Leuko, I am your Prototype: A Prototypical Detection Framework for Fine-Grained Laryngeal Lesion Stratification

Lorenzo Federici[1], Maria Chiara Fiorentino[2], Riccardo Rosati[3], Claudio Sampieri[4], Giorgio Peretti[5], Francesco Mora[5], Elisa Bellini[5], Leonardo S. De Mattos[6], Primo Zingaretti[1], and Sara Moccia[2]

[1] Dept. of Information Engineering, Università Politecnica delle Marche, Ancona, Italy
l.federici@pm.univpm.it
[2] Dept. of Innovative Technologies in Medicine & Dentistry, Università degli Studi G. d'Annunzio, Chieti-Pescara, Italy
[3] Dept. of Political Science, Communication, and International Relations, University of Macerata, Macerata, Italy
[4] Dept. of Otorhinolaryngology, Hospital Clinic, Barcelona, Spain
[5] Dept. of Surgical Sciences and Integrated Diagnostics, University of Genova, Italy
[6] Dept. of Advanced Robotics, Istituto Italiano di Tecnologia, Genova, Italy

Abstract. Narrow-Band Imaging (NBI) endoscopy plays a central role in laryngeal lesion screening, yet fine-grained risk stratification remains challenging due to strong visual overlap, vascular variability, and contextual bias. While single-stage detectors achieve reliable localization, they often struggle to structure lesion representations for robust classification, particularly for morphologically ambiguous and underrepresented categories as Leukoplakia. We propose a dual-branch real-time detection framework that integrates prototypical learning into a YOLO-based architecture to explicitly organize the latent space around clinically meaningful lesion prototypes. A multi-scale feature fusion module and an auxiliary prototype head are jointly optimized through a two-stage training strategy with bootstrap stabilization and exponential moving average prototype updates. Experiments conducted on a retrospective NBI cohort and an independent prospective dataset demonstrate that our approach significantly improves fine-grained stratification, achieving superior macro F1-score compared to state-of-the-art CNN- and Transformer-based detectors, while preserving competitive detection performance and real-time efficiency. The source code is publicly available at https://github.com/lorenzo-federici/Prototypical-YOLO.

Keywords: Laryngeal lesion screening · Prototype learning · Laryngeal cancer · Laryngeal lesion localization.

1 Introduction

Endoscopic examination using Narrow-Band Imaging (NBI) is the cornerstone of laryngeal lesion screening. With NBI, laryngeal lesions may present a large

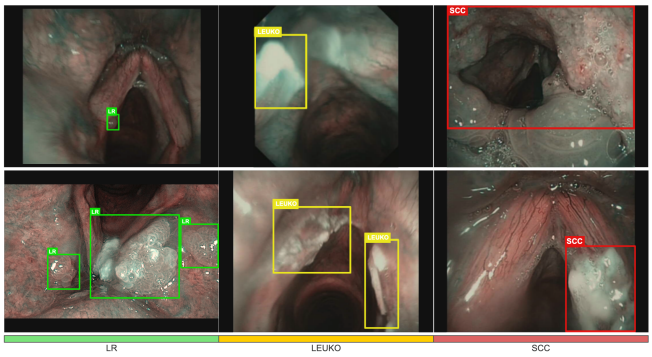

Fig. 1. Visual challenges of laryngeal lesions in NBI endoscopy. Two representative examples per class are shown. Fine-grained discrimination is complicated by pronounced scale variability (from minute lesions to large masses), acquisition artifacts (specular reflections, saliva, motion blur, blood), and wide areas of healthy mucosa. Importantly, substantial visual overlap exists across categories.

variety of visual characteristics in terms of vascular patterns and macroscopic appearance, ranging from Squamous Cell Carcinoma (SCC) and leukoplakia (LEUKO), commonly considered high-risk (HR) lesions, to low-risk (LR) lesions, such as polyps and cysts. The ability to stratify these lesions is fundamental, as different lesions imply different follow-up (from mere observation to surgery), but currently represents an open challenge due to shared characteristics among the different lesions [13]. Deep learning (DL)–based diagnostic systems have recently shown promising results in laryngeal lesion analysis [5, 18] and single-stage detectors, such as YOLO-based architectures [1, 2, 4, 14, 17], prove themselves to be able balance detection performance and real-time inference, a key requirement in clinical settings [1]. While lesion localization performance has progressively stabilized over time, limited effort has been devoted to improving classification performance of the localized lesions. Existing approaches predominantly focus on HR lesions, particularly SCC [12]. This gap can be partially attributed to the scarcity of publicly available datasets embedding both HR and LR lesions, which reflects data imbalance found in clinical practice: patients with SCC are more likely to undergo endoscopic examination, as symptoms often do not remain unnoticed or clinically underestimated [3].

