Efficient Polyp Segmentation via Low-Rank Adaptation: A Cross-Dataset Study on Colonoscopy Images

Nafisa Binte Ghulam Kibria[1], Md Kishor Morol[2], Ramisa Anjum Oishi[3], Tamzid Tanvi Alam[2], and Diya Rahman[4]

[1] Department of Computer Science and Engineering
Islamic University of Technology
Gazipur 1704, Bangladesh
nafisa21@iut-dhaka.edu
[2] ELITE Research Lab
New York, USA
kishoremorol@gmail.com, tamzid.t.alam@gmail.com
[3] California Lutheran University
California, United States
roishi@callutheran.edu
[4] American International University-Bangladesh
drahman024@gmail.com

Abstract. Accurate polyp segmentation in colonoscopy images is critical for early colorectal cancer detection. yet existing deep learning models often suffer from high computational costs and limited generalization ability. To address these challenges, this paper proposes a parameter-efficient polyp segmentation framework based on Low-Rank Adaptation (LoRA) integrated into a ConvNeXt-Large encoder with a U-Net decoder. By freezing the pretrained backbone and fine-tuning only lightweight LoRA adapters and the decoder, the proposed model significantly reduces trainable parameters while preserving strong representational capacity. A composite loss function combining Focal, Dice, and Tversky losses is utilized to address imbalance and prioritize clinically relevant recall. Extensive experiments are conducted on three public colonoscopy datasets-Kvasir-SEG, CVC-ClinicDB, and ETIS-LaribPolypDB-including rigorous cross-dataset evaluations. The proposed approach achieves state-of-the-art performance on Kvasir-SEG, with a Dice score of 0.9199, while training only 19.82% of total parameters. Compared to full fine-tuning, the LoRA-based model offers approximately 3× faster training with comparable or superior segmentation accuracy. Furthermore, Grad-CAM analysis confirms that the model focuses on clinically meaningful polyp regions. These results demonstrate that LoRA provides an effective and practical solution for efficient, robust, and interpretable polyp segmentation in real-world clinical settings.

Polyp segmentation, LoRA, ConvNeXt, U-Net, Cross-Dataset Generalization, Explainable AI

1 Introduction

Colorectal cancer (CRC) is one of the most common and deadly cancers worldwide, ranking third in incidence and second in cancer-related deaths globally . More than 1.9 million new CRC cases and nearly 900,000 deaths are reported every year [13, 15]. CRC usually develops through the well-known adenoma-carcinoma sequence, where untreated precancerous polyps gradually evolve into invasive cancer over time . Early detection and removal of these polyps during colonoscopy significantly reduce both CRC incidence and mortality [19, 9]. For this reason, colonoscopy is considered the gold standard for CRC screening. However, its effectiveness largely depends on the accurate detection and segmentation of all polyps during the procedure.

Manual polyp segmentation is highly challenging because polyps vary greatly in size, shape, texture, and color. In addition, colonoscopy images often suffer from poor lighting, motion blur, specular reflections, and occlusions caused by mucosal folds [1, 16]. Flat and sessile polyps are particularly difficult to identify because they closely resemble surrounding healthy tissue [12]. Studies have shown that adenoma miss rates during routine colonoscopy can reach up to 26%, with even higher miss rates for flat lesions [8]. Furthermore, manually annotating colonoscopy images and videos is time-consuming, labor-intensive, and subject to inter-observer variability [2]. Clinicians must examine thousands of frames and precisely outline polyp boundaries, which can lead to inconsistencies due to fatigue and subjective interpretation. These challenges can affect diagnosis and treatment quality, especially in healthcare settings with limited expert availability.

Deep learning has emerged as a highly effective approach for automated polyp detection and segmentation, significantly improving the accuracy and reliability of computer-aided diagnosis systems in gastrointestinal endoscopy. Among the most widely adopted approaches, convolutional neural network (CNN)-based encoder-decoder architectures, particularly U-Net and its variants, have demonstrated strong performance in medical image segmentation tasks due to their ability to learn hierarchical spatial representations from endoscopic images. These architectures effectively capture low-level texture information as well as high-level semantic features, enabling consistent and objective segmentation of polyp regions while reducing the dependence on manual interpretation by clinicians. In addition, skip connections used in encoder-decoder frameworks help preserve fine-grained spatial information, which is important for accurately delineating polyp boundaries.

