RESTRICTED PYRAMID GRAPH LEARNING FOR LARYNGOPHARYNGEAL TUMOR DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate detection of laryngo-pharyngeal tumors from endoscopic images is critical for early diagnosis, yet remains challenging because lesions are often small, low-contrast, and visually similar to surrounding mucosa or imaging artifacts. Classifying laryngo-pharyngeal endoscopic images requires both local lesion detail and surrounding anatomical context. Lesion-guided methods provide a region of interest, but their encoders usually process subsequent feature maps without an explicit spatial relation between coarse and fine evidence. As a result, background responses re-enter during fine-scale refinement, contaminating attention and obscuring subtle lesion boundaries. To bridge this gap, we present RPGNet, which combines mask-supervised dual-view learning with coarse-to-fine feature routing. An adapted SAM2 generates lesion-focused views and provides spatial supervision for attention learning, while a shared GreedyViG encoder extracts complementary global and local representations without feeding lesion masks to the classifier. A Restricted-Pyramid Network adaptively allocates the retained feature budget according to regional complexity, preserving informative details while suppressing redundant background. On the internal FAHSYSU test set, RPGNet reaches 94.85% accuracy and 92.55% macro-F1. On three held-out external cohorts, its macro-F1 exceeds the strongest result in the comparison table by 3.41, 3.38, and 9.58 percentage points on SAHSYSU, NHSMU, and FAHSU, respectively. These findings support spatially constrained multiscale selection while motivating further work on minority-class recognition. Code is available at `https://anonymous.4open.science/r/Q7M4V9K2D8R6X3TP/`.

1 INTRODUCTION

Laryngo-pharyngeal tumors (LPT) occur in anatomically critical regions responsible for breathing, swallowing, and phonation, making early and accurate diagnosis essential for patient survival and quality of life (Mattavelli et al., 2024). Computer-assisted analysis of laryngoscopic images is therefore of great clinical value, yet it remains highly challenging. Early lesions are often small, low-contrast, and visually similar to mucosal folds or inflammation; diagnostically critical boundaries and textures are intertwined with surrounding anatomical context; and variations in endoscopic devices and acquisition protocols introduce background patterns spuriously correlated with disease labels. A robust model must consequently balance lesion focus, contextual preservation, and cross-center stability.

Existing approaches address these challenges along three main directions. First, endoscopic pretraining and multi-view representation learning improve data efficiency and capture global-local information (Boers et al., 2024; Hu et al., 2024; Liu et al., 2025b; Lin et al., 2025). However, they primarily optimize image- or video-level representations and do not explicitly control which lesion subregions are retained across feature scales. Second, lesion-prior-guided methods employ segmentation priors, attention, or multi-task supervision to localize lesions (Ma et al., 2024; Ling et al., 2024; Li et al., 2024; Wei et al., 2024; 2026; Fontanella et al., 2023). While these approaches effectively narrow the search space, localization is typically followed by conventional hierarchical encoding along fixed stage-wise paths. Third, efficient hierarchical, sparse, or graph-based models reduce redundancy or improve long-range interaction (Norouzi et al., 2024; Lou et al., 2025; Munir et al., 2024; Zhu et al., 2024; Liu et al., 2023), but their selection rules remain task-agnostic and are not conditioned on lesion evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These directions leave a common and critical gap: *localization identifies **where** a lesion is, but rarely constrains **which** hierarchical features should be selected for subsequent refinement.* As illustrated in Figure 1, once a lesion crop or attention map is obtained, propagating the full candidate region through a fixed network hierarchy allows background responses to re-enter at finer stages, causing attention contamination. Global pooling can further mix diagnostically important boundaries with weakly relevant tissue. This problem is particularly pronounced for small or heterogeneous laryngopharyngeal lesions, where only a subset of the candidate region provides reliable evidence.

To bridge this gap, we propose RPGNet, a lesion-guided framework that makes localization continuously constrain hierarchical feature selection. An adapted SAM2 first estimates lesion support and constructs a lesion-focused RGB view. During training, SAM2-derived masks supervise the spatial attention head on the global view only; the local branch receives no mask-attention supervision. The estimated masks are used to construct the lesion-focused local RGB view, but are not concatenated with either view or supplied as classifier channels. No manually annotated masks are required at inference. The two views are encoded by a shared GreedyViG-B backbone (Munir et al., 2024), which captures spatially separated lesion structures and provides a multi-resolution feature pyramid while avoiding the repeated k-nearest-neighbor search of conventional vision graph networks.

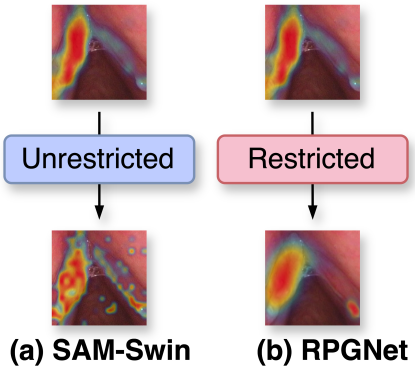


Figure 1: **Motivation of lesion-conditioned routing.** Comparison of attention heatmaps across two adjacent layers. (a) SAM-Swin propagates attention without spatial constraints, causing attention contamination at finer scales. (b) RPGNet restricts inter-layer attention to spatially corresponding regions, maintaining a focused lesion region across scales.

On top of this feature pyramid, we introduce the **Restricted-Pyramid Network** to perform Region-Feature Refinement (RFR). Unlike fixed hierarchical paths, RFR models the global pyramid as a spatial parent-child graph. Starting from the coarsest level, it selects lesion-related regions and restricts the candidates at each finer level to their geometrically corresponding descendants. A Dynamic-k budget further adapts the number of retained features to regional complexity: homogeneous regions retain fewer features, whereas irregular lesions preserve more detailed evidence. Finally, a **Lesion-Aware Fusion** module integrates the restricted-pyramid representation with global and local descriptors via Consistency-Difference Fusion (CDF), capturing their shared and complementary information for classification.

We evaluate RPGNet on the internal FAHSYSU test set and three external cohorts (SAHSYSU, NHSMU, and FAHSU). RPGNet achieves the highest Accuracy and F1-score among all compared methods on all four datasets, and the highest Precision on every external dataset. Relative to the strongest competing score for each external cohort and metric, it improves accuracy by 1.64–2.73, macro precision by 3.05–8.15, and macro-F1 by 3.38–9.58 percentage points. These results support our central hypothesis: lesion localization is most effective when it constrains not only the initial region of interest, but also the subsequent selection and refinement of hierarchical features.

The main contributions of this work are summarized as follows:

- We propose RPGNet, the first framework to progressively constrain the receptive field from coarse to fine layers. By explicitly centering the learning process on lesion-specific information, this design effectively mitigates background contamination and significantly improves detection performance.
- We design the Restricted-Pyramid Network (RPN) to bridge the disconnect between spatial localization and feature selection. By imposing parent-constrained spatial routing over multiscale features, it prevents background contamination from propagating into finer scales.
- We design a Lesion-Aware Fusion module to address the complementarity between global context, local details, and restricted-pyramid evidence. Within this module, a Consistency-

Difference Fusion (CDF) mechanism explicitly models shared and discrepant information across views, thereby yielding a highly discriminative representation for final prediction.

- Extensive experiments on one internal and three external datasets demonstrate that RPGNet achieves superior performance and robust generalization, particularly on unseen external datasets.

2 Related Work

2.1 Endoscopic Representation Learning

Endoscopy-specific pretraining improves data efficiency and transfer relative to generic natural-image initialization (Boers et al., 2024; Wang et al., 2025). M2CRL learns from global and local *video* views using masked modeling and contrastive objectives (Hu et al., 2024); its aim is transferable temporal representation learning rather than lesion-guided image classification. Such pretraining complements spatial control at the downstream classifier. Cross-center changes in illumination, image quality, and acquisition conditions remain substantial sources of performance variation (Jaspers et al., 2024; Kusters et al., 2025).

2.2 Lesion Priors for Classification

Segmentation foundation models provide automatically estimated regions that can guide medical-image analysis (Kirillov et al., 2023; Ma et al., 2024; Wu et al., 2025; Chen et al., 2024; Gong et al., 2024; Hu et al., 2026; Huang et al., 2025; Liang et al., 2025; Mazurowski et al., 2023; Ji et al., 2024). In laryngo-pharyngeal endoscopy, SAM-FNet fuses lesion-focused and global information, and SAM-Swin combines SAM2-derived lesion views with two Swin branches and lesion-aware feature enhancement (Wei et al., 2024; 2026). Multi-task methods similarly use segmentation to supervise discriminative features (Ling et al., 2024; Li et al., 2024). RPGNet retains the benefit of a local view and auxiliary spatial supervision, then adds an explicit spatial rule for selecting finer global features from previously selected coarse locations.