Laryngeal lesion stratification poses a number of challenges, including intra- and inter lesion variability (Fig. 1), that single-stage detectors, which are currently optimized via standard bounding-box regression and cross-entropy losses

without explicitly enforcing lesion-aware feature representations, may not be able to tackle. As DL networks tend to exploit contextual correlations [7], and detection models are particularly sensitive to surrounding context [9], lesion embeddings in endoscopic frames, where healthy mucosa may dominate (Fig. 1), encode mixed contextual and pathological cues.

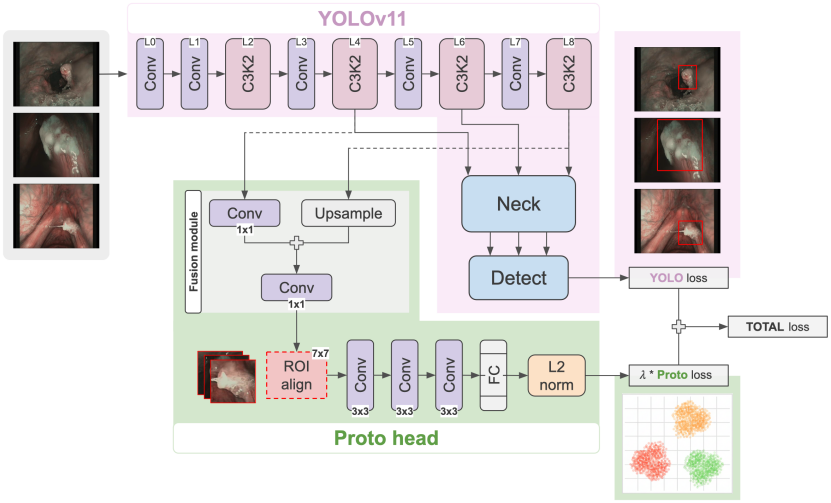

Fig. 2. Proposed dual-branch detection framework. Multi-scale features from the YOLO11n backbone are combined in a dedicated fusion module and fed to our auxiliary prototype (proto) head. Here, RoIAlign works on localised lesion patches and map them into an L2-normalized space clustered around three lesion prototypes (SCC, LEUKO, LR). The framework is optimized end-to-end with a joint loss.

Prototype-based learning has emerged as an effective strategy to explicitly structure latent spaces, promoting intra-class compactness, inter-class separability, and reduced reliance on spurious context [21, 6, 16]. Despite the promising results in other fields of medical image analysis, its integration into real-time laryngeal lesion detection frameworks, where localization and risk stratification must be jointly optimized, remains unexplored. To move the state of the art in laryngeal lesion stratification forward, this work makes four main contributions: (i) we propose a single-stage lesion detection framework augmented with a multi-scale prototypical learning branch that explicitly structures the latent space, towards lesion-aware representation; (ii) we introduce a prototype-based distance constraint enforcing intra-class compactness and inter-class separability, mitigating contextual bias and promoting lesion-centered feature learning for fine-grained classification, (iii) we design a joint end-to-end optimization strategy integrating detection and prototypical regularization without compromising real-time performance; (iv) we curate two dedicated NBI laryngeal datasets collected in the actual clinical practice, attenuating the lack of HR/LR benchmarks in our domain.

