



Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO



Chiara Baldini ^{a,b}, Lucia Migliorelli ^{c,a}, Daniele Berardini ^c, Muhammad Adeel Azam ^{a,b}, Claudio Sampieri ^{d,e}, Alessandro Ioppi ^f, Rakesh Srivastava ^g, Giorgio Peretti ^{h,i}, Leonardo S. Mattos ^a

^a Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genova, Italy

^b Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genova, Italy

^c Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

^d Department of Experimental Medicine (DIMES), University of Genoa, Genova, Italy

^e Department of Otolaryngology, Hospital Clínic, Barcelona, Spain

^f Department of Otorhinolaryngology-Head and Neck Surgery, S. Chiara Hospital, Azienda Provinciale per i Servizi Sanitari (APSS), Trento, Italy

^g Sushrut Institute of Plastic Surgery & Super specialty Hospital, Lucknow, India

^h Unit of Otorhinolaryngology – Head and Neck Surgery, IRCCS Ospedale Policlinico San Martino, Genova, Italy

ⁱ Department of Surgical Sciences and Integrated Diagnostics (DISC), University of Genoa, Genova, Italy

ARTICLE INFO

Keywords:

Laryngeal lesions
Endoscopy
Deep learning
Super resolution
Real-time assistance
Decision support system

ABSTRACT

Background and Objective: Laryngeal Cancer (LC) constitutes approximately one third of head and neck cancers. Detecting early-stage lesions in this anatomical region is crucial for achieving a high survival rate. However, it poses significant diagnostic challenges owing to the varied appearance of lesions and the need for precise characterization for appropriate clinical management. Conventional diagnostic approaches rely heavily on endoscopic examination, which often requires expert interpretation and may be limited by subjective assessment. Deep learning (DL) approaches offer promising opportunities for automating lesion detection, but their efficacy in handling multi-modal imaging data and accurately localizing small lesions remains a subject of investigation. Furthermore, the clinical domain may largely benefit from the deployment of efficient DL methods that can ensure equitable access to advanced technologies, regardless of the availability of resources that can often be limited. In this study, a DL-based approach, named SRE-YOLO, was introduced to provide real-time assistance to less-experienced personnel during laryngeal assessment, by automatically detecting lesions at different scales from endoscopic White Light (WL) and Narrow-Band Imaging (NBI) images.

Methods: During the training, the SRE-YOLO integrates a YOLOv8 nano (YOLOv8n) baseline with a Super-Resolution (SR) branch to enhance lesion detection. This last component is decoupled during inference to preserve the low computational demand of the YOLOv8n baseline. The evaluation was conducted on a multi-center dataset, encompassing diverse laryngeal pathologies and acquisition modalities.

Results: The SRE-YOLO method improved the Average Precision ($AP_{@IoU=0.5}$) in lesion detection by 5% with respect to the YOLOv8n baseline, while maintaining the inference speed of 58.8 Frames Per Second (FPS). Comparative analyses against state-of-the-art DL methods highlighted the efficacy of the SRE-YOLO approach in balancing detection accuracy, computational efficiency, and real-time applicability.

Conclusions: This research underscores the potential of SRE-YOLO in developing efficient DL-driven decision support systems for real-time detection of laryngeal lesions at different scales from both WL and NBI endoscopic data.

1. Introduction

Head and Neck Cancer (HNC) refers to different malignant tumors that develop in or around the oral cavity, paranasal sinuses, pharynx, and larynx. It is the seventh most common cancer globally, accounting

for an estimated 890,000 new cases and 450,000 deaths per year [1]. Smoking, alcohol intake, and infection with sexually transmitted Human PapillomaVirus (HPV) have been identified as significant risk factors for HNC [2].

* Corresponding author at: Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genova, Italy.
E-mail address: chiara.baldini@iit.it (C. Baldini).

Laryngeal Cancer (LC) impacts the area around the vocal folds and constitutes about one third of HNCs [3]. Transnasal digital flexible endoscopes have proven to be highly effective for inspecting the upper aerodigestive tract, playing a pivotal role in the early identification of LC [4]. This outpatient procedure involves otolaryngologists using the endoscope – mainly in white light (WL) modality – which is inserted through the nose and directed towards the throat, to explore laryngeal tissues [5]. During the examination, when otolaryngologists detect suspect lesions, these are further investigated with a biopsy to determine if they are malignant. Should the biopsy confirm malignancy, a tailored treatment plan is devised for the patient.

Biopsy is a cost and time-demanding procedure, often requiring to be undertaken under general anesthesia in the operative room with related risks for the patient. Moreover, it is an often inefficient procedure, with a non-irrelevant rate of false negatives causing substantial delays in patients' treatment. For these reasons, the literature underscores the importance of adopting prevention-focused approaches to mitigate the need for such interventions [6]. Following this paradigm, enhanced optical techniques such as Narrow Band Imaging (NBI), which highlights the superficial blood vessel networks by filtering standard light into blue and green colors (415 nm and 540 nm, respectively), are used in addition to WL to improve the visualization and characterization of HNC tissues [4,7,8]. The use of NBI can indeed help clinicians have an enhanced awareness about the tissue histology. Therefore, especially for early-stage vocal fold malignancies, biopsy can be omitted sometimes.

However, in facilities that are not specialized in treating HNC, the full potential of these advanced imaging modalities may not be realized, owing to the requirement for clinicians with specialized experience to detect and assess lesions [9]. The need for substantial investment in training clinicians to use NBI systems for early lesion's detection hampers their widespread adoption [10]. Moreover, NBI still remains an operator-dependent technique constrained by intrinsic limitations of human beings regarding attention, visual inspection capabilities, and the capacity to simultaneously process vast amounts of information. In this scenario, computer-assisted systems based on Deep Learning (DL) may provide support to physicians and less-experienced personnel during both their training and in the procedural phase. In a recent publication, we have reviewed the potential impact of this technology in several settings of HNC endoscopy [11]. The explored methods for detecting laryngeal lesions have demonstrated potential, but are limited by their focus on only one type of imaging (WL or NBI). Furthermore, current approaches do not consider factors, such as different lesion sizes relative to the endoscope field of view, that affect the accuracy of detection. Indeed, the effectiveness of these DL methods increases when the lesion is close to the camera, suggesting the need for improvements to ensure reliable detection especially when the lesion is small.

To detect small object in images, researchers in related fields leverage complex DL methods alongside high-resolution input [12,13]. However, such methods need high computational power, which leads to considerable hardware expenses during their implementation. This may raise issues with respect to the deployment of such decision support systems in actual clinical practice [14].

Given the premises, in this study we propose a DL approach for laryngeal-lesion detection from frames acquired by flexible or rigid endoscopes in WL or NBI optical mode: the Super-Resolution Enhanced-You Only Look Once network (SRE-YOLO).

Inspired by the work in related fields of research [15,16], the proposed SRE-YOLO implements a Convolutional Neural Network (CNN) that integrates a decoupled Super Resolution (SR) branch with the goal of enhancing the detection capabilities, particularly for small-sized lesions. The SR branch is used during the training phase and is subsequently discarded during inference. This enhances lesion-detection accuracy without adding to the computational load of the detection architecture. The innovative contents of the proposed research are:

- The integration of a decoupled SR branch within the YOLOv8 nano (YOLOv8n) detector to enhance a clinical decision support system. This approach improves the CNN's ability particularly on small-sized laryngeal lesions without increasing its computational complexity during inference, and therefore assuring real-time performances for clinical integration.
- The proposed SRE-YOLO was trained and tested on a dataset collected from three different centers across various regions of the world. This dataset comprises WL and NBI frames from patients with either benign or malignant lesions, as well as healthy subjects. Images were acquired using both flexible and rigid endoscopes: flexible endoscopes are primarily used in ambulatory settings due to their ability to access difficult areas, while rigid endoscopes are favored in surgical contexts where high accuracy and stability are required. This diverse range of tools and imaging modalities enhances the training of the SRE-YOLO, enabling its effective application in varied clinical environments.

The remainder of this paper is structured as follows: Section 2 presents an overview of the existing research in computer-aided endoscopic systems for supporting LC detection. Section 3 provides information on the data and the proposed DL approach. Section 4 describes the experimental setup. Section 5 presents the achieved results, which are discussed in Section 6. Finally, Section 7 summarizes the findings of this work and presents the potential next steps needed to bring the technology towards clinical use.

2. Related work

In the investigation of LC-detection performance, [17] explored how Faster Region-based Convolutional Neural Network (R-CNN), Single Shot Detector (SSD), and YOLOv3 localize visible malignant tumors on a limited-sized dataset with only 54 WL endoscopic images. Building on previous methodologies, [18] introduced a ResNet-101-based strategy with an extensive dataset comprising 3223 images to detect vocal cords or epiglottis, regardless of their normal or benign state. Information on the endoscopic light associated with the images was not provided by the authors. Furthering this line of research, [22] employed a Faster R-CNN to detect LC instances and benign lesions (vocal polyps, cysts, nodules, leukoplakia, granuloma, and papilloma) within 2179 endoscopic images. This latter study expanded the data collection across six different hospitals, using different endoscopic systems, but keeping a focus on the WL optical modality. [20] proposed a Mask R-CNN with ResNet-101 as backbone to detect laryngeal masses using 1224 acquisitions in WL mode. Likewise, authors in [23] used YOLOv4 for the detection of benign lesions in 2183 WL images, while [24] proposed a DL approach that implemented YOLOv5 to detect carcinoma, normal tissues, and anomalies in larynx using 4488 WL frames. Similarly, authors in [19] implemented a RetinaNet to detect laryngopharyngeal cancer from 2400 NBI frames acquired in healthy controls and subjects with cancer.

Recognizing the constraint of using a single imaging modality, authors in [21] integrated for the first time both WL and NBI-based data and trained a YOLOv5 to detect laryngeal carcinoma in 657 frames.

Then, more recently, the comparison of four state-of-the-art detection models was presented in [25] by examining their ability to detect benign and suspicious malignancy lesions in 8172 videoframes acquired by means of a laryngoscope with stroboscopic light.

While all these studies demonstrated the effectiveness of the DL approaches in localizing laryngeal lesions, they predominantly used a single-imaging modality in their analyses. This narrow focus may limit the comprehensiveness of their findings. The use of NBI was recommended by [27,28] due to its advantage in identifying early-stage lesions and reducing the occurrence of unnecessary biopsies. However, the authors agreed that WL endoscopy is the most widely used diagnostic tool in the assessment of lesions of the larynx, as visual evaluation of vascular patterns in NBI images is challenging and highly

Table 1

A comprehensive overview of Deep Learning (DL)-based assistive support systems for detecting laryngeal lesions in endoscopic images.