Despite these advantages, conventional CNN-based methods still exhibit several important limitations. CNNs primarily rely on local receptive fields and stacked convolutional operations, which restrict their ability to model long-range spatial dependencies and global contextual relationships within an image. This limitation becomes particularly problematic in challenging colonoscopy scenarios where polyps exhibit irregular shapes, low contrast, blurred boundaries, specular highlights, or strong visual similarity with surrounding mucosal tissues. Small or flat polyps are especially difficult to segment accurately because their visual char-

acteristics can be easily overlooked when only local features are considered. As a result, traditional CNN-based models may fail to fully capture the broader contextual information required for robust and precise segmentation performance.

Recently, transformer-based models and hybrid CNN-transformer architectures have also gained significant attention in medical image analysis. Transformers enable improved feature representation and enhanced segmentation accuracy. Hybrid architectures further combine the strengths of CNNs and transformers by integrating local feature extraction with global contextual learning, resulting in more discriminative representations for complex medical imaging tasks [11]. These models have shown promising performance improvements in polyp segmentation by better handling variations in polyp size, shape, texture, and illumination conditions.

Recently, parameter-efficient fine-tuning (PEFT) techniques, originally designed for large language models, have attracted attention in medical imaging because they allow large pretrained models to be adapted using only a small number of trainable parameters [5, 4]. One of the most effective PEFT methods is Low-Rank Adaptation (LoRA) [5], which reduces computational complexity by representing weight updates with low-rank matrices. When applied to medical image segmentation, LoRA helps adapt powerful pretrained vision models to specialized tasks while reducing overfitting on small datasets. Motivated by these advantages, this study proposes a parameter-efficient polyp segmentation framework that integrates LoRA with a ConvNeXt-based U-Net architecture. The main contributions of this work are:

1. A LoRA-enhanced ConvNeXt-Large encoder integrated with a fully trainable U-Net decoder, reducing trainable parameters (80.18% reduction) while maintaining competitive performance.
2. Cross-dataset evaluation across Kvasir-SEG, CVC-ClinicDB, and ETIS-Larib PolypDB, demonstrating robust generalization under distribution shifts.
3. A multi-component loss combining Focal, Dice, and Tversky losses to handle class imbalance and prioritize high recall for clinically relevant polyp detection.
4. Empirical evidence of faster training compared to full fine-tuning, along with Gradcam Analysis to improve interpretability and support clinical decision-making.

2 Literature Review

Deep learning has largely replaced traditional handcrafted methods for medical image segmentation due to its superior ability to model complex visual patterns. U-Net established a foundational framework through its encoder-decoder architecture with skip connections, enabling effective fusion of spatial and semantic features.

Several U-Net variants have enhanced segmentation accuracy through architectural refinements. U-Net++ reduced the semantic gap using dense skip connections, while Attention U-Net emphasized salient regions via attention gates.

Residual and dense designs, such as ResU-Net and DenseU-Net, further improved feature propagation. AttDenseUNet combined dense blocks with attention mechanisms, achieving a Dice score of 91.48% and a Jaccard index of 84.30% on the Kvasir-SEG dataset [7].

Transformer-based models introduced global context modeling into segmentation. TransUNet integrated transformer encoders with U-Net decoders, while Swin Transformer improved efficiency through hierarchical feature representations. SwinSAM further combined Swin Transformers with the Segment Anything Model, reporting 91.3% Dice and 87.2% IoU on Kvasir-SEG [3].

Hybrid CNN-Transformer architectures have demonstrated improved robustness by combining local feature learning with global context modeling. seUNet-Trans achieved strong results on CVC-ClinicDB and ISIC 2018 datasets [11]. UViT-SEG incorporated squeeze-excitation and dual attention modules, achieving an 83.5% Dice score on Kvasir-SEG [10]. DATTNet modeled spatial and channel dependencies through dual attention and showed robust performance across multiple datasets, including Kvasir-SEG [20].

