



Published in final edited form as:

Laryngoscope. 2026 June ; 136(6): 2511–2518. doi:10.1002/lary.70396.

Exploring Real-Time Tracking of Vocal Fold Polyps in Video-Stroboscopy using Deep Learning

Sanjana Kaza, MS^{*,1}, Abhinata S. Mohanty, MS^{*,2}, Aisha Serpedin, BA³, Jenny Yau, BA³, Yeo E. Kim, BA³, Rachel B. Kutler, BA³, Olivier Elemento⁶, Lucian Sulica³, Pegah Khosravi, PhD^{**,4,5}, Anaïs Rameau, MD, MS, MPhil, FACS^{**,3}

¹Cornell Tech, New York, NY, USA

²Center for Computational and Integrative Biology, Rutgers University, Camden, NJ, USA

³Department of Otolaryngology-Head and Neck Surgery, Sean Parker Institute for the Voice, Weill Cornell Medicine, New York, New York, USA

⁴Department of Clinical Sciences, College of Medicine, University of Central Florida, Orlando, FL, USA

⁵Institute for Artificial Intelligence, Department of Computer Science, University of Central Florida, Orlando, FL, USA

⁶Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA

Abstract

Objective: To develop and evaluate a deep learning object detection system for identifying vocal fold polyps in stroboscopic video frames using You Only Look Once (YOLO), and to assess the added benefit of temporal tracking on detection performance.

Methods: A retrospective dataset of 12,742 annotated frames from 55 laryngoscopy video recordings was annotated with bounding boxes identifying vocal fold polyps. Pre-trained YOLO11 and YOLO12 models were fine-tuned to detect the polyps in the frames. A temporal tracking algorithm was further developed to propagate missed detections across adjacent frames.

Results: YOLO12 outperformed YOLO11 across all metrics. On the hold-out test set, YOLO12 reached a precision of 83.1% and an F1 score of 67.6%, with a mean average precision at 0.5 (mAP@0.5) of 64.1%. By comparison, YOLO11 achieved a precision of 67.3% and an F1 score of 56.2%, with a mAP@0.5 of 56.0%. Incorporating temporal tracking increased mAP@0.5 to 70.4% with YOLO 12, while maintaining a detection speed of 21.4 frames per second (fps), close to real time (30 fps).

Corresponding Author: Anaïs Rameau, MD MSc MPhil MS FACS, Department of Otolaryngology– Head and Neck Surgery, Sean Parker Institute for the Voice, Weill Cornell Medicine, 240 East 59th Street, New York, NY 10022. anr2783@med.cornell.edu.

*Co-first authors

**Co-last authors

Meeting Information: This work was presented as a podium presentation at the American Laryngological Association Society Annual Meeting at COSM, New Orleans, USA; May, 2025.

Conflict of Interest Statement: Anaïs Rameau is a medical advisor for SoundHealth Systems Inc. and Pentax Medical.

Conclusions: Using YOLO 12 for vocal fold polyp detection in stroboscopy was enhanced with temporal tracking, achieving a mAP@0.5 of 70.4% with near real time performance. These results demonstrate the potential of real-time AI-assisted detection of vocal fold lesions.

Level of Evidence: 4

Keywords

stroboscopy; artificial intelligence; object detection; vocal fold polyp

INTRODUCTION

With recent advances in artificial intelligence (AI) and machine learning (ML) specifically, computer vision has improved performance in classification, segmentation, and object detection tasks in medical imaging and can now assist clinicians in clinical decision-making.^{1,2} Computer vision has been widely applied to video-endoscopy of the upper airway tract (UAT) to detect, classify, and segment pathologies in the nasopharynx, oropharynx, and larynx.^{3–9} It holds the promise to reduce interrater variability in interpretation of UAT video-endoscopy and enhance reliability across varied clinical settings, thus improving collective diagnostic performance.

Object detection is a computer vision technique that helps identify and localize anatomical regions or pathology. Object detection consists of two specific tasks: object localization and object classification.^{10–12} Object localization identifies specific objects through bounding boxes or other markers, while object classification bins the objects into a predefined class or category, such as benign or malignant lesion.^{10–12} One of the earliest techniques used for object detection was the sliding window technique. This was followed by higher performing methods involving Convolution Neural Networks (CNNs) like R-CNN, Fast R-CNN, and Faster CNN.^{13–15} You Only Look Once (YOLO) was developed in 2015 and demonstrated better real-time performance than all previous algorithms in object detection.¹⁶ YOLO enables real-time object detection by performing classification and localization in a single forward pass, a critical feature for video applications.¹⁶

Interest and research in YOLO-based systems has been rapidly growing in gastrointestinal endoscopy, particularly for real-time detection of colorectal polyps. Numerous studies have demonstrated its effectiveness, with models achieving detection recall rate as high as 98%.^{17–21} These models have proven effective not only in detecting large polyps but also in identifying small, flat, and difficult-to-visualize lesions that may be missed with routine evaluation.²² Systematic and scoping reviews of AI in colonoscopy have highlighted YOLO as an effective architecture for polyp detection, particularly where rapid inference and deployment are critical.^{23–25} These reviews also report trends toward improved adenoma detection rates in YOLO-based systems compared with other computer aided detection (CAdE) platforms. Several FDA approved CAdE algorithms for colorectal polyp detection are currently in commercial use through systems such as GI genius (Medtronic), SKOUT (Iterative Scopes), MAGENTIQ-COLO (MAGNETIQ EYE) and CADDIE (Olympus), though their reliance on YOLO is not confirmed.