Work	Objective	Limitations
[17]	Employing Faster R-CNN, SSD, YOLOv3 for tumor localization.	(i) Limited dataset size; (ii) WL frames only; (iii) Overlooking considerations on algorithm's efficiency. (iv) Not encompassing benign lesion detection.
[18]	Employing ResNet-101 to detect vocal cords or epiglottis states.	(i) Lack of information on the endoscopic light-type used; (ii) Overlooking considerations on algorithm's efficiency.
[19]	Employing RetinaNet to detect laryngopharyngeal cancer.	(i) NBI frames only; (ii) Overlooking considerations on algorithm's efficiency. (iii) Not encompassing benign lesion detection.
[20]	Employing Mask R-CNN and ResNet-101 to detect laryngeal masses.	(i) WL frames only; (ii) Overlooking considerations on algorithm's efficiency.
[21]	Integration of WL and NBI data using YOLOv5 on frames for carcinoma detection.	(i) Limited dataset size; (ii) Overlooking considerations on algorithm's efficiency. (iii) Not encompassing benign lesion detection.
[22]	Employing Faster R-CNN to detect LC and benign lesions.	(i) WL frames only; (ii) Overlooking considerations on algorithm's efficiency.
[23]	Employing YOLOv4 for detecting benign lesions.	(i) WL frames only; (ii) Overlooking considerations on algorithm's efficiency; (iii) Not encompassing malignant detection.
[24]	Employing YOLOv5 to detect carcinoma and other anomalies.	(i) WL frames only; (ii) Overlooking considerations on algorithm's efficiency
[25]	Comparing four models on frames with stroboscopic light for the detection of lesions.	(i) Focused on state-of-the-art DL models' comparison; (ii) Overlooking considerations on algorithm's efficiency.
[26]	Comparative analysis of multiple YOLO variants on high-definition WL video endoscope for nasopharyngeal carcinoma.	(i) WL frames only; (ii) Computational demands high; (iii) Challenges in detecting small lesions effectively. (iv) Not encompassing benign lesion detection.

dependent on the examiner's experience. With such a view, training the detector with both WL and NBI data types is paramount for a thorough assessment of the larynx. Furthermore, there is a gap in discussions surrounding the efficiency of the DL approaches. Nevertheless, to enable these DL-based decision support systems to gain broader acceptance and implementation while reducing the digital divide, it is crucial to equally prioritize both efficiency and effectiveness [29,30].

Following this paradigm, in [26] a comparative analysis was conducted for an affine task, namely the detection of nasopharyngeal carcinoma using a high-definition video endoscope in WL mode. The performance of SSD, Faster RCNN, YOLOv6 medium variant, YOLOv7, and YOLOv8 large variant (YOLOv8l) was compared. All the YOLO variants improved the ability in terms of inferred frames per second by at least 30 points. Among these, YOLOv8l, which is the most computational demanding, had superior performance. Lighter networks (which were evaluated based solely on the number of trainable parameters and inference speed), failed in predicting cancerous lesions particularly when the cancerous lesion covered a tiny portion of the entire endoscopic frame. Table 1 shows a summary of the state of the art.

Therefore, the domain presents two significant needs. The first involves the development and deployment of efficient DL approaches that align with the ethical principle of distributive justice in technology, particularly within the clinical domain [30]. This demands a commitment to ensuring equitable access to advanced technological solutions, especially in healthcare centers that may be resource-constrained. The second need focuses on the improvement of network architectures' predictive abilities, specifically their capacity to accurately detect small lesions within both NBI and WL frames.

To tackle the current technological needs, this work introduces the SRE-YOLO approach. This method uses a YOLOv8n baseline, enhanced with a SR branch.

Image SR is commonly used to enhance the detection of small objects by magnifying the image size prior to the application of object-detection methods [31]. Similar techniques involve Generative Adversarial Network (GAN)-based workflows combining image SR with object detection [32]. Although these strategies are effective in detecting small objects, their computational needs – due to the employment

of multiple DL architectures – raise concerns about their feasibility for widespread clinical use. To solve this issue, here the SR is removed during the inference phase. This approach has the potential to boost the detection efficacy particularly on small lesions while preserving the computational complexity – and the inference speed – of the YOLOv8n.

3. Material and methods

3.1. Dataset

The dataset used to train, test and validate our approach is a collection of datasets coming from 3 different clinical centers: ENDO-LC 1, Laryngoscope8, and ENDO-LC 2. ENDO-LC 1 was collected and annotated by our clinical partners at San Martino Hospital, University of Genova, in Genova, Italy.¹ Endoscopic videos were captured during the outpatient assessment using a high-definition video flexible naso pharynx laryngoscope² through a transnasal route or during intraoperative examinations with rigid endoscopes coupled with high-definition camera.³ The extraction of frames from recorded videos was conducted by our clinical partners, following the standard procedure used in clinical practice. Specifically, they consistently saved at least 5 relevant images in WL and 5 in NBI per patient, using a video player software. It counts 2037 images with the size of 1072 × 1920 or 1088 × 1440 from healthy subjects and patients with laryngeal lesions, of which 864 WL and 506 NBI were acquired with flexible endoscopes, while rigid tools captured 475 WL and 192 NBI frames. Data included malignant lesions, i.e., Squamous Cell Carcinoma (SCC), and benign lesions such as cysts, granuloma, leukoplakia, papilloma, polyps, and Reinke's edema.

¹ This study was conducted following the principles of the Declaration of Helsinki and the local Institutional Review Board approval was obtained (CER Liguria: 169/2022).

² HD Video Rhino-laryngoscope Olympus ENF-VH, Olympus Medical System Corporation, Tokyo, Japan

³ 0°, 30°, or 70° rigid telescopes with HD camera head connected to a Visera Elite CLV-S190 light source, Olympus Medical System Corporation, Tokyo, Japan

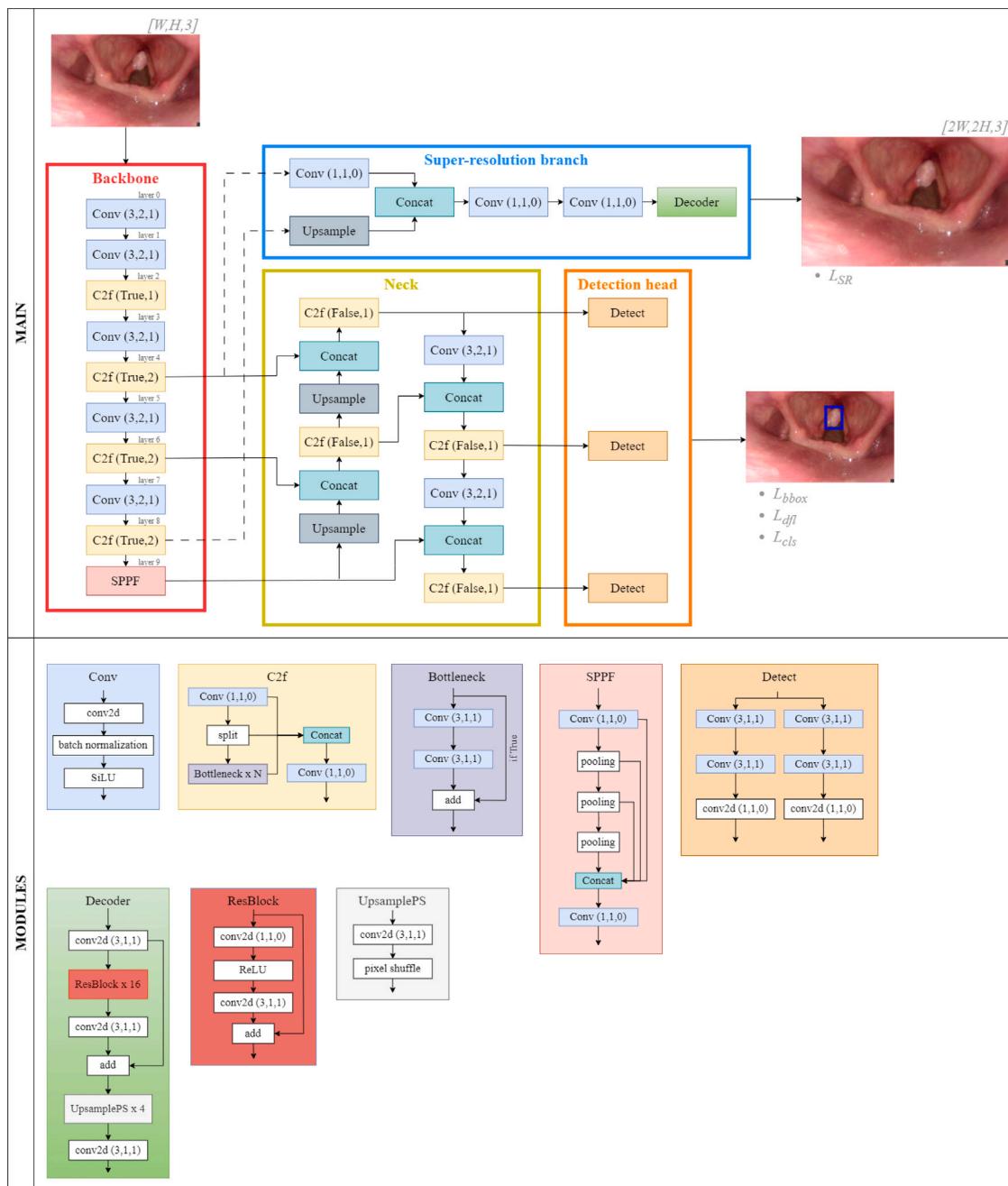


Fig. 1. Overview of the proposed SRE-YOLO: the ground-truth frame resized to $H \times W$ is given in input to the YOLOv8 backbone. Two feature maps at different scales are extracted from the 4th and 8th layers in the configuration 4–8 of the Super-Resolution (SR) branch, and used to reconstruct a $2H \times 2W$ image via an encoder-decoder-based process. In parallel, the YOLO head is trained to provide the bounding boxes of the identified lesions. Below, each module included in the proposed method is graphically described.

ENDO-LC 2 consists of 70 WL and 63 NBI frames with size 1080×1728 of both healthy patients and patients with abnormalities. A high-definition video flexible naso pharynx laryngoscope² was used for the acquisition of frames at the ENT Center, Voice & Airway Clinic of the Sushrut Institute of Plastic Surgery & Super Specialty Hospital, Lucknow, India.

