S & M 3822

Towards Boundary More Precise Detection: Surrounding-to-aggregating Deep Learning in Videoscope Imaging

Huang Yangyiyi,[1*] Jinchao Ge,[2] Weiming Fan,[3] YiQun Zheng,[5] and Changting Lin[3,4]

[1]Department of Otolaryngology, Second Affiliated Hospital, Zhejiang University School of Medicine,
Hangzhou, Zhejiang 310009, China
[2]the School of Computer and Mathematical Sciences, University of Adelaide, Adelaide 5000, Australia
[3]Innovation Center for Smart Medical Technologies & Devices, Binjiang Institute of Zhejiang University,
Hangzhou, Zhejiang 310053, China
[4]College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310058, China
[5]Hangzhou Tuya Information Technology Co., Ltd, Hangzhou, Zhejiang 310030, China

(Received May 14, 2024; accepted July 3, 2024)

Keywords: laryngeal cancer, deep learning, narrow-band imaging, computer-assisted image interpretation, videolaryngoscopy

The assessment of early laryngeal cancer and pre-neoplastic lesions is subjective and depends on doctors' experience, leading to missed diagnoses in primary institutions. Our objective was to develop and validate a deep learning algorithm for the real-time identification of early laryngeal cancer and pre-neoplastic lesions, aiming to enhance diagnostic accuracy. The challenge observed in the domain of deep learning arises from overlooking contextual information. In response, we introduce in this paper a learning methodology that advances from acknowledging the surrounding context to integrating it, providing a resolution to this problem. Initially, we introduce side-aware features to capture relevant characteristics. Subsequently, we employ a rectangular selection technique for accurately determining regions of interest. To assess the effectiveness of our approach in object detection, we perform evaluations on a clinical dataset. Our deep learning approach exhibits robust performance in discriminating cancer. The images were randomly divided into training (80%), testing (10%), and validation (10%) sets. The testing was performed on a laryngoscope dataset consisting of 1123 samples. When compared with other advanced detection models, our methodology surpassed them, demonstrating superior results in laryngoscope detection, including *mAP*, accuracy, recall, and *F*1 *score*. In this study, we identified a learning method conducive to polyp detection in video laryngoscopy under both white-light and narrow-band imaging. The promising detection performance holds the potential to improve diagnostic proficiency and decrease the likelihood of missed diagnoses among primary otolaryngologists.

1. Introduction

Laryngeal cancer stands out as one of the most prevalent tumors in the head and neck.[(1,2)] The possibility of early diagnosis and treatment permits larynx-preserving approaches and

[*]Corresponding author: e-mail: 2317155@zju.edu.cn

https://doi.org/10.18494/SAM5140

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

correlates with an improved prognosis, contributing to a heightened five-year survival rate.(3) However, early laryngeal cancer and pre-neoplastic lesions frequently present as subtle mucosal alterations during white-light transnasal flexible electronic endoscopy, resulting in overlooked diagnoses even by seasoned endoscopists.(4) Lesions featuring diverse grades of dysplasia may display analogous mucosal white plaques under conventional white-light endoscopy. The histological diversity of lesions further complicates the process of diagnosis and treatment planning.(5) Recently developed image-enhanced endoscopic techniques, such as narrow-band imaging(6) and iSCAN,(7) offer a more detailed visualization of epithelial and subepithelial microvessel patterns. Owing to a relatively prolonged learning curve, the detection and diagnosis of laryngeal cancer and pre-neoplastic lesions pose considerable challenges in primary institutions or among inexperienced physicians.(8,9)

Recent studies have consistently shown the exceptional capabilities of deep learning algorithms in image detection and recognition,(10–13) aiding physicians in the detection and identification of specific lesions. This, in turn, enhances the accuracy and efficiency of disease diagnosis.

An artificial intelligence assistance system is crucial for laryngoscope examinations, aiding clinical physicians in detecting frequently overlooked cancers and providing characterizations.(14,15) Deep learning models and methods consistently exhibit remarkable performance in AI-assisted diagnostics, especially when applied to extensive datasets.(16,17) With the ongoing evolution of deep learning, various object detection techniques, which are adaptable for endoscopic image analysis, have emerged.(18–20) Pacal *et al.*(21) introduced a single-stage regression-based object detector utilizing the You-Only-Look-Once (YOLO) convolutional neural network (CNN) architecture for polyp detection, employing bounding boxes in endoscopic images. Furthermore, Eelbode *et al.*(22) enhanced polyp segmentation accuracy by integrating a temporal architecture, implementing a recurrent neural network (RNN) atop a CNN. Shin *et al.*(23) adopted a region-based object detection approach using Faster R-CNN, complemented by post-learning methods for localizing polyp lesion positions. A recent innovative approach(24) reported the use of 3D-CNNs for polyp detection, introducing spatiotemporal features through consecutive frames of polyps to enhance detection performance.

In this manuscript, we present SA-Detect, a method specifically designed for detecting tissue frames in laryngoscope videos, with a primary focus on cancer detection. This approach relies on learning regional edge features to precisely define bounding boxes. The model takes video data as input, performs initial data preprocessing to enhance detection performance, and then extracts image features using a CNN. SA-Detect is subsequently applied to obtain side-aware features for detection and localization. The implementation of SA-Detect demonstrates a significant improvement in precision, surpassing other detectors by 91.4% on the clinical dataset. These findings highlight SA-Detect's potential as a seamlessly integrable solution with minimal computational overhead, contributing to enhanced boundary delineation.