Recent studies have explored alternative and efficient architectures. VMAXL-UNet integrated visual state-space models with xLSTM modules to capture long-range dependencies with linear complexity, achieving competitive Dice scores on ISIC, Kvasir-SEG, and ClinicDB datasets [21]. Multi-branch designs such as D2HU-Net and UMF-Net enhanced multi-scale feature fusion and reported strong Dice and IoU performance on Kvasir-SEG [18, 17].

Despite these advances, parameter-efficient fine-tuning methods remain underexplored in medical image segmentation. In particular, systematic evaluation of Low-Rank Adaptation (LoRA) for segmentation is limited, especially regarding cross-dataset generalization and computational efficiency. Moreover, many models exhibit performance degradation on unseen datasets. To address these challenges, the present work proposes a LoRA-integrated ConvNeXt-based U-Net, emphasizing parameter efficiency, training speed, and robust cross-dataset generalization for practical clinical deployment.

3 Methodology

The figure 1 illustrates the experimental pipeline adopted in this study. The input datasets are first preprocessed through resizing, normalization, and stratified data splitting. Data augmentation techniques are applied to improve robustness and reduce overfitting. Model training is conducted using a combined loss function with learning rate scheduling and early stopping. Finally, cross-dataset evaluation is performed to assess generalization performance across different benchmark datasets.

3.1 Dataset description

In this study, three publicly available colonoscopy datasets were utilized for polyp segmentation. The Kvasir-SEG dataset[6], comprising 1,000 colonoscopy

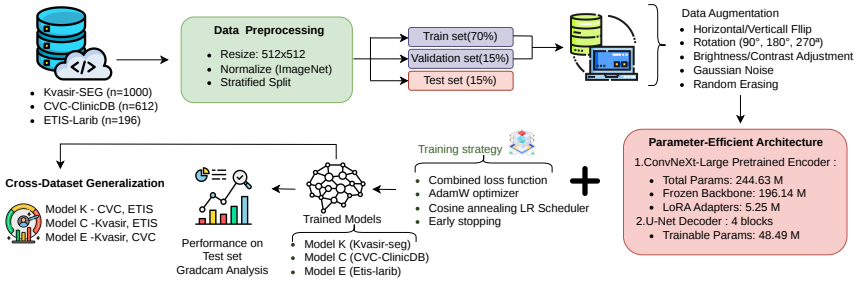

Fig. 1. Experimental workflow for training and cross-dataset evaluation of Polyp Segmentation Using ConvNext-UNet classifier

images with expert-annotated pixel-level ground truth masks, was employed for model training. This dataset includes substantial variability in polyp size, shape, texture, and illumination conditions, closely reflecting real-world clinical scenarios. To evaluate the robustness and cross-dataset generalization capability of the proposed method, independent testing was conducted on the ETIS-Larib dataset and the CVC-ClinicDB dataset.

The ETIS-Larib dataset[14] contains 196 high-resolution colonoscopy images, each presenting at least one polyp with precise pixel-level annotations. It is characterized by challenging visual conditions, including low contrast between polyps and surrounding mucosa, specular highlights, and large variations in polyp morphology. The CVC-ClinicDB dataset[1] consists of 612 colonoscopy images with corresponding expert-labeled binary masks and exhibits diverse imaging conditions, such as variations in illumination, scale, and polyp appearance. No images from the ETIS-Larib or CVC-ClinicDB datasets were used during training, ensuring an unbiased evaluation of cross-dataset performance.

3.2 Dataset Preprocessing

All images and corresponding ground truth masks were preprocessed using a unified pipeline. Images were resized to 512x512 pixels using bilinear interpolation, while nearest-neighbor interpolation was applied to masks to preserve segmentation boundaries. The masks were converted to single-channel grayscale images and binarized using a threshold of 127, where values above the threshold were labeled as polyp regions and the remaining pixels as background. Input images were normalized using ImageNet mean and standard deviation values to ensure compatibility with the pretrained ConvNeXt-Large encoder. The dataset was split into training (70%), validation (15%), and test (15%) sets using a fixed random seed of 42.