2 Methodology

Model backbone and multi-scale feature fusion. We here use YOLO11 nano (YOLO11n) [10], a fully end-to-end single-stage detector that provides a

favorable trade-off between detection accuracy and computational efficiency, enabling real-time inference in clinical endoscopic workflows. Let $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ denote an input frame. YOLO11n backbone extracts hierarchical feature representations defined as $\mathbf{F}_l = \mathcal{B}_l(\mathbf{I})$, with $l \in \{4, 8\}$, where $\mathcal{B}_l(\cdot)$ represents the backbone up to layer l, and $\mathbf{F}_l \in \mathbb{R}^{H_l\times W_l\times C_l}$ denotes the corresponding feature map. Shallow layers ($L4$) preserve high spatial resolution and fine structural details, whereas deeper layers ($L8$) encode stronger semantic abstraction. To jointly exploit spatial resolution and semantic information, we rely on a multi-scale feature fusion module where channel dimensions are first matched: $\tilde{\mathbf{F}}_l = \phi_l(\mathbf{F}_l)$, where $\phi_l(\cdot)$ denotes a 1×1 convolution. The deeper feature map ($\tilde{\mathbf{F}}_8$) is upsampled to match the spatial resolution of the shallow representation ($\tilde{\mathbf{F}}_8$) The fused representation ($\mathbf{F}_{\text{fuse}}$) is obtained via channel-wise concatenation: $\tilde{\mathbf{F}}_4 \oplus \hat{\mathbf{F}}_8$, where $\oplus$ denotes concatenation along the channel dimension. $\mathbf{F}_{\text{fuse}}$ is fed to the prototype head.

Prototypical Learning Branch & Joint Optimization. To explicitly structure the lesion representation space, we introduce an auxiliary Prototypical (Proto) head (Fig. 2). Given $\mathbf{F}_{\text{fuse}}$, RoIAlign maps localized lesions to fixed-size lesion-specific regions $\mathbf{R} \in \mathbb{R}^{7\times 7\times C}$. Each region is mapped into a d-dimensional embedding ($z \in \mathbb{R}^d$), which is then ℓ_2-normalized. The embedding space is organized around three learnable class prototypes ($\mathbf{p}_{\text{SCC}}, \mathbf{p}_{\text{LEUKO}}, \mathbf{p}_{\text{LR}} \in \mathbb{R}^d$). For the i-th embedding (z_i) for which the corresponding ground-truth (GT) label (y_i) is known, we compute cosine similarities with all prototypes and apply a temperature-scaled cross-entropy loss defined as

$$\mathcal{L}_{proto} = -\log \frac{\exp((z_i \cdot \mathbf{p}_{y_i})/\tau)}{\sum_{j=1}^{C} \exp((z_i \cdot \mathbf{p}_j)/\tau)} \tag{1}$$

where $C = 3$ is the number of classes and τ controls the sharpness of the similarity distribution. This formulation encourages lesion embeddings to align with their corresponding clinical prototype while maintaining inter-class separation.

Training is performed in two stages. During an initial bootstrap phase using a number of epochs (E_{bs}) identified via an ablation study, the framework is optimized using only the detection loss. This ensures the backbone learns a structured and discriminative feature space preventing prototypes from being initialized on random, uninformative noise. After this phase, class prototypes are initialized as the normalized mean embedding of GT lesion patches for each class $c \in \{\text{SCC}, \text{LEUKO}, \text{LR}\}$, i.e.,

$$\bar{z}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} z_i^{(c)}, \quad \mathbf{p}_c = \frac{\bar{z}_c}{\|\bar{z}_c\|_2} \tag{2}$$

During joint training, prototypes are updated via an Exponential Moving Average (EMA) according to

$$p_c \leftarrow \frac{mp_c + (1-m)\mu_c}{\|mp_c + (1-m)\mu_c\|_2} \tag{3}$$

Table 1. Ablation study. Comparison of our proposed architecture on RETRO against the YOLO11nano baseline under different backbone tuning strategies (*Full*, *Frozen*, *LoRA*). *Yolo* and *Proto* refer to the standard detection and proposed prototype classification heads. E_{bs}: bootstrap epochs. Best results in **bold**.

