VNU Journal of Science: Comp. Science & Com. Eng. Vol. 42, No. 1 (2026) 36–48

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

LEAF: Learning Endoscopy with Ablation-Aware Features for Colorectal Cancer Classification

Quang Huy Nguyen, Xuan Hoang Pham, Van Khanh Tran*

Thai Nguyen University of Information and Communication Technology, Quyet Thang, Thai Nguyen, Vietnam

Received 06th February 2026;
Revised 7th March 2026; Accepted 26th March 2026

Abstract: Colorectal cancer screening pipelines increasingly rely on automatic video analysis to assist endoscopists, yet existing systems typically specialize in a single task and generalize poorly outside their training domains. This paper presents *LEAF* (Learning Endoscopy with Ablation-aware Features), a backbone-agnostic multi-task learning framework that jointly predicts fine-grained lesion classes, pathological severity, and anatomical regions from still frames. On the public CRCCD_V1 dataset, LEAF achieves 93.86% accuracy and 93.78% F1 when fine-tuned from ImageNet, outperforming the strongest single-task baselines by up to 2.1 absolute points. When trained from scratch, LEAF reaches 77.44% accuracy on the external HyperKvasir benchmark, surpassing all single-task counterparts by 1.9–6.4 points. Our contributions include (i) a flexible hard-parameter-sharing architecture with empirically calibrated loss weights that generalizes across backbone choices, (ii) a comprehensive benchmark covering EfficientNet, ResNet, DenseNet, and Swin Transformer baselines, and (iii) a cross-dataset evaluation protocol demonstrating improved generalization. The framework delivers consistent accuracy gains over single-task learners while maintaining computational efficiency suitable for clinical deployment.

Keywords: Medical image analysis, endoscopic imaging, colorectal cancer, deep learning, multi-task learning.

1. Introduction

Colorectal cancer (CRC) is among the most common and deadly cancers worldwide, with rising incidence particularly in low- and middle-income regions [1]. Colonoscopy remains the gold standard for early detection, but its effectiveness is limited by operator variability and practical constraints [2].

Recent advances in artificial intelligence (AI), particularly deep learning (DL), have demonstrated remarkable success in computer-aided detection (CADe) and diagnosis (CADx) for gastrointestinal en [3]. Convolutional neural networks (CNNs) [4], and vision transformers

*Corresponding author.
E-mail address: tvkhanh@ictu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6949

have been widely adopted to identify polyps, classify lesions, and assess malignancy risk directly from endoscopic images or videos [5, 6]. Several prospective clinical trials have shown that AI-assisted colonoscopy can significantly increase adenoma detection rates while maintaining real-time performance [7–9]. However, most existing systems address only a single task, despite the fact that clinicians simultaneously consider disease type, severity, and anatomical location during interpretation.

Multi-task learning (MTL) provides a principled framework to address this limitation by jointly learning multiple related tasks with shared representations [10]. In medical imaging, MTL has been shown to improve generalization, robustness, and data efficiency by leveraging complementary supervisory signals [11]. However, the design of clinically meaningful auxiliary tasks remains an open challenge, especially in endoscopy, where domain knowledge plays a crucial role in defining task relationships.

In this work, we propose a multi-task learning framework for colorectal endoscopy image classification that jointly addresses three clinically motivated tasks: (1) fine-grained disease classification with 14 categories, (2) pathological severity stratification into four levels, and (3) anatomical region classification into three gastrointestinal segments. Unlike prior studies that define auxiliary tasks heuristically, our task formulation is explicitly grounded in gastrointestinal oncology knowledge, aiming to disentangle confounded visual cues related to disease type and anatomical location. Our contributions can be summarized as follows:

- We introduce a clinically informed multi-task learning formulation for colorectal endoscopy image analysis, jointly modeling disease class, pathological severity, and anatomical location.
- We develop a unified deep learning architecture with shared feature extraction and task-specific heads, compatible with both CNN and transformer backbones.
- We conduct extensive experiments on the CRCCD dataset and evaluate cross-dataset generalization on HyperKvasir, demonstrating consistent improvements over strong single-task baselines.

The proposed approach advances the development of clinically reliable AI systems for gastrointestinal endoscopy and highlights the importance of task design in multi-task medical image learning.

2. Related works

2.1. AI for Colorectal Endoscopy

Deep learning has rapidly transformed colorectal endoscopy analysis, particularly in CADe and CADx systems. Early studies leveraged CNN-based architectures to detect polyps from colonoscopy videos, achieving performance comparable to expert endoscopists [5]. Subsequent works demonstrated real-time lesion detection and classification in clinical settings, significantly improving adenoma detection rates [7, 9].

Beyond detection, CADx systems aim to characterize lesion pathology, such as differentiating neoplastic from non-neoplastic polyps. Byrne et al. proposed a real-time AI system for optical biopsy, achieving high accuracy in prospective trials [6]. Similar approaches have employed deep CNNs to classify colorectal lesions using narrow-band imaging and white-light endoscopy [12].

Nevertheless, most existing methods focus on a single prediction target and do not explicitly model the hierarchical or correlated nature of clinical attributes inherent in endoscopic interpretation.

2.2. Deep Learning Architectures in Endoscopic Imaging

CNN backbones such as ResNet [13], DenseNet [14], and EfficientNet [15] have been widely adopted for endoscopic image classification due to their strong representational capacity and efficient parameterization. More recently, transformer-based architectures, including Vision Transformer (ViT) [16] and Swin Transformer [17], have shown promising results in medical image analysis by capturing long-range dependencies.