To improve object detection consistency and detection across frames with YOLO, temporal tracking has been implemented in some colorectal studies. Temporal tracking refers to the process of following objects across consecutive video frames by leveraging motion and spatial continuity. In colonoscopy, spatio-temporal models have improved the identification of small polyps by analyzing their appearance and movement across frames.²⁶ Similarly, software tools such as Polyp-Alert made use of tracking to assist endoscopists in real-time colonoscopies for improved polyp detection.²⁷

Given the success of YOLO in colorectal polyp detection studies, its application to other forms of video-based medical imaging, such as stroboscopy and laryngoscopy, presents a promising avenue for real-time lesion detection and clinical decision support. Several teams have notably explored the potential of YOLO in laryngeal cancer and benign vocal fold lesion detection.^{28–30} The latest YOLO model used in the laryngology literature is YOLOv8n.²⁸ In this study, we explore the latest versions of YOLO (YOLO11 and YOLO12) for detection and localization of laryngeal polyps in video-stroboscopy frames. The task focuses on identifying the presence of polyps on the vocal folds, rather than diagnosing or classifying amongst different pathological subtypes. To enhance the detection of smaller or less distinct laryngeal polyps, we also implemented a temporal tracking module on top of the YOLO detection models. To our knowledge, this is the first study to integrate YOLO models with temporal tracking for the detection of laryngeal lesions across frames in video-laryngoscopy.

METHODS

Dataset and Annotations

This project was approved by the Weill Cornell Medicine Institutional Review Board (Protocol 19–05020151). A retrospective dataset of stroboscopic videos was collected at the Sean Parker Institute for the Voice. Exams and diagnoses were performed by three laryngologists on adult patients in a clinic setting using a Pentax VNL-1590sTi naso-pharyngo-laryngoscope connected to a Pentax Laryngeal Strobe (Model 9400). Recordings were collected through this integrated system. The dataset for this study, comprising 55 videos, was obtained as a subset of this larger database and consisted of 33 videos with laryngeal polyps from consecutive distinct individuals and 22 videos from consecutive healthy controls. A deep learning-based informative frame classifier was used to extract frames. Informative frames were defined as images showing a full view of the vocal folds and of sufficient visual quality to highlight them. Using this classifier, 12,742 informative frames were extracted from the 55 videos.³¹ Each polyp was annotated with a bounding box for the presence of laryngeal polyps. Annotations were made by a medical student (AS) under continuous supervision of the senior laryngologist in the team (AR). While formal inter-rater reliability statistics were not computed, the supervised training approach ensured annotation quality through iterative feedback and correction. The annotations followed the YOLO format with a single class label (“polyp”) and bounding box coordinates. The dataset was divided into training (69%), validation (11%), and test (20%) sets, ensuring that no patient’s frames appeared in more than one set (Figure 1). The training set included 4,788 polyp frames (23 individuals with vocal fold polyps) and 4,031 normal frames (13

individuals with normal vocal folds). The validation set included 951 polyp frames (4 individuals with vocal fold polyps) and 419 normal frames (2 individuals with normal vocal folds). The test set included 1,028 polyp frames (6 individuals with vocal fold polyps) and 1,525 normal frames (7 individuals with normal vocal folds).

YOLO Detection Models

We fine-tuned two YOLO-based object detection models for polyp detection, YOLO11 and YOLO12, pretrained on the Common Objects in Context (COCO) dataset.^{32–34} For preprocessing, input images were resized to 1024×1024 pixels and normalized by dividing pixel values by 255, which scales the original pixel values (ranging from 0 to 255) to a range between 0 and 1. To improve model generalization and mitigate overfitting, we employed multiple augmentation strategies using the built-in YOLO methods.³⁵ These included mosaic augmentation, label smoothing, dropout, and various random transformations (flips and angle variations), collectively ensuring the final models could handle the range of polyp sizes and visual contexts encountered in the stroboscopic exams.

Both YOLO models were configured for single-class polyp detection. For training optimization, we employed Rectified Adam (RAdam) as the optimization algorithm with a starting learning rate of 5×10^{-5} , a final learning rate factor of 0.01, and an initial warm-up period of 500 steps (during which the learning rate increased linearly from zero to the base rate) to stabilize early training. RAdam was selected for its adaptive learning rate capabilities, which stabilize training in noisy medical datasets. A default loss function combining weighted cross-entropy and distributional localization loss was applied to ensure balanced learning across the polyp and the background regions.

Training continued for up to 100 epochs with early stopping if there was no improvement in the validation set's F1 score, which is the combination of precision and recall, occurred for 40 consecutive epochs. The final model weights were selected from the epoch achieving the highest validation performance. Once trained, the models were applied to each frame in the test set independently, without using any temporal information. During inference, the models' output bounding box coordinates and confidence scores for each detection.

Temporal Tracking Pipeline

To leverage temporal consistency in laryngoscopy videos, we developed a tracking module that operates on the YOLO detections. The pipeline was designed to improve detection continuity by propagating polyp identifications across consecutive frames where the YOLO model might have momentarily missed polyps due to visual obstructions or subtle visual presentations.