2. Data, Materials, and Methods

2.1 Data acquisition

The dataset comprises patients who underwent laryngeal endoscopy from August 2020 to December 2022 at the Endoscopy Center within the Otolaryngology Department, the Second Affiliated Hospital of Zhejiang University School of Medicine. Endoscopic images were captured using iSCAN Defina EPK-3000 series endoscopes (Pentax Corporation, Japan). Inclusion criteria stipulated that the lesion be situated in the glottis area with corresponding postoperative pathological results. Exclusion criteria encompassed the following: (1) poor image quality (e.g., occlusion, blurring and out of focus) and (2) patients with a history of head and neck radiation therapy. The dataset comprised 239 patients with a total of 1123 images.

2.2 Image detection network

Building upon YOLOv5 with CSP Darknet as the backbone, our innovative architectural framework significantly enhances localization capabilities by efficiently propagating robust responses from low-level patterns, highlighted in red. These responses, often corresponding to edges or specific instance parts, contribute substantially to precise instance localization. As illustrated in Fig. 1, our proposed SA-Detect relies on a fusion of CNNs and a Feature Pyramid Network (FPN) for object detection. The backbone network employs convolutional operations to extract object-related information, spanning multiscale visual features. Subsequently, the neck network consolidates these multiscale features from the backbone network, concluding with classification and discrimination scores for distinct objects. The input image's spatial resolution is set at 300 × 300 pixels. We employ Layers 1, 2, and 3 as feature levels, each with dimensions of 76 × 76 × 255, 38 × 38 × 255, and 19 × 19 × 255, respectively.

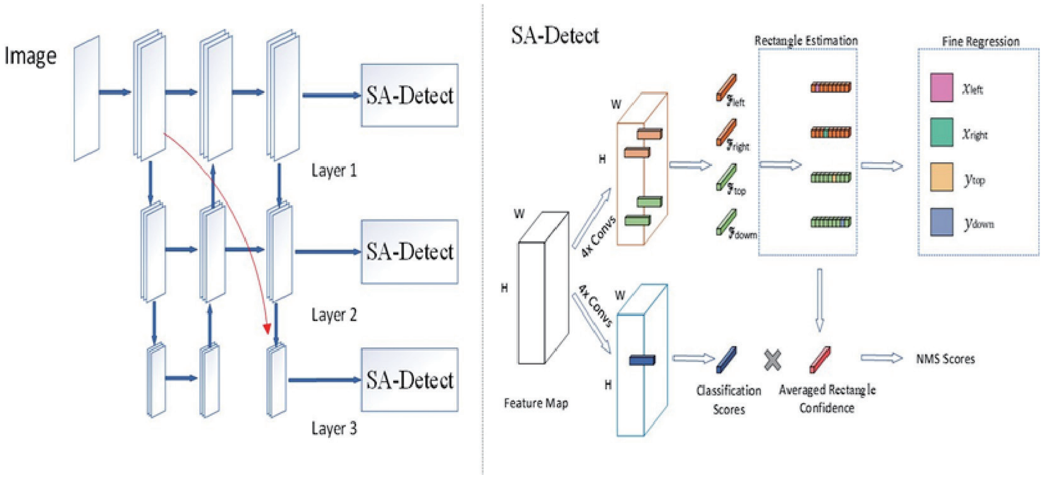

Fig. 1. (Color online) Overview of SA-Detect. We first trained an image network with CSP Darknet to create a feature vector describing the input image. Following this, the SA-Detect framework combines features from regions of interest to produce side-aware features. The classification network is tasked with identifying cancerous regions.

2.3 Surrounding-to-aggregating learning

In tasks involving object detection, frameworks predominantly depend on bounding box regression for object localization. The conventional approach often focuses on predicting object centers and sizes,(25–27) potentially constraining overall detection performance, especially in scenarios with substantial displacements and variances between anchor points and target objects. For example, when an anchor point is positioned at the edge of a target object, its detection might be impacted by relative scaling factors, leading to background features outweighing those related to the object. To overcome this limitation, we present SA-Detect, which relies on side-aware boundary features.

SA-Detect comprises CNN and RNN components, taking FPN features as input. After four convolutional layers, the model learns classification and localization features separately, effectively capturing directional information about regions of interest—a crucial step for generating side-aware features within SA-Detect. When presented with a candidate bounding box, denoted as (R_{left}, R_{right}, R_{top}, R_{down}), we expand this candidate region by a scale factor α (where $\alpha > 1$) to encompass the entire object. The candidate region is then divided into $2k$ rectangular regions along the x- and y-axes, with each boundary corresponding to k rectangles. In the rectangle selection stage, a binary classifier based on side-aware features determines whether each boundary lies inside one of the side rectangles or is closest to a particular side rectangle. Additionally, a regressor predicts the offset from the centerline of the selected rectangle to the true boundary, enabling finer regression. Finally, confidence estimation for rectangles indicates the reliability of predicted positions. To maintain more accurate bounding boxes during nonmaximum suppression (NMS),(28,29) localization confidence guides the process by computing average confidence scores for the four boundaries of the rectangles. Multiclass classification scores are adjusted by multiplying them with the average localization confidence for candidate ranking during NMS. This scoring mechanism plays a crucial role in preserving the best boxes with high classification confidence and precise localization.