Hybrid and comparative studies suggest that while transformers excel in global context modeling, CNNs remain competitive in scenarios with limited training data, such as medical imaging [18]. This motivates backbone-agnostic frameworks that can leverage different architectures under a unified learning paradigm.

2.3. Multi-Task Learning in Medical Imaging

Multi-task learning has a long history in machine learning [10] and has gained renewed interest in medical imaging. Prior studies have applied MTL to jointly learn segmentation and classification [19], disease grading and localization [20], and multi-organ analysis [21]. These works consistently report improved performance and robustness compared to single-task counterparts.

In gastrointestinal endoscopy, however, MTL remains underexplored. Existing approaches often define auxiliary tasks without explicit clinical grounding or limit evaluation to a single dataset, leaving questions of generalization unanswered. Furthermore, few studies investigate how task design influences feature disentanglement between anatomical and pathological factors.

2.4. Datasets and Cross-Dataset Generalization

Public endoscopy datasets such as HyperKvasir [22] have facilitated large-scale benchmarking of AI models. However, domain shifts across datasets—due to differences in equipment, protocols, and patient populations—pose significant challenges. Recent works emphasize the importance of cross-dataset evaluation to assess real-world robustness [23].

In this context, leveraging auxiliary tasks that encode domain-invariant clinical semantics offers a promising direction for improving generalization, which we explicitly explore in this study.

3. Methodology

3.1. Problem Formulation

Given an endoscopic RGB image $\mathbf{x} \in \mathbb{R}^{3\times224\times224}$ sampled from CRCCD_V1 [24], the goal is to predict three labels: (1) a fine-grained lesion class $y^{(1)} \in \{1, \ldots, 14\}$, (2) a pathological severity label $y^{(2)} \in \{1, \ldots, 4\}$, and (3) an anatomical region label $y^{(3)} \in \{1, 2, 3\}$. Let $f(\mathbf{x}; \theta)$ denote the shared backbone with parameters θ, and $g_t(\cdot; \phi_t)$ denote the task-specific head for task $t \in \{1, 2, 3\}$. The network outputs logits

$$\mathbf{z}^{(t)} = g_t(f(\mathbf{x}; \theta); \phi_t) \in \mathbb{R}^{C_t}, \tag{1}$$

where C_t denotes the number of classes for task t (i.e., $C_1 = 14$, $C_2 = 4$, $C_3 = 3$). These logits are transformed into probability distributions via the softmax function:

$$p_i^{(t)} = \frac{\exp(z_i^{(t)})}{\sum_{j=1}^{C_t} \exp(z_j^{(t)})}, \tag{2}$$

where $p_i^{(t)}$ represents the prSedicted probability for class i in task t, and the final prediction is obtained as $\hat{y}^{(t)} = \arg\max_i p_i^{(t)}$.

3.2. Architecture

The proposed LEAF framework is designed to be backbone-agnostic, enabling flexible integration with various deep learning architectures. The backbone extracts a feature vector from the input image, which is then

passed to three task-specific linear heads with dropout, each tailored to a task-specific number of classes. Unlike fully decoupled networks, the hard-sharing design enforces a common representation that captures texture, color, and spatial cues relevant across tasks.

Figure 1 illustrates the overall pipeline of the proposed multi-task learning framework, detailing the data preprocessing module, the backbone feature extractor, and the multi-head configuration designed for simultaneous disease classification ($y^{(1)}$), severity assessment ($y^{(2)}$), and region localization ($y^{(3)}$).

Table 1 summarizes the architectural specifications of the backbone models evaluated in this study, including input dimensions, feature vector sizes, parameter counts, and output shapes for each task head.

3.3. Loss Function and Optimization

Training minimizes a weighted sum of cross-entropy losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}(\mathbf{z}^{(1)}, y^{(1)}) + \lambda_2 \mathcal{L}_{\text{CE}}(\mathbf{z}^{(2)}, y^{(2)}) + \lambda_3 \mathcal{L}_{\text{CE}}(\mathbf{z}^{(3)}, y^{(3)}), \tag{3}$$

with $(\lambda_1, \lambda_2, \lambda_3) = (1.0, 0.1, 0.5)$ selected after grid search. The standard cross-entropy loss for task t is defined as:

$$\mathcal{L}_{\text{CE}}(\mathbf{z}^{(t)}, y^{(t)}) = -\log\left(\frac{\exp(z^{(t)}_{y^{(t)}})}{\sum_{j=1}^{C_t}\exp(z^{(t)}_j)}\right) = -\log p^{(t)}_{y^{(t)}}. \tag{4}$$

To reduce overconfidence and improve generalization, we apply label smoothing [25] with smoothing factor $\alpha = 0.1$. The smoothed target distribution $\tilde{\mathbf{y}}^{(t)}$ is constructed as:

$$\tilde{y}^{(t)}_i = \begin{cases} 1 - \alpha + \frac{\alpha}{C_t} & \text{if } i = y^{(t)}, \\ \frac{\alpha}{C_t} & \text{otherwise,} \end{cases} \tag{5}$$

and the label-smoothed cross-entropy loss becomes:

$$\mathcal{L}_{\text{LS}}(\mathbf{z}^{(t)}, y^{(t)}) = -\sum_{i=1}^{C_t} \tilde{y}^{(t)}_i \log p^{(t)}_i. \tag{6}$$