The tracker operated in a forward-only manner, as backward propagation risked reinforcing false positives and introducing more noise than benefit. The tracking workflow consisted of two main steps: frame-level similarity evaluation and object-level matching. First, the algorithm assessed frame-to-frame similarity using Oriented FAST and Rotated BRIEF (ORB), which identified distinctive points in each frame to measure how visually similar two frames are. Frames deemed visually similar proceeded to the object matching step.

In the object-level matching phase, the goal was to recover polyps that were missed by YOLO by checking if a previously detected polyp region continued into the next frame. Template matching was first used to localize the most similar region in frame $t+1$, by comparing the bounding box from frame t within a constrained spatial window. This helped account for minor shifts in scope position. To confirm that the tracked region corresponded to the same polyp, a pixel-level similarity check was then applied.

To ensure precision, both frame-level and object-level similarity conditions must have been met before propagation occurred. Furthermore, propagation was constrained to at most one frame gap, preventing any detection from being extended across multiple uncertain frames. The tracking was also restricted at the patient level to avoid any false detections.

The performance of the tracking module depended on several key parameters. We optimized these parameters to maximize the F1 score. The main optimized parameters included, (1) the frame similarity threshold, i.e. the minimum number of matched ORB features required to consider frames similar, (2) the pixel similarity threshold, minimum score needed to confirm that the tracked polyp region matches the original YOLO detected polyp and (3) the detection confidence threshold, minimum confidence score for initial YOLO detections.

Optimal parameters were determined separately for YOLO11 and YOLO12 models. The pixel similarity threshold was set at 0.99 for both YOLO11 and YOLO12. Frame similarity thresholds were established at 50 and 15 matched features for YOLO11 and YOLO12, respectively, while confidence thresholds were set at 0.107 and 0.196. These values reflect the trade-off between recovering missed detections and avoiding false positives. The code with explanation is available at <https://github.com/Rameau-Lab/PolypDetection>.

Evaluation Metrics

Performance was assessed using precision, recall, F1 score, mean Average Precision (mAP) and fps. Precision measured the proportion of detected polyps that were correct, while recall quantified the proportion of actual polyps successfully detected. F1 score provided a balanced measure accounting for both false positives and negatives. mAP@0.5 evaluated bounding boxes localization precision using a 50% Intersection over Union (IoU) threshold, a measure of how much the predicted and ground truth boxes overlap, with 50% overlap used as the threshold. While mAP@0.5:0.95 averaged precision across multiple IoU thresholds. Frame per second (fps) indicates the number of frames analyzed by the system per second. AI Statement: During the preparation of this work, the authors used ChatGPT-4o to assist with information retrieval, shortening and proofreading, as well as code debugging and assistance. After using this tool, all content has been reviewed and edited by the authors, and they assume full responsibility for it.

RESULTS

Demographic details are summarized in Table 1, including age, sex and smoking history.

Detection and Tracking Performance

Without temporal tracking, YOLO12 detected polyps with a precision of 83.1% and a recall of 56.9%, corresponding to an F1 score of 67.6%. The mAP@0.5 for YOLO12 was 64.1%, with mAP@0.5:0.95 at 29.4% (Table 2). By comparison, YOLO11 achieved 67.3% precision and 48.3% recall, with an F1 score of 56.2%. Its mAP@0.5 was 56.0%, and mAP@0.5:0.95 was 25.5% (Table 2).

Integrating the temporal tracking module led to performance improvements for both models. For YOLO12, tracking increased recall from 56.9% to 57.9%, with a slight decrease in precision from 83.1% to 82.4% (Table 2). The resulting F1 score improved to 68.0%. YOLO12's mAP@0.5 showed a promising increase from 64.1% to 70.4% after tracking. For YOLO11, tracking increased recall from 48.3% to 49.2% and increased precision from 67.3% to 72.6%, improving its F1 score from 56.2% to 58.7%. YOLO11's mAP@0.5 improved from 56.0% to 61.3% (Table 2).

Bootstrap resampling was performed ($n=1000$) to estimate 95% confidence intervals for all metrics. YOLO12 achieved an F1 score of 68.0% (95% CI: 65.6% - 70.4%), which is significantly higher than YOLO11's 58.7% (95% CI: 56.1% - 61.3%). The non-overlapping bootstrapped intervals confirmed statistically meaningful gains for YOLO12 over YOLO11. Full metrics are reported in Figure 2.

Using two NVIDIA A40 GPUs, YOLO12 achieved 21.5 fps for detection, 617.0 fps for tracking, and a combined pipeline performance of 21.4 fps. For comparison, YOLO11 achieved 16.9 fps for detection, 597.7 fps for tracking, and 16.8 fps for the combined pipeline.

Grad-CAM Analysis

To further investigate the failed cases of the YOLO11 and YOLO12 models, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) on cases resulting in false negatives (Figure 3). Grad-CAM is a visualization technique that highlights important regions in an image that a convolutional neural network focuses on when making predictions.³⁶ As shown in the first example (Figure 3) where both YOLO11 and YOLO12 failed to detect the polyp, the Grad-CAM heatmaps reveal that neither model effectively focused on the region containing the ground truth bounding box, although YOLO12 did show some activation near the polyp region. In the second example (Figure 3), YOLO11 failed to detect while YOLO12 succeeded. Grad-CAM for YOLO11 showed that the model was concentrating on non-relevant parts of the vocal fold, thereby missing the lesion area entirely, whereas YOLO12 provided a more localized concentration near the lesion.