2.4 Loss design

We have formulated a singular task objective for training SA-Detect, integrating the confidence loss $\mathcal{L}_{conf}$, rectangle loss $\mathcal{L}_{rec}$, and fine regression loss $\mathcal{L}_{reg}$. Recognizing the impact of distinguishing samples of varying difficulty on detection outcomes, we address the potential imbalance in scenarios with numerous easily distinguishable negative samples that may lead to the neglect of positive samples throughout the entire training process. To alleviate this issue, we introduce a modulation factor. Consequently, $\mathcal{L}_{conf}$ is expressed as

$$\mathcal{L}_{conf}\left(x\right)=-\left(1-x\right)^{\tau}\log\left(x\right), \tag{1}$$

where $-(1-x)$ amplifies the loss ratio for challenging-to-distinguish samples. We have replaced the conventional bounding box regression loss with the rectangle loss $\mathcal{L}_{rec}$ and included the fine regression loss $\mathcal{L}_{reg}$. Specifically, the rectangle loss employs the Cross-Entropy Loss,[30] whereas the fine regression loss utilizes the Smooth L1 Loss.[31] In summary, the overall loss function is denoted as

$$\mathcal{L} = 1.5\left(\mathcal{L}_{rec} + \mathcal{L}_{reg}\right). \tag{2}$$

2.5 Surrounding-to-aggregating learning

We comprehensively evaluated our proposed model using various metrics derived from the counts of true-positive (*TP*), true-negative (*TN*), false-positive (*FP*), and false-negative (*FN*) samples.

The Intersection over Union (*IoU*) metric quantifies the overlap between the predicted bounding box and the ground truth bounding box. A *TP* is registered when the *IoU* is greater than or equal to 0.5. The *IoU* can be mathematically expressed as

$$IoU = \frac{\text{Predicted result} \cap Ground\ true}{\text{Predicted result} \cup Ground\ true}. \tag{3}$$

The mean Average Precision (*mAP*) serves as a comprehensive metric for assessing the performance of object detection models. It is defined as

$$mAP = \frac{\sum_{i=0}^{n} AP_i}{n}, \tag{4}$$

where *n* represents the number of classes.

Precision (*P*) evaluates the accuracy of positive predictions and can be expressed as

$$Precision = \frac{TP}{TP + FP}. \tag{5}$$

Recall (*R*) measures the model's ability to correctly identify all relevant instances (true positives) and can be calculated as

$$Recall = \frac{TP}{TP + FN}. \tag{6}$$

The *F*1 *Score*, offering a balance between precision and recall, is computed as

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{7}$$

3. Results

3.1 Equations

We performed a comparative analysis involving SA-Detect and various prior and contemporary approaches. Table 1 provides a summary of their components and performances characteristics based on the clinical dataset. The performance metrics presented in this study are obtained from released implementations and models.

In this study, we systematically assessed the overall performance of our methodology in detecting diverse abnormalities within a clinical dataset. The outcomes, illustrated in Table 1, highlight the exceptional performance of our detection model in multiclass tasks, achieving notable metrics with a recall of 87.6%, a precision of 91.4%, and an F1 score of 89.5%. These metrics validate the efficacy of our proposed network in multi-object detection. Note that each region category occupies a distinct proportion of the image, leading to a notable presence of negative regions that may impact the precise prediction of positive regions. Nevertheless, SA-Detect has demonstrated its adaptability in addressing challenges presented by imbalanced data. In particular, SA-Detect yields outstanding results in both precision and *mAP* scores. In contrast, other methods frequently neglect correlations between instances and disregard the balance between positive and negative samples. Additionally, SA-Detect surpasses all other competing methods, showing a significant increase of 1.5% in *mAP* and a 3.7% boost in accuracy compared with the second-best method. Note that YOLO-based methods often struggle to capture contextual information, whereas FCOS exhibits the ability to grasp extensive global contextual information. The experimental results emphasize the benefits of side-aware boundary feature learning in object detection. Figure 2 effectively illustrates the challenges posed by images with diverse colors and textures. In comparison with alternative models, our approach adeptly recognizes cancer categories and accurately labels cancerous regions.

Table 1
Tracking performance based on the clinical dataset.

Model	Dataset	*mAP*@.5	*P*	*Recall*	*F*1
YOLOv4[32]	clinical dataset	83.6	83.8	78.8	81.2
YOLOv5s	clinical dataset	86.4	81.7	84.0	82.8
CornerNet[33]	clinical dataset	87.9	87.8	86.3	87.0
YOLOv5n	clinical dataset	86.5	85.5	81.1	83.2
YOLOv5x	clinical dataset	86.2	82.4	81.2	81.8
YOLOv5s	clinical dataset	88.0	85.3	83.6	84.4
YOLOv5m	clinical dataset	88.1	81.2	84.0	82.6
YOLOv5l	clinical dataset	88.5	85.0	83.8	84.4
FCOS[34]	clinical dataset	89.2	87.7	86.9	87.3
PP-YOLOE[35]	clinical dataset	89.8	88.3	87.1	87.6
PE-YOLO[36]	clinical dataset	88.3	86.4	85.7	85.8
SA-Detect (Ours)	clinical dataset	90.7	91.4	87.6	89.5