3.3 Data Augmentation

Data augmentation was applied during training to improve robustness and reduce overfitting. Augmentations were applied with a probability of 0.6 and included horizontal and vertical flipping, as well as random rotations of 90, 180, or 270 degrees. Brightness and contrast adjustments were performed using scaling factors in the range [0.8, 1.2] to account for illumination variability. To simulate imaging noise, additive Gaussian noise with zero mean and a standard deviation of 5 was introduced with a probability of 0.3. All augmentations were applied consistently to images and masks and were performed on-the-fly during training to ensure efficient memory usage.

3.4 Proposed Architecture

The proposed model adopts a parameter-efficient encoder-decoder design that integrates aConvNeXt-Large encoder with a fully trainable U-Net decoder for polyp segmentation. This architecture aims to retain the strong representational capability of pretrained models while enabling efficient task-specific adaptation.

LoRA-Enhanced ConvNeXt-Large Encoder The encoder was initialized with ConvNeXt-Large weights pretrained on ImageNet-1K. It follows a four-stage hierarchical structure with channel dimensions of [192, 384, 768, 1536]. To enable efficient fine-tuning, Low-Rank Adaptation (LoRA) was applied to convolutional and linear layers while keeping the original pretrained weights frozen.

For convolutional layers, LoRA was implemented using lightweight 1×1 convolution-based down- and up-projection modules. Given a frozen weight W_0, the adapted output is computed as

$$h = W_0 x + \frac{\alpha}{r} B(A(x)), \tag{1}$$

where r denotes the rank and α is a scaling factor. In all experiments, $r = 16$ and $\alpha = 32$ were used. The LoRA layers were initialized such that the model initially behaves identically to the pretrained backbone, ensuring stable optimization.

U-Net Decoder A symmetric U-Net decoder with four upsampling stages was employed to recover spatial details. Each decoder block consists of transposed convolution for $2\times$ upsampling, followed by convolution, batch normalization, and ReLU activation. Skip connections were introduced between corresponding encoder and decoder stages to preserve fine-grained spatial information.

The number of channels was progressively reduced from 1536 to 96 across the decoder. A final segmentation head composed of a 3×3 convolution and a 1×1 convolution produced the binary segmentation output, which was upsampled to the original input resolution of 512×512. All decoder parameters were trained from scratch.

3.5 Loss Function Design

A combined loss function was employed to address key challenges in polyp segmentation, including class imbalance, small target regions, and the clinical need for high recall. The overall objective integrates Focal Loss, Dice Loss, and Tversky Loss, each contributing complementary supervision during training.

1)Focal Loss was used to mitigate the strong class imbalance between background and polyp pixels. By down-weighting well-classified samples, this loss emphasizes hard and misclassified pixels.

$$\mathcal{L}_{focal} = -\alpha(1-p_t)^{\gamma}\log(p_t), \tag{2}$$

where p_t denotes the predicted probability of the true class. The parameters were set to $\alpha = 0.25$ and $\gamma = 2.0$. The loss was computed at the pixel level using softmax probabilities and averaged over the batch.

2)Dice Loss was incorporated to directly optimize region-level overlap between predicted and ground truth masks, which is critical for segmentation tasks.

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon}, \tag{3}$$

where p_i and g_i represent predicted and ground truth pixel values, respectively, and $\epsilon = 10^{-7}$ ensures numerical stability. The loss was computed using the foreground (polyp) class probabilities.

3)Tversky Loss was adopted to explicitly favor recall by assigning higher penalty to false negatives, which is important in clinical polyp detection. The loss is defined as:

$$\mathcal{L}_{tversky} = 1 - \frac{\sum_i p_i g_i + \epsilon}{\sum_i p_i g_i + \alpha\sum_i (1-g_i)p_i + \beta\sum_i g_i(1-p_i) + \epsilon}, \tag{4}$$

with $\alpha = 0.3$ and $\beta = 0.7$, emphasizing recall over precision.