2.3 Hierarchical and Sparse Visual Models

Hierarchical vision models encode broad semantics and spatial detail at different resolutions (Liu et al., 2024; Shi et al., 2024; Hatamizadeh et al., 2023; Hatamizadeh & Kautz, 2025). Token merging and sparse cross-layer connections reduce redundant computation or reuse features (Norouzi et al., 2024; Lou et al., 2025; Bolya et al., 2023; Duggal et al., 2025); their selection criteria are generally task-independent. GreedyViG constructs efficient within-scale spatial relations without repeated global nearest-neighbor graph search (Munir et al., 2024). We use it as a shared encoder and organize its global multiscale outputs into a parent-child pyramid. The routing restriction is applied across scales and is learned on features trained with lesion-mask supervision. This differs from generic token reduction in both the spatial candidate rule and its clinical training signal.

3 Methodology

3.1 Overview

Given an endoscopic image X, the SAM2-GLLM procedure of Wei et al. (2026) estimates a region mask X^m and constructs a lesion-focused RGB view X^l. The estimated mask is used as auxiliary supervision during training and to construct the local view; no manually annotated mask is required at inference. A shared GreedyViG-B encoder (Munir et al., 2024) processes X and X^l. It produces global features $\{F_i^g\}_{i=1}^4$ and a deep local feature F_4^l, where $i = 1$ is the finest level and $i = 4$ the coarsest. Only the global pyramid enters the Restricted-Pyramid Network (Figure 2). Its four routed readouts form F^{tr}, which is fused with global and local descriptors for three-class prediction.

The two sources of spatial guidance have distinct roles. The estimated mask supervises the global attention head; it is not concatenated with classifier features. The routing module instead uses learned feature scores and the parent-child geometry of the pyramid to determine which finer locations may

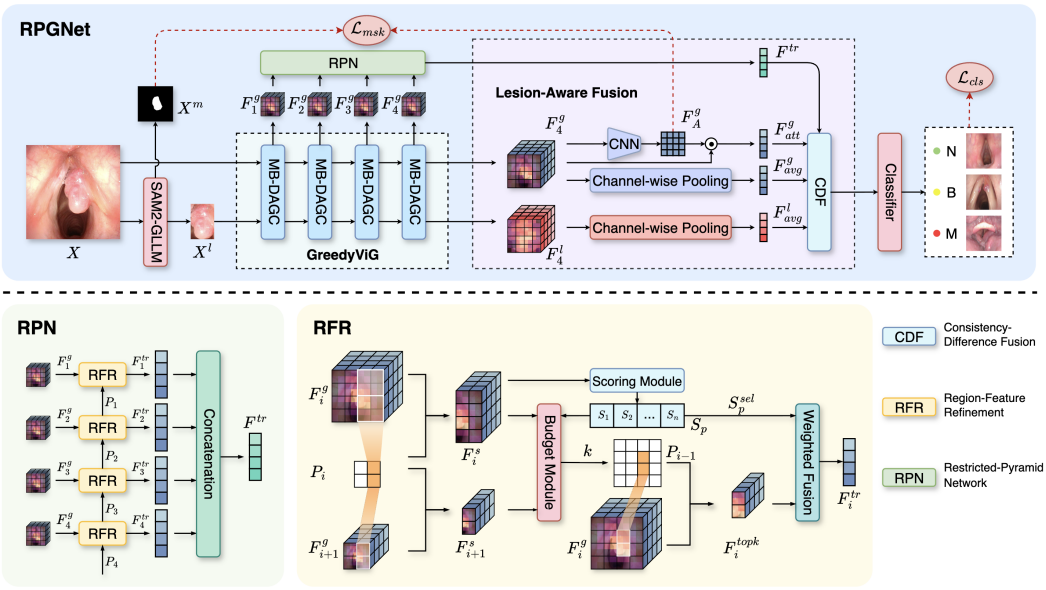


Figure 2: **Overview of RPGNet.** SAM2-GLLM produces a lesion mask X^m and a lesion-focused view X^l. A shared GreedyViG encoder extracts global pyramid features $\{F_i^g\}$ and a local feature F_4^l. The Restricted-Pyramid Network performs coarse-to-fine selection via Region-Feature Refinement (RFR), where a Scoring Module and a Budget Module decide which and how many descendants to retain, yielding F^{tr}. Lesion-Aware Fusion then combines F^{tr} with global and local descriptors via Consistency-Difference Fusion (CDF) for classification.

contribute. This separation makes the mechanism explicit without assuming that the segmentation mask directly gates every level.

3.2 Restricted-Pyramid Network

Let $F_i^g \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ and let $T_i(v) \in \mathbb{R}^D$ denote a projected token at location v of grid $\mathcal{G}_i$. Adjacent grids differ by a factor of two along each spatial axis, so each node $p \in \mathcal{G}_{i+1}$ has four children $\mathcal{C}_i(p) \subset \mathcal{G}_i$. The coarsest level starts with the full grid as one candidate region. At each subsequent level, only children of retained parents are considered. If P_{i+1} is the selected parent set, Region-Feature Refinement (RFR) constructs

$$P_i = \bigcup_{p\in P_{i+1}} \big\{v \in \mathcal{C}_i(p) : g_{i,p,v} = 1\big\}, \tag{1}$$

where $g_{i,p,v}$ is a binary retention gate. The union establishes spatial correspondence between the coarse and fine evidence; unvisited fine locations do not enter the routed readout.

For a candidate region $\mathcal{C}_i(p)$, a scoring head predicts logits $s_{i,p,v}$. A separate budget head receives the parent token $c_{i,p}$ and the mean and maximum of the candidate tokens. It predicts a distribution over a discrete set of allowed budgets:

$$\begin{aligned} \xi_{i,p} &= \big[c_{i,p}; \mathrm{mean}_{v\in\mathcal{C}_i(p)}\, T_i(v); \max_{v\in\mathcal{C}_i(p)} T_i(v)\big], \\ \pi_{i,p} &= \mathrm{softmax}\big(\phi_i^{\mathrm{bud}}(\xi_{i,p})\big), \qquad k_{i,p} = \mathcal{B}_i\big[\arg\max_q \pi_{i,p,q}\big]. \end{aligned} \tag{2}$$

The root uses $\mathcal{B}_4 = \{4, 8, 12, 16, 24, 32\}$; each finer parent uses $\mathcal{B}_i = \{0, 1, 2, 3, 4\}$. The gate keeps the $k_{i,p}$ highest-scoring candidates. A ranked straight-through surrogate supplies gradients to this discrete operation during training (Appendix 6.3). The budget is input dependent; we do not assume an explicit lesion-complexity target.

Retained tokens receive within-region weights $\omega_{i,p,v} = \mathrm{softmax}_{u\in\mathcal{C}_i(p)}(s_{i,p,u})_v$. We average duplicate contributions if a location is reached more than once, set the weight of an unvisited location

to zero, and normalize the remaining weights across the level:

$$\tilde{a}_i(v) = \mathrm{mean}_{p:v\in\mathcal{C}_i(p)}\big(g_{i,p,v}\omega_{i,p,v}\big), \quad a_i(v) = \frac{\tilde{a}_i(v)}{\sum_{u\in\mathcal{G}_i}\tilde{a}_i(u)+\epsilon}, \quad F_i^{\mathrm{tr}} = \sum_{v\in\mathcal{G}_i} a_i(v)T_i(v). \tag{3}$$

The four readouts are concatenated as $F^{\mathrm{tr}} = [F_4^{\mathrm{tr}}; F_3^{\mathrm{tr}}; F_2^{\mathrm{tr}}; F_1^{\mathrm{tr}}]$. This representation retains selected coarse semantics and their corresponding fine structure while limiting the fine-level candidate set.

3.3 Lesion-Aware Fusion

The deep global feature retains anatomical context, whereas the local view magnifies a candidate lesion region. We extract global average-pooled descriptors $F^g_{\mathrm{avg}} = \mathrm{GAP}(F^g_4)$ and $F^l_{\mathrm{avg}} = \mathrm{GAP}(F^l_4)$. A convolutional head produces a spatial attention map on the global feature, yielding

$$F^g_A = \mathrm{sigmoid}\big(\mathrm{CNN}_{\mathrm{att}}(F^g_4)\big), \qquad F^g_{\mathrm{att}} = \frac{\sum_v F^g_A(v)F^g_4(v)}{\sum_v F^g_A(v)+\epsilon}. \tag{4}$$

A learned scalar $\alpha = \mathrm{sigmoid}(\eta)$ mixes the two global summaries:

$$F^g = (1-\alpha)F^g_{\mathrm{avg}} + \alpha F^g_{\mathrm{att}}. \tag{5}$$

Consistency-Difference Fusion (CDF) then represents the views jointly through their original descriptors, absolute difference, and elementwise product, together with the routed pyramid:

$$z = \big[F^g; F^l_{\mathrm{avg}}; |F^g - F^l_{\mathrm{avg}}|; F^g \odot F^l_{\mathrm{avg}}; F^{\mathrm{tr}}\big]. \tag{6}$$

The final prediction is

$$p(y=c \mid X) = \big[\mathrm{softmax}(W_{\mathrm{cls}}z + b_{\mathrm{cls}})\big]_c, \qquad c \in \{\mathrm{normal}, \mathrm{benign}, \mathrm{malignant}\}. \tag{7}$$

3.4 Training Objective

For image label y, label-smoothed target $q_{y,c}$, and class weight w_c, the classification loss is

$$\mathcal{L}_{\mathrm{cls}} = -\sum_c w_c q_{y,c} \log p(y=c \mid X). \tag{8}$$

The estimated mask X^m is resized to the global attention grid, giving $M = \mathrm{Resize}(X^m, H_4, W_4)$. It supervises F^g_A with binary cross-entropy:

$$\mathcal{L}_{\mathrm{msk}} = -\frac{1}{|\mathcal{G}_4|}\sum_{v\in\mathcal{G}_4}\left[M(v)\log F^g_A(v) + (1-M(v))\log(1-F^g_A(v))\right]. \tag{9}$$

The first stage trains the dual-view classifier without F^{tr} using $\mathcal{L}_{\mathrm{cls}} + 0.8\mathcal{L}_{\mathrm{msk}}$. The second stage attaches the routing representation and minimizes $\mathcal{L}_{\mathrm{cls}}$ for the final classifier, with the shared encoder and view descriptors frozen. Implementation details are in Appendix 6.1.