The Laryngoscope8 dataset, which is publicly available⁴ [33], initially comprised 3057 WL laryngeal frames. These frames were categorized into eight groups based on either the normal appearance of the larynx or the presence of various conditions such as glottic cancer, Reinke's edema, granuloma, vocal cord leukoplakia, vocal cord

cysts, vocal cord nodules, and vocal cord polyps. The images were collected during head and neck surgery procedures at the Department of Otorhinolaryngology of the Sixth Medical Center of PLA General Hospital in Beijing, China. As the majority of Laryngoscope8 data was classified as normal larynx, we selected a subset of 1700 images for our study to keep the dataset balanced. It includes 239 frames obtained using flexible tools, with the remaining frames captured with rigid instruments.

To combine the datasets cohesively and ensure uniform standards, 5 laryngologists with more than 5 years of experience in the field, first conducted a pre-evaluation of all frames. This step is aimed at selecting frames with relevant information while excluding those with a bad quality. Exclusion criteria included underexposure, blurred frames, and the presence of saliva or specular reflections, as in our previous

⁴ <https://github.com/greenyin/Laryngoscope8.git>.

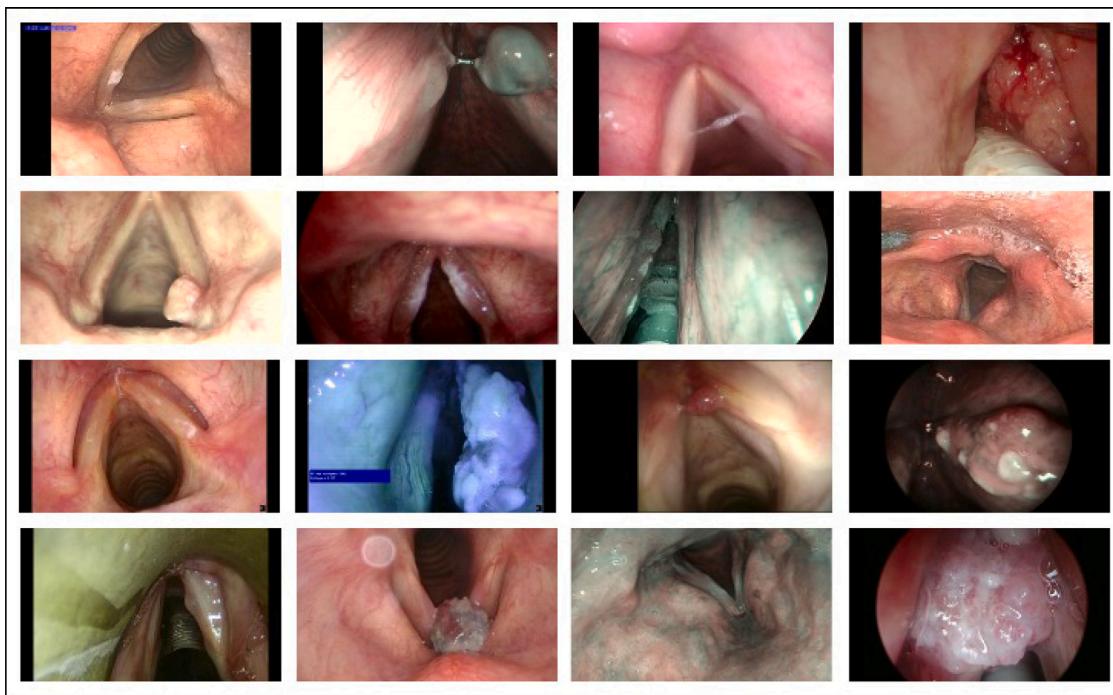


Fig. 2. Challenges inherent in the collected dataset include lesions of varying sizes (first row), images captured under different light intensities (second row), heterogeneous lesion types (third row), and several acquisition settings (last row). Additionally, some images exhibit occlusions caused by medical instruments or noise elements like light reflections or bleedings.

work [34]. Following this preliminary screening, they used the CVAT online tool⁵ to manually annotate each frame. This process involved drawing a bounding box around each identified abnormality on the vocal cords and labeling it as “lesion”. This class includes a range of abnormalities, from SCC to benign conditions such as cysts, granuloma, leukoplakia, papilloma, polyps, Reinke’s edema, and nodules. Fig. 2 shows the diversity of data collected from multiple centers, while also demonstrating the inherent challenges including: different lesions’ sizes, light conditions, and partial occlusion occurrences.

An external dataset, referred to as ENDO-LC ext, consisting of 149 frames in WL and NBI modalities, was collected and manually annotated by our clinical partners from the San Martino Hospital, University of Genova, Genova, Italy. This dataset includes both small lesions (bounding box area $< 32 \times 32$) and medium lesions ($32 \times 32 < \text{bounding box area} < 96 \times 96$) [35]. It is worth noting that the patients included in this dataset differ from those in the ENDO-LC 1, Laryngoscope8, and ENDO-LC 2.

3.2. Proposed SRE-YOLO

3.2.1. YOLOv8-nano baseline

The YOLOv8 model for object detection is available in several sizes, including nano, small, medium, large, and extra large. Each version differs in terms of network depth and width, and is optimized to offer a specific balance of speed, size, and accuracy. The YOLOv8n is the smallest version and highly prioritizes speed and efficiency over accuracy, making it especially suitable for real-time use in environments with limited resources. The architecture of YOLOv8n can be partitioned into three main components: the backbone, the neck, and the head.

The backbone is responsible for analyzing input images to extract both shallow and deep features. As in [36,37], it is based on the CSPDarknet53 architecture, which integrates Cross Stage Partial (CSP) connections [38] into the Darknet53 network. This network consists

of multiple convolutional CSP-sized stages (referred to as C2F stages), each composed of groups of residual bottleneck blocks. Between each stage, there is an interleaving 3×3 convolutional layer with a padding of 1 and a stride of 2. The configuration of these stages defines the main structure of YOLOv8, and the YOLOv8n comprises four stages with 1,2,2 and 1 bottleneck blocks, respectively. The backbone ends with the Spatial Pyramid Pooling Fast (SPPF) module, an advanced variant of the traditional SPP module [39]. SPPF enhances the YOLOv8n capability to detect objects of various sizes and shapes, by pooling the feature map at different scales and fusing them.

The neck of the YOLOv8n processes features from three different levels of the backbone within a Path Aggregation Network (PANet)-like structure [39] using C2F stages. This structure, aggregating high- and low-level features in a bidirectional way, enriches spatial details and significantly enhances the detection of smaller objects.

The head of YOLOv8n outputs detections across small, medium and large scales. Each scale features two 3×3 convolutional layers, followed by two parallel 1×1 convolutions for bounding box coordinates regression (i.e., x, y, width (W) and height (H)) and object classification into the *lesion* category. The YOLOv8n predictions are refined with the Non-Maximum Suppression (NMS) algorithm to retain only the most promising candidate bounding boxes.

3.2.2. Super-resolution branch

Detecting small objects in real-time applications remains a challenge. Here, we propose to address this problem with an additional high-resolution branch appended to the YOLOv8n structure. By taking as input feature maps obtained from YOLOv8n, the branch aims to reconstruct the super-resolution frame comparable with the ground-truth image. The SRE-YOLO exploits the feature maps provided by the convolutional layers of the backbone and applies two consecutive steps of encoding and decoding (Fig. 1). The output of the eighth block of the backbone is upsampled and then combined with the output from the fourth block of the same backbone. By passing through two convolutional operations with kernel and stride equal to 1 to reduce the dimensionality, the encoded feature maps are decoded via a *deep*

⁵ <https://www.cvat.ai/>.

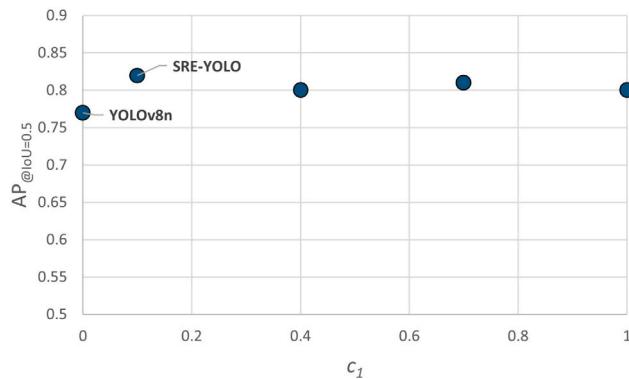


Fig. 3. Impact of different values of c_1 on the Average Precision ($AP_{@IoU=0.5}$) of SRE-YOLO. This parameter regulates the contribution of the Super Resolution (SR) reconstruction loss to the overall loss.

structure that evokes the Enhanced Deep Super-Resolution residual network (EDSR) [40]. The output of the additional SR branch ($X_{SRoutpu}$) corresponds with a high-resolution frame of size $2W \times 2H$ of the input given to the backbone, whose dimensions are $W \times H$. The similarity between the ground-truth frame (X) and $X_{SRoutpu}$ is computed in terms of L1 loss (Eq. (1)) and used as a learning component during the training phase.

$$L_{SR} = \|X - X_{SRoutpu}\|_1 \quad (1)$$

Taking into consideration the L_{bbox} , L_{dfl} , and L_{cls} loss terms implemented in [37], the modified loss for the SRE-YOLO model used during the training can be represented as:

$$Loss = c_1 \cdot L_{SR} + c_2 \cdot L_{bbox} + c_3 \cdot L_{dfl} + c_4 \cdot L_{cls} \quad (2)$$

where c_1 was set to 0.1 after a proper tuning (Fig. 3), and the YOLO default values of 7.5, 1.5, and 0.5 were chosen for c_2 , c_3 and c_4 , respectively.

4. Experimental protocol

4.1. Data preparation

As the data from the three datasets had different sizes, the images were resized to 1280×1280 and this size was used as ground truth for the SR branch ($2H \times 2W$, Fig. 1). In parallel, a low-resolution version of the frame was computed by downsampling it to 640×640 ($H \times W$, Fig. 1). This latter size serves as input to the baseline architectures.

Since several frames were extracted from the same patient, the split of the overall internal dataset was executed by following a patient-level strategy to prevent biases outcomes, resulting in a training, validation, and test ratio of 89:7:4 (Table 2).

On-the-fly training-data augmentation was applied to improve data variability, namely space, translation, scaling, horizontal flipping, and mosaic transformations were implemented following the work in [41].

To validate the SRE-YOLO architecture, we performed additional experiments using a hold-out training strategy, creating distinct splits for training, testing, and validation as shown in Table 3. For each hold-out split, we conducted further testing using the ENDO-LC ext dataset.