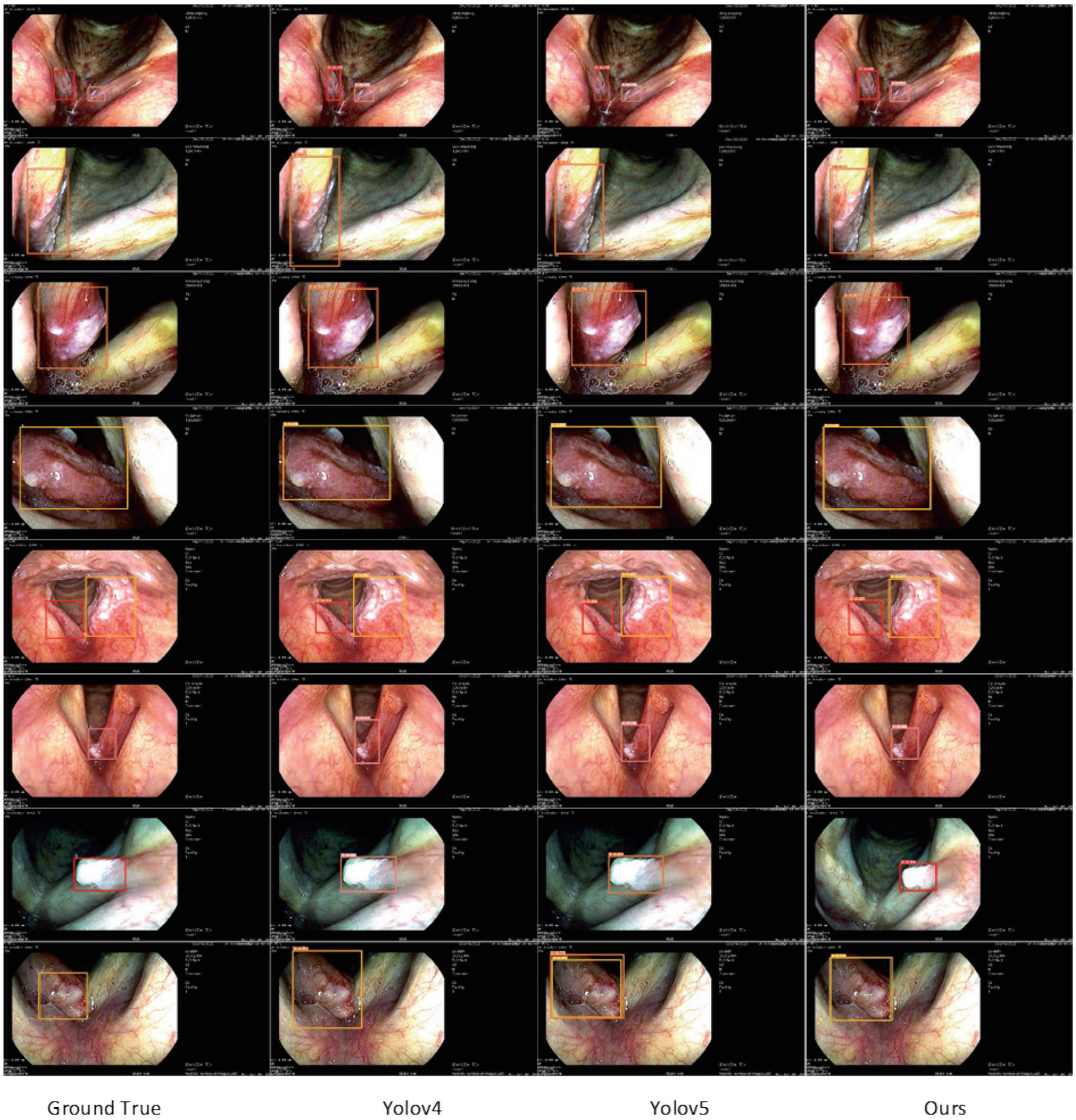

Fig. 2. (Color online) Examples of detection output from a challenging clinical dataset. The identities for each object are denoted on boxes. 0, 1, 2, and 3 represent the chronic inflammation of mucosa or squamous hyperplasia, low-grade dysplasia, high-grade dysplasia, and invasive carcinoma, respectively.

Figure 3 presents the confusion matrix of the training data for the SA-Detect model. Clearly, in the multiclass object detection task, the model demonstrates exceptional overall accuracy across all four classes, achieving accuracies for the chronic inflammation of mucosa or squamous hyperplasia, low-grade dysplasia, high-grade dysplasia, and invasive carcinoma.

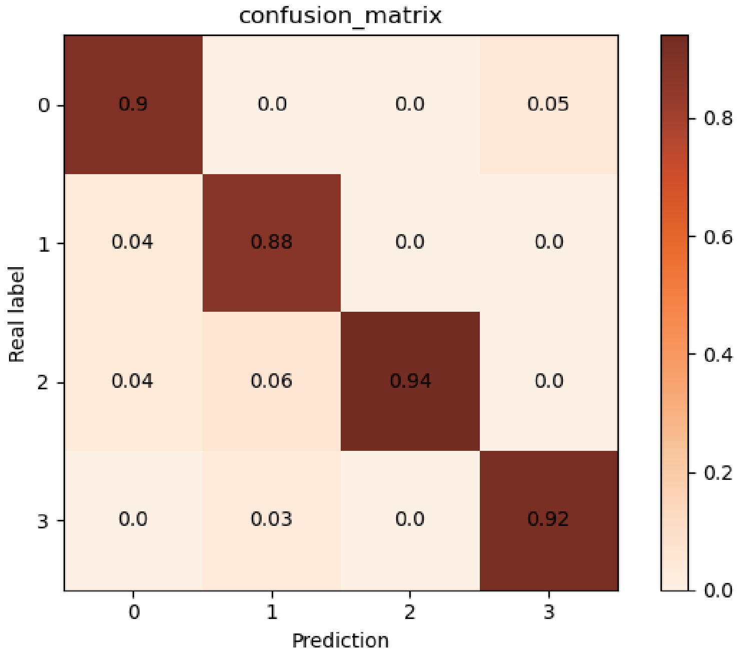

Fig. 3. (Color online) Confusion matrix of SA-Detect. 0, 1, 2, and 3 represent the chronic inflammation of mucosa or squamous hyperplasia, low-grade dysplasia, high-grade dysplasia, and invasive carcinoma, respectively.