4 Experiments

4.1 Datasets

We evaluate three-class classification on 39,590 endoscopic images from four centers. FAHSYSU contributes 25,256 images for training, validation, and internal testing. Its splits are made at the subject level, with all images from a subject assigned to the same split. SAHSYSU (2,788 images), NHSMU (6,684), and FAHSU (4,862) are held out as external test cohorts. FAHSYSU and SAHSYSU include white-light imaging (WLI) and narrow-band imaging (NBI), whereas NHSMU and FAHSU contain WLI images only. The per-class and imaging-modality counts for all four cohorts are summarized in Table 1. Evaluation metrics are computed per image.

Table 1: Dataset statistics. FAHSYSU is used for training, validation, and internal testing; SAHSYSU, NHSMU, and FAHSU are external test sets.

	FAHSYSU		SAHSYSU		NHSMU		FAHSU	
	NBI	WLI	NBI	WLI	NBI	WLI	NBI	WLI
Normal	695	7,310	9	2,186	0	5,716	0	4,374
Benign	1,488	3,332	27	234	0	592	0	447
Malignant	5,954	6,477	99	233	0	376	0	41
Total	8,137	17,119	135	2,653	0	6,684	0	4,862

4.2 Baselines and Evaluation Protocol

We compare RPGNet with general image classifiers ResNet101, ViT, and SwinV2 (He et al., 2016; Dosovitskiy et al., 2021; Liu et al., 2022); medical-image methods DKA, LCA, and ASMIL (Zhu et al., 2026; Liu et al., 2025a; Ye et al., 2026); and the lesion-guided endoscopic classifier SAM-Swin (Wei et al., 2026). The comparator scores come from reruns on the same subject-disjoint FAHSYSU split and evaluation on the same external cohorts, rather than from numbers reported in the cited papers. RPGNet alone uses the shared GreedyViG-B encoder. RPGNet is trained in two stages: 30 epochs of dual-view learning, followed by 15 epochs of routing adaptation with the encoder and view descriptors frozen. We select its checkpoint using FAHSYSU validation macro-F1; optimization settings are given in Appendix 6.1. We report image-level accuracy, macro precision, macro recall, macro-F1, and recall for each class. Macro metrics weight the three classes equally, whereas accuracy follows the observed class distribution. Definitions of the evaluation metrics are provided in Appendix 6.2.

4.3 Comparative Experiment Results

Table 2 reports image-level results for RPGNet and seven comparators on the internal test split and three external cohorts. Each gain below is measured against the strongest non-RPGNet score for the same cohort and metric; the comparator therefore differs across some columns.

Internal FAHSYSU. RPGNet attains 94.85% accuracy, 93.96% macro precision, 91.54% macro recall, and 92.55% macro-F1. SAM-Swin is the strongest comparator on all four aggregate metrics, with RPGNet ahead by 0.33, 0.21, 0.33, and 0.25 percentage points, respectively. Normal and malignant recall also increase slightly, from 96.54% to 97.46% and from 98.35% to 98.43%. Benign recall remains tied at 78.73%, so the internal aggregate gains do not extend to that class.

External SAHSYSU. RPGNet reaches 94.48% accuracy and 88.68% macro-F1, exceeding ASMIL, the strongest comparator for those metrics, by 1.73 and 3.41 points. Its 87.40% macro precision is 3.05 points above ResNet101, while its 90.11% macro recall is 1.72 points above SAM-Swin. RPGNet also has the highest normal, benign, and malignant recalls at 96.26%, 80.08%, and 93.98%, respectively. These class-wise leads are narrower: 0.13 points over LCA for normal, 0.39 over SAM-Swin for benign, and 0.61 over ASMIL for malignant recall.

External NHSMU. RPGNet records 94.70% accuracy, 85.61% macro precision, 88.93% macro recall, and 87.21% macro-F1. Relative to SAM-Swin, the strongest comparator on each aggregate metric, these are gains of 1.64, 5.36, 0.73, and 3.38 points, respectively. RPGNet achieves the highest normal recall (96.69%) and benign recall (75.68%), but its malignant recall is 94.41%, 3.46 points below SAM-Swin's 97.87%. The higher aggregate scores therefore do not imply an improvement for every class.

External FAHSU. This cohort contains 4,374 normal, 447 benign, and 41 malignant images (Table 1). RPGNet reaches 87.84% accuracy, 2.73 points above DKA, and 54.04% macro precision and 56.35% macro-F1, 8.15 and 9.58 points above SAM-Swin. Its 73.74% macro recall exceeds SAM-Swin by only 0.59 points. Normal recall is highest at 93.92%, and both RPGNet and SAM-Swin attain 100.00% malignant recall. However, RPGNet's benign recall is 27.29%, 6.94 points below SAM-Swin's 34.23%; this class remains difficult despite the aggregate gains.

Table 2: Performance comparison on the FAHSYSU internal test split and three external cohorts (%). Precision, recall, and F1-score are macro-averaged; the last three columns report class-wise recall. Blue and green shading mark the highest and second-highest distinct values within each cohort and metric, with ties sharing the highest shading.

Dataset	Method	Overall results				Recall for different classes		
		Accuracy	Precision	Recall	F1-score	Normal	Benign	Malignant
FAHSYSU	ViT Dosovitskiy et al. (2021)	89.71	86.74	85.40	85.99	91.58	69.64	94.99
	SwinV2 Liu et al. (2022)	88.99	84.89	85.56	85.15	93.99	70.98	91.71
	ResNet101 He et al. (2016)	91.12	89.00	86.71	87.68	94.52	69.72	95.89
	DKA Zhu et al. (2026)	86.40	81.89	81.09	81.31	92.06	59.72	91.49
	LCA Liu et al. (2025a)	90.47	87.11	86.77	86.85	95.70	71.14	93.47
	ASMIL Ye et al. (2026)	90.65	87.57	87.61	87.58	93.55	75.56	93.71
	SAM-Swin Wei et al. (2026)	94.52	93.75	91.21	92.30	96.54	78.73	98.35
	RPGNet	**94.85**	**93.96**	**91.54**	**92.55**	**97.46**	**78.73**	**98.43**
SAHSYSU	ViT Dosovitskiy et al. (2021)	90.53	78.81	81.58	80.12	94.31	63.98	86.45
	SwinV2 Liu et al. (2022)	90.24	78.78	81.22	79.80	94.17	66.67	82.83
	ResNet101 He et al. (2016)	92.72	84.35	85.00	84.51	95.85	66.67	92.47
	DKA Zhu et al. (2026)	88.74	76.84	78.81	77.60	93.03	62.07	81.33
	LCA Liu et al. (2025a)	92.36	83.36	83.07	83.16	96.13	60.92	92.17
	ASMIL Ye et al. (2026)	92.75	83.19	87.63	85.27	94.81	74.71	93.37
	SAM-Swin Wei et al. (2026)	91.28	80.14	88.39	83.63	92.39	79.69	93.07
	RPGNet	**94.48**	**87.40**	**90.11**	**88.68**	**96.26**	**80.08**	**93.98**
NHSMU	ViT Dosovitskiy et al. (2021)	91.65	77.84	86.24	81.56	93.67	70.10	94.95
	SwinV2 Liu et al. (2022)	88.69	70.02	85.83	76.03	89.85	74.83	92.82
	ResNet101 He et al. (2016)	92.82	78.73	86.93	82.23	95.05	68.41	97.34
	DKA Zhu et al. (2026)	86.94	65.27	75.14	68.16	91.17	44.09	90.16
	LCA Liu et al. (2025a)	90.74	73.50	81.26	76.41	94.12	56.59	93.09
	ASMIL Ye et al. (2026)	92.89	80.12	88.05	83.69	94.72	73.14	96.28
	SAM-Swin Wei et al. (2026)	93.06	80.25	88.20	83.83	94.94	71.79	**97.87**
	RPGNet	**94.70**	**85.61**	**88.93**	**87.21**	**96.69**	**75.68**	94.41
FAHSU	ViT Dosovitskiy et al. (2021)	74.19	39.75	57.74	38.80	79.45	23.04	70.73
	SwinV2 Liu et al. (2022)	71.62	37.75	59.81	39.13	76.57	22.37	80.49
	ResNet101 He et al. (2016)	70.36	38.82	59.41	36.49	75.65	17.23	85.37
	DKA Zhu et al. (2026)	85.11	44.61	64.68	45.81	92.36	13.87	87.80
	LCA Liu et al. (2025a)	75.32	40.38	67.89	41.85	79.90	28.64	95.12
	ASMIL Ye et al. (2026)	72.93	38.35	61.91	38.36	77.98	22.37	85.37
	SAM-Swin Wei et al. (2026)	80.67	45.89	73.15	46.77	85.23	**34.23**	100.00
	RPGNet	**87.84**	**54.04**	**73.74**	**56.35**	**93.92**	27.29	**100.00**