We optimize the model using AdamW optimizer [26] with experiment-specific initial learning rates η_0 summarized in Table 4. The AdamW update rule for parameter θ at iteration k is:

$$\theta_{k+1} = \theta_k - \eta_k\left(\frac{\hat{\mathbf{m}}_k}{\sqrt{\hat{\mathbf{v}}_k} + \epsilon} + \lambda\theta_k\right), \tag{7}$$

where $\hat{\mathbf{m}}_k$ and $\hat{\mathbf{v}}_k$ are bias-corrected estimates of the first and second moments of the gradients, λ is the weight decay coefficient, and ϵ is a small constant for numerical stability. The learning rate follows a cosine annealing schedule:

$$\eta_k = \eta_{\min} + (\eta_0 - \eta_{\min}) \cdot \frac{1 + \cos(\pi k/K)}{2}, \tag{8}$$

where $\eta_{\min}$ is the minimum learning rate, K is the total number of training iterations, and k is the current iteration. To prevent gradient explosion, we apply gradient clipping with maximum norm $\gamma = 1.0$:

$$\tilde{\nabla}\mathcal{L} = \begin{cases} \nabla\mathcal{L} & \text{if } \|\nabla\mathcal{L}\| \le \gamma, \\ \gamma \cdot \frac{\nabla\mathcal{L}}{\|\nabla\mathcal{L}\|} & \text{otherwise.} \end{cases} \tag{9}$$

Both ImageNet-pretrained [27] and scratch initializations are considered; the latter uses longer schedules (80 epochs) to compensate for the lack of transferred features.

3.4. Evaluation Metrics

Accuracy measures the proportion of correctly classified instances among all test samples. It is defined as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C}(TP_i + TN_i)}{\sum_{i=1}^{C}(TP_i + TN_i + FP_i + FN_i)}, \tag{10}$$

Precision quantifies the fraction of positive predictions that are actually correct. For each class i, precision is calculated as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{11}$$

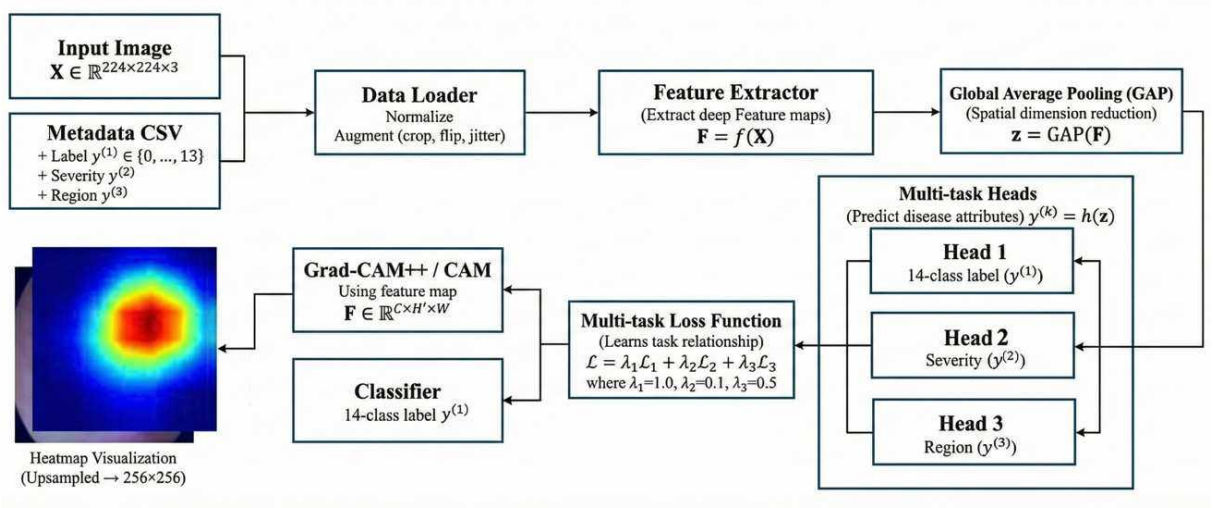

Figure 1. Schematic illustration of the proposed Multi-task Learning framework. The architecture extracts features to simultaneously predict disease class ($y^{(1)}$), severity ($y^{(2)}$), and region ($y^{(3)}$) via a weighted loss function, while employing Grad-CAM++ for visual interpretability.

Table 1. Architectural specifications of models

Backbone	**Input Shape**	**Feature Dim.**	**Parameters (M)**	**Task Outputs**
EfficientNet-B0	(3, 224, 224)	1280	4.03	(14)
EfficientNet-B1	(3, 224, 224)	1280	6.54	(14)
EfficientNet-B2	(3, 224, 224)	1408	7.73	(14)
EfficientNet-B3	(3, 224, 224)	1536	10.73	(14)
ResNet-50	(3, 224, 224)	2048	23.55	(14)
ResNet-101	(3, 224, 224)	2048	42.54	(14)
DenseNet-121	(3, 224, 224)	1024	6.98	(14)
Swin Transformer	(3, 224, 224)	768	27.54	(14)
LEAF-EfficientNet B1 (Ours)	(3, 224, 224)	1280	6.54	(14, 4, 3)
LEAF-EfficientNet B2 (Ours)	(3, 224, 224)	1408	7.73	(14, 4, 3)
LEAF-Swin Transformer (Ours)	(3, 224, 224)	768	27.54	(14, 4, 3)

and the overall precision is obtained by taking the weighted average:

$$\text{Precision} = \sum_{i=1}^{C} w_i \cdot \text{Precision}_i, \tag{12}$$

where w_i represents the weight for class i, and N is the total number of samples.