4.2. Training settings

The SRE-YOLO was implemented using PyTorch and all the experiments were conducted with a Dell Precision 7820 equipped with a 48 GB NVIDIA RTX A6000 GPU. After a comprehensive hyperparameter

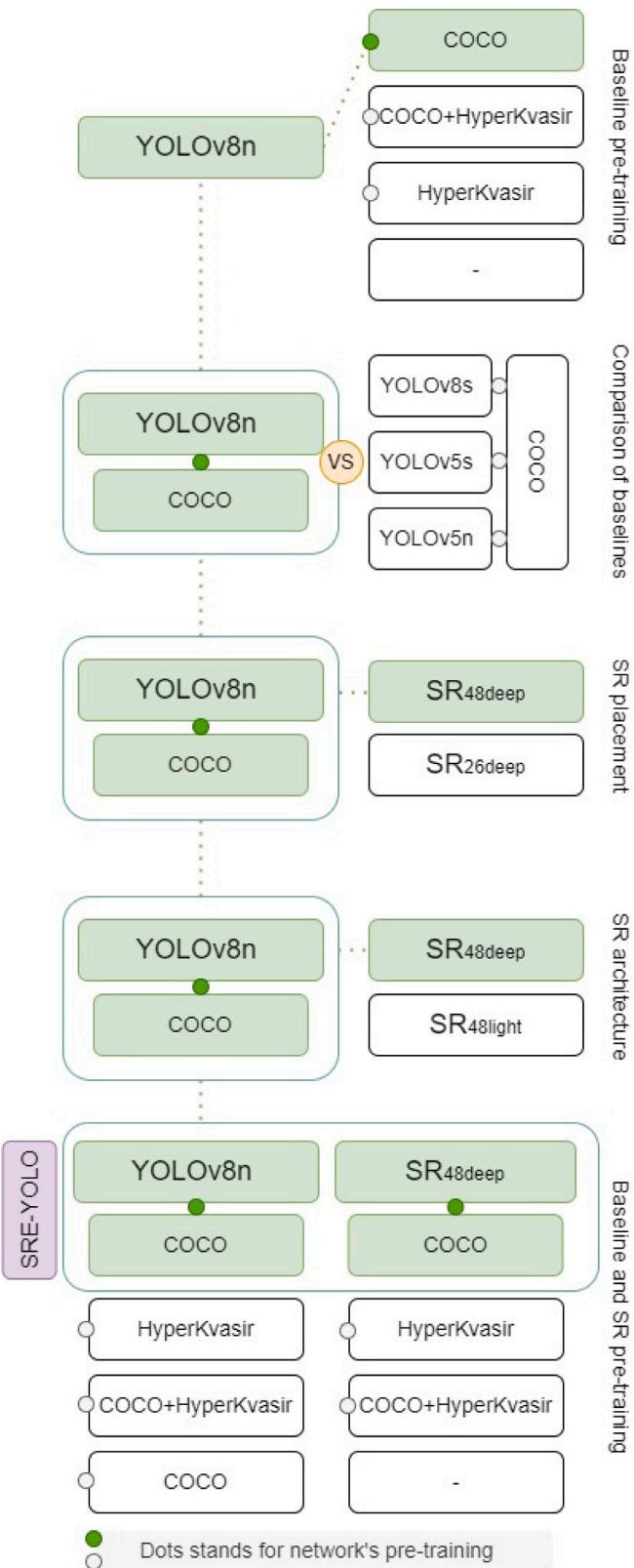


Fig. 4. Graphic illustration of the comparisons against state-of-the-art approaches and ablation studies implemented in this work. For each comparison study, the green color indicates the best model.

Table 2

Frames from the three different datasets were filtered and annotated by our clinical partners, subsequently organized into internal training, validation, and test sets.

Dataset	Optical modality	Number of patients	Number of frames
ENDO-LC 1	WL, NBI	233	2037
Laryngoscope8	WL	1297	1700
ENDO-LC 2	WL, NBI	63	155
Training set	WL, NBI	188 ENDO-LC 1; 1164 Laryngoscopes; 60 ENDO-LC 2	3452 (1816 lesion; 1636 healthy)
Validation set	WL, NBI	36 ENDO-LC 1; 47 Laryngoscopes; 3 ENDO-LC 2	269 (lesion)
Test set	WL, NBI	9 ENDO-LC 1; 86 Laryngoscope8	171 (lesion)

Table 3

Data and patient distribution of the three splits generated from the Random Hold-Out method. Notably, the training, validation, and test sets of **Table 2** are the same as those used in Split 1. The ENDO-LC ext dataset was used for external validation, both as part of the hold-out strategy and to assess the impact of incorporating the SR branch into the baseline model.

Split	Optical modality	Number of patients	Number of frames
Split1 (training)	WL, NBI	188 ENDO-LC 1; 1164 Laryngoscopes; 60 ENDO-LC 2	3452 (1816 lesion; 1636 healthy)
Split1 (validation)	WL, NBI	36 ENDO-LC 1; 47 Laryngoscopes; 3 ENDO-LC 2	269 (lesion)
Split1 (test)	WL, NBI	9 ENDO-LC 1; 86 Laryngoscope8	171 (lesion)
Split2 (training)	WL, NBI	188 ENDO-LC 1; 1164 Laryngoscopes; 60 ENDO-LC 2	3452 (1816 lesion; 1636 healthy)
Split2 (validation)	WL, NBI	36 ENDO-LC 1; 46 Laryngoscopes; 3 ENDO-LC 2	269 (lesion)
Split2 (test)	WL, NBI	9 ENDO-LC 1; 87 Laryngoscope8	171 (lesion)
Split3 (training)	WL, NBI	188 ENDO-LC 1; 1164 Laryngoscopes; 60 ENDO-LC 2	3452 (1816 lesion; 1636 healthy)
Split3 (validation)	WL, NBI	30 ENDO-LC 1; 97 Laryngoscopes; 3 ENDO-LC 2	269 (lesion)
Split3 (test)	WL, NBI	15 ENDO-LC 1; 36 Laryngoscope8	171 (lesion)
ENDO-LC ext	WL, NBI	73	149 (lesion)

tuning, we opted for AdamW optimizer with an initial learning rate (r_0) equal to 0.05, and a minimal learning rate of $0.2 \cdot r_0$.⁶

We trained the SRE-YOLO for 100 epochs using a batch size of 16. However, early stopping after 50 epochs without improvements to the validation metrics was implemented to prevent overfitting.

We applied pre-training strategies by initially training the SRE-YOLO on the COCO dataset. As the image resolution of the COCO dataset is 640×480 , the same data preprocessing described in Section 4.1 was used to obtain the ground truth for the SR component of the model.

4.3. Performance metrics

To assess the model's performance, we calculated the Average Precision (*AP*) using an Intersection over Union (*IoU*) threshold of 0.50. *AP* is calculated using the Area Under the Curve (*AUC*) of the *Precision* (*Prec*) \times *Recall* (*Rec*) plot. The formulas for calculating *Prec* and *Rec* are as follows:

$$Prec = \frac{TP}{TP + FP} \quad (3)$$

$$Rec = \frac{TP}{TP + FN} \quad (4)$$

TP represents the correct detections made by the DL approach. This was computed by considering the *IoU* between the ground-truth bounding box and the detected one. Specifically, if the *IoU* exceeds the threshold and the class confidence score exceeds 0.001, the detection is considered correct [41]. False Negative (*FN*) indicates ground-truth lesions not detected by the DL approach, while False Positive (*FP*), on the other hand, occurs when the model incorrectly identifies a lesion that is not present in the ground truth.

We measured the inference speed in terms of Frames Per Second (FPS) and the number of billion floating point operations (GFLOPs) to compare the SRE-YOLO real-time applicability and computational complexity with the other tested DL approaches. We carried out inference speed tests using the same hardware described in Section 4.2.

⁶ The codes will be made available upon request by contacting the corresponding author.

4.4. Ablation studies and comparison with other architectures

Comparisons against state of the art were conducted to evaluate the efficacy of the proposed DL approach in improving lesion-detection accuracy while avoiding any escalation in computational load. Alongside these investigations, we conducted ablation studies as illustrated in Fig. 4. For all the experiments conducted, we used the training settings described in Section 4.2.

4.4.1. Pre-training of the baseline

The first experiment dealt with the pre-training of the YOLOv8n baseline. Indeed, while the COCO dataset provides a rich variety of visual features and scenarios, it lacks specific representations of medical contexts, particularly in the narrow and homogeneous domains of endoscopic images. Consequently, we augmented our pre-training process by including the HyperKvasir dataset⁷ [42], which has gastrointestinal examinations that closely mimic the textural and color patterns found in laryngeal endoscopy.

The data of the HyperKvasir were collected using standard endoscopy equipment from Olympus (Olympus Europe, Germany) and Pentax (Pentax Medical Europe, Germany) at the Department of Gastroenterology, Bærum Hospital, Vestre Viken Hospital Trust, Norway. Despite the large size of the dataset, it includes corresponding bounding boxes (stored in JSON files) only for 1000 JPEG compressed images from the polyp class. Frame sizes are varied, ranging from 332×487 to 1920×1072 pixels.

Despite the datasets limitations, this inter-domain pre-training approach can potentially enhance the ability of the DL model to generalize across similar medical imaging tasks [43]. The pre-training exposes the model to domain-specific variations, thereby reducing overfitting to non-medical imaging features that are present in the COCO dataset.

Therefore, we tested the performance of the YOLOv8n by (i) initializing its weights using the standard Kaiming initialization (regarding this, we used the symbol “-” in Fig. 4 and in the subsequent tables), (ii) pre-training on the COCO dataset, (iii) pre-training on the HyperKvasir dataset and (iv) pre-training on the COCO and fine-tuning on the HyperKvasir (COCO + HyperKvasir).

⁷ <https://github.com/simula/hyper-kvasir>.

4.4.2. Comparison of baselines

We conducted a comparison of baselines by evaluating different baseline models with varying complexities to determine their effectiveness. These models included the baseline YOLOv8n and YOLOv8-small (YOLOv8s). We also assessed the performance of YOLOv5-small (YOLOv5-s), which was used in [24], and YOLO-v5-nano (YOLOv5n). Medium or larger versions were excluded from this investigation due to the requirements of maintaining the complexity as limited as possible while ensuring real-time performance. Indeed, as highlighted in [26] medium and larger versions are characterized by a greater number of GFLOPs and lower inference speeds.

4.4.3. Super resolution placement

To investigate the influence of feature-map depth from the shared backbone on detection performance, we compared the performance of the *configuration 4–8* with that of the *configuration 2–6*. Here, *configuration 4–8* refers to YOLOv8n+SR_{48deep}, where the SR branch processed feature maps from the fourth and eighth backbone layers. Likewise, *configuration 2–6* refers to YOLOv8n+SR_{26deep}, where the SR branch processed feature maps from the second and sixth backbone layers.