***Note**: denotes the highest value, and denotes the second-highest distinct value in each dataset and column.*

Across the three external cohorts, RPGNet leads all four aggregate metrics, with larger gains in macro precision and macro-F1 than in macro recall. The malignant-recall deficit on NHSMU and benign-recall deficit on FAHSU limit any claim of uniform class-wise improvement. These single-run, image-level results describe the evaluated cohorts but do not quantify variability across subjects or new centers.

4.4 Ablation Study

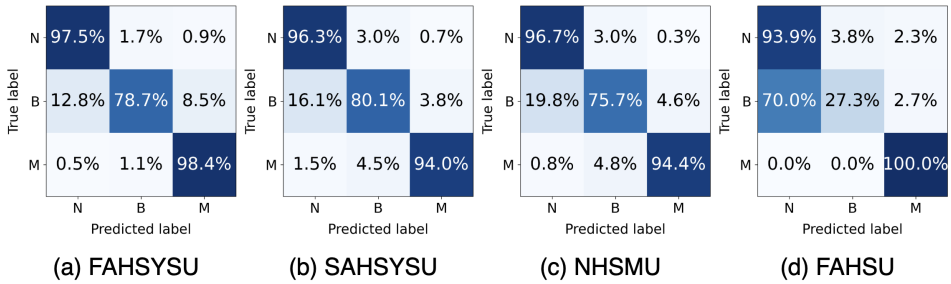
Table 3 evaluates Consistency-Difference Fusion (CDF), auxiliary mask supervision, and the Restricted-Pyramid Network (RPN) on the same four cohorts. M1 is the shared-encoder dual-view baseline; M2 and M3 add CDF and mask supervision separately, M4 combines them, and the full RPGNet adds RPN with Dynamic-k.

The isolated components have different effects. On FAHSYSU, CDF alone reduces macro-F1 from 88.36% to 85.45%, whereas mask supervision alone raises it to 89.89%. The latter also lowers FAHSU macro-F1 from 49.59% to 45.82%. Combining both components in M4 improves on M1 by 2.70, 3.01, 0.46, and 0.59 macro-F1 points across FAHSYSU, SAHSYSU, NHSMU, and FAHSU, although its 83.93% on NHSMU remains below M3's 84.27%.

Table 3: Ablation of CDF, mask-attention loss, and RPN on four cohorts. Entries are accuracy and macro-F1 (%); the full model adds RPN with Dynamic-k to M4.

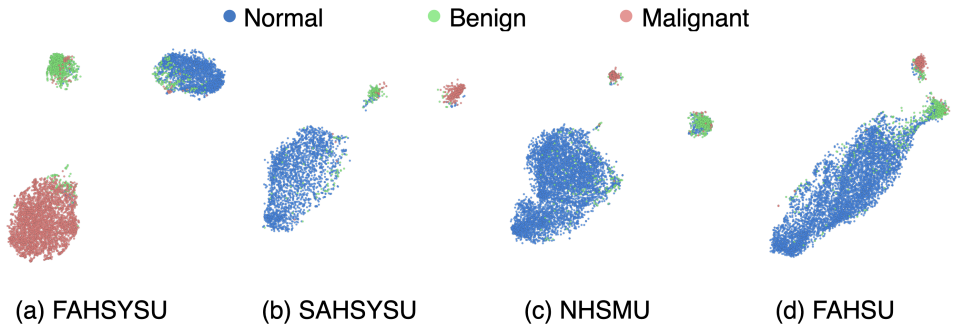
Variant	CDF	L_{msk}	RPN	FAHSYSU		SAHSYSU		NHSMU		FAHSU	
				Acc	F1	Acc	F1	Acc	F1	Acc	F1
M1				91.58	88.36	91.61	83.31	93.31	83.47	82.52	49.59
M2	✓			90.07	85.45	92.54	84.70	93.82	82.82	85.36	48.15
M3		✓		92.56	89.89	92.83	84.75	93.15	84.27	81.04	45.82
M4	✓	✓		93.66	91.06	93.54	86.32	93.16	83.93	85.99	50.18
RPGNet	✓	✓	✓	**94.85**	**92.55**	**94.48**	**88.68**	**94.70**	**87.21**	**87.84**	**56.35**

Note: *indicates the best performance (bold), and* *indicates the second-best performance.*



(a) FAHSYSU (b) SAHSYSU (c) NHSMU (d) FAHSU

Figure 3: Row-normalized confusion matrices for FAHSYSU, SAHSYSU, NHSMU, and FAHSU (N, normal; B, benign; M, malignant).



(a) FAHSYSU (b) SAHSYSU (c) NHSMU (d) FAHSU

Figure 4: Two-dimensional t-SNE projections for FAHSYSU, SAHSYSU, NHSMU, and FAHSU. Blue, green, and red points denote normal, benign, and malignant images.

Adding RPN with Dynamic-k to M4 raises macro-F1 by 1.49, 2.36, 3.28, and 6.17 points across the four cohorts, yielding the best result in each column. This comparison supports the combined routing design. The table does not include a fixed-budget RPN or an RPN-only variant, so it cannot isolate the effect of Dynamic-k from the rest of the routing module.

4.5 Visualization and Interpretability

Class-wise error patterns. Figure 3 shows normal and malignant recall above 93% in all four cohorts. Benign recall is 78.7%, 80.1%, and 75.7% on FAHSYSU, SAHSYSU, and NHSMU, but falls to 27.3% on FAHSU. In the latter cohort, 70.0% of benign images are classified as normal, although all 41 malignant images are classified correctly. Aggregate accuracy therefore obscures the difficulty of benign recognition.

Feature-space organization. In Figure 4, normal and malignant points form largely separate groups on FAHSYSU, with benign points near both. SAHSYSU and NHSMU show broad normal clouds and smaller peripheral groups. On FAHSU, the normal cloud approaches a benign-rich group, con-

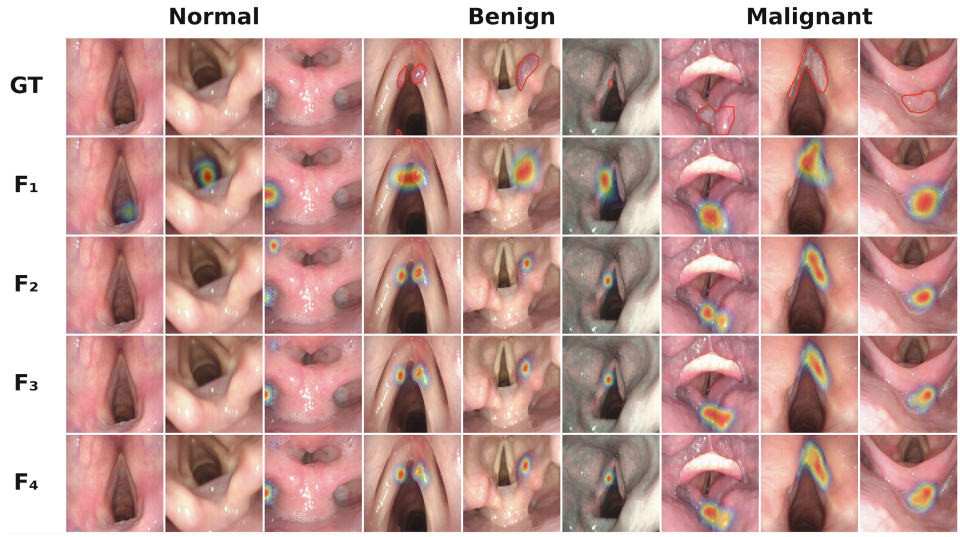


Figure 5: Selected normal, benign, and malignant examples. The GT row shows reference images and red lesion contours where available; rows F_1–F_4 show activation overlays at four feature stages.

sistent with the benign-to-normal errors in Figure 3. These projections are qualitative; the held-out performance is reported in Table 2.