Method	Configuration		Detection		Classification (F1-score)				
	Backbone	Head	mAP_{50}	mAP_{95}	LR	Leuko	SCC	Macro	Weighted
Baseline									
YOLO11nano		Yolo	0.47	0.25	0.77	0.26	0.69	0.57	0.68
Proposed									
$E_{bs} = 0$	Full	Yolo	0.44	0.23	0.80	0.36	0.68	0.61	0.71
		Proto	—	—	0.88	0.35	0.35	0.53	0.62
$E_{bs} = 20$	Full	Yolo	**0.50**	**0.27**	0.80	0.30	0.69	0.60	0.70
		Proto	—	—	0.85	0.30	0.79	0.65	0.78
	Frozen	Yolo	0.43	0.23	0.73	0.28	0.64	0.55	0.64
		Proto	—	—	0.85	0.13	0.58	0.53	0.67
	LoRA	Yolo	0.45	0.25	0.79	0.33	0.70	0.61	0.70
		Proto	—	—	**0.90**	0.15	0.66	0.57	0.73
$E_{bs} = 40$	Full	Yolo	0.47	**0.27**	0.80	0.28	0.71	0.60	0.71
		Proto	—	—	0.86	**0.42**	**0.80**	**0.69**	**0.79**
	Frozen	Yolo	0.44	0.24	0.75	0.29	0.68	0.57	0.68
		Proto	—	—	0.87	0.14	0.65	0.55	0.71
	LoRA	Yolo	0.44	0.25	0.82	0.27	0.73	0.60	0.73
		Proto	—	—	0.89	0.16	0.68	0.58	0.74

where μ_c is the batch centroid and $m \in [0, 1]$ is the momentum parameter. The final objective loss to be minimized ($\mathcal{L}_{total}$) is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{YOLO} + \lambda \mathcal{L}_{proto} \tag{4}$$

with λ controlling the influence of the prototypical constraint.

Implementation details. The framework was implemented in PyTorch using the Ultralytics library [15]. All models were trained on an NVIDIA A100 GPU for 100 epochs with a batch size of 4, utilizing the standard Ultralytics optimizer and data augmentation protocols. Input images were resized to 640×640. To preserve the spatial integrity of the learned embeddings, Mosaic and MixUp were explicitly disabled after the bootstrap epoch. Prototypical hyperparameters were optimized via grid search by selecting the configuration that minimized the validation $\mathcal{L}_{total}$. The chosen values were: EMA momentum $m = 0.90$ (search space: 0.2–0.99), temperature $\tau = 0.80$ (0.2–1.0), and loss weight $\lambda = 0.5$ (0.2–1.0).

Datasets. To train and evaluate the proposed framework, we assembled a dataset of laryngeal NBI endoscopic images collected during routine outpatient examinations using a high-definition flexible transnasal nasopharyngolaryngoscope (Olympus ENF-VH, Olympus Medical Systems, Tokyo, Japan). As data were acquired in real-world clinical settings, the images exhibit substantial vari-

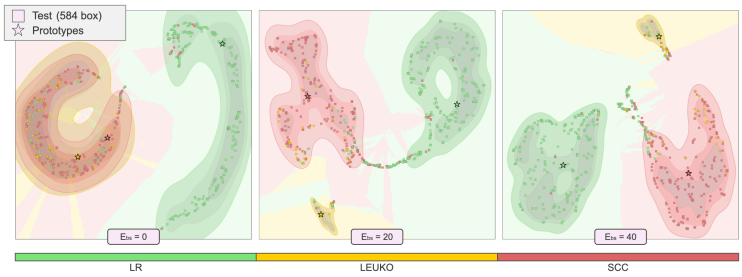

Fig. 3. Impact of the bootstrap phase on the embedding space. This t-SNE visualization maps the distribution of *training* bounding box-level features via Kernel Density Estimation (KDE) contours, overlaid with *test set* box features (squares) and learned *prototypes* (stars). The shaded background illustrates the corresponding decision boundaries. Models trained with higher E_{bs} progressively disentangles the classes, resulting in compact clusters and stable prototype placement.

ability, including non-uniform illumination, motion artifacts, complex anatomical backgrounds, and subtle lesion boundaries. For each patient, 2–10 high-quality representative NBI frames were manually identified by clinicians for annotation. All selected images were annotated by three expert laryngologists (at least five years of experience in NBI interpretation). Ambiguous cases were resolved by majority consensus. Data were split at the patient level to prevent information leakage.

The retrospective cohort, referred to as RETRO (June 2020–August 2022), was split into training (1661 images, 331 patients), validation (424 images, 82 patients), and internal test (525 images, 101 patients) sets. Class proportions were consistent across splits, with LR representing approximately 50% of cases, LEUKO 10%, and SCC 40%. An independent prospective cohort, PROS (July 2024; 327 images, 64 patients), served as an external clinical test set to evaluate real-world generalization, with an overall class distribution of 22% LR, 16% LEUKO, and 62% SCC.