The final training objective was computed as a weighted sum of the three losses:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{focal} + \lambda_2\mathcal{L}_{dice} + \lambda_3\mathcal{L}_{tversky}, \tag{5}$$

where the weights were set to $\lambda_1 = 1.0$, $\lambda_2 = 1.5$, and $\lambda_3 = 1.0$. A higher weight was assigned to Dice Loss due to its direct alignment with the evaluation metric.

This multi-task formulation utilizes complementary strengths of each loss component. Focal Loss handles pixel-level classification with attention to hard examples, Dice Loss directly optimizes the primary evaluation metric through region-level overlap, and Tversky Loss fine-tunes the model's behavior toward higher recall.

3.6 Training Strategy

A thorough training strategy was designed to effectively optimize the architecture while addressing the computational constraints and convergence challenges.

Table 1. Comparison of previous works on the Kvasir-SEG dataset, where (-) represents a missing value; bold numbers indicate the highest score.

Year	Model	Acc.	Dice	IoU	Prec.	Recall	F1
2024	seUNet-Trans	–	91.90	85.00	92.60	91.20	–
2024	UViT-Seg	–	83.50	80.10	81.10	81.00	81.60
2024	AttDenseUnet	86.31	91.48	84.30	–	–	–
2024	DATTNet	96.50	89.10	81.10	–	–	–
2025	D2HU-Net	–	89.90	81.70	–	–	–
2025	UMF-Net	–	90.90	85.70	92.90	91.60	89.70
2025	VMAXL-UNet	–	91.81	84.87	–	–	–
2026	SwinSAM	–	91.30	87.20	–	–	–
2026	**Ours**	**97.58**	**91.99**	**87.34**	**94.24**	**91.72**	**92.36**

The strategy encompasses differential learning rates, cosine annealing scheduling, and mixed precision training to achieve optimal performance within limited GPU memory budgets.

The AdamW optimizer was employed for gradient-based optimization due to its adaptive learning rate capabilities and improved weight decay regularization compared to standard Adam. A higher learning rate (1×10^{-4} with no weight decay) for LoRA parameters was used. It motivated by their initialization state and adaptation requirements. Since LoRA adapters begin from zero-initialized or near-zero states, they require larger learning rates to rapidly acquire task-specific features during early training. In contrast, the decoder parameters, randomly initialized, benefit from a more conservative learning rate (5×10^{-5} with weight decay of 0.01) to ensure stable convergence. The absence of weight decay for LoRA parameters prevents the regularization from counteracting the low-rank adaptation process, while weight decay on decoder parameters provides necessary regularization to prevent overfitting.

Learning rate scheduling was implemented using a cosine annealing strategy. This schedule gradually reduces the learning rate following a cosine curve, promoting smooth convergence and allowing fine-grained optimization in later training stages. Mixed precision training was used to reduce memory usage and accelerate training while preserving numerical stability. During the forward pass, compute-intensive operations such as convolutions and matrix multiplications were executed in FP16, whereas numerically sensitive operations, including loss computation and batch normalization, were maintained in FP32.

4 Results

The proposed LoRA-ConvNeXt U-Net was evaluated using a multi-dataset experimental protocol to assess both in-domain segmentation accuracy and cross-dataset generalization capabilities. Three independent training scenarios were conducted, where each of the three colonoscopy polyp datasets-Kvasir-SEG,

CVC-ClinicDB, and ETIS-LaribPolypDB-was used as a training source and validated on the remaining two datasets. All experiments were conducted using identical configurations.

4.1 Comparison with Previous Work

Table 1 presents a comparative evaluation of the proposed method against several recent state-of-the-art polyp segmentation approaches on the Kvasir-SEG dataset. The compared methods include CNN-based, transformer-based, and hybrid architectures published between 2024 and 2026. As observed, the proposed ConvNeXt–U-Net with LoRA achieves the best overall performance across all reported evaluation metrics, demonstrating its effectiveness for accurate and reliable polyp segmentation.