DISCUSSION

We demonstrated the feasibility of near real-time laryngeal polyp detection in video-stroboscopy using deep learning with the latest YOLO models. YOLO12 outperformed YOLO11 across all metrics, with an F1 score of 68.0% and mAP@0.5 of 70.4%. Furthermore, differences in the tuned tracking parameters, particularly the frame similarity thresholds, reflected that, while both models had missed detections during motion or angle

changes, YOLO11 was less stable and necessitated stricter temporal matching to maintain tracking reliability.

YOLO12's higher performance is consistent with its architectural advancements, which includes the incorporation of enhanced attention mechanisms and improved feature aggregation capabilities.^{34,37} The inclusion of the temporal tracking module modestly improved performance with slight increases in the F1 score and the mAP metrics. Prior research in endoscopic video analysis has noted that single-frame detectors can be susceptible to momentary misdetections and cannot leverage temporal context.³⁸ YOLO11's bigger leap in performance with temporal tracking compared to YOLO12 may be explained by a lower baseline performance with more room for improvement compared to YOLO12, with its stronger base performance with less room for relative gain.

Grad-CAM has been employed by previous researchers to explain the performance of a model.^{39,40} In our case, the visualizations reinforced the observed performance differences between YOLO12 and YOLO11. YOLO12 showed more concentrated activation near lesions, whereas YOLO11 displayed scattered attention. (Figures 2, 3). These shortcomings may arise from differences in model architecture. However, it is important to interpret these results cautiously, as Grad-CAM provides only a coarse approximation of model attention and may not always reflect the true reasoning behind a prediction.⁴¹

Despite these promising results, YOLO performance in vocal fold polyp detection remained behind reported performances in polyp detection for colonoscopy. Recent colonoscopy models have achieved precision and recall exceeding 90%, with F1 scores and mAP often above 85%.^{42,43} For example, a YOLOv8-based system recently reported a precision of 95.6% and an F1 score of 92.4% in real-time colorectal polyp detection.⁴³ Colonoscopy offers a more favorable landscape for lesion detection. The colon provides a relatively uniform tubular anatomy with consistent lighting, fewer obstructions, and slower flexible endoscope movements facilitated by patient sedation. Polyps typically present as protruding lesions against a homogenous background and persist across multiple frames, enabling better learning of visual features.⁴⁴

In contrast to the colon, the UAT anatomy, specifically that of the larynx, is structurally complex. The vocal folds, epiglottis, and surrounding tissues produce variable lighting, shadowing, and movement artifacts in the awake patient. The rapid laryngeal movements during phonation, breath, cough, gagging, or swallowing create motion blur and reduced visibility of subtle vocal fold lesions, such as polyps. Frames may contain occlusions, inconsistent viewpoints, and fewer clear, repeated views of the same lesion, limiting opportunities for the model to detect a polyp across frames. These factors make video-laryngoscopy a more challenging task for polyp detection compared to colonoscopy.

In colonoscopy, AI-assisted systems have improved adenoma detection rates by offering a second observer effect, prompting endoscopists to re-examine areas flagged by AI.⁴⁵ A similar paradigm could benefit laryngoscopy, where vocal fold polyps and other vocal fold lesions may be subtle and go undetected by the otolaryngologist. An object detection AI algorithm, such as the one we created with YOLO, could flag polyps, potentially improving

diagnostic yield during routine examinations, and assisting in-office procedural interventions such as laser ablation.

Several limitations of this study must be acknowledged. The dataset, while relatively large in frame count, comprised only 55 patients, limiting the diversity of polyp presentations and anatomical variability. The tracking module used was effective but rudimentary, relying on ORB-based visual similarity and template matching. While it helped improve recall, it may propagate false positives and lacks the accuracy and precision of deep learning models. The moderate F1 score indicates that some detections are still missed or incorrectly flagged, limiting consideration for immediate clinical deployment. Our pipeline detects lesions at 21.4 fps compared to the 30 fps rate of clinical laryngoscopic and stroboscopic examinations. While suitable for near real-time applications, it does not yet achieve true real-time performance. Achieving full real-time integration within a stroboscopic workflow will require additional components such as strobe light synchronization, synthetic cycle reconstruction, and live video augmentation, all of which represent important directions for future development.

Future work will focus on expanding the dataset to include a larger and more diverse cohort of patients with varied laryngeal pathologies. Cross-institutional collaborations will be essential for this effort, as they facilitate access to heterogeneous patient populations, diverse clinical practices, and a range of imaging hardware. Such collaborations are critical for enhancing model generalizability across demographics, healthcare settings, and for validating performance on diverse cohorts and equipment types. Advanced temporal modeling, such as the use of three-dimensional convolutional neural networks or transformer-based architectures, may further enhance detection accuracy by integrating optical flow, motion, and contextual information across frames. Developing multi-class detection frameworks capable of distinguishing between multiple lesion types, such as nodules, cysts, and malignancies, would more accurately reflect the clinical heterogeneity observed in laryngoscopy, as previously explored by Baldini et al.²⁸ Making use of Grad-CAM alternatives such as HiResCAM could also provide better insights into model interpretation, allowing for more informed training adjustments. Ultimately, these advancements could pave the way for accurate, real-time, AI-assisted diagnostic tools to support otolaryngologists in the early detection and management of vocal fold lesions.