4.4.4. Super resolution architecture

We explored the configuration of the SR branch. The *deep* variant of the SR branch was compared against its *light* akin. In the latter, the encoder comprised two 1×1 convolutional layers that supplemented the backbone. The decoding process had three transposed convolutional layers, each with a kernel size of 4×4 and a stride of 2. The main difference between the *deep* and *light* structures of the decoder dealt with the respective presence or absence of the multiple residual blocks, and the upsampling methodology used. EDSR in the *deep* version used multiple residual blocks with short connections to make the optimization easier and removed batch normalization to retain the range flexibility and reduce memory usage. Here, upsampling was performed via pixel shuffle operations, i.e., efficient sub-pixel convolution layers which learnt an array of filters to upscale the low-resolution feature maps into the high-resolution output. Thus, we compared the results of the YOLOv8n+SR_{48deep} with those of YOLOv8n+SR_{48light} with both baselines pre-trained on COCO dataset.

4.4.5. Baseline and super-resolution branch pre-training

In the final experiment, both the SR branch and the YOLOv8n baseline underwent studies to assess the best pre-training configuration. Specifically, the configurations tested included: (i) both architectures pre-trained on the COCO dataset, proposed as SRE-YOLO; (ii) both architectures pre-trained on COCO and subsequently fine-tuned on the HyperKvasir dataset; (iii) both architectures exclusively pre-trained on the HyperKvasir dataset; (iv) YOLOv8n baseline pre-trained on COCO while the SR architecture with weights initially set according to the Kaming initialization standard. Similar to the COCO dataset, the HyperKvasir dataset was also preprocessed by resizing the frames as described in Section 4.1 to prepare the ground truth for pre-training the SR branch.

For point (ii), we followed two different strategies, since the HyperKvasir dataset available online includes a labeled portion of the data for segmentation and detection purposes (1000 frames) and a larger collection of unlabeled frames. Starting from the model's pre-trained on COCO (both the baseline and the SR), we compared the effect of the pre-training obtained using only the labeled portion or both the unlabeled – in a self-supervised manner – and labeled data from the HyperKvasir dataset. To this specific end, (ii-a) we initialized the baseline and the SR branch with COCO weights, and then we trained the SR branch – due to the lack of ground truth for the YOLO detector training – with 10,000 frames from the HyperKvasir_{unlabeled} portion. This pre-training phase updates the SR branch weights using the unlabeled data in a self-supervised manner. The (supervised) detection-related loss terms are set to zero, as they require bounding-box annotations that

Table 4

Results from the YOLOv8n baseline pre-training. Efficacy performance was assessed via the Average Precision (AP) computed using an Intersection over Union (IoU) threshold of 0.50 ($AP_{@IoU=0.5}$).

Model	Pre-training baseline	$AP_{@IoU=0.5}$
YOLOv8n	–	0.71
YOLOv8n	HyperKvasir	0.72
YOLOv8n	COCO+HyperKvasir	0.75
YOLOv8n	COCO	0.77

Table 5

Results from the baselines – namely, YOLOv8n, YOLOv8s, YOLOv5n, YOLOv5s – comparison. Efficacy performance was assessed via the Average Precision (AP) computed using an Intersection over Union (IoU) threshold of 0.50 ($AP_{@IoU=0.5}$). Efficiency performance was assessed via the billion floating point operations (GFLOPs) and the Frames Per Second (FPS).

Model	Pre-training baseline	$AP_{@IoU=0.5}$	GFLOPs	FPS
YOLOv5s	COCO	0.71 (−6%)	15.8	46.7
YOLOv5n	COCO	0.70 (−7%)	4.1	64.9
YOLOv8s	COCO	0.77	28.6	37.9
YOLOv8n	COCO	0.77	8.2	58.8

are unavailable. Consequently, only the SR loss is activated. Following this, (ii-b) we fine-tuned the entire model using only the labeled subset of the HyperKvasir dataset (HyperKvasir_{labeled}). This fine-tuning phase, built on the model pre-trained with unlabeled data, uses labeled data to enhance lesion performance through supervised learning (HyperKvasir_{unlabeled+labeled}).

5. Results

5.1. Ablation studies and comparison with other architectures

5.1.1. Pre-training of the baseline

Table 4 summarizes the results obtained by YOLOv8n baseline with a different pre-training. The table specifically shows how YOLOv8n performs on the test set under different scenarios: (i) with weights initialized using standard Kaiming initialization, (ii) after pre-training on the HyperKvasir dataset, (iii) after pre-training on COCO followed by fine-tuning on the HyperKvasir dataset, and (iv) after pre-training only on the COCO dataset. The YOLOv8n pre-trained only on the COCO dataset overcomes the others, giving the highest $AP_{@IoU=0.5}$ of 0.77.

5.1.2. Comparison of baselines

Table 5 shows the results from the comparison of baselines (YOLOv8n, YOLOv8s, YOLOv5n, YOLOv5s). The YOLOv8n outperforms YOLOv5s and YOLOv5n. Indeed, the $AP_{@IoU=0.5}$ is reduced by 7% for YOLOv5s and 6% for YOLOv5n when compared to YOLOv8n. However, YOLOv8n and YOLOv8s achieve the same $AP_{@IoU=0.5}$ of 0.77.

When assessing the performance efficiency, YOLOv8n and YOLOv5n have the lowest GFLOPs (equal to 8.2 and 4.1, respectively) and highest FPS (equal to 58.8 and 64.9, respectively) with respect to YOLOv5s (GFLOPs = 15.8, FPS = 46.7) and YOLOv8s (GFLOPs = 28.6, FPS = 37.9). Thus, when seeking the optimal trade-off between efficiency and effectiveness, YOLOv8n emerged as the optimal choice.

5.1.3. Super resolution placement

Table 6 compares the performance of the SR branch in its *configuration 4–8* (YOLOv8n+SR_{48deep}) with that of the *configuration 2–6* (YOLOv8n+SR_{26deep}). The architectures differ in the feature maps in input from the backbone. YOLOv8n+SR_{48deep} outperforms its akin of 2 percentage points achieving a $AP_{@IoU=0.5}$ equal to 0.82.

5.1.4. Super resolution architecture

In **Table 7**, the results from the *deep* variant of the SR branch were compared against those of the *light* one. The architectures differ

in encoder-decoder structure. The YOLOv8n+SR_{48deep} outperforms in 2 percentage points the performance of YOLOv8n+SR_{48light}, which achieves $AP_{@IoU=0.5}$ equal to 0.80.

5.1.5. Baseline and super-resolution branch pre-training

Table 8 displays the outcomes of assessing the most effective pre-training configuration for both the baseline (YOLOv8n) and the SR branch (SR_{48deep}). The lowest performance is achieved when both the baseline and SR are pre-trained on the HyperKvasir ($AP_{@IoU=0.5} = 0.72$). The performance of pre-trained architectures on COCO and fine-tuned on HyperKvasir are slightly better, with an $AP_{@IoU=0.5}$ equal to 0.80 when only the labeled portion of this dataset was considered, and $AP_{@IoU=0.5}$ equal to 0.79 when the model was fine-tuned first on the unlabeled HyperKvasir frames and then on those provided by the official repository with the corresponding labels. Pre-training the baseline only on the COCO dataset enhances performance, achieving an $AP_{@IoU=0.5}$ of 0.82, regardless of the pre-training applied to the SR branch.

5.2. Validation on the external dataset and hold-out strategy

We assessed the enhancement in performance and its generalization capabilities of the baseline YOLOv8n model through integration of the SR branch on the ENDO-LC ext set. The outcomes (**Table 9**) compare the performance of the SRE-YOLO, YOLOv8n+SR_{48light}, and YOLOv8n+

SR_{26deep} models against their baseline YOLOv8n. The ENDO-LC ext set, which includes data from patients not part of ENDO-LC 1, Laryngoscope8, or ENDO-LC 2, revealed that SRE-YOLO enhances mainly the small-sized lesion detection (+15% of $AP_{@IoU=0.5}$), overcoming the baseline even on external frames. The other models also demonstrated improved detection of small lesions relative to the baseline, while gains for medium-sized lesions were modest.

We conducted additional experiments comparing the performance of the proposed SRE-YOLO with its baseline, YOLOv8n. We employed the Random Hold-Out method, generating multiple random splits of the data and performing three separate tests, each with unique training, validation, and test sets (while preserving the patient-level split procedure). The results of this validation are presented in **Table 10**; as visible for all the 3 splits the proposed method (mean $AP_{@IoU=0.5} = 0.82$) exceeds the baseline (mean $AP_{@IoU=0.5} = 0.79$). Moreover, we evaluated the performance of each split on the ENDO-LC ext dataset. The test showed that the proposed SRE-YOLO outperformed the baseline YOLOv8n, achieving a mean $AP_{@IoU=0.5}$ across the splits of 0.86 compared to the baseline's mean $AP_{@IoU=0.5}$ of 0.80.

5.3. Efficiency metrics and qualitative results

Fig. 5 displays a comparison of performances between the proposed SRE-YOLO and baseline models, including YOLOv5n, YOLOv5s, YOLOv8s, and YOLOv8n. The two scatterplots illustrate the relation between the $AP_{@IoU=0.5}$ and efficiency-related metrics, i.e., FPS (upper plot) and GFLOPs (lower plot). The SRE-YOLO (purple circle), achieves an $AP_{@IoU=0.5}$ equal to 0.82, it operates at 58.8 FPS (as shown in the upper plot), and requires 8.2 GFLOPs (as reported in the lower plot). Such results are compared with those of YOLOv8n and YOLOv8-s (blue circle and triangle, respectively), and of YOLOv5n and YOLOv5s (orange circle and triangle, respectively). The figure shows that the SRE-YOLO architecture outperforms its counterparts, particularly the YOLOv8n (i.e., SRE-YOLO baseline), in terms of quantitative metrics, while keeping the same effectiveness of YOLOv8n.

Fig. 6 presents a qualitative comparison of the performance of the SRE-YOLO and YOLOv8n in detecting laryngeal lesions of varying apparent sizes, using frames obtained with different tools and optical modalities.

We also investigated in **Fig. 7** the performance of the proposed SRE-YOLO against its baseline on frames with small lesions, frames with medium lesions, and frames with large lesions (bounding box area $> 96 \times 96$) considering the same object size categorization proposed in [35].