Specifically, the proposed model attains the highest Accuracy of 97.58%, outperforming DATTNet by 1.08 percentage points. In terms of overlap-based metrics, the model achieves a Dice score of 91.99% and an IoU of 87.34%, which are the highest among all compared methods. Although the improvement in Dice score over seUNet-Trans and VMAXL-UNet is relatively small, the proposed approach consistently provides superior IoU performance, indicating more precise boundary localization and region agreement between predicted and ground-truth masks.

The proposed method also achieves the highest Precision (94.24%), Recall (91.72%), and F1 score (92.36%), reflecting a strong balance between false positive and false negative predictions. High precision indicates that the model effectively suppresses incorrect polyp predictions, while the strong recall demonstrates its ability to accurately capture polyp regions. The improved F1 score further confirms the robustness and stability of the segmentation performance.

Overall, the results demonstrate that integrating a ConvNeXt-based encoder with LoRA adaptation enables the network to learn discriminative and transferable representations while maintaining parameter efficiency. This contributes to improved segmentation accuracy and stronger generalization compared with existing methods on the Kvasir-SEG benchmark dataset.

A quantitative comparison with previously reported polyp segmentation methods on the Kvasir-SEG dataset is presented in Table 1. As shown in the table, the proposed ConvNeXt–U-Net with LoRA consistently achieves the best performance across all reported evaluation metrics. In particular, the highest Accuracy, Dice score, IoU, Precision, Recall, and F1 score are obtained, indicating a strong improvement over existing CNN-, transformer-, and hybrid-based approaches. These results demonstrate that the proposed method provides more accurate and reliable segmentation performance while maintaining robustness across different evaluation criteria.

4.2 In-Domain Performance

Table 2 presents the in-domain test results for all three datasets using the 70/15/15 train-validation-test split. The LoRA-enhanced model demonstrated

Table 2. In-domain test performance. Each model was trained and evaluated on the same dataset using a 70/15/15 train-validation-test split. All models used LoRA with $r = 16$, $\alpha = 32$, and 48.5M trainable parameters.

Dataset	Soft Dice	IoU	F1 Score	Recall	Precision	Accuracy
Kvasir-Seg	**0.9199**	**0.8734**	0.9236	0.9172	0.9424	0.9758
CVC-Clinic	0.9127	0.8632	**0.9261**	**0.9274**	0.9323	0.9891
ETIS-Larib	0.8756	0.8294	0.9207	0.9007	**0.9441**	**0.9930**
Average	0.9027	0.8553	0.9235	0.9151	0.9396	0.9860

Soft Dice is computed directly from the predicted probability maps, whereas F1 is computed after binarizing predictions using a threshold of 0.4.

consistently strong performance across all datasets while training only 48.49M parameters (19.82% of total model parameters).

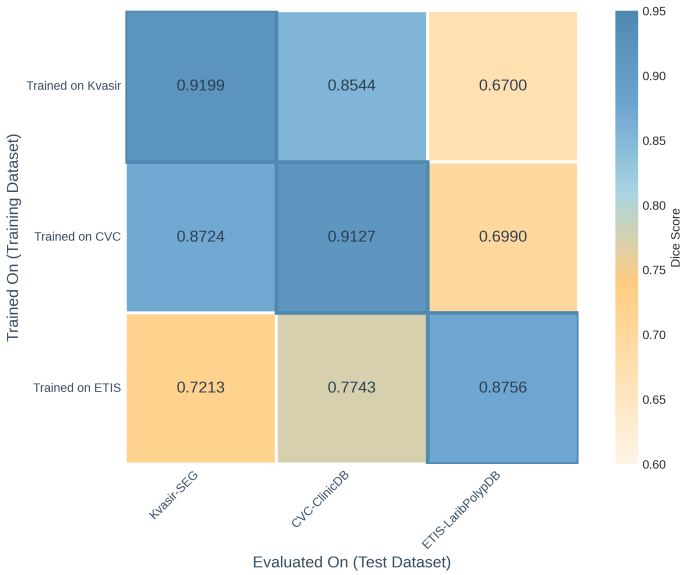

Fig. 2. 3×3 cross-dataset Dice score matrix for the proposed Model. Diagonal cells (blue borders) indicate in-domain performance, while off-diagonal cells represent cross-dataset generalization

Across all three datasets, the average Dice score reached 0.9027 with an average IoU of 0.8553, indicating consistent and reliable performance. The minimal variation in Dice scores (standard deviation: 0.0222) confirms the robustness of the LoRA approach across diverse colonoscopy imaging protocols and polyp characteristics.