3.2 Ablation study

As shown in Table 2, to further investigate the contribution of the SA-Detect module to overall performance, we conducted a series of ablation studies on the clinical dataset. The top-down feature augmentation consistently enhances *mAP* and precision, confirming the efficacy of low-level feature information (region edges). We introduce the Confidence Class (CC) to adjust classification scores, inherently reflecting the model's confidence in boundary localization. The application of CC further enhances AP performance. As depicted in Fig. 4, we applied Surrounding-to-Aggregating Learning (SAL) to bring the detection boxes closer to reality. This method significantly improved performance from 85.7 to 90.7%, emphasizing the crucial role of feature learning for each boundary in object detection.

3.3 Boosting existing detection

The performance of SA-Detect with various detectors is presented in Table 3. We utilized Faster R-CNN and Mask R-CNN as two-stage detectors. Following the recent convention for single-stage methods, SA-Detect was incorporated after their backbone networks. This integration resulted in improvements, increasing the *mAP* of Faster R-CNN from 87.6 to 89.4% and the *mAP* of Mask R-CNN from 86.9 to 88.6%. For single-stage detectors, applying SA-Detect to YOLOv3 consistently enhanced performance. SA-Detect closely aligned with the labels in terms of box accuracy. The visualization of detection results before and after integrating SA-Detect into the model can be found in Fig. 5.

Table 2
Effectiveness of various designs. FA, CC, and SAL denote Feature Augmentation, Confidence Class, and Surrounding-to-Aggregating Learning, respectively.

SAL	CC	FA	*mAP@.5*	*P*	*Recall*	*F*1
			85.7	83.5	81.3	82.4
✓			88.5	88.6	85.2	86.9
✓	✓		89.1	89.0	86.6	87.8
✓	✓	✓	90.7	91.4	87.6	89.5

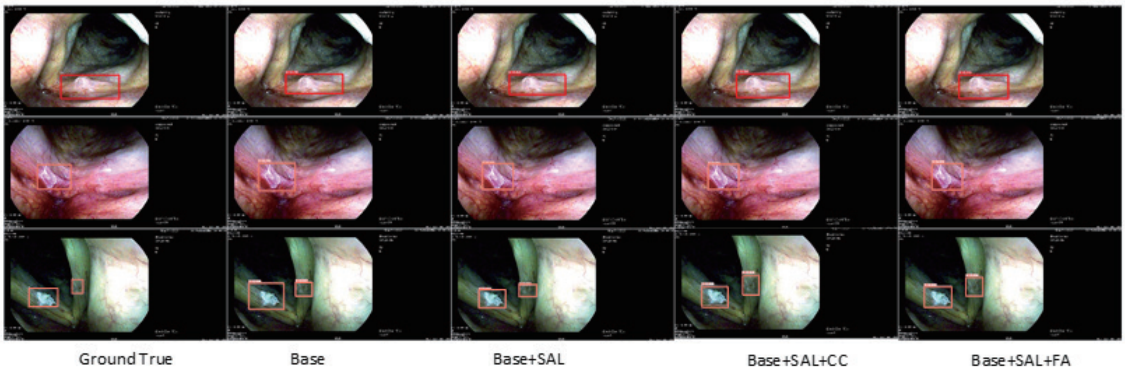

Fig. 4. (Color online) Comparison of qualitative results between FA, CC, and SAL based on the clinical dataset. 0, 1, 2, and 3 represent the chronic inflammation of mucosa or squamous hyperplasia, low-grade dysplasia, high-grade dysplasia, and invasive carcinoma, respectively.

Table 3
Performance of the proposed SA-Detect using different backbone networks.

Model	SA-Detect	*mAP@.5*	*P*	*Recall*
Faster R-CNN(37)		87.6	84.1	83.9
Mask R-CNN(38)		86.9	85.5	82.2
YOLOv3(39)		85.9	85	83.3
Faster R-CNN	✓	89.4	87.1	85.6
Mask R-CNN	✓	88.6	86.9	83.7
Yolov3	✓	88.0	85.6	84.2

4. Discussion

In this study, we developed a deep learning algorithm for detecting laryngeal lesions in laryngoscope images. In practical clinical settings, patients with suspicious lesions detected by conventional white-light endoscopy in outpatient offices are recommended to undergo enhanced endoscopy examinations, such as iSCAN or NBI endoscopy, to further screen potential lesions and predict the pathological type of lesions for treatment planning.(40) Physicians need to search for erythema, leukoplakia, and subtle changes in intraepithelial capillary loops during examination. The main goals of physicians are twofold: (1) to detect as many lesions as possible without missing any and (2) to accurately infer the pathological type of lesions. Achieving these