3 Experiments and results

Ablations. As reported in Table 1, we conduct an ablation study to assess: (i) the contribution of the prototypical branch (*Proto head* vs. standard *Yolo head*); (ii) the impact of the bootstrap phase ($E_{\text{bs}} \in 0, 20, 40$); and (iii) the backbone adaptation strategy (Full, Frozen, LoRA [8]) after Bootstrap. Applying prototype regularization from scratch destabilizes training (Macro F1: 0.53), whereas introducing a bootstrap phase enables the detection loss to establish a stable feature space. At $E_{\text{bs}} = 40$, the *Proto head* achieves the best classification performance (Macro F1: 0.69) while preserving detection stability (mAP_{95}: 0.27), confirming effective task decoupling and improved representation disentanglement (Fig. 3). Crucially, fine-grained discrimination requires full

Table 2. Comparison with State-of-the-Art. Detection and classification performance on RETRO and PROS datasets. *Yolo* and *Proto* refer to the standard and proposed prototype heads of our dual-branch model. Best results are in **bold**. Stars indicate whether the proposed Proto is significantly better than SOTA approaches (*: $\alpha = 0.05$).

Method	RETRO						PROS						Params
	Detection		Class. (F1-score)				Detection		Class. (F1-score)				
	mAP_{50}	mAP_{95}	LR	Leuko	SCC	Macro	mAP_{50}	mAP_{95}	LR	Leuko	SCC	Macro	
YOLO5nano	**0.50**	**0.27**	0.75*	0.25*	0.65*	0.55*	**0.40**	**0.19**	0.55	0.30*	0.70	0.52	**2.51M**
YOLO8nano	0.41	0.20	0.59*	0.21*	0.54*	0.45*	0.35	0.17	0.44*	0.27*	0.63*	0.45*	3.01M
YOLO11nano	0.47	0.25	0.77*	0.26*	0.69*	0.57*	0.35	0.17	0.55	0.26*	0.67*	0.49	2.59M
RT-DETR	0.46	0.25	0.84	0.15*	**0.85**	0.61	0.27	0.14	**0.64**	0.16*	**0.86**	0.55	32.81M
DINO-DETR	0.40	0.18	0.64*	0.27*	0.61*	0.50*	0.35	0.17	0.61	0.20*	0.54*	0.45*	47.00M
Proposed (*Yolo*)	0.47	**0.27**	0.80*	0.29*	0.71*	0.60*	0.31	0.16	0.57	0.21*	0.71	0.50	7.45M
Proposed (*Proto*)	—	—	**0.86**	**0.42**	0.80	**0.69**	—	—	0.57	**0.40**	0.73	**0.57**	

backbone fine-tuning. Both Frozen and LoRA adaptations fail to adequately separate morphologically ambiguous lesions, particularly LEUKO (F1: 0.14–0.16 at $E_{\text{bs}} = 40$), demonstrating that low-rank or fixed updates are insufficient to reshape the embedding geometry, in low-data regime. End-to-end optimization is therefore necessary to allow the prototypical loss to reorganize the latent space, achieving the highest LEUKO F1-score (0.42).

Comparisons with the state-of-the-art. We evaluate the proposed dual-branch framework against established CNN-based detectors (YOLOv5n, v8n, v11n) and Transformer-based models (RT-DETR [22], DINO-DETR [20]). In laryngeal endoscopy, YOLO architectures are widely adopted for their balance between accuracy and efficiency; we specifically select the nano variants to ensure real-time intra-procedural applicability. As summarized in Table 2, our framework significantly outperforms existing methods in F1 while maintaining competitive detection performance. On the RETRO dataset, the Proto head achieves a macro-average F1 of 0.69, surpassing the YOLO11n baseline (0.57) and even the parameter-heavy RT-DETR (0.61). This superior macro-performance is primarily driven by a substantial gain for frames with leukoplakia. In clinics, leukoplakia typically represents a challenging minority class characterized by extreme morphological ambiguity. While standard detectors struggle to characterize these lesions, our framework integrating the Proto head achieves a marked improvement, reaching an F1-score of 0.42. These quantitative findings are qualitatively supported by the detection results in Fig. 4. While YOLO and Transformers struggle with low-contrast false negatives and high localization variance, respectively, our framework delivers stable, precise predictions. By incorporating a Proto head, we eliminate redundant detections and class confusion (e.g., LR vs. SCC), significantly outperforming baseline architectures on LEUKO lesions. Our findings suggest that prototypical regularization effectively mitigates the bias toward majority classes (SCC and LR) by anchoring features to a dedicated clinical