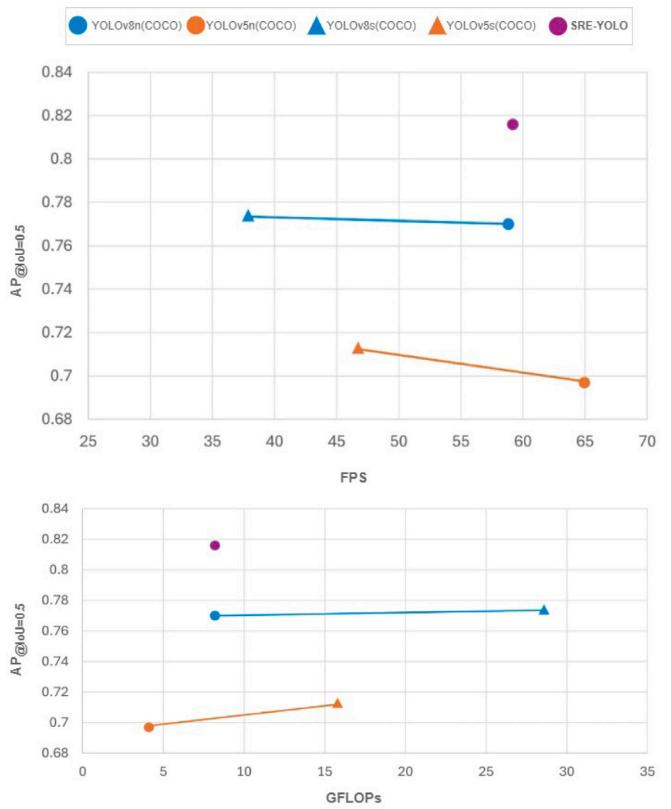


Fig. 5. Scatter plots of the efficacy-efficiency performance from the baselines comparison (i.e., YOLOv5n, YOLOv5s, YOLOv8s, YOLOv8n) against our proposed SRE-YOLO. The plot above shows the Frame Per Second (FPS) against the Average Precision (AP) computed using an Intersection over Union (IoU) threshold of 0.50 ($AP_{@IoU=0.5}$). The graph below presents the billion floating point operations (GFLOPs) against the $AP_{@IoU=0.5}$.

Table 6

Results from different placement of the Super Resolution (SR) branch. We compared the performance of the *configuration 4–8* (YOLOv8n+SR_{48deep}) with that of the *configuration 2–6* (YOLOv8n+SR_{26deep}). In the first architecture, the SR branch processes feature maps from the fourth and eighth layers of the YOLOv8n backbone, whereas in the second architecture, it processes those from the second and sixth layers. We assessed the performance via the Average Precision (AP) computed using an Intersection over Union (IoU) threshold of 0.50 ($AP_{@IoU=0.5}$).

Model	Pre-training baseline	AP _{@IoU=0.5}
YOLOv8n+SR _{26deep}	COCO	0.80
YOLOv8n+SR _{48deep}	COCO	0.82

Table 7

Results from the comparison of the Super-Resolution (SR) branch architecture. We tested the performance of the proposed *deep* (YOLOv8n+SR_{48deep}) version with that of the *light* one (YOLOv8n+SR_{48light}). We assessed the performance via the Average Precision (AP) computed using an Intersection over Union (IoU) threshold of 0.50 ($AP_{@IoU=0.5}$).

Model	Pre-training baseline	AP _{@IoU=0.5}
YOLOv8n+SR _{48light}	COCO	0.80
YOLOv8n+SR _{48deep}	COCO	0.82

6. Discussion

Endoscopic examination represents a crucial tool for the detection and assessment of vocal fold lesions. However, the diagnostic accuracy of this evaluation is strictly influenced by the laryngologists' level of expertise and effort requested to analyze the endoscopic videos, which can lead to missed detections and misdiagnoses. Advanced optical

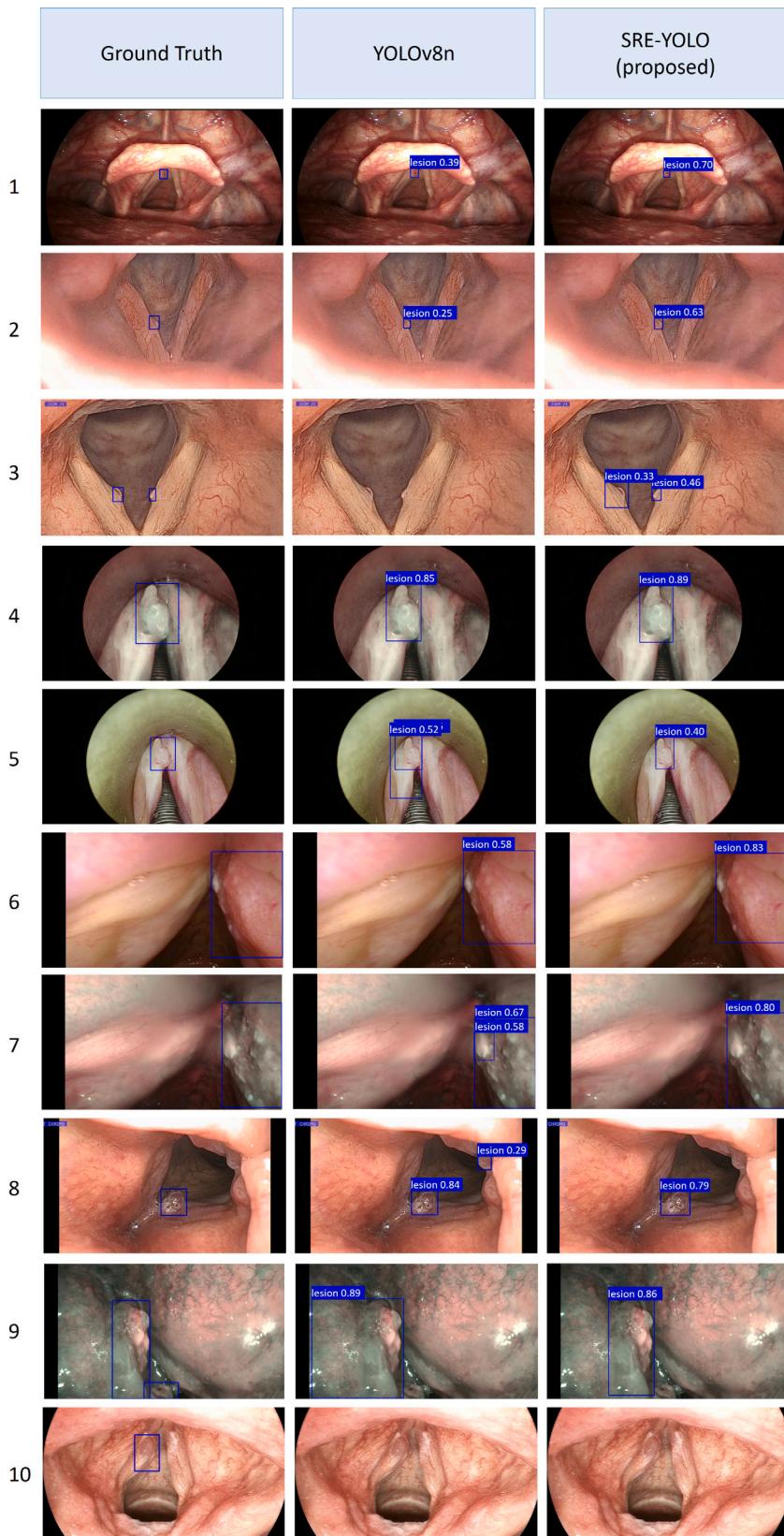


Fig. 6. Samples for qualitative results. Row 1–2: Left and right vocal fold polyps. Row 3: Bilateral vocal fold polyps. Row 4–5: Intraoperative Narrow-Band Imaging (NBI) of a left vocal fold polyp and the corresponding White Light (WL)-imaging view. Row 6–7: Squamous Cell Carcinoma (SCC) captured with both WL and NBI modalities. Row 8: Left vocal fold polyp. Row 9: NBI view of a SCC arising from the left laryngeal vestibular fold. Row 10: Model failure case in detecting a left vocal fold polyp. The second and third columns illustrate the detection performance of the YOLOv8n (i.e., the baseline) and the proposed SRE-YOLO.

Table 8

Analysis of the pre-training effect on detection performance. The subscripts *labeled* and *unlabeled* were included to improve readability and indicate whether the Super-Resolution (SR) branch pre-training was conducted (i) on the labeled portion or (ii) on the unlabeled and subsequently labeled portion of the HyperKvasir dataset.

Model	Pre-training baseline	Pre-training SR	AP _{@IoU=0.5}
YOLOv8n+SR _{48deep}	HyperKvasir _{labeled}	HyperKvasir _{labeled}	0.72
YOLOv8n+SR _{48deep}	COCO	COCO+HyperKvasir _{unlabeled+labeled}	0.79
YOLOv8n+SR _{48deep}	COCO+HyperKvasir _{labeled}	COCO+HyperKvasir _{labeled}	0.80
YOLOv8n+SR _{48deep}	COCO	-	0.82
SRE-YOLO	COCO	coco	0.82

Table 9

Performance evaluation in terms of the Average Precision (*AP*) computed using an Intersection over Union (*IoU*) threshold of 0.50 (*AP_{@IoU=0.5}*) for the proposed SRE-YOLO, YOLOv8n+SR_{48light}, and YOLOv8n+SR_{26deep} models on the ENDO-LC ext test set, divided into small (bounding box area < 32 × 32) and medium-sized (32 × 32 < bounding box area < 96 × 96) lesions.

Model	Small lesions	Medium lesions
YOLOv8n	0.66	0.80
YOLOv8n+SR _{48light}	0.82	0.77
YOLOv8n+SR _{26deep}	0.74	0.82
SRE-YOLO	0.80	0.82

Table 10

Performance evaluation (in terms of the Average Precision (*AP*) computed using an Intersection over Union (*IoU*) threshold of 0.50 (*AP_{@IoU=0.5}*)) of the baseline YOLOv8n and the SRE-YOLO architecture (**bold**) across different training and testing splits, including results on the external ENDO-LC dataset for small and medium lesions.

Train ↓ Test →	Split 1	Split 2	Split 3	ENDO-LC ext small lesions	ENDO-LC ext medium lesions
Split 1	0.77	-	-	0.66	0.80
	0.82			0.80	0.82
Split 2	-	0.81	-	0.82	0.82
		0.83		0.95	0.85
Split 3	-	-	0.79	0.86	0.85
			0.82	0.88	0.84

technologies, such as the NBI, have enhanced the detection of early-stage LC, but they still present challenges such as the need for extensive training and experience to result in a successful outcome.

DL algorithms may be employed to automatically identify lesions in endoscopic frames and assist less-experienced clinicians during examinations. However, current state-of-the-art methods tend to be ineffective for small lesions. In addition, they lack validation across different imaging modalities, and often overlook computational efficiency, which is a critical aspect when considering the integration of the technology into clinical practice.