Stage-wise activation. Figure 5 shows two examples per class across F_1–F_4. In the lesion examples, broad F_1 responses become more concentrated near the marked regions in later rows; early hotspots in the normal examples become less apparent. These selected cases illustrate stage-wise behavior without quantifying localization quality across the datasets.

Mask-loss sensitivity. Figure 6 shows that all four metrics improve overall from $\lambda_{\text{mask}} = 0.1$ to 0.8, where they reach their highest displayed values. They decline at 0.9 and partially recover at 1.0, so larger weights do not yield a monotonic gain in this sweep.

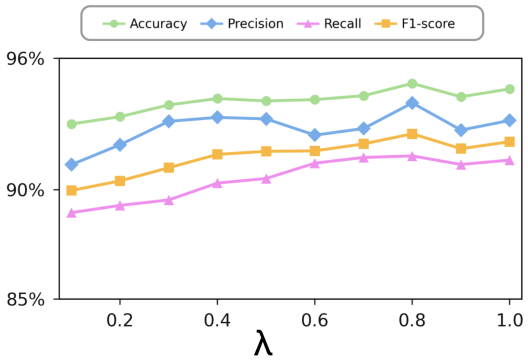


Figure 6: Sensitivity to the mask-loss coefficient λ_{mask} from 0.1 to 1.0.

5 Conclusion

RPGNet combines lesion-focused dual-view learning with spatially restricted routing through a global feature pyramid. An estimated lesion mask supervises global attention during initial training; at each routing level, selected coarse nodes delimit the candidate fine locations, and an input-dependent budget determines how many to retain. Consistency-Difference Fusion combines these routed features with whole-image and local descriptors. On the FAHSYSU internal test set, RPGNet attains 94.85% accuracy and 92.55% macro-F1. It also records the highest accuracy and macro-F1 in the comparison table on the three held-out cohorts, with macro-F1 gains of 3.41, 3.38, and 9.58 percentage points over the strongest competing scores on SAHSYSU, NHSMU, and FAHSU. The ablation supports the combined RPN and Dynamic-k addition, but does not separate their individual effects. Despite the aggregate gains, FAHSU benign recall remains 27.29%, below SAM-Swin's 34.23%. Better minority-class recognition, repeated-run uncertainty estimates, and an isolated test of adaptive budgets are the next priorities before clinical deployment claims can be made.

USAGE OF LLMS

We used large language models (LLMs) solely for polishing the writing, such as improving sentence fluency and fixing grammar issues. No content generation, idea creation, or experimental analysis was conducted by LLMs.

ETHICS STATEMENT

This study uses previously released research datasets and involves no new patient recruitment or intervention. Dataset-specific access conditions and de-identification follow the source releases.

REPRODUCIBILITY STATEMENT

The model architecture is introduced in detail with equations and figures in the main text. All the implementation details are included in the Appendix 6.1, including dataset descriptions, metrics of each task, model configurations, and experiment settings. Code is available at this repository: `https://anonymous.4open.science/r/Q7M4V9K2D8R6X3TP/`

REFERENCES

Tim G. W. Boers, Kiki N. Fockens, Joost A. van der Putten, Tim J. M. Jaspers, Carolus H. J. Kusters, Jelmer B. Jukema, Martijn R. Jong, Maarten R. Struyvenberg, Jeroen de Groof, Jacques J. Bergman, Peter H. N. de With, and Fons van der Sommen. Foundation models in gastrointestinal endoscopic AI: Impact of architecture, pre-training approach and data efficiency. *Medical Image Analysis*, 98:103298, 2024. doi: 10.1016/j.media.2024.103298. URL `https://doi.org/10.1016/j.media.2024.103298`.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL `https://iclr.cc/virtual/2023/poster/12173`.

Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, and Quanzheng Li. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024. doi: 10.1016/j.media.2024.103310. URL `https://doi.org/10.1016/j.media.2024.103310`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Shivam Duggal, Phillip Isola, Antonio Torralba, and William Freeman. Adaptive length image tokenization via recurrent allocation. In *International Conference on Learning Representations*, 2025. URL `https://iclr.cc/virtual/2025/poster/28458`.

Alessandro Fontanella, Antreas Antoniou, Wenwen Li, Joanna Wardlaw, Grant Mair, Emanuele Trucco, and Amos Storkey. ACAT: Adversarial counterfactual attention for classification and detection in medical imaging. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10153–10169. PMLR, 2023. URL `https://proceedings.mlr.press/v202/fontanella23a.html`.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024. doi: 10.1016/j.media.2024.103324. URL `https://doi.org/10.1016/j.media.2024.103324`.

Ali Hatamizadeh and Jan Kautz. MambaVision: A hybrid mamba-transformer vision backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.

25261–25270. IEEE/CVF, 2025. doi: 10.1109/CVPR52734.2025.02352. URL https://doi.org/10.1109/CVPR52734.2025.02352.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12633–12646. PMLR, 2023. URL https://proceedings.mlr.press/v202/hatamizadeh23a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Jihong Hu, Yinhao Li, Rahul Kumar Jain, Lanfen Lin, and Yen-Wei Chen. Spa: Leveraging the sam with spatial priors adapter for enhanced medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 30(2):993–1005, 2026. doi: 10.1109/JBHI.2025.3526174. URL https://doi.org/10.1109/JBHI.2025.3526174.

Kai Hu, Ye Xiao, Yuan Zhang, and Xieping Gao. Multi-view masked contrastive representation learning for endoscopic video analysis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47987–48014. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-1521. URL https://doi.org/10.52202/079017-1521.

Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, Chen Gong, and Dong Liang. Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(5):2295–2306, 2025. doi: 10.1109/TMI.2025.3530097. URL https://doi.org/10.1109/TMI.2025.3530097.

Tim J. M. Jaspers, Tim G. W. Boers, Carolus H. J. Kusters, Martijn R. Jong, Jelmer B. Jukema, Albert J. de Groof, Jacques Bergman, Peter H. N. de With, and Fons van der Sommen. Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies. *Medical Image Analysis*, 94:103157, 2024. doi: 10.1016/j.media.2024.103157. URL https://doi.org/10.1016/j.media.2024.103157.

Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, 21(4):617–630, 2024. doi: 10.1007/s11633-023-1385-0. URL https://doi.org/10.1007/s11633-023-1385-0.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003. IEEE/CVF, 2023. doi: 10.1109/ICCV51070.2023.00371. URL https://doi.org/10.1109/ICCV51070.2023.00371.

Carolus H. J. Kusters, Tim J. M. Jaspers, Tim G. W. Boers, Martijn R. Jong, Jelmer B. Jukema, Kiki N. Fockens, Albert J. de Groof, Jacques Bergman, Fons van der Sommen, and Peter H. N. de With. Will transformers change gastrointestinal endoscopic image analysis? a comparative analysis between CNNs and transformers, in terms of performance, robustness and generalization. *Medical Image Analysis*, 99:103348, 2025. doi: 10.1016/j.media.2024.103348. URL https://doi.org/10.1016/j.media.2024.103348.

Jiajia Li, Pingping Zhang, Xia Yang, Lei Zhu, Teng Wang, Ping Zhang, Ruhan Liu, Bin Sheng, and Kaixuan Wang. SSM-Net: Semi-supervised multi-task network for joint lesion segmentation and classification from pancreatic EUS images. *Artificial Intelligence in Medicine*, 154:102919, 2024. doi: 10.1016/j.artmed.2024.102919. URL https://doi.org/10.1016/j.artmed.2024.102919.

Pengchen Liang, Leijun Shi, Bin Pu, Renkai Wu, Jianguo Chen, Lixin Zhou, Lite Xu, Zhuangzhuang Chen, Qing Chang, and Yiwei Li. Mambasam: A visual mamba-adapted sam framework for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 29(8):5824–5835, 2025. doi: 10.1109/JBHI.2025.3544548. URL https://doi.org/10.1109/JBHI.2025.3544548.

Yuena Lin, Yiyuan Wang, Gengyu Lyu, Yongjian Deng, Haichun Cai, Huibin Lin, Haobo Wang, and Zhen Yang. Enhance multi-view classification through multi-scale alignment and expanded boundary. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/28081.

Yating Ling, Yuling Wang, Wenli Dai, Jie Yu, Ping Liang, and Dexing Kong. Mtanet: Multi-task attention network for automatic medical image segmentation and classification. *IEEE Transactions on Medical Imaging*, 43(2):674–685, 2024. doi: 10.1109/TMI.2023.3317088. URL https://doi.org/10.1109/TMI.2023.3317088.

Mingzhou Liu, Ching-Wen Lee, Xinwei Sun, Xueqing Yu, Yu Qiao, and Yizhou Wang. Learning causal alignment for reliable disease diagnosis. In *International Conference on Learning Representations*, 2025a. URL https://iclr.cc/virtual/2025/poster/28327.

Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14430. IEEE/CVF, 2023. doi: 10.1109/CVPR52729.2023.01386. URL https://doi.org/10.1109/CVPR52729.2023.01386.