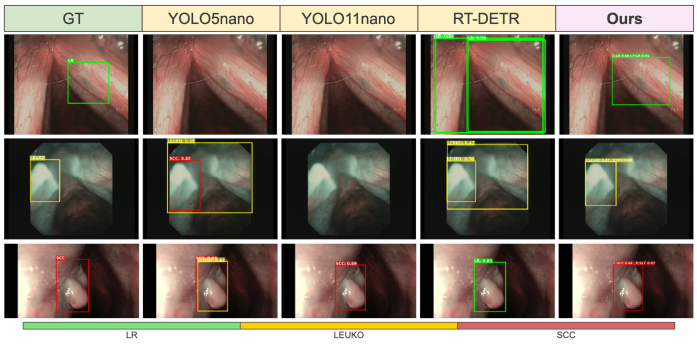

Fig. 4. Qualitative comparison of detection results. Bounding box colors correspond to the clinical classes (Green: LR, Yellow: LEUKO, Red: SCC). While SOTA models frequently struggle with missed detections (*YOLO11nano*), overlapping false positives (*YOLO5nano*), or severe class confusion (*RT-DETR*), our proposed framework provides accurate localization and robust classification that closely aligns with the GT.

prototype, capturing nuanced patterns that standard cross-entropy losses fail to structure. Crucially, this structural advantage is maintained on the PROS cohort, where our model preserves a leading Average F1-score of 0.57. Even under the domain shifts inherent to real-world endoscopic equipment and illumination, our framework continues to provide the most robust classification for the LEUKO class (0.40) compared to all other baselines. This demonstrates that anchoring features to clinical prototypes promotes the learning of pathology-aware representations that are significantly more invariant to contextual noise than traditional data-driven paradigms. Finally, despite the integration of the auxiliary Prototypical branch, our model remains highly efficient with only 7.45M parameters, suitable for deployment in diverse clinical settings.

Statistical Analysis. As competing models may correctly localize different subsets of lesions, we discard proposal-level paired statistical tests. As reported in Table 2, we evaluated the statistical significance of the performance differentials by treating our approach against each SOTA architecture as independent pairwise comparison. Based on the non-parametric image-level bootstrap resampling (B=10,000), the empirical 95% confidence intervals confirm that the proposed architecture achieves a statistically significant improvement ($p<0.05$) in pure categorical discrimination. Most notably, the stratified per-class analysis highlights the clinical robustness of our approach on the most challenging and underrepresented LEUKO class, where Proto consistently yields a statistically significant F1-Score improvement over all methods across both datasets.

Conclusion. We presented a dual-branch real-time detection framework that integrates prototypical learning into a YOLO-based architecture to improve fine-grained laryngeal lesion stratification in NBI endoscopy. By explicitly structuring

the embedding space around clinically meaningful prototypes, our approach enhances discrimination—particularly for the under-represented and challenging leukoplakia class—while preserving competitive localization performance and real-time feasibility. Despite promising results on both retrospective and prospective cohorts, limitations include the restricted number of lesion categories and the prototype granularity constrained to three clinical classes. Future developments will explore the extension to a broader spectrum of lesions, the adoption of hierarchical or sub-prototypical structures to better model intra-class heterogeneity [19], and the integration of complementary imaging modalities, such as White Light endoscopy [11], to enrich contextual and morphological information. These directions may further improve robustness in complex and multi-lesion clinical scenarios.

