibits good Grad-CAM results by focusing heavily on important anatomical landmarks for vocal fold exploration. Furthermore, our distilled model also performs well in feature extraction in both visible and invisible vocal fold images compared to scratch model.

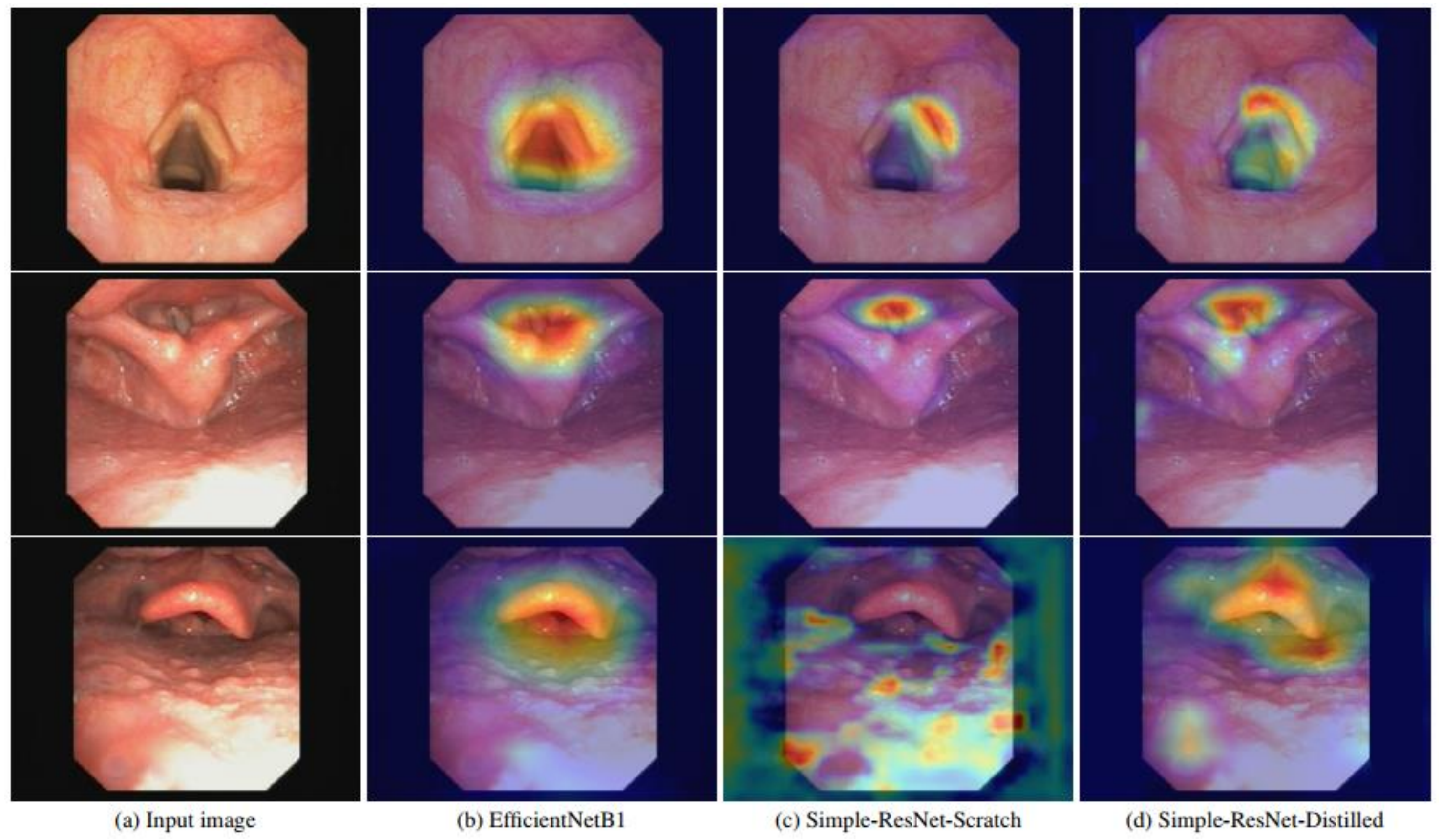

Figure 4. Grad-CAM visualization on our laryngoscopy image dataset

Conclusions

- ➢We introduce a new dataset about the laryngeal endoscopic images for detecting vocal folds.
- ➢We evaluate some backbones to choose a suitable model for applying knowledge distillation.
- ➢We propose a solution of AI assistance on smartphones to create a portable and smart laryngoscope system.

Contact

- Name: Thao Thi Phuong Dao
- Organisation:
  - John von Neumann Institute - University of Science, Viet Nam National University, Vietnam
  - Thong Nhat Hospital, Vietnam
- Email: Thao.dao2020@ict.jvn.edu.vn

References

1. Rosen, C. & Murry, T. Diagnostic laryngeal endoscopy.. Otolaryngologic Clinics Of North America. 33 4 pp. 751-8 (2000)
2. Samlan, R. & Kunduk, M. Visual Documentation of the Larynx. Cummings Otolaryngology: Head And Neck Surgery. 1 pp. 808-813 (2020)
3. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C. & Ma, K. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. 2019 IEEE/CVF International Confer-ence On Computer Vision (ICCV). pp. 3712-3721 (2019)