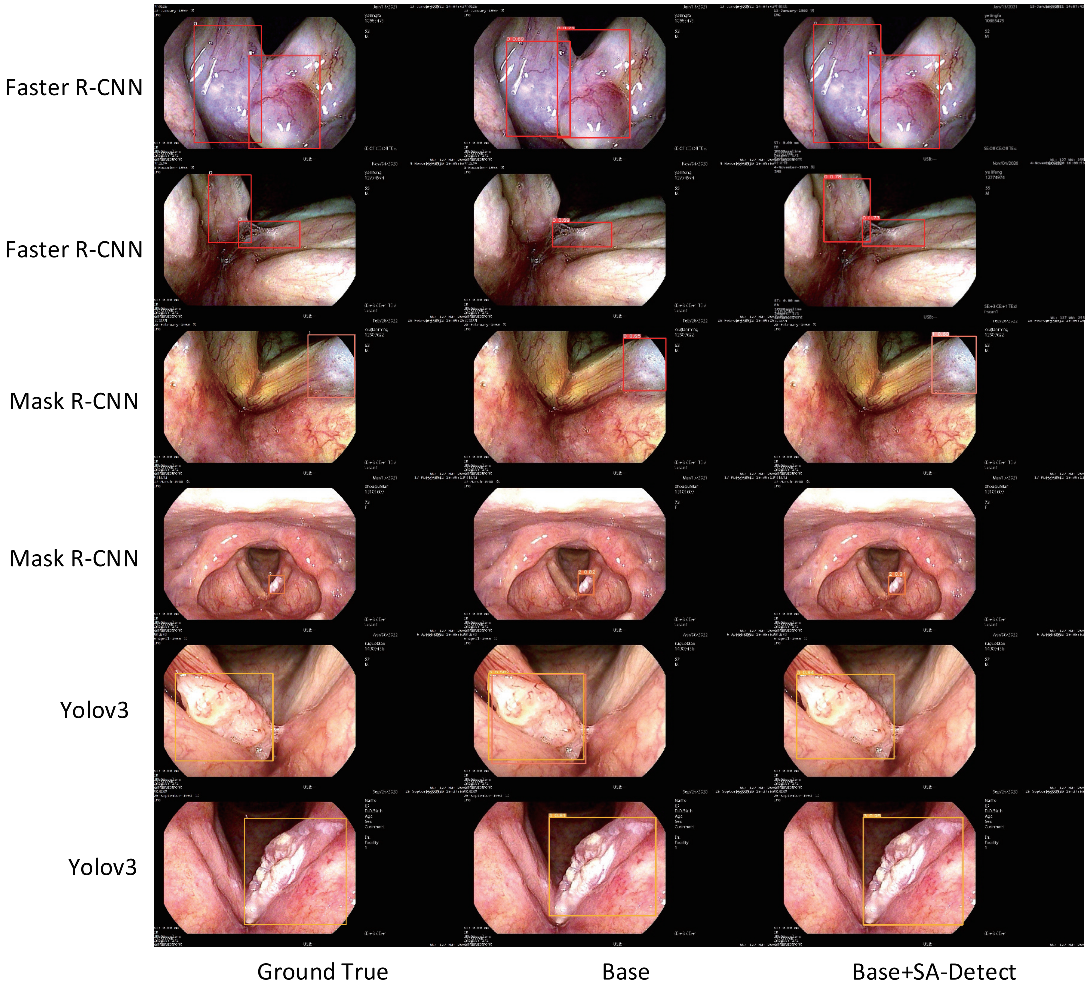

Fig. 5. (Color online) Visualizations of various backbone networks before and after implementing SA-Detect. 0, 1, 2, and 3 represent the chronic inflammation of mucosa or squamous hyperplasia, low-grade dysplasia, high-grade dysplasia, and invasive carcinoma respectively..

goals demands that physicians maintain sustained attention, sharp insight, and profound experience. However, meeting any of these requirements is challenging, especially for physicians in primary institutions with less experience. With the assistance of our proposed model, physicians can identify lesions that might have been easily missed in the past. Even physicians with limited experience in identifying intraepithelial capillary loops can correctly identify laryngeal cancer and pre-neoplastic lesions. This tool significantly alleviates the burden on physicians.

Despite previous research endeavors to develop artificial intelligence systems for detecting laryngoscope images, the diagnostic performance and processing speed have not met the requirements for real-time applications. One primary goal of this study is to explore the feasibility of SA-Detect in real-time video laryngoscope image diagnosis, addressing the practical needs of supporting medical professionals in their diagnoses. In practical terms, validating video streams is feasible. Given that laryngoscope examinations typically occur within a range of 2–30 ms, with an average time of 38 ms, the YOLO model can analyze a video frame in just 33 ms. The original YOLOv5 comprises 2.6 million parameters, whereas YOLOv5 with SA-Detect applied reduces this to 2 million parameters. Consequently, SA-Detect diminishes the computational complexity of the model, resulting in an acceleration of detection.

Moreover, we performed an analysis of the average number of positive boxes per image at various *IoU* thresholds (e.g., $IoU \geq 0.5$). SA-Detect consistently yielded a reduced number of positive boxes across all thresholds, especially for higher *IoU* thresholds. Note that AP is impacted not solely by localization accuracy but also by classification precision, and SA-Detect demands bounding boxes with high *IoU*. To augment *mAP* further, additional efforts should be focused on other facets.