Recall indicates the proportion of actual positive cases that are correctly identified by the model. For class i, recall is defined as:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{13}$$

and the weighted average recall across all classes is:

$$\text{Recall} = \sum_{i=1}^{C} w_i \cdot \text{Recall}_i. \tag{14}$$

F1-score provides a balanced measure by combining precision and recall through their harmonic mean. The F1-score for class i is:

$$\mathrm{F1}_i = \frac{2 \cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}, \tag{15}$$

and the overall F1-score is computed as:

$$\text{F1-score} = \sum_{i=1}^{C} w_i \cdot \mathrm{F1}_i. \tag{16}$$

In the above formulations, TP_i, FP_i, TN_i, and FN_i represent the number of true positives, false positives, true negatives, and false negatives for class i, respectively, and $w_i = n_i/N$ corresponds to the empirical proportion of samples belonging to class i (with n_i samples and N total samples), yielding weighted macro metrics that reflect class imbalance.

Inference latency reports the average wall-clock time required to process a single image on the evaluation GPU:

$$t_{\text{infer}} = \frac{1}{N} \sum_{j=1}^{N} \left(t_j^{\text{stop}} - t_j^{\text{start}} \right), \tag{17}$$

where t_j^{start} and t_j^{stop} denote the timestamps (in milliseconds) immediately before and after the j-th forward pass.

Frames per second (FPS) is the reciprocal of the latency expressed in seconds:

$$\text{FPS} = \frac{1}{t_{\text{infer}}/1000}. \tag{18}$$

GFLOPs measure the floating-point operations needed for one forward pass, normalized by 10^9:

$$\text{GFLOPs} = \frac{1}{10^9} \sum_{\ell} 2 \cdot H_\ell W_\ell C_\ell^{\text{in}} C_\ell^{\text{out}} k_\ell^2, \tag{19}$$

where the sum runs over convolutional or fully connected layers with spatial size $H_\ell \times W_\ell$, input/output channel counts C_ℓ^{in} and C_ℓ^{out}, and kernel size k_ℓ. Reporting latency, FPS, and GFLOPs alongside accuracy metrics highlights the efficiency–accuracy trade-off of each backbone.

4. Experiments

4.1. Dataset and Experimental Setup

We train all models on the CRCCD_V1 dataset [24] and evaluate on the HyperKvasir test set [22]. All models are implemented using PyTorch [28] and the timm library [29]. CRCCD_V1 supplies balanced supervision for all three tasks: every lesion class contributes 500 training and 125 testing frames, and each class maps deterministically to one severity bucket and one anatomical region. Table 2 details the distribution across the 14 lesion identities together with their associated auxiliary targets.

For cross-dataset validation we retain the seven HyperKvasir categories that overlap with CRCCD_V1 labels (Table 3).

Cross-dataset evaluation on HyperKvasir assesses domain generalization without target-domain fine-tuning.

4.2. Training Hyperparameters and Data Augmentation

Table 4 reports the global optimization settings and augmentation operators applied across all experiments.

The tri-task output $(14, 4, 3)$ is instantiated for all multi-task LEAF variants, matching the number of classes for Task 1 (lesion classification), Task 2 (pathological severity), and Task 3 (anatomical localization), respectively. For each task t, the head architecture consists of a dropout layer followed by a linear projection:

$$\mathbf{h}^{(t)} = \text{Dropout}(\mathbf{f}; p = 0.2), \tag{20}$$

$$\mathbf{z}^{(t)} = \mathbf{W}^{(t)} \mathbf{h}^{(t)} + \mathbf{b}^{(t)}, \tag{21}$$

where $\mathbf{f} \in \mathbb{R}^d$ is the feature vector extracted from the backbone (with d being the feature dimension, which varies depending on the backbone architecture), $\mathbf{h}^{(t)} \in \mathbb{R}^d$ is the feature vector after dropout, $\mathbf{W}^{(t)} \in \mathbb{R}^{C_t \times d}$ and $\mathbf{b}^{(t)} \in \mathbb{R}^{C_t}$ are the learnable weight matrix and bias vector

Table 2. CRCCD_V1 label distribution across lesion classes, severity groups, and anatomical regions

Class	Severity	Region	Train	Test
ASCENDING_COLON_ADENOCARCINOMA	Malignant Cancer	Colon	500	125
CASIGMOID_COLON	Malignant Cancer	Colon	500	125
CECUM	Normal Anatomy	Colon	500	125
COLONADENOCARCINOMA	Malignant Cancer	Colon	500	125
ESOPHAGITIS	Benign Disease	Upper GI	500	125
HEMORRHOID_SUSPECTED_CA_COLON	Benign Disease	Recto Anal	500	125
ILEOCECAL_GROWTH	Suspicious PreCancer	Colon	500	125
POLYPS	Suspicious PreCancer	Colon	500	125
PYLORUS	Normal Anatomy	Upper GI	500	125
RECTADENOCARCINOMA	Malignant Cancer	Recto Anal	500	125
RECTGROWTH	Suspicious PreCancer	Recto Anal	500	125
SQUAMOUS_CELL_ANAL_CARCINOMA	Malignant Cancer	Recto Anal	500	125
ULCERATIVE_COLITIS	Benign Disease	Colon	500	125
Z_LINE	Normal Anatomy	Upper GI	500	125

Table 3. HyperKvasir categories overlapping with CRCCD_V1 and their sample counts (total images)

Category	Images
Polyp	1028
Ulcerative Colitis	851
Hemorrhoids	6
Cecum	1009
Z-line	932
Pylorus	999
Esophagitis	663

for task t, respectively. The dropout operation is defined as:

$$\begin{aligned} \text{Dropout}(\mathbf{x}; p) &= \mathbf{x} \odot \mathbf{m}, \\ m_i &\sim \text{Bernoulli}(1-p). \end{aligned} \tag{22}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{m}$ is a binary mask sampled from a Bernoulli distribution with success probability $1 - p$. The dropout technique [30] serves as a regularization method to prevent overfitting.