In light of the above, this work proposed the SRE-YOLO, an architecture that integrates a decoupled SR branch within the YOLOv8n baseline. The SR branch, which is discarded during testing, improves architecture's ability to detect small lesions by helping the network fine-tune its parameters towards a more detailed feature extraction, while keeping the same computational complexity of the baseline.

Observing the results in both Table 5 and Fig. 5, YOLOv8n pre-trained on COCO emerged as the baseline model that detains the best trade-off between efficiency and efficacy. Indeed, it achieves the same performance as YOLOv8s (in terms of *AP_{@IoU=0.5}*) by reducing the number of GFLOPs and FPS. These quantitative results obtained from YOLOv8n and YOLOv8s, apparently similar despite the greater complexity of the second model, may depend on the fact that the dataset on which we are testing the approaches is very limited in size to appreciate improvements in *AP_{@IoU=0.5}* of percentage units.

Additionally, YOLOv8n demonstrated enhanced localization accuracy compared to YOLOv5n and YOLOv5s, with a performance increase of +7% and +6% in terms of *AP_{@IoU=0.5}*. The enhanced detection performance can be attributed to various improvements implemented in the YOLOv8 architecture. These include (i) the integration of C2F modules, which incorporate more skip connections and additional split

operations compared to those in the YOLOv5 model, facilitating the concatenation of different features, and (ii) the usage of a mosaic approach for data augmentation, which introduces spatial and contextual variations, forcing the model to acquire more robust features.

The SR branch, added to the YOLOv8n to implement the SR reconstruction of the original frame, boosted the detection, particularly for small lesions (Figs. 6 and 7, Tables 10, 8, 9). Among all the enhanced models examined, integrating the deep SR structure at the 4th and 8th layers of the backbone led to a 5% increase in *AP_{@IoU=0.5}* over the baseline, and a 2% improvement if compared with the other 2–6 and *light* SR configurations.

The configuration 4–8 deep variant of the network benefits from the higher level of features extracted by the shared backbone and processed through the EDSR decoder. This improves the network ability to capture a higher-definition representation of the semantic content of the frame while searching for objects to detect [44]. Furthermore, according to the reference work [15], a lightweight and simple structure of the SR decoder, like the one used in our *light* configuration, may not be capable of decoding and reconstructing the frames properly.

The pre-training of the SRE-YOLO architecture positively influenced the detection performance, indicating the network acquired a basic and transferable understanding of data characteristics. This prior knowledge was shown advantageous for the complexity of our task, which involves highly variable image features but is limited by the low availability of training datasets. Supported by the results obtained during the first experiments on YOLOv8n pre-training (Table 4), we believe that the impact of the pre-training or fine-tuning based on the HyperKvasir dataset was insufficient to achieve a satisfactory performance as a consequence of its limited size (1000 frames). Regarding the self-supervised pre-training on the unlabeled portion of the HyperKvasir dataset followed by fine-tuning on the labeled frames, we believe that, despite the larger number of frames available, the suboptimal performance may be due to limitations inherent in the data. Specifically, the self-supervised training of the SR branch relies solely on interpolated data obtained by resizing images to 1280 × 1280 pixels. This resizing introduces a vast number of similar pixels that may not be informative for accurately distinguishing lesion features from the underlying tissue, owing to the homogeneity of the initial pixels. This may hamper the network's generalization capabilities. Conversely, opting to pre-train on the COCO dataset can be more exemplary due to its breadth and diversity. As the results were equal among the models with only the baseline and with both YOLO and SR components pre-trained on COCO, we proposed the SR-enhanced model fully pre-trained on COCO for the lesion detection task. The improvement achieved by pre-training also the SR branch on COCO – mainly in terms of overfitting mitigation – will be more evident when larger datasets, than the one available for this work, are considered during future developments.

The advantage of the proposed SRE-YOLO relies on performance improvement without adding supplementary computational complexity during inference, which avoids affecting real-time applicability. During the testing phase, with the SR branch deactivated, frames were predicted with GFLOPs and FPS comparable to those of the YOLOv8n baseline, but with enhanced detection accuracy (Fig. 5).

Examples of improved detection predictions from test frames are illustrated in Fig. 6. In the first column, the original frames were coupled with the bounding boxes annotated by clinicians, while the second and

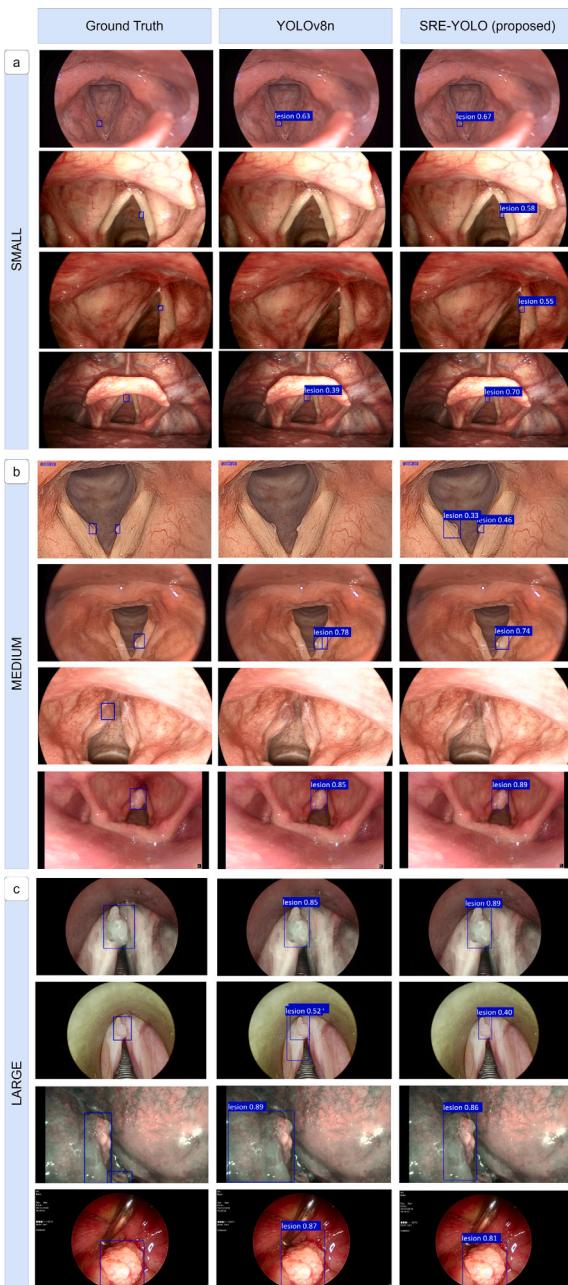


Fig. 7. Qualitative results divided by lesion size. (a) Small-sized lesions ($<32 \times 32$ pixels): right vocal fold polyps in the first three rows, and a left vocal fold polyp in the last row. (b) Medium-sized lesions ($32 \times 32 < \text{bounding box} < 96 \times 96$): bilateral vocal fold polyps in the first row, left vocal fold polyp in the second and third (failure event) rows, and white-light view of squamous cell carcinoma in the last row. (c) Large-sized lesions ($>96 \times 96$): in the first and second rows an intraoperative narrow-band-imaging and white-light view of a left vocal fold polyp, a narrow-band-imaging frame with squamous cell carcinoma arising from the left laryngeal vestibular fold, and a squamous cell carcinoma case in the intraoperative environment. In order, the columns illustrate the ground-truth bounding boxes, the YOLOv8n (i.e., the baseline), and SRE-YOLO predictions.

third columns reported the YOLOv8n and SRE-YOLO outputs. Except for the frames in the last row, for which all the detection models failed because of the structural and color similarity of the lesion with the surrounding tissue, the SRE-YOLO model outperformed the baseline model. It exhibited fewer false positives concerning small lesions and

achieved more precise localizations. This appears evident in the first six rows where the introduction of the SR branch has impacted the confidence of the proposed method in detecting small-sized lesions.

Furthermore, these results are aligned with Tables 9, 10 and Fig. 7, in which the proposed model's superiority in detecting lesions was further demonstrated.

Despite the good results, this study still has some limitations. For example, the size of the training dataset was small due to the requirement to select as SR ground truth only the frames acquired with height or width >1280 . In addition, the proposed model currently lacks a step to discern the severity of the detected lesion, and thus is not able to distinguish between benign and malignant cases. We plan to address these limitations by enhancing the training dataset with more high resolution images, and by training the network to recognize lesion severity.

7. Conclusions

Endoscopic examination represents the gold standard method to diagnose laryngeal lesions. Nonetheless, the detection of tissue abnormalities can be challenging, particularly when the lesions are small or when clinicians lack the necessary expertise. The literature underscores that systems empowered by DL may offer support for clinical decision-making. Yet, to fully harness the potential of these assistive systems, it is crucial to design algorithms that achieve both enhanced performance and computational efficiency. To address this challenge, this study introduces the SRE-YOLO model. This approach employs a YOLOv8n integrated with a SR branch. The network was designed and trained to detect laryngeal lesions across three different datasets featuring WL and NBI frames. The SR branch is concurrently trained with the baseline YOLOv8n model, but is excluded during the testing phase. This augments lesion detection capabilities while keeping the same computational efficiency and inference speed of its akin without SR, i.e., the YOLOv8n. Future improvements of the proposed research will include: (i) the collection of a larger and more diverse dataset, with a broader range of lesion types, (ii) the deployment of the architecture on a low-cost edge device like Nvidia Jetson Nano [45], (iii) the use of knowledge distillation methods to further optimize computational demands [46].