Acknowledgment. This work was funded by the Horizon Europe project AIR-CARE, 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.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azam, M.A., Sampieri, C., Ioppi, A., Africano, S., Vallin, A., Mocellin, D., Fragale, M., Guastini, L., Moccia, S., Piazza, C., et al.: Deep learning applied to white light and narrow band imaging videolaryngoscopy: toward real-time laryngeal cancer detection. The Laryngoscope **132**(9), 1798–1806 (2022)
2. Baldini, C., Migliorelli, L., Berardini, D., Azam, M.A., Sampieri, C., Ioppi, A., Srivastava, R., Peretti, G., Mattos, L.S.: Improving real-time detection of laryngeal lesions in endoscopic images using a decoupled super-resolution enhanced yolo. Computer Methods and Programs in Biomedicine (March 2025). https://doi.org/10.1016/j.cmpb.2024.108539
3. Carrard, V.C., Van der Waal, I.: A clinical diagnosis of oral leukoplakia; a guide for dentists. Medicina oral, patologia oral y cirugia bucal **23**(1), e59 (2017)
4. Cen, Q., Pan, Z., Li, Y., Ding, H.: Laryngeal tumor detection in endoscopic images based on convolutional neural network. In: IEEE 2nd International Conference on Electronic Information and Communication technology. pp. 604–608. IEEE (2019)
5. Du, S., Guo, J., Huang, D., Liu, Y., Zhang, X., Lu, S.: Diagnostic accuracy of deep learning-based algorithms in laryngoscopy: a systematic review and meta-analysis. European Archives of Oto-Rhino-Laryngology **282**(1), 351–360 (2025)
6. Fiorentino, M.C., Migliorelli, G., Villani, F.P., Frontoni, E., Moccia, S.: Contrastive prototype federated learning against noisy labels in fetal standard plane detection. International Journal of Computer Assisted Radiology and Surgery **20**(7), 1431–1439 (2025)
7. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)

8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
9. Jamali, M., et al.: Context in object detection: a systematic literature review. Artificial Intelligence Review (2025)
10. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
11. Kuroki, T., Maeda, Y., Kudo, S.E., Ogata, N., Takabayashi, K., Takenaka, K., Kawashima, J., Kawabata, Y., Iwasaki, S., Shiina, O., et al.: Combination of white-light imaging–based and narrow-band imaging–based artificial intelligence models during colonoscopy in patients with ulcerative colitis. Journal of Crohn's and Colitis **19**(1), jjaf014 (2025)
12. Liu, R., Zhou, Y., Wang, R., Chen, X., Yang, Y., Jiang, H., Xie, K., Ning, Y., Deng, Y., Yu, Q., et al.: Multimodal deep learning for laryngeal squamous cell carcinoma staging using ct and laryngoscopy. European Radiology pp. 1–12 (2026)
13. Ni, X., He, S., Xu, Z., Gao, L., Lu, N., Yuan, Z., Lai, S., Zhang, Y., Yi, J., Wang, X., et al.: Endoscopic diagnosis of laryngeal cancer and precancerous lesions by narrow band imaging. The Journal of Laryngology & Otology **125**(3), 288–296 (2011)
14. Nie, X., Zhang, X., Wang, D., Liu, Y., Xing, L., Liu, W.: Laryngeal cancer diagnosis based on improved yolov8 algorithm. Machine Learning: Science and Technology **6**(1), 015011 (2025)
15. Sapkota, R., Karkee, M.: Ultralytics yolo evolution: An overview of yolo26, yolo11, yolov8 and yolov5 object detectors for computer vision and pattern recognition. arXiv preprint arXiv:2510.09653 (2025)
16. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
17. Wellenstein, D.J., Woodburn, J., Marres, H.A., van den Broek, G.B.: Detection of laryngeal carcinoma during endoscopy using artificial intelligence. Head & Neck **45**(9), 2217–2226 (2023)
18. Xiong, H., Lin, P., Yu, J.G., Ye, J., Xiao, L., Tao, Y., Jiang, Z., Lin, W., Liu, M., Xu, J., et al.: Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. EBioMedicine **48**, 92–99 (2019)
19. Xu, W., Huang, H., Gong, Y., Yu, L., Wu, Q., Zhang, J.: Hierarchical multi-prototype discrimination: Boosting support-query matching for few-shot segmentation. IEEE Transactions on Multimedia (2025)
20. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations
21. Zhang, Y., Li, K., Shi, J., Zhang, H.: A multi-orthogonal prototype contrastive learning framework with feature compression for imbalanced skin lesion classification. Applied Soft Computing p. 114342 (2025)
22. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2024)