4.3. Baselines and Ablations

We train single-task baselines for EfficientNet B0–B3 [15], ResNet-50/101 [13], DenseNet-121 [31], and Swin Transformer [17] under identical training configurations. Ablation variants include: (i) Task1 only, (ii) Task1+Task2, (iii) Task1+Task3, and (iv) the full tri-task model. Both scratch and pretrained initializations are evaluated.

4.4. Results and Analysis

4.4.1. Comparison with Baseline Methods

Table 5 summarizes the cross-dataset performance after fine-tuning on CRCCD_V1. LEAF attains up to 93.86% accuracy and 93.78% F1 on HyperKvasir, exceeding the best single-task baselines by 0.66–2.1 absolute points. The tri-task head does not introduce noticeable computational overhead, enabling throughput above 20 FPS.

Table 6 presents a comprehensive comparison of our proposed LEAF method against state-of-the-art single-task baseline models. All

Table 4. Unified hyperparameters and augmentation operators

Optimization	
Optimizer	AdamW
Scheduler	Cosine annealing ($T_{\max} = 10$) with a linear warm-up of 5 epochs
Learning rate	7×10^{-5} (EfficientNet-B1/B2 tri-task fine-tune); 1×10^{-4} (Swin Transformer tri-task fine-tune); 3×10^{-4} (EfficientNet-B1/B2 tri-task scratch); 1×10^{-4} (Task 1-only ablation fine-tune); 7×10^{-5} (Task 1+2 and Task 1+3 ablations, EfficientNet/ResNet/DenseNet baselines); 1×10^{-4} (Task 1-only ablation scratch)
Batch size	16
Total epochs	30 (fine-tune) / 80 (scratch)
Label smoothing α	0.1
Gradient clipping	$\|\nabla\|_2 \leq 1.0$
Data augmentation	
Random resized crop	scale $\in [0.8, 1.0]$, aspect ratio $\in [0.75, 1.33]$
Horizontal / vertical flip	$p = 0.5$ / $p = 0.3$
Random rotation	$\pm 30^\circ$
Color jitter	$\pm 20\%$ brightness/contrast/saturation, hue ± 0.1
Random affine	translation $\pm 10\%$, shear $\pm 10^\circ$
Random grayscale	$p = 0.1$
Random erasing	$p = 0.2$, area range $[2\%, 33\%]$

Table 5. Performance comparison of our proposed method against baseline models on the HyperKvasir test set after fine-tuning. All accuracy-related metrics are reported as percentages

Method	Acc.	Prec.	Rec.	F1	Infer. (ms)	FPS	GFLOPs
LEAF-EfficientNet B1 (Ours)	**93.09**	**93.22**	**93.24**	**93.08**	**48.47**	**20.63**	**0.57**
LEAF-EfficientNet B2 (Ours)	**93.86**	**93.90**	**93.59**	**93.78**	**49.17**	**20.34**	**0.66**
LEAF-Swin Transformer (Ours)	**93.28**	**93.42**	**93.44**	**93.35**	**105.71**	**9.46**	**4.37**
EfficientNet-B3	93.20	93.79	93.20	93.45	64.24	15.56	0.96
EfficientNet-B2	92.75	93.35	92.75	93.32	51.83	19.29	0.66
EfficientNet-B1	92.60	93.01	92.47	92.74	51.39	19.46	0.57
EfficientNet-B0	91.09	91.40	91.10	91.00	35.25	28.37	0.38
ResNet-50	91.75	91.54	91.91	91.67	88.34	11.32	4.13
ResNet-101	90.30	90.45	90.30	89.97	149.84	6.67	7.86
DenseNet-121	90.96	91.29	91.08	90.78	88.66	11.28	2.83
Swin Transformer	93.00	92.87	93.01	93.02	107.32	9.13	4.37

models are evaluated on the HyperKvasir test set after training from scratch on CRCCD_V1, ensuring fair comparison under identical training conditions.

When trained from scratch (Table 6), LEAF achieves up to 77.44% accuracy, outperforming all single-task baselines by 1.9–6.4 absolute points, confirming that auxiliary supervision improves representation quality even without external pretraining.

Table 6. Performance comparison of our proposed method against baseline models on the HyperKvasir test set. All accuracy-related metrics are reported as percentages

Method	Acc.	Prec.	Rec.	F1	Infer. (ms)	FPS	GFLOPs
LEAF-EfficientNet B1 (Ours)	**77.44**	**78.12**	**77.43**	**77.34**	**58.87**	**16.99**	**0.57**
LEAF-EfficientNet B2 (Ours)	**76.60**	**78.37**	**76.58**	**77.15**	**48.26**	**20.72**	**0.66**
LEAF-Swin Transformer (Ours)	**73.52**	**77.33**	**73.54**	**74.88**	**102.11**	**9.79**	**4.37**
EfficientNet-B3	74.70	75.93	74.73	74.99	62.96	15.88	0.96
EfficientNet-B2	74.11	76.19	74.11	74.88	49.26	20.30	0.66
EfficientNet-B1	70.54	72.52	70.54	71.14	47.77	20.93	0.57
EfficientNet-B0	70.18	72.87	70.18	70.55	32.78	30.51	0.38
ResNet-50	73.80	76.00	73.73	74.16	80.99	12.35	4.13
ResNet-101	71.94	72.78	72.02	71.94	134.80	7.42	7.86
DenseNet-121	74.42	75.89	74.40	74.52	88.48	11.30	2.83
Swin Transformer	73.40	77.07	73.25	74.85	102.44	9.76	4.37