Xueli Liu, Wentao Gong, Xiao Chen, Zhen Li, Yinlong Liu, Li Wang, Quan Liu, Xicai Sun, Xiaofeng Liu, Xinrong Chen, Yuxuan Shi, and Hongmeng Yu. Vision-language foundation model for generalizable nasal disease diagnosis using unlabeled endoscopic records. *Pattern Recognition*, 165:111646, 2025b. doi: 10.1016/j.patcog.2025.111646. URL https://doi.org/10.1016/j.patcog.2025.111646.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-3273. URL https://doi.org/10.52202/079017-3273.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11999–12009. IEEE/CVF, 2022. doi: 10.1109/CVPR52688.2022.01170. URL https://doi.org/10.1109/CVPR52688.2022.01170.

Meng Lou, Yunxiang Fu, and Yizhou Yu. SparX: A sparse cross-layer connection mechanism for hierarchical vision mamba and transformer networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19104–19114. AAAI Press, 2025. doi: 10.1609/aaai.v39i18.34103. URL https://doi.org/10.1609/aaai.v39i18.34103.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. doi: 10.1038/s41467-024-44824-z. URL https://doi.org/10.1038/s41467-024-44824-z.

Davide Mattavelli, Gunnar Wichmann, Davide Smussi, Alberto Paderno, Maria Serrahima Plana, Ricard Nin Mesia, Micaela Compagnoni, Alessandro Medda, Susanna Chiocca, Stefano Calza, Yinxiu Zhan, Carla Rognoni, Rosanna Tarricone, Erika Stucchi, Luigi Lorini, Cristina Gurizzan, Ksenia Khelik, Eivind Hovig, Andreas Dietz, Cesare Piazza, and Paolo Bossi. Is precision medicine the solution to improve organ preservation in laryngeal/hypopharyngeal cancer? a position paper by the preserve research group. *Frontiers in Oncology*, 14:1433333, 2024. doi: 10.3389/fonc.2024.1433333. URL https://doi.org/10.3389/fonc.2024.1433333.

Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918, 2023. doi: 10.1016/j.media.2023.102918. URL https://doi.org/10.1016/j.media.2023.102918.

Mustafa Munir, William Avery, Md. Mostafijur Rahman, and Radu Marculescu. GreedyViG: Dynamic axial graph construction for efficient vision GNNs. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 6118–6127. IEEE/CVF, 2024. doi: 10.1109/CVPR52733.2024.00585. URL https://doi.org/10.1109/CVPR52733.2024.00585.

Narges Norouzi, Svetlana Orlova, Daan De Geus, and Gijs Dubbelman. ALGM: Adaptive local-then-global token merging for efficient semantic segmentation with plain vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15773–15782. IEEE/CVF, 2024. doi: 10.1109/CVPR52733.2024.01493. URL https://doi.org/10.1109/CVPR52733.2024.01493.

Yuheng Shi, Minjing Dong, and Chang Xu. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25687–25708. Neural Information Processing Systems Foundation, Inc., 2024. doi: 10.52202/079017-0808. URL https://doi.org/10.52202/079017-0808.

Zhao Wang, Chang Liu, Lingting Zhu, Tongtong Wang, Shaoting Zhang, and Qi Dou. Improving foundation model for endoscopy video analysis via representation learning on long sequences. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3526–3536, 2025. doi: 10.1109/JBHI.2025.3532311. URL https://doi.org/10.1109/JBHI.2025.3532311.

Jia Wei, Yun Li, Meiyu Qiu, Hongyu Chen, Xiaomao Fan, and Wenbin Lei. SAM-guided fusion network for laryngo-pharyngeal tumor detection. In *2024 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 2667–2672. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10822832. URL https://doi.org/10.1109/BIBM62325.2024.10822832.

Jia Wei, Yun Li, Xiaomao Fan, Wenjun Ma, Meiyu Qiu, Hongyu Chen, and Wenbin Lei. Sam-swin: Sam-driven dual-swin transformers with adaptive lesion enhancement for laryngo-pharyngeal tumor detection. *Medical Image Analysis*, 109:103906, 2026. doi: 10.1016/j.media.2025.103906. URL https://doi.org/10.1016/j.media.2025.103906.

Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 102:103547, 2025. doi: 10.1016/j.media.2025.103547. URL https://doi.org/10.1016/j.media.2025.103547.

Linfeng Ye, Shayan Mohajer Hamidi, Zhixiang Chi, Guang Li, Mert Pilanci, Takahiro Ogawa, Miki Haseyama, and Konstantinos Plataniotis. ASMIL: Attention-stabilized multiple instance learning for whole-slide imaging. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010847.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

Ziquan Zhu, Hanruo Zhu, Si-Yuan Lu, Xiang Li, Yanda Meng, Yunxiao Zhang, Gaojie Jin, Lu Yin, Lijie Hu, Di Wang, Lu Liu, and Tianjin Huang. Dual-kernel adapter: Expanding spatial horizons for data-constrained medical image analysis. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10008841.

6 APPENDIX

6.1 IMPLEMENTATION DETAILS

We initialize the shared GREEDYVIG-B encoder from pretrained weights and disable its optional dual-relation stages. Stage-1 dual-view training uses AdamW for up to 30 epochs with a batch size of 32 and 256×256 inputs. The backbone and newly introduced modules use learning rates of 1.7×10^{-4} and 2.7×10^{-4}, respectively; one warm-up epoch is followed by cosine decay to 7.2×10^{-7}. Weight decay is 0.01. Class weighting is computed as $w_c = \mathrm{clip}((N/(3n_c))^{0.5}, 0.6, 1.8)$, and the mask-attention loss is weighted by 0.8. Training augmentation uses color jitter of strength 0.07 and StyleMix with probability 0.3. The random seed is 42. Hyperparameters and the checkpoint were selected using macro-F1 on the FAHSYSU validation split. The subject-disjoint FAHSYSU test split and the SAHSYSU, NHSMU, and FAHSU external cohorts were not used for hyperparameter tuning or checkpoint selection; they were used only for final evaluation.

Table 4 summarizes the Stage-1 configuration. All reported percentages are computed from saved per-sample predictions; Macro-F1 is the unweighted mean of the three class F1 scores.

Table 4: Stage-1 dual-view hyperparameter settings.

Hyperparameter	Value
Optimizer	AdamW
Backbone learning rate	1.7×10^{-4}
New-module learning rate	2.7×10^{-4}
Batch size / maximum epochs	32 / 30
Input size	256×256
Warm-up / schedule	1 epoch / cosine decay
Minimum learning rate	7.2×10^{-7}
Weight decay	0.01
Label smoothing	0.1
Class-weight exponent	0.5
Mask-attention loss coefficient	0.8
Random seed	42

6.2 EVALUATION METRICS

We report classification results separately for FAHSYSU, SAHSYSU, NHSMU, and FAHSU. Let $\mathcal{C} = \{\text{normal}, \text{benign}, \text{malignant}\}$, and let $M_{c,c'}$ be the number of images with ground-truth class c predicted as c'. For each class c, $\mathrm{TP}_c = M_{c,c}$, $\mathrm{FP}_c = \sum_{c' \in \mathcal{C} \setminus \{c\}} M_{c',c}$, and $\mathrm{FN}_c = \sum_{c' \in \mathcal{C} \setminus \{c\}} M_{c,c'}$.

- **Accuracy** is the proportion of correctly classified images in each dataset:

$$\mathrm{Accuracy} = \frac{\sum_{c \in \mathcal{C}} M_{c,c}}{\sum_{c,c' \in \mathcal{C}} M_{c,c'}}. \tag{10}$$

- **Macro precision** is the unweighted mean of the class-wise precision scores:

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad \text{Macro-Precision} = \frac{1}{3} \sum_{c \in \mathcal{C}} P_c. \tag{11}$$

- **Macro recall** is the unweighted mean of the class-wise recall scores:

$$R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \qquad \text{Macro-Recall} = \frac{1}{3} \sum_{c \in \mathcal{C}} R_c. \tag{12}$$

- **Macro F1-score** is the unweighted mean of the class-wise F1-scores:

$$F1_c = \frac{2 P_c R_c}{P_c + R_c}, \qquad \text{Macro-F1} = \frac{1}{3} \sum_{c \in \mathcal{C}} F1_c. \tag{13}$$

- **Class-wise recall** reports R_{normal}, R_{benign}, and $R_{\text{malignant}}$ separately.

A ratio with a zero denominator is set to zero. All scores are reported as percentages.