CONCLUSION

This study presents the first approach to near real-time deep learning-based laryngeal benign lesion detection using YOLO-based object detection with a temporal tracking system. YOLO12, enhanced with a simple tracking pipeline, demonstrated good accuracy in identifying and localizing vocal fold polyps in stroboscopic videos, demonstrating potential for AI-assisted diagnosis. However, performance lagged that reported in colonoscopy, likely due to the anatomical complexity of the larynx and combined motion of the flexible laryngoscope and the vocal folds. With larger and more diverse datasets, advanced temporal modeling, and multi-class detection capabilities, future AI systems could assist otolaryngologists with the detection and identification of varied vocal fold pathologies in real-time, reducing the likelihood of missed lesions amongst less experienced clinicians.

Financial Support or Funding:

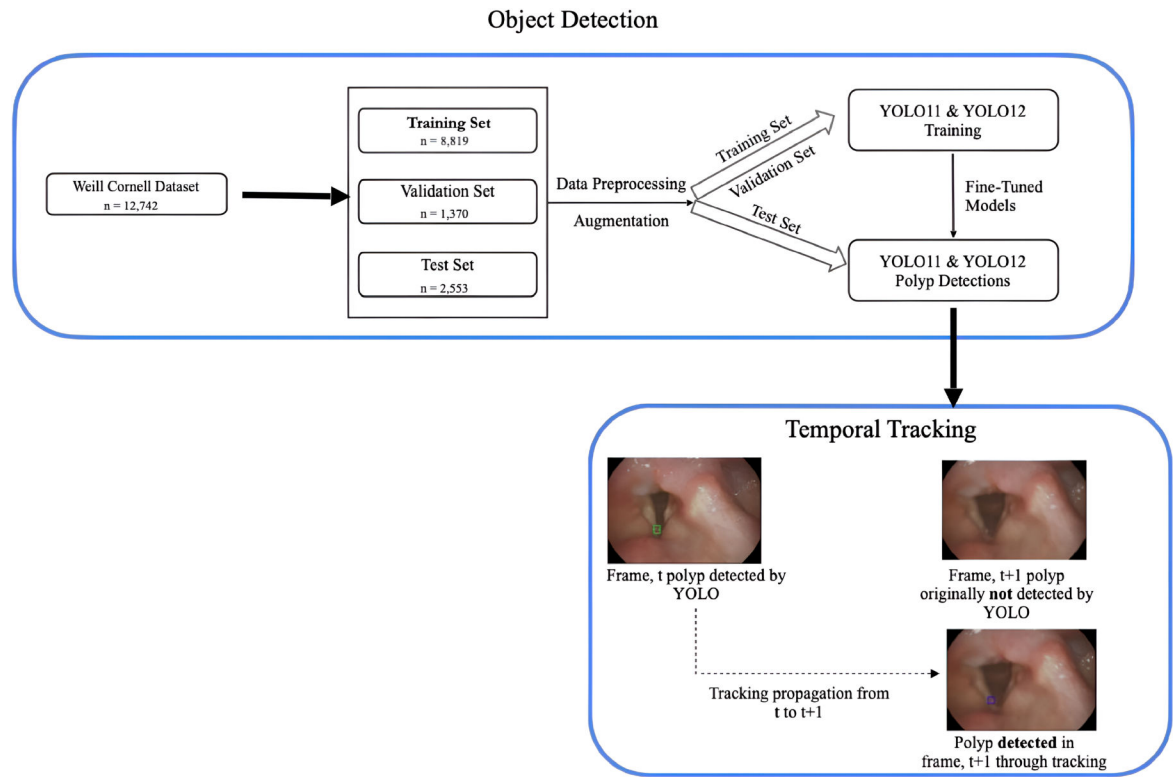
Anaïs Rameau was supported by a Paul B. Beeson Emerging Leaders Career Development Award in Aging (K76 AG079040) from the National Institute on Aging and by the Bridge2AI award (OT2 OD032720) from the NIH Common Fund.

REFERENCES

1. Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med*. 2021;4(1):5. doi:10.1038/s41746-020-00376-2 [PubMed: 33420381]
2. Elyan E, Vuttipittayamongkol P, Johnston P, et al. Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. *Artif Intell Surg*. 2022;2(1):24–45.
3. Yao P, Witte D, Gimonet H, et al. Automatic classification of informative laryngoscopic images using deep learning. *Laryngoscope Investig Otolaryngol*. 2022;7(2):460–466. doi:10.1002/lio2.754
4. Kim YE, Serpedin A, Periyakoil P, German D, Rameau A. Sociodemographic reporting in videomicroscopy research: a review of practices in otolaryngology - head and neck surgery. *Eur Arch Otorhinolaryngol*. 2024;281(11):6047–6056. doi:10.1007/s00405-024-08659-0 [PubMed: 38704768]
5. Yao P, Usman M, Chen YH, et al. Applications of artificial intelligence to office laryngoscopy: a scoping review. *Laryngoscope*. 2022;132(10):1993–2016. doi:10.1002/lary.30113 [PubMed: 34582043]
6. Xiong H, Lin P, Yu JG, et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*. 2019;48:92–99. doi:10.1016/j.ebiom.2019.08.027 [PubMed: 31594753]
7. Ren J, Jing X, Wang J, et al. Automatic recognition of laryngoscopic images using a deep-learning technique. *Laryngoscope*. 2020;130(11):E686–E693. doi:10.1002/lary.28544 [PubMed: 32068890]
8. Bensoussan Y, Vanstrum EB, Johns MM 3rd, Rameau A. Artificial intelligence and laryngeal cancer: from screening to prognosis: a state of the art review. *Otolaryngol Head Neck Surg*. 2023;168(3):319–329. doi:10.1177/01945998221149713 [PubMed: 35787073]
9. Yao P, Usman M, Chen YH, et al. Applications of artificial intelligence to office laryngoscopy: a scoping review. *Laryngoscope*. 2022;132(10):1993–2016. doi:10.1002/lary.29886 [PubMed: 34582043]
10. Yang R, Yu Y. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front Oncol*. 2021;11:638182. doi:10.3389/fonc.2021.638182
11. Albuquerque C, Henriques R, Castelli M. Deep learning-based object detection algorithms in medical imaging: systematic review. *Heliyon*. 2025;11(1):e00000.
12. Ragab MG, Abdulkadir SJ, Muneer A, et al. A comprehensive systematic review of YOLO for medical object detection (2018 to 2023). *IEEE Access*. 2024;12:57815–57836. doi:10.1109/ACCESS.2024.3381234
13. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017:2961–2969. doi:10.1109/ICCV.2017.322
14. Girshick R. Fast R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015:1440–1448. doi:10.1109/ICCV.2015.169
15. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–1149. doi:10.1109/TPAMI.2016.2577031 [PubMed: 27295650]
16. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:779–788. doi:10.1109/CVPR.2016.91
17. Yang X, Song E, Ma G, et al. Yolo-ob: an improved anchor-free real-time multiscale colon polyp detector in colonoscopy. *Biomed Signal Process Control*. 2025;103:107326. doi:10.1016/j.bspc.2024.107326