Qualitative Grad-CAM++ visualizations (Figure 2) reveal that LEAF concentrates on mucosal boundaries and vascular patterns that align with clinician annotations, whereas single-task baselines often highlight specular highlights or background mucosa. The improved attention supports the quantitative gains observed in Tables 5 and 6 and underscores the benefit of jointly reasoning about lesion identity, severity, and anatomical context.

To further evaluate the reliability of our framework, we analyzed the prediction confidence across different backbones on representative samples. (Figure 3) illustrates a comparative visualization for a 'Polyp' case. While all LEAF variants correctly identify the lesion and focus on the relevant mucosal area (as shown by the Grad-CAM++ heatmaps), the LEAF-Swin Transformer demonstrates superior certainty with a confidence score of 91.51%, compared to 88.37% for EfficientNet-B2 and 84.27% for EfficientNet-B1. This suggests that the Transformer-based backbone, when integrated into our multi-task framework, provides more robust feature representations for ambiguous boundary regions.

4.4.2. Ablation Study

Table 7 presents ablation results comparing different task combinations. Removing either auxiliary head degrades performance: excluding Task 2 (severity) or Task 3 (region) drops accuracy by 0.99 and 1.22 points, respectively, while the single-task baseline trails the full model by 1.11 points. These results justify the selected loss weights $(1.0, 0.1, 0.5)$ and demonstrate that the additional supervision acts as an effective regularizer.

5. Conclusion and Future Works

This work introduced LEAF, a backbone-agnostic tri-head multi-task learning framework that unifies lesion classification, pathological severity grading, and anatomical localization for colorectal endoscopy. The framework can be instantiated with various deep learning architectures, including EfficientNet variants and vision transformers, demonstrating consistent improvements across different backbone choices. Extensive cross-dataset experiments

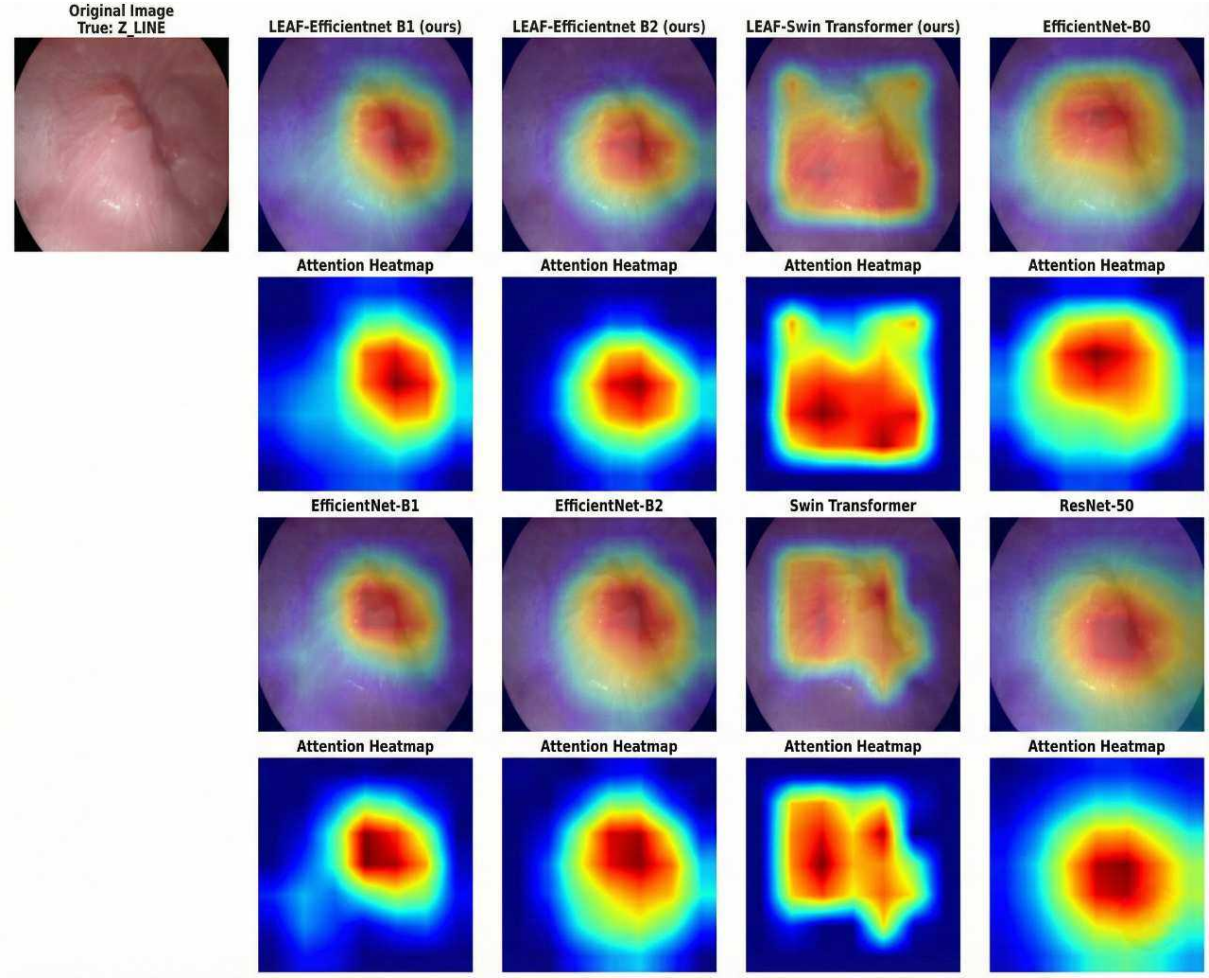

Figure 2. Comparison of class-discriminative attention maps between fine-tuned LEAF and baseline networks on ulcerative colitis images.