6.3 RFR ALGORITHM

Algorithm 1 End-to-end Region-Feature Refinement (RFR)

Require: Global pyramid $\{F_i^g\}_{i=1}^4$, where $F_i^g \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$; temperature τ; adjacent grids satisfy $H_i = 2H_{i+1}$ and $W_i = 2W_{i+1}$ for $i = 1, 2, 3$
Ensure: Restricted-pyramid representation $F^{\text{tr}} \in \mathbb{R}^{B\times 4D}$

```
for i = 1 to 4 do
    N_i ← H_i W_i
    G_i ← Grid(H_i, W_i)
    X_i ← Tokenize(F_i^g) ∈ R^{B×N_i×C_i}
    T_i ← Proj_i(X_i) ∈ R^{B×N_i×D}
end for
P_5 ← {r}                                              ▷ r is an initialization sentinel
for i = 4 downto 1 do
    if i = 4 then
        C_4(r) ← G_4
        c_{4,r} ← MeanTokens(T_4) ∈ R^{B×D}
        m_4 ← N_4
    else
        m_i ← 4
    end if
    for all p ∈ P_{i+1} do
        if i < 4 then
            C_i(p) ← SpatialChildren(p, G_{i+1}, G_i)
            c_{i,p} ← TokenAt(T_{i+1}, p) ∈ R^{B×D}
        end if
        F_{i,p}^s ← Gather(T_i, C_i(p)) ∈ R^{B×m_i×D}
        s_{i,p} ← ScoringModule_i(F_{i,p}^s) ∈ R^{B×m_i}
        ξ_{i,p} ← [c_{i,p}; Mean_cand(F_{i,p}^s); Max_cand(F_{i,p}^s)] ∈ R^{B×3D}
        (E[K_{i,p}], k_{i,p}) ← BudgetModule_i(ξ_{i,p})        ▷ E[K_{i,p}] ∈ R^B, k_{i,p} ∈ Z^B_{≥0}
        (g_{i,p}^hard, g̃_{i,p}) ← RankedSTE(s_{i,p}, E[K_{i,p}], k_{i,p}; τ)
        ω_{i,p} ← Softmax(s_{i,p}) ∈ R^{B×m_i}
        a_{i,p} ← ω_{i,p} ⊙ g̃_{i,p} ∈ R^{B×m_i}
        if i > 1 then
            I_i(p) ← SelectedNodes(C_i(p), g_{i,p}^hard)
        end if
    end for
    if i > 1 then
        P_i ← UnionSupport({I_i(p) : p ∈ P_{i+1}})
    end if
    ω̄_i ← DeduplicatedScatterMean({(C_i(p), a_{i,p}, g_{i,p}^hard) : p ∈ P_{i+1}}; G_i) ∈ R^{B×N_i}
    z_i ← WeightedReadout(T_i, ω̄_i) ∈ R^{B×D}
    F_i^tr ← z_i
end for
F^tr ← Concat(F_4^tr, F_3^tr, F_2^tr, F_1^tr) ∈ R^{B×4D}
return F^tr
```

Algorithm 1 first projects the four global feature maps into a shared D-dimensional token space. At the coarsest level, no parent region is available, so the full grid $\mathcal{G}_4$ serves as the initial candidate region. The sentinel r only marks this initialization, and $c_{4,r}$, obtained by mean pooling T_4, is used as budget context; it is not itself read out or included in F^{tr}. At each level, RFR gathers

Table 5: Model efficiency comparison.

Stage	Model	Params (M)	GFLOPs	Train (ms)	Infer (ms)
Stage1	SAM-Swin	184.22	45.5	636.46	146.54
	RPGNet	**29.44**	**13.53**	**333.11**	**87.53**
Stage2	SAM-Swin	184.25	45.5	648.80	145.12
	RPGNet	**31.98**	**15.13**	**117.23**	**97.11**

the candidate features associated with the current parent regions, scores them, and uses the parent context and pooled candidate statistics to predict a Dynamic-k budget. RankedSTE applies hard Top-k selection in the forward pass while providing surrogate gradients during backpropagation. Selected locations are merged across parents, with duplicate contributions averaged on the spatial grid. A weighted readout then produces one D-dimensional vector per level. Finally, concatenating the vectors from F_4^g through F_1^g yields the restricted-pyramid representation F^{tr}.

6.4 Model Complexity Analysis

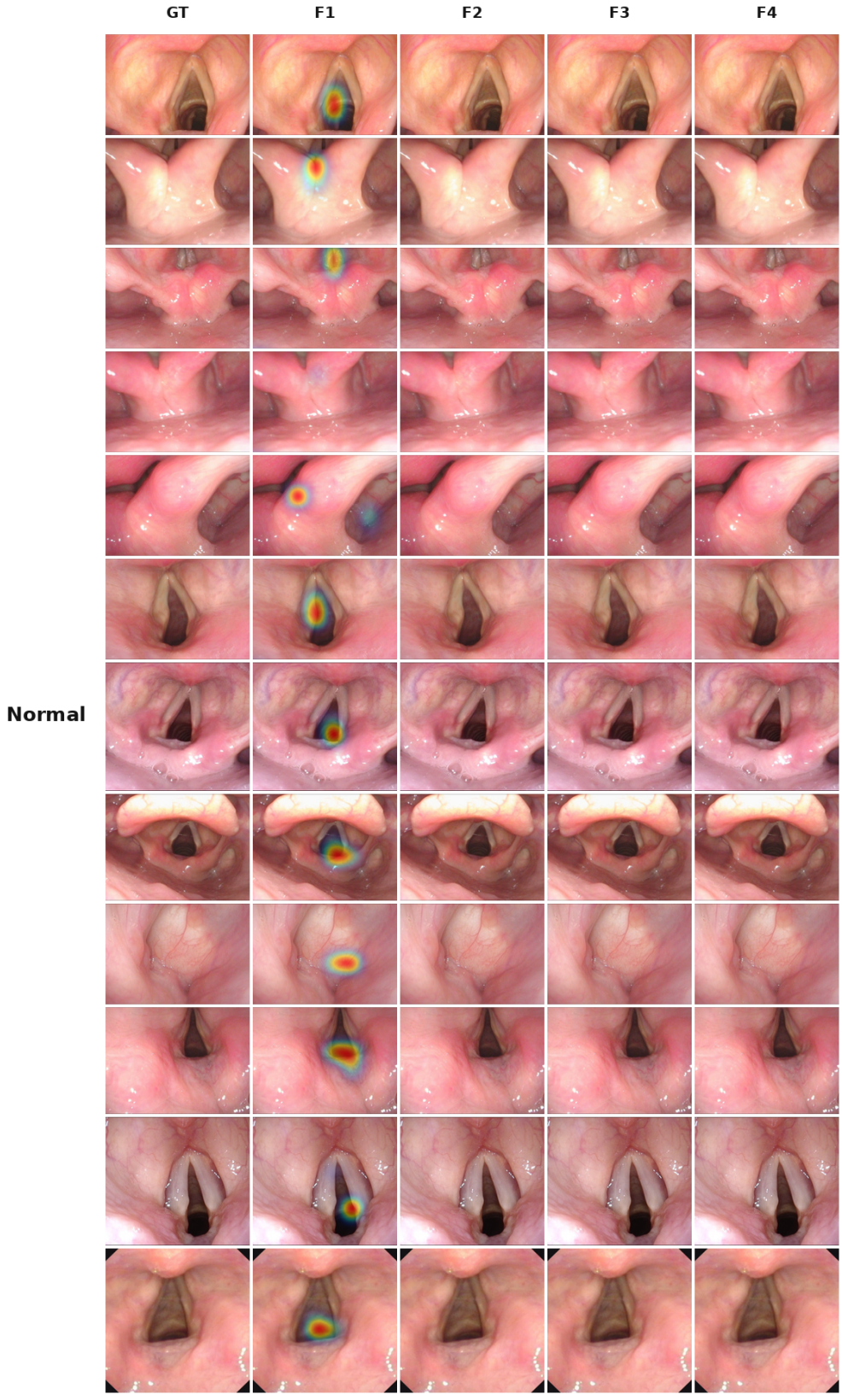
RPGNet comprises 29.44 million parameters and requires 13.53 GFLOPs in Stage 1, and comprises 31.98 million parameters and requires 15.13 GFLOPs in Stage 2, for one paired global–local sample at 256×256 resolution. Runtime measurements were obtained on a single NVIDIA RTX 5880 Ada Generation GPU using mixed-precision computation and a batch size of 32. As reported in Table 5, the training and inference times are 333.11 ms and 87.53 ms in Stage 1, and 117.23 ms and 97.11 ms in Stage 2, respectively.