18. Pacal I, Karaman A, Karaboga D, et al. An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets. *Computers in Biology and Medicine*. 2022;141:105031. doi:10.1016/j.combiomed.2021.105031
19. Liu Y. Real-time multi-class detection of colorectal polyps based on the colon-YOLO network. *Alexandria Engineering Journal*. 2025;127:595–605. doi:10.1016/j.aej.2025.05.045
20. Pacal I, Karaboga D. A robust real-time deep learning based automatic polyp detection system. *Computers in Biology and Medicine*. 2021;134:104519. doi:10.1016/j.combiomed.2021.104519 [PubMed: 34090014]
21. Amer A, Hussein A, Ahmadvand N, et al. Generalized polyp detection from colonoscopy frames using proposed edf-yolo8 network. In: Ali S, van der Sommen F, Papież BW, Ghatwary N, Jin Y, Kolenbrander I, eds. *Cancer Prevention, Detection, and Intervention*. Cham: Springer Nature Switzerland; 2025:124–132. doi:10.1007/978-3-031-73376-5_12
22. Wan J, Zhu W, Chen B, Wang L, Chang K, Meng X. CRH-YOLO for precise and efficient detection of gastrointestinal polyps. *Sci Rep*. 2024;14(1):30033. doi:10.1038/s41598-024-81842-9 [PubMed: 39627309]
23. Soleymanjahi S, Huebner J, Elmansy L, et al. Artificial intelligence-assisted colonoscopy for polyp detection : a systematic review and meta-analysis. *Ann Intern Med*. 2024;177(12):1652–1663. doi:10.7326/ANNALS-24-00981 [PubMed: 39531400]
24. Hassan C, Spadaccini M, Mori Y, et al. Real-time computer-aided detection of colorectal neoplasia during colonoscopy : a systematic review and meta-analysis. *Ann Intern Med*. 2023;176(9):1209–1220. doi:10.7326/M22-3678 [PubMed: 37639719]
25. Spadaccini M, Menini M, Massimi D, et al. Ai and polyp detection during colonoscopy. *Cancers (Basel)*. 2025;17(5):797. doi:10.3390/cancers17050797 [PubMed: 40075645]
26. Wu L, Hu Z, Ji Y, Luo P, Zhang S. Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. July 2021. doi:10.48550/arXiv.2107.03609
27. Wang Y, Tavanapong W, Wong J, Oh JH, de Groen PC. Polyp-Alert: Near real-time feedback during colonoscopy. *Computer Methods and Programs in Biomedicine*. 2015;120(3):164–179. doi:10.1016/j.cmpb.2015.04.002 [PubMed: 25952076]
28. Baldini C, Migliorelli L, Berardini D, Azam MA, Sampieri C, Ioppi A, Srivastava R, Peretti G, Mattos LS. Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced YOLO. *Computer Methods and Programs in Biomedicine*. 2025 Mar 1;260:108539.
29. Wellenstein DJ, Woodburn J, Marres HA, van den Broek GB. Detection of laryngeal carcinoma during endoscopy using artificial intelligence. *Head & Neck*. 2023 Sep;45(9):2217–26. [PubMed: 37377069]
30. Kim GH, Hwang YJ, Lee H, Sung ES, Nam KW. Convolutional neural network-based vocal cord tumor classification technique for home-based self-prescreening purpose. *BioMedical Engineering OnLine*. 2023 Aug 18;22(1):81. [PubMed: 37596652]
31. Yao P, Witte D, German A, et al. A deep learning pipeline for automated classification of vocal fold polyps in flexible laryngoscopy. *Eur Arch Otorhinolaryngol*. 2024;281(4):2055–2062. doi:10.1007/s00405-023-08190-8 [PubMed: 37695363]
32. Ultralytics. Coco. <https://docs.ultralytics.com/datasets/detect/coco>. Accessed April 14, 2025.
33. Ultralytics. Yolo11. <https://docs.ultralytics.com/models/yolo11>. Accessed April 21, 2025.
34. Ultralytics. Yolo12. <https://docs.ultralytics.com/models/yolo12>. Accessed April 14, 2025.
35. Ultralytics. Train. <https://docs.ultralytics.com/modes/train>. Accessed April 14, 2025.
36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. December 2019. doi:10.48550/arXiv.1610.02391.
37. Watt) FSN (Tech. # yolo12: redefining object detection with an attention-centric architecture. Medium. February 20visualizations comparing YOLO11 and YOLO12 model 25. <https://medium.com/@felixsam922/yolo12-redefining-object-detection-with-an-attention-centric-architecture-a0029fac00a3>. Accessed April 14, 2025.