In this investigation, we present an AI-assisted real-time laryngoscope detection approach that utilizes lateral feature learning. SA-Detect showcases its efficiency across diverse images and videos in different models. Our approach exhibits exceptional laryngoscope detection performance, showcasing superior accuracy, recall, and *mAP*@0.5. Technically, we highlight the importance of side-aware feature learning, specifically emphasizing boundary content for precise localization. We underscore the pivotal role of side-aware features in the realm of video image detection, particularly in endoscopy.

5. Conclusions

In this paper, we introduced side-aware boundary learning for real-time laryngoscope detection, aiming to narrow the gap between top-level and bottom-level feature layers for the efficient transmission of informative cues. Our learning approach, centered on side-aware features, placed a crucial emphasis on capturing boundary information aligned with ground truth labels. Notably, our proposed SA-Detect model demonstrated excellent performance in terms of accuracy, recall, and calculation time. It has the potential to provide real-time assistance in the diagnosis of laryngeal cancer and pre-neoplastic lesions, thereby reducing the likelihood of missed diagnoses. This tool enables primary institutions and junior physicians to diagnose and treat laryngeal cancer and pre-neoplastic lesions with expertise, leading to enhanced clinical outcomes for patients.

References

1 M. Falco, C. Tammaro, T. Takeuchi, A. M. Cossu, G. Scafuro, S. Zappavigna, A. Itro, R. Addeo, M. Scrima, A. Lombardi, F. Ricciardiello, C. Irace, M. Caraglia, and G. Misso: Cancers **14** (2022) 1716. https://doi.org/10.3390/cancers14071716

2 A. M. Cossu, L. Mosca, S. Zappavigna, G. Misso, M. Bocchetti, F. De Micco, L. Quagliuolo, M. Porcelli, M. Caraglia, and M. Boccellino: Int. J. Mol. Sci. **20** (2019) 3444. https://doi.org/10.3390/ijms20143444

3 G. Marioni, R. Marchese-Ragona, G. Cartei, F. Marchese, and A. Staffieri: Cancer Treatment Reviews **32** (2006) 504. https://doi.org/10.1016/j.ctrv.2006.07.002

4 M. A. Zwakenberg, G. B. Halmos, J. Wedman, B. F. van Der Laan, and B. E. Plaat: Laryngoscope **131** (2021) E2222. https://doi.org/10.1002/lary.29361

5 J. Chen, Z. Li, T. Wu, and X. Chen: Laryngoscope Investigative Otolaryngology **8** (2023) 508. https://doi.org/10.1002/lio2.1049

6 N. Mohamed, R. L. Almutairi, S. Abdelrahim, R. Alharbi, F. M. Alhomayani, B. M. Elamin Elnaim, A. A. Elhag, and R. Dhakal: Cancers **16** (2023) 181. https://doi.org/10.3390/cancers16010181

7 R. Gabbiadini, F. D'Amico, A. De Marco, M. Terrin, A. Zilli, F. Furfaro, M. Allocca, G. Fiorino, and S. Danese: J. Clin. Med. **11** (2022) 509. https://doi.org/10.3390/jcm11030509

8 H. Irjala, N. Matar, M. Remacle, and L. Georges: Eur. Archiv. Oto-Rhino-Laryngology **268** (2011) 801. https://doi.org/10.1007/s00405-011-1516-z

9 X. Ni, G. Wang, F. Hu, X. Xu, L. Xu, X. Liu, X. Chen, L. Liu, X. Ren, Y. Yang, L. Guo, Y. Gu, J. Hou, J. Zhang, Y. Yang, B. Xing, J. Ren, and H. Guo: Clin. Otolaryngology **44** (2019) 729. https://doi.org/10.1111/coa.13361

10 A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun: Nature **547** (2017). https://doi.org/10.1038/nature21056

11 B. Zhang, Z. Jin, and S. Zhang: Lancet Digital Health **3** (2021) e410. https://doi.org/10.1016/S2589-7500(21)00108-4

12 S. Foersch, M. Eckstein, D. C. Wagner, F. Gach, A. C. Woerl, J. Geiger, C. Glasner, S. Schelbert, S. Schulz, S. Porubsky, A. Kreft, A. Hartmann, A. Agaimy, and W. Roth: Annals of Oncology **32** (2021) 1178. https://doi.org/10.1016/j.annonc.2021.06.007

13 J. Yu, T. Ma, H. Chen, M. Lai, Z. Ju, and Y. Xu: IEEE Trans. Syst. Man Cybern.: Syst. **53** (2023) 7099. https://doi.org/10.1109/TSMC.2023.3290205

14 B. A. Tama, G. Kim, S. W. Kim, and S. Lee: Clin. Exp. Otorhinolaryngology **13** (2020) 326. https://doi.org/10.21053/ceo.2020.00654

15 G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and P. Baldi: Gastroenterology **155** (2018) 1069. https://doi.org/10.1053/j.gastro.2018.06.037

16 K. Kumar, P. Kumar, D. Deb, M. L. Unguresan, and V. Muresan: Healthcare **11** (2023) 207. https://doi.org/10.3390/healthcare11020207

17 J. Yu, T. Ma, Y. Fu, H. Chen, M. Lai, C. Zhuo, and Y. Xu: Comput. Med. Imaging Graphics **107** (2023) 102230. https://doi.org/10.1016/j.compmedimag.2023.102230

18 M. Żurek, K. Jasak, K. Niemczyk, and A. Rzepakowska: J. Clin. Med. **11** (2022) 2752. https://doi.org/10.3390/jcm11102752

19 M. Misawa, S. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda, and K. Mori: Gastrointestinal Endoscopy **93** (2021) 960. https://doi.org/10.1016/j.gie.2020.07.060