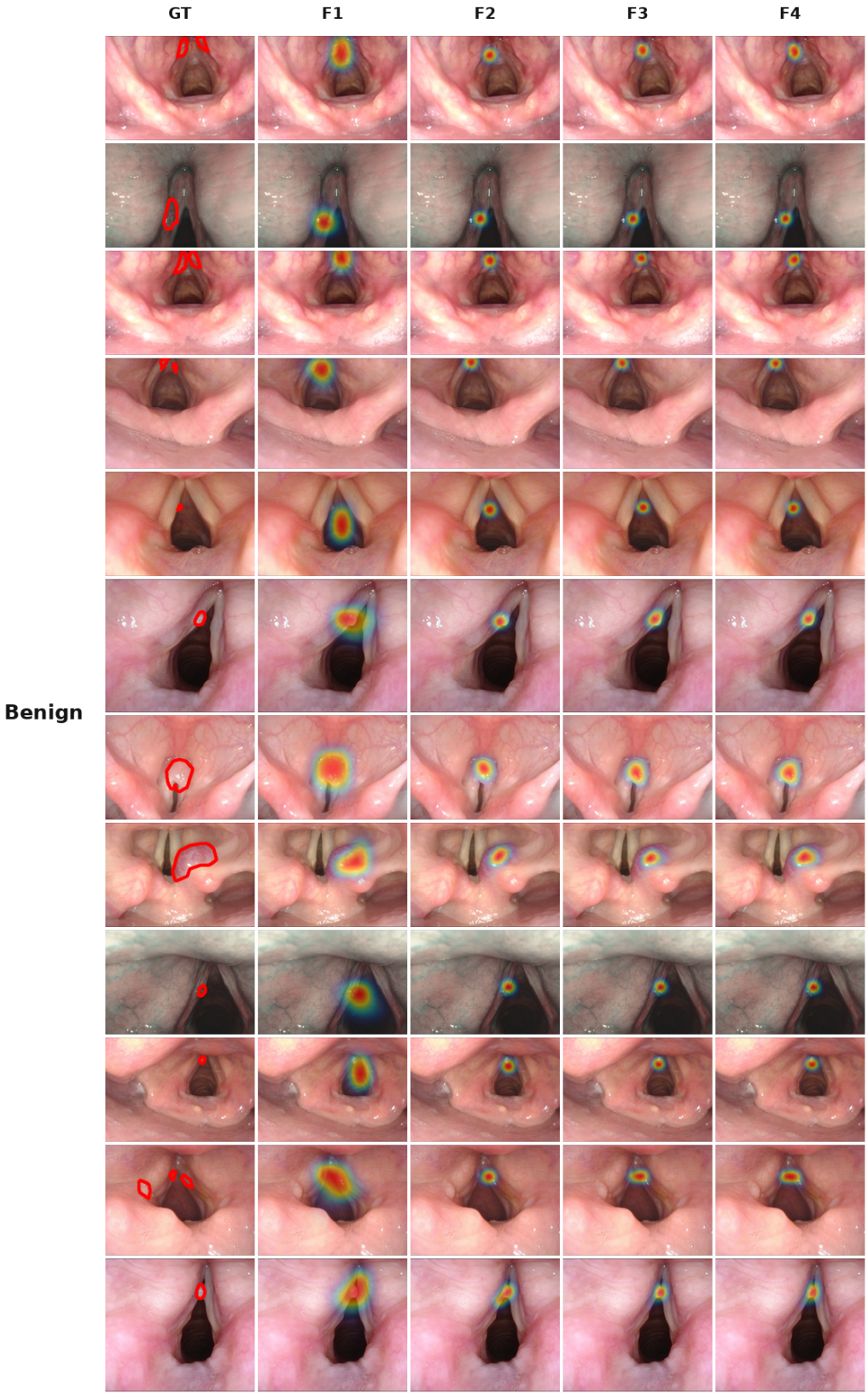
6.5 Datasets

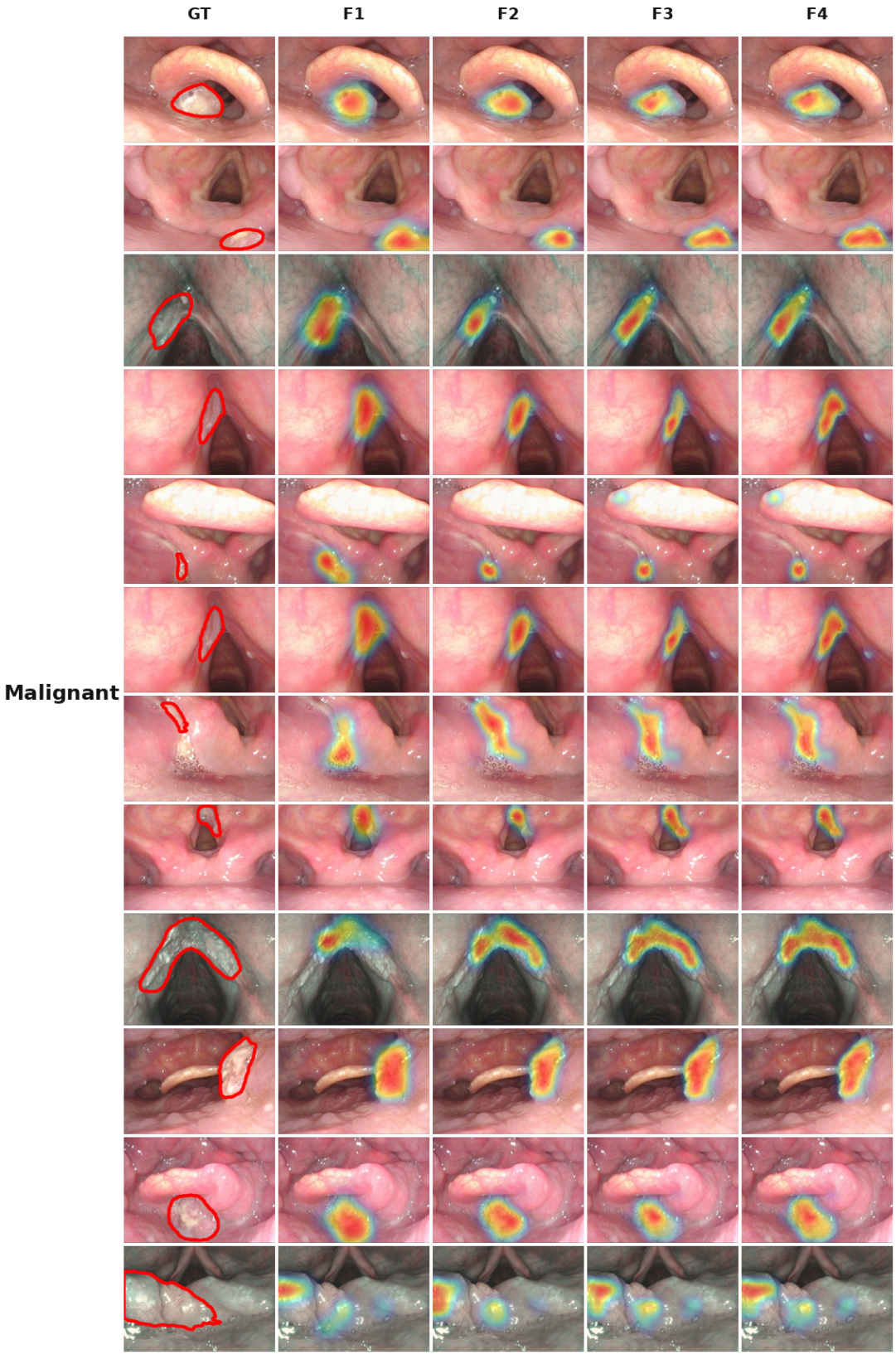
We evaluate RPGNet on four laryngoscopic image datasets provided by four medical centers: the First Affiliated Hospital of Sun Yat-sen University (FAHSYSU), the Sixth Affiliated Hospital of Sun Yat-sen University (SAHSYSU), Nanfang Hospital of Southern Medical University (NHSMU), and the Second People's Hospital of Shenzhen (FAHSU). Following the protocol used for SAM-Swin (Wei et al., 2026), FAHSYSU serves as the internal dataset and is used for training, validation, and internal testing, whereas the remaining three datasets are held out exclusively for external testing and are not used for model selection. Each image is labeled as normal, benign, or malignant; the source-directory class named `tumor` is referred to as malignant throughout this paper. Images were acquired using either White Light Imaging (WLI) or Narrow-Band Imaging (NBI). NBI emphasizes superficial microvasculature, whereas WLI preserves the conventional mucosal appearance. FAHSYSU and SAHSYSU contain images from both modalities, while NHSMU and FAHSU contain WLI images only. Table 1 summarizes the class and modality distributions, with all counts referring to images rather than patients. The pronounced class imbalance in FAHSU, which contains only 41 malignant images, makes it a particularly challenging external cohort. SAM2 predictions and lesion-focused RGB views are derived from the original images and therefore do not constitute additional samples.

- **FAHSYSU:** FAHSYSU is the internal dataset and contains 25,256 images, including 8,137 NBI images and 17,119 WLI images. It is used for model training, validation, and internal testing.
- **SAHSYSU:** SAHSYSU is an external evaluation dataset containing 2,788 images, including 135 NBI images and 2,653 WLI images. It is not used for training, hyperparameter tuning, or model selection, and is reserved for external testing.
- **NHSMU:** NHSMU is an external evaluation dataset containing 6,684 images, all acquired in WLI mode. It is used only for external testing.
- **FAHSU:** FAHSU is an external evaluation dataset containing 4,862 images, all acquired in WLI mode. It is used only for external testing.

6.6 Case Study


GT
F1
F2
F3
F4
Normal


GT
F1
F2
F3
F4
Benign


GT
F1
F2
F3
F4
Malignant