38. Ye H, Miao H, Liu J, Wang D, Li H. Temporal context framework for endoscopy artefact segmentation and detection. In: ; 2022. <https://openreview.net/forum?id=dBQOJVkc1D>. Accessed April 14, 2025.
39. Kim JH, Jung HS, Lee SE, Hou JU, Kwon YS. Improving difficult direct laryngoscopy prediction using deep learning and minimal image analysis: a single-center prospective study. *Sci Rep*. 2024;14(1):14209. doi:10.1038/s41598-024-65060-x [PubMed: 38902319]
40. Xia M, Jin C, Zheng Y, et al. Deep learning-based facial analysis for predicting difficult videolaryngoscopy: a feasibility study. *Anaesthesia*. 2024;79(4):399–409. doi:10.1111/anae.16194 [PubMed: 38093485]
41. Draelos RL, Carin L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. November 2021. doi:10.48550/arXiv.2011.08891
42. Urban G, Tripathi P, Alkayali T, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*. 2018;155(4):1069–1078.e8. doi:10.1053/j.gastro.2018.06.037 [PubMed: 29928897]
43. Lalinia M, Sahafi A. Colorectal polyp detection in colonoscopy images using YOLO-V8 network. *SIVIP*. 2024;18(3):2047–2058. doi:10.1007/s11760-023-02835-1
44. Misawa M, Kudo SE, Mori Y, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology*. 2018;154(8):2027–2029.e3. doi:10.1053/j.gastro.2018.04.003 [PubMed: 29653147]
45. Wang P, Liu XG, Kang M, et al. Artificial intelligence empowers the second-observer strategy for colonoscopy: a randomized clinical trial. *Gastroenterol Rep (Oxf)*. 2023;11:goac081. doi:10.1093/gastro/goac081

**Figure 1:**

Simplified workflow of the entire pipeline, from dataset preparation to final tracking detections.

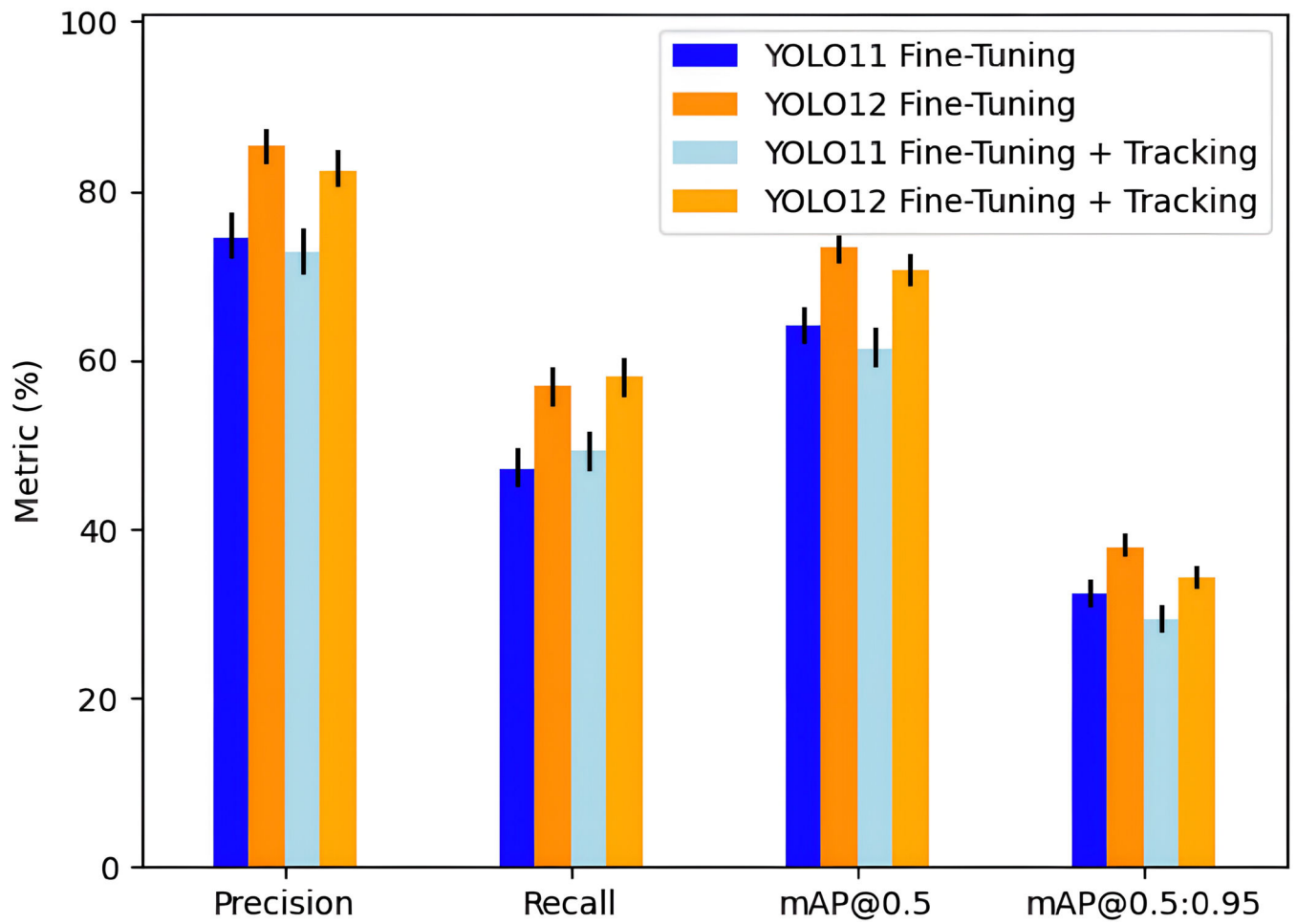
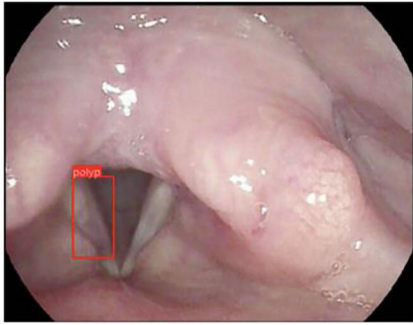
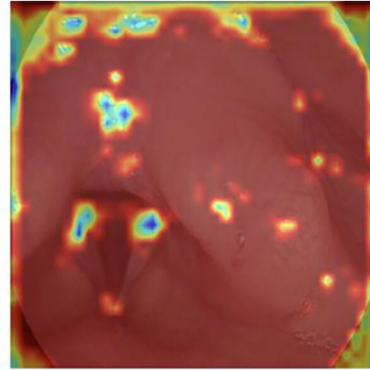