CRediT authorship contribution statement

Chiara Baldini: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lucia Migliorelli:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Daniele Berardini:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Muhammad Adeel Azam:** Data curation. **Claudio Sampieri:** Resources, Data curation. **Alessandro Ioppi:** Resources, Data curation. **Rakesh Srivastava:** Resources. **Giorgio Peretti:** Supervision, Resources, Data curation. **Leonardo S. Mattos:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Leonardo S. Mattos reports financial support was provided by RAISE, Robotics and AI for Socio-economic Empowerment (ECS00000035). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded by the European Union – NextGenerationEU and by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project “RAISE - Robotics and AI for Socio-economic Empowerment” (ECS00000035) and Horizon Europe project AIRCARE, GA no. 101137426. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

- [1] A. Barsouk, J.S. Aluru, P. Rawla, K. Saginala, A. Barsouk, Epidemiology, risk factors, and prevention of head and neck squamous cell carcinoma, *Med. Sci. (Basel)* 11 (2) (2023).
- [2] A.K. Dhull, R. Atri, R. Dhankhar, A.K. Chauhan, V. Kaushal, Major risk factors in head and neck cancer: A retrospective analysis of 12-year experiences, *World J. Oncol.* 9 (3) (2018) 80–84.
- [3] A. Koroulakis, M. Agarwal, Laryngeal cancer, in: StatPearls, StatPearls Publishing, 2022.
- [4] N. Davaris, S. Voigt-Zimmermann, S. Kropf, C. Arens, Flexible transnasal endoscopy with white light or narrow band imaging for the diagnosis of laryngeal malignancy: diagnostic value, observer variability and influence of previous laryngeal surgery, *Eur. Arch. Otorhinolaryngol.* 276 (2) (2019) 459–466.
- [5] C.A. Rosen, M.R. Amin, L. Sulica, C.B. Simpson, A.L. Merati, M.S. Courey, M.M. Johns, G.N. Postma, Advances in office-based diagnosis and treatment in laryngology, *Laryngoscope* 119 Suppl 2 (S2) (2009) S185–212.
- [6] M. Źurek, K. Jasak, K. Niemczyk, A. Rzepakowska, Artificial intelligence in laryngeal endoscopy: Systematic review and meta-analysis, *J. Clin. Med.* 11 (10) (2022) 2752.
- [7] A. Iandelli, C. Sampieri, F. Marchi, A. Pennacchi, A.L.C. Carobbio, P. Lovino Camerino, M. Filastro, G. Parrinello, G. Peretti, The role of peritumoral depapillation and its impact on narrow-band imaging in oral tongue squamous cell carcinoma, *Cancers (Basel)* 15 (4) (2023).
- [8] I. Vilaseca, M. Valls-Mateus, A. Nogués, E. Lehrer, M. López-Chacón, F.X. Avilés-Jurado, J.L. Blanch, M. Bernal-Sprekelsen, Usefulness of office examination with narrow band imaging for the diagnosis of head and neck squamous cell carcinoma and follow-up of premalignant lesions, *Head Neck* 39 (9) (2017) 1854–1863.
- [9] M. Źurek, A. Rzepakowska, E. Osuch-Wójcikiewicz, K. Niemczyk, Learning curve for endoscopic evaluation of vocal folds lesions with narrow band imaging, *Braz. J. Otorhinolaryngol.* 85 (6) (2019) 753–759.
- [10] S. Thakur, U. Patnaik, S.K. Singh, K. Sahai, R. Chugh, G.P.S. Gahlot, A comparison of the efficacy of narrow band imaging and contact endoscopy in an early diagnosis of squamous malignancies of the upper aerodigestive tract, *Med J. Armed Forces India* 79 (Suppl 1) (2023) S250–S257.
- [11] C. Sampieri, C. Baldini, M.A. Azam, S. Moccia, L.S. Mattos, I. Vilaseca, G. Peretti, A. Ioppi, Artificial intelligence for upper aerodigestive tract endoscopy and laryngoscopy: A guide for physicians and state-of-the-art review, *Otolaryngol. Head Neck Surg.* 169 (4) (2023) 811–829.
- [12] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al., Intermage: Exploring large-scale vision foundation models with deformable convolutions, 2022, arXiv preprint [arXiv:2211.05778](https://arxiv.org/abs/2211.05778).
- [13] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L.M. Ni, H.-Y. Shum, Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022, arXiv preprint [arXiv:2203.03605](https://arxiv.org/abs/2203.03605).
- [14] R.T. Sutton, D. Pincock, D.C. Baumgart, D.C. Sadowski, R.N. Fedorak, K.I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, *NPJ Digit. Med.* 3 (1) (2020) 17.
- [15] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, Q. Du, SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery, *IEEE Trans. Geosci. Remote Sens.* 61 (2023) 1–15, [http://dx.doi.org/10.1109/TGRS.2023.3258666](https://dx.doi.org/10.1109/TGRS.2023.3258666).
- [16] D. Berardini, L. Migliorelli, A. Galdelli, M.J. Marín-Jiménez, Edge artificial intelligence and super-resolution for enhanced weapon detection in video surveillance, *Eng. Appl. Artif. Intell.* 140 (2025) 109684.
- [17] Q. Cen, Z. Pan, Y. Li, H. Ding, Laryngeal tumor detection in endoscopic images based on convolutional neural network, in: 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT), IEEE, 2019.
- [18] B. Luan, Y. Sun, C. Tong, Y. Liu, H. Liu, R-FCN based laryngeal lesion detection", in: 2019 12th International Symposium on Computational Intelligence and Design (ISCID), IEEE, 2019.
- [19] A. Inaba, K. Hori, Y. Yoda, H. Ikematsu, H. Takano, H. Matsuzaki, Y. Watanabe, N. Takeshita, T. Tomioka, G. Ishii, S. Fujii, R. Hayashi, T. Yano, Artificial intelligence system for detecting superficial laryngopharyngeal cancer with high efficiency of deep learning, *Head Neck* 42 (9) (2020) 2581–2592.
- [20] G.H. Kim, E.-S. Sung, K.W. Nam, Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network, *BioMed. Eng. OnLine* 20 (1) (2021) 1–10.
- [21] M.A. Azam, C. Sampieri, A. Ioppi, S. Africano, A. Vallin, D. Mocellin, M. Fragale, L. Guastini, S. Moccia, C. Piazza, et al., Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection, *Laryngoscope* 132 (9) (2022) 1798–1806.
- [22] P. Yan, S. Li, Z. Zhou, Q. Liu, J. Wu, Q. Ren, Q. Chen, Z. Chen, Z. Chen, S. Chen, et al., Automated detection of glottic laryngeal carcinoma in laryngoscopic images from a multicentre database using a convolutional neural network, *Clin. Otolaryngol.* 48 (3) (2023) 436–441.
- [23] G.H. Kim, Y.J. Hwang, H. Lee, E.-S. Sung, K.W. Nam, Convolutional neural network-based vocal cord tumor classification technique for home-based self-prescreening purpose, *BioMed. Eng. OnLine* 22 (1) (2023) 81.
- [24] D.J. Wellenstein, J. Woodburn, H.A. Marres, G.B. van den Broek, Detection of laryngeal carcinoma during endoscopy using artificial intelligence, *Head Neck* 45 (9) (2023) 2217–2226.
- [25] A.M. Bur, T. Zhang, X. Chen, H. Kavookjian, S. Kraft, O. Karadaghy, N. Farrokhan, C. Mussatto, J. Penn, G. Wang, Interpretable computer vision to detect and classify structural laryngeal lesions in digital flexible laryngoscopic images, *Otolaryngol. Head Neck Surg.* (2023).
- [26] Z. He, K. Zhang, N. Zhao, Y. Wang, W. Hou, Q. Meng, C. Li, J. Chen, J. Li, Deep learning for real-time detection of nasopharyngeal carcinoma during nasopharyngeal endoscopy, *iScience* 26 (2023) 107463, <http://dx.doi.org/10.1016/j.isci.2023.107463>.
- [27] H. Klimza, J. Jackowska, W. Pietruszewska, A. Rzepakowska, M. Wierzbicka, The narrow band imaging as an essential complement to white light endoscopy in recurrent respiratory papillomatosis diagnostics and follow-up process, *Otolaryngol. Pol.* 76 (1) (2021) 1–5.
- [28] C. Saraniti, E. Chianetta, G. Greco, N. Mat Lazim, B. Verro, The impact of narrow-band imaging on the pre- and intra-operative assessments of neoplastic and preneoplastic laryngeal lesions. a systematic review, *Int. Arch. Otorhinolaryngol.* 25 (3) (2021) e471–e478.
- [29] S. Tiribelli, Inequalities and artificial intelligence, *Filos. Morale/Moral Philos.* (3) (2023).
- [30] S. Tiribelli, A. Monnot, S.F. Shah, A. Arora, P.J. Toong, S. Kong, Ethics principles for artificial intelligence-based telemedicine for public health, *Am. J. Public Health* 113 (5) (2023) 577–584.
- [31] B. Na, G.C. Fox, Object detection by a super-resolution method and a convolutional neural networks, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 2263–2269.
- [32] S.M.A. Bashir, Y. Wang, Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network, *Remote Sens.* 13 (9) (2021) 1854.
- [33] L. Yin, Y. Liu, M. Pei, J. Li, M. Wu, Y. Jia, Laryngoscope8: Laryngeal image dataset and classification of laryngeal disease based on attention mechanism, *Pattern Recognit. Lett.* 150 (C) (2021) 207–213, <http://dx.doi.org/10.1016/j.patrec.2021.06.034>.
- [34] S. Moccia, G.O. Vanone, E. De Momi, A. Laborai, L. Guastini, G. Peretti, L.S. Mattos, Learning-based classification of informative laryngoscopic frames, *Comput. Methods Programs Biomed.* 158 (2018) 21–30.
- [35] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common objects in context, 2015, [arXiv:1405.0312](https://arxiv.org/abs/1405.0312), URL <https://arxiv.org/abs/1405.0312>.
- [36] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, Scaled-yolov4: Scaling cross stage partial network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13029–13038.
- [37] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxxy, Lorna, Z. Yifu, C. Wong, A. V. D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, M. Jain, ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, 2022, <http://dx.doi.org/10.5281/zenodo.7347926>.
- [38] C.-Y. Wang, H.-Y.M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, I.-H. Yeh, CSPNet: A new backbone that can enhance learning capability of CNN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391.
- [39] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (9) (2015) 1904–1916.
- [40] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.
- [41] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLO, 2023, URL <https://github.com/ultralytics/ultralytics>.

- [42] H. Borgli, V. Thambawita, P.H. Smetsrud, S. Hicks, D. Jha, S.L. Eskeland, K.R. Randel, K. Pogorelov, M. Lux, D.T.D. Nguyen, et al., HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, *Sci. Data* 7 (1) (2020) 283.
- [43] M. Cherti, J. Jitsev, Effect of pre-training scale on intra- and inter-domain, full and few-shot transfer learning for natural and X-Ray chest images, in: 2022 International Joint Conference on Neural Networks, IJCNN, 2022, pp. 1–9, <http://dx.doi.org/10.1109/IJCNN55064.2022.9892393>.
- [44] Y. Chen, X. Zhu, Y. Li, Y. Wei, L. Ye, Enhanced semantic feature pyramid network for small object detection, *Signal Process., Image Commun.* 113 (2023) 116919.
- [45] L. Migliorelli, A. Cacciatore, V. Ottaviani, D. Berardini, R.L. Dellaca', E. Frontoni, S. Moccia, TwinEDA: a sustainable deep-learning approach for limb-position estimation in preterm infants' depth images, *Med. Biol. Eng. Comput.* 61 (2) (2023) 387–397.
- [46] L. Serrador, F.P. Villani, S. Moccia, C.P. Santos, Knowledge distillation on individual vertebrae segmentation exploiting 3D U-Net, *Comput. Med. Imaging Graph.* (2024) 102350.