Table 7. Ablation study analyzing the contribution of different task combinations. All models use a fine-tuned EfficientNet-B2 backbone and are evaluated on the HyperKvasir test set

Task Configuration	Acc.	Prec.	Rec.	F1
Task 1 + Task 2 + Task 3 (Full)	**93.86**	**93.90**	**93.59**	**93.78**
Task 1 + Task 3	92.87	93.06	93.87	92.81
Task 1 + Task 2	92.64	92.52	92.85	92.62
Task 1 only (Baseline)	92.75	93.35	92.75	93.32

demonstrated that multi-task sharing provides consistent gains over single-task EfficientNet, ResNet, DenseNet, and Swin Transformer baselines. When instantiated with EfficientNet-B2, LEAF preserves a lightweight 0.66 GFLOP footprint and 22 FPS throughput. Fine-tuned LEAF variants improve HyperKvasir accuracy from 93.20% (best baseline) to up to 93.86%,

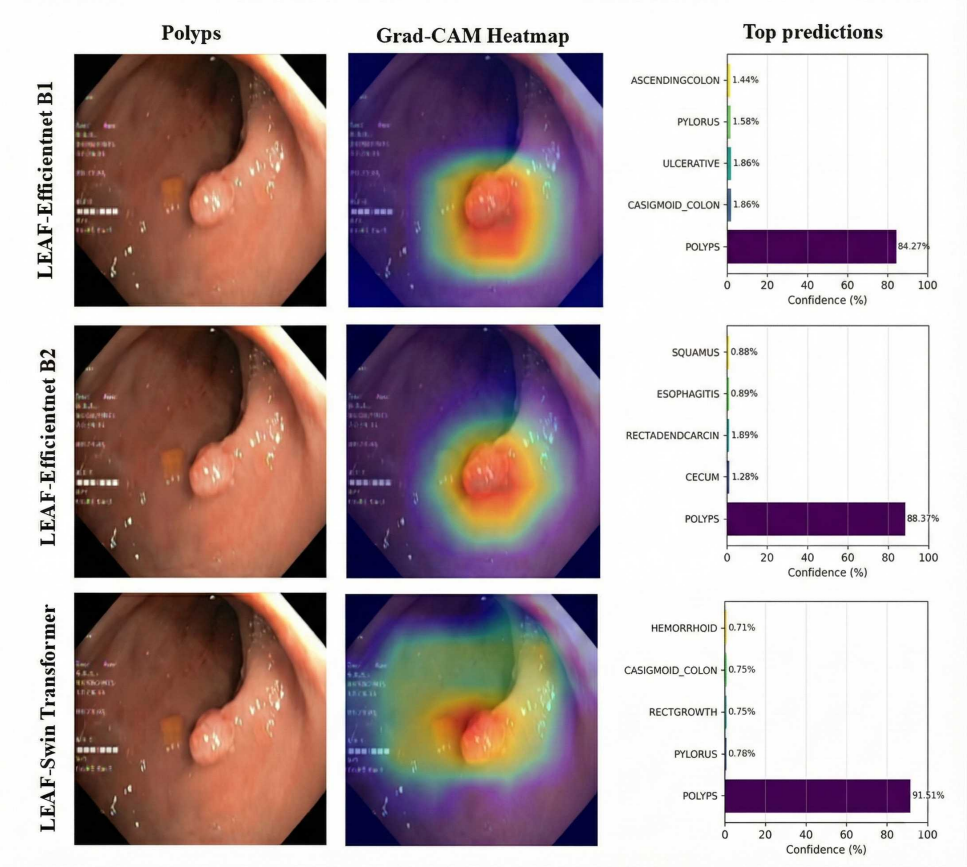

Figure 3. Visualization of prediction confidence and attention maps for a Polyp instance across different LEAF backbones. From top to bottom: LEAF-EfficientNet B1, LEAF-EfficientNet B2, and LEAF-Swin Transformer. The columns display the original input frame, the Grad-CAM++ heatmap focusing on the lesion, and the top-5 classification probabilities. The LEAF-Swin Transformer exhibits the highest confidence (91.51%) for the correct class.

and the scratch setting still yields a 1.9–6.4 point advantage, confirming the robustness of the auxiliary supervision across different backbone architectures. Ablation studies further highlighted that severity and region heads jointly regularize the backbone, leading to sharper Grad-CAM++ attention and better generalization, regardless of the underlying feature extractor.

Future work will explore (i) adaptive loss weighting strategies that respond to per-task uncertainty, (ii) semi-supervised or self-supervised pretraining on unlabeled colonoscopy videos, and (iii) temporal extensions that ingest short video clips to capture motion cues. Integrating model-based uncertainty into the reporting interface and validating LEAF prospectively in multi-center clinical trials are additional priorities toward real-world deployment.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA: A Cancer Journal for Clinicians, Vol. 71, No. 3, 2021, pp. 209–249.