Figure 2:
Bootstrapped detection performance (mean $\pm$ 95% confidence intervals) for YOLO11 and YOLO12, with and without temporal tracking.

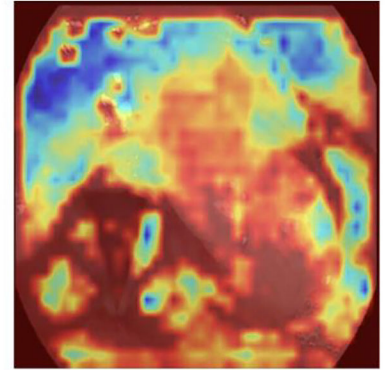
(a)



Ground Truth Annotated Frame



Grad-CAM Visualization for YOLO11

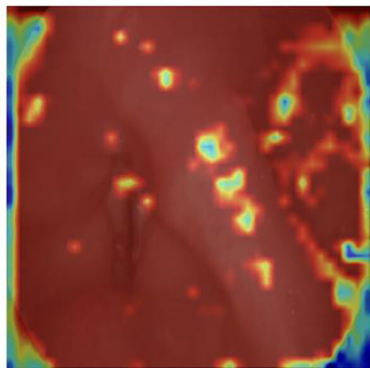


Grad-CAM Visualization for YOLO12

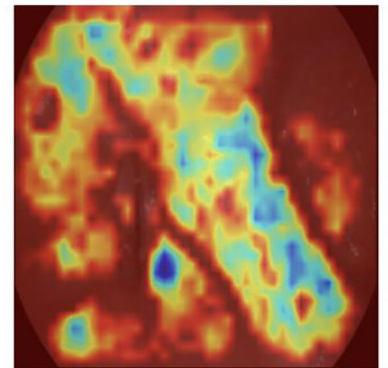
(b)



Ground Truth Annotated Frame



Grad-CAM Visualization for YOLO11



Grad-CAM Visualization for YOLO12

Figure 3:

Grad-CAM visualizations comparing YOLO11 and YOLO12 model attention.

(a) A case where both YOLO11 and YOLO12 failed to detect the polyp (false negative).

(b) A case where YOLO11 failed to detect the polyp but YOLO12 succeeded.

In both panels, the blue regions indicate areas ignored by the model, while warmer red regions highlight the regions of focus.

Table 1:
Demographic Characteristics of Overall Cohort (N = 55)

Characteristic	Healthy N = 22	Polyp N = 33
Age		
0–18	1	0
19–30	1	7
31–45	8	5
46–60	4	15
60+	8	6
Sex		
Male	4	21
Female	18	12
Occupation		
Routine	13	14
High Intensity	1	6
Performer	6	10
Unknown	2	3
Smoking History		
Yes	5	9
No	17	24

Table 2: Comparison of full detection performance metrics for YOLO11 and YOLO12, before and after applying tracking.

Model	Precision (IoU=0.5)	Recall (IoU=0.5)	F1 Score (IoU=0.5)	mAP@0.5	mAP@0.5:0.95
YOLO11 Fine-Tuning	67.3	48.3	56.2	56.0	25.5
YOLO11 Fine-Tuning + Tracking	72.6	49.2	58.7	61.3	29.3
YOLO12 Fine-Tuning	83.1	56.9	67.6	64.1	29.4
YOLO12 Fine-Tuning + Tracking	82.4	57.9	68.0	70.4	34.0