20 J. Jeong, E. M. Song, C. Ha, H. J. Lee, J. E. Koo, D. H. Yang, N. Kim, and J. S. Byeon: Sci. Rep. **10** (2020) 8379. https://doi.org/10.1038/s41598-020-65387-1

21 I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun: Comput. Biol. Med. **141** (2022) 105031. https://doi.org/10.1016/j.compbiomed.2021.105031

22 T. Eelbode, I. Demedts, R. Bisschops, P. Roelandt, C. Hassan, E. Coron, P. Bhandari, H. Neumann, O. Pech, A. Repici, and F. Maes: Gastrointestinal Endoscopy **89** (2019) AB618. https://doi.org/10.1016/j.gie.2019.03.1075

23 Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham: IEEE Access **6** (2018) 40950. https://doi.org/10.1109/ACCESS.2018.2856402

24 L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng: IEEE J. Biomed. Health. Inf. **21** (2016) 65. https://doi.org/10.1109/JBHI.2016.2637004

25 K. Nguyen, N. T. Huynh, P. C. Nguyen, K. D. Nguyen, N. D. Vo, and T. V. Nguyen: Electronics **9** (2020) 583. https://doi.org/10.3390/electronics9040583

26 S. Oh, A. Chang, A. Ashapure, J. Jung, N. Dube, M. Maeda, D. Gonzalez, and J. Landivar: Remote Sens. **12** (2020) 2981. https://doi.org/10.3390/rs12182981

27 C. B. Murthy, M. F. Hashmi, N. D. Bokde, and Z. W. Geem: Appl. Sci. **10** (2020) 3280. https://doi.org/10.3390/app10093280

28 Z. Xu, X. Xu, L. Wang, R. Yang, and F. Pu: Remote Sens. **9** (2017) 1312. https://doi.org/10.3390/rs9121312

29 J. Ding, J. Zhang, Z. Zhan, X. Tang, and X. Wang: Remote Sens. **14** (2020) 844. https://doi.org/10.3390/rs14030663

30 B. Jin, P. Liu, P. Wang, L. Shi, and J. Zhao: Entropy **22** (2020) 844. https://doi.org/10.3390/e22080844

31 L. Wei, C. Zheng, and Y. Hu: Remote Sens. **15** (2023) 1350. https://doi.org/10.3390/rs15051350

32 A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao: arXiv preprint **2004** (2020) 10934. https://doi.org/10.48550/arXiv.2004.10934

33 H. Law and J. Deng: arXiv preprint (2018) 734. https://doi.org/10.48550/arXiv.1808.01244

34 Z. Tian, C. Shen, H. Chen, and T. He: International Conference on Computer Vision (ICCV, 2019) 9627. https://doi.org/10.1109/ICCV.2019.00972

35 S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai: arXiv preprint **2203** (2022) 16250. https://doi.org/10.48550/arXiv.2203.16250

36 X. Yin, Z. Yu, Z. Fei, W. Lv, and X. Gao: Artificial Neural Networks and Machine Learning – ICANN 2023 (2023) 163. https://doi.org/10.1007/978-3-031-44195-0_14
37 R. Girshick: International Conference on Computer Vision (ICCV, 2015) 1440. https://doi.org/10.1109/ICCV.2015.169
38 K. He, G. Gkioxari, P. Dollár, and R. Girshick: International Conference on Computer Vision (ICCV, 2017) 2961. https://doi.org/10.1109/ICCV.2017.322
39 J. Redmon, and A. Farhadi: arXiv preprint **1804** (2018) 02767. https://doi.org/10.48550/arXiv.1804.02767
40 J. Galli, S. Settimi, D. A. Mele, A. Salvati, E. Schiavi, C. Parrilla, and G. Paludetti: J. Clin. Med. **10** (2021) 1224. https://doi.org/10.3390/jcm10061224

About the Authors

Huang Yangyiyi received his B.S. degree from the College of Life Sciences of Zhejiang University, China, and his M.D. degree from Zhejiang University School of Medicine, China. He completed his residency training at the Second Affiliated Hospital of Zhejiang University School of Medicine and has been serving as an attending physician at the same hospital and a secretary of the residency training in the department. His research interests are in the application of artificial intelligence in otolaryngology-head and neck surgery diagnosis and treatment, and the application of VR and 3D printing in medical teaching otolaryngology-head and neck surgery. (2317155@zju.edu.cn)]

Jinchao Ge received her master's degree in Software Engineering from Zhejiang University of Technology, Hangzhou, China, in 2020. She is currently pursuing her Ph.D. at the University of Adelaide, Australia. Her research interests include computer vision and deep learning. (jinchao.ge@adelaide.edu.au)

Weiming Fan holds a bachelor’s degree from Luoyang University of Technology, China, and a master’s degree from Shenyang Ligong University, China. He is currently a research assistant in Binjiang Institute of Zhejiang University, China. His research interests revolve around deep-learning-based detection methods in the field of medicine. (weiming@stu.sylu.edu.cn)

YiQun Zheng is currently a senior researcher of Hangzhou TUYA Information Technology Co., Ltd., focusing on AI and IoT solution. (eqin@tuya.com)

Changting Lin received his Ph.D. degree in computer science from Zhejiang University, China, in 2018. His is currently a researcher in Binjiang Institute of Zhejiang University, China. His research interests include AI and AI security. (linchangting@zju.edu.cn)