[2] S. Q. INDICATORS, Quality Indicators for Colonoscopy, Am J Gastroenterol, Vol. 101, 2006, pp. 873–885.

[3] H. Lee, J.-W. Chung, S.-C. Yun, S. W. Jung, Y. J. Yoon, J. H. Kim, B. Cha, M. A. Kayasseh, K. O. Kim, Validation of Artificial Intelligence Computer-Aided Detection on Gastric Neoplasm in Upper Gastrointestinal Endoscopy, Diagnostics, Vol. 14, No. 23, 2024, pp. 2706.

[4] K. O'shea, R. Nash, An Introduction to Convolutional Neural Networks, ArXiv Preprint ArXiv:1511.08458 (2015).

[5] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, P. Baldi, Deep Learning Localizes and Identifies Polyps in Real Time with 96% Accuracy in Screening Colonoscopy, Gastroenterology, Vol. 155, No. 4, 2018, pp. 1069–1078.

[6] M. F. Byrne, N. Chapados, F. Soudan, C. Oertel, M. L. Pérez, R. Kelly, N. Iqbal, F. Chandelier, D. K. Rex, Real-Time Differentiation of Adenomatous and Hyperplastic Diminutive Colorectal Polyps during Analysis of Unaltered Videos of Standard Colonoscopy Using a Deep Learning Model, Gut, Vol. 68, No. 1, 2019, pp. 94–100.

[7] A. Ahmad, A. Wilson, A. Haycock, A. Humphries, K. Monahan, N. Suzuki, S. Thomas-Gibson, M. Vance, P. Bassett, K. Thiruvilangam, et al., Evaluation of a Real-Time Computer-Aided Polyp Detection System during Screening Colonoscopy: AI-DendoscopistsETECT Study, Endoscopy, Vol. 55, No. 04, 2023, pp. 313–319.

[8] W.-N. Liu, Y.-Y. Zhang, X.-Q. Bian, L.-J. Wang, Q. Yang, X.-D. Zhang, J. Huang, Study on Detection Rate of Polyps and Adenomas in Artificial-Intelligence-Aided Colonoscopy, Saudi Journal of Gastroenterology, Vol. 26, No. 1, 2020, pp. 13–19.

[9] J. Li, J. Lu, J. Yan, Y. Tan, D. Liu, Artificial Intelligence Can Increase the Detection Rate of Colorectal Polyps and Adenomas: a Systematic Review and Meta-Analysis, European Journal of Gastroenterology & Hepatology, Vol. 33, No. 8, 2021, pp. 1041–1048.

[10] R. Caruana, Multitask Learning, Machine Learning, Vol. 28, No. 1, 1997, pp. 41–75.

[11] S. Ruder, An Overview of Multi-Task Learning in Deep Neural Networks, arXiv Preprint arXiv:1706.05098 (2017).

[12] Y. Mori, S.-e. Kudo, T. M. Berzin, M. Misawa, K. Takeda, Computer-Aided Diagnosis for Colonoscopy, Endoscopy, Vol. 49, No. 08, 2017, pp. 813–819.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[14] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks 2017, pp. 4700–4708.

[15] M. Tan, Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 6105–6114.

[16] A. Dosovitskiy, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv Preprint arXiv:2010.11929 (2020).

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[18] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al., Big Self-Supervised Models Advance Medical Image Classification 2021, pp. 3478–3488.

[19] Q. Xu, Y. Zeng, W. Tang, W. Peng, T. Xia, Z. Li, F. Teng, W. Li, J. Guo, Multi-Task Joint Learning Model for Segmenting and Classifying Tongue Images Using a Deep Neural Network, IEEE Journal of Biomedical and Health Informatics, Vol. 24, No. 9, 2020, pp. 2481–2489.

[20] J. Kamiri, G. Wambugu, A. Oirere, Multi-Task Deep Learning in Medical Image Processing: A Systematic Review, International Journal of Computing Sciences Research, Vol. 9, , DOI: 10.25147/ijcsr.2017.001.1.226 (01 2025).

[21] G. Dai, D. Dai, C. Wang, Q. Tang, M. Hamilton, H. Chen, Y. Zhang, Multi-Task Learning Network for Medical Image Analysis Guided by Lesion Regions and Spatial Relationships of Tissues, IEEE Transactions on Circuits and Systems for Video Technology (2025).

[22] H. Borgli, K. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. Stensland, C. Garcia-Ceja, T. de Lange, P. T. Schmidt, M. Haugstvedt, M. Kampffmeyer, A. Riegler, P. Halvorsen, M. Riegler, HyperKvasir, a Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy, Scientific Data, Vol. 7, No. 1, 2020, pp. 283.

[23] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, H.-I. Suk, Domain Generalization for Medical Image Analysis: A Review, Proceedings of the IEEE (2024).

[24] A. Dash, S. Dash, S. Padhy, G. K. Pati, K. U. Singh, CRCCD_V1 (Colorectal Cancer Classification and Detection), DOI: 10.17632/2pybr7f7yc.2 (2024).

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[26] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: Proceedings of the International Conference on Learning Representations, 2019.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, Vol. 115, No. 3, 2015, pp. 211–252.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[29] R. Wightman, PyTorch Image Models, https://github.com/rwightman/pytorch-image-models, gitHub repository (2021).

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, Vol. 15, No. 1, 2014, pp. 1929–1958.

[31] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.