Table 3. Cross-Dataset Generalization Performance. Each row shows a model trained on one dataset and evaluated on the other two datasets (complete sets).

Trained On	Evaluated On	Size	Soft Dice	IoU	F1	Recall	Prec.	Acc.
Kvasir-SEG	CVC-Clinic	612	0.8544	0.7868	0.8544	0.8643	0.9111	0.9772
	ETIS-Larib	196	0.6700	0.6036	0.6700	0.7242	0.8153	0.9813
CVC-Clinic	Kvasir-SEG	1000	0.8724	0.8049	0.8718	0.8794	0.8943	0.9597
	ETIS-Larib	196	0.6990	0.6188	0.7039	0.8062	0.6888	0.9789
ETIS-Larib	Kvasir-SEG	1000	0.7213	0.6180	0.7301	0.8993	0.6626	0.9075
	CVC-Clinic	612	0.7743	0.6956	0.7907	0.7819	0.8653	0.9672

4.3 Cross-Dataset Generalization

Cross-dataset generalization was evaluated to assess robustness across different clinical settings, imaging protocols, and dataset distributions. The complete 3×3 evaluation matrix is illustrated in Figure 2 and summarized in Table 3. Strong bidirectional generalization was observed between Kvasir-SEG and CVC-ClinicDB, indicating high compatibility between these two datasets despite being collected at different institutions. In contrast, performance degradation was consistently observed when transferring to or from ETIS-LaribPolypDB, reflecting its more challenging characteristics, including smaller polyps and higher visual variability. This performance degradation is likely caused by several dataset-specific factors. ETIS-LaribPolypDB contains a larger number of flat and sessile polyps with low contrast against the surrounding mucosa, making them more difficult to detect compared to the mainly pedunculated polyps present in Kvasir-SEG and CVC-ClinicDB. In addition, the ETIS images were originally collected for wireless capsule endoscopy research, resulting in differences in illumination and image resolution that are not included in the training data distribution. These factors create a stronger distribution shift, suggesting that dedicated domain adaptation techniques or ETIS-specific data augmentation strategies may be required to achieve stronger generalisation on this dataset. Models trained on ETIS demonstrated higher sensitivity when evaluated on other datasets, whereas models trained on Kvasir-SEG or CVC-ClinicDB exhibited more balanced behavior. Overall, the results confirm that the proposed model maintains clinically meaningful segmentation performance under cross-dataset distribution shifts, while highlighting ETIS-LaribPolypDB as a substantially distinct domain within colonoscopy imaging

4.4 Comparison with Full Fine-Tuning

To validate the effectiveness of the LoRA-based parameter-efficient approach, a comprehensive comparison was conducted against full fine-tuning of the ConvNeXt U-Net architecture on the Kvasir-SEG dataset. The comparison was designed to assess both computational efficiency and segmentation performance under controlled conditions.

Table 4. Fair Comparison: LoRA vs. Full Fine-Tuning on the Kvasir-SEG dataset. All results are reported after 50 epochs to ensure equal training conditions.

Metric	**Full Fine-Tuning (50 epochs)**	**LoRA (50 epochs)**
Segmentation Performance (Kvasir-SEG Test Set)		
Dice Score	0**.9121**	0.9112 (-0.09%)
IoU (Jaccard)	0.8610	**0.8654** (+0.44%)
F1 Score	0.9121	**0.9156** (+0.35%)
Recall	**0.9204**	0.9087 (-1.17%)
Precision	0.9258	**0.9420** (+1.62%)
Accuracy	0.9728	**0.9755** (+0.27%)
Model Parameters		
Total Parameters	**239.36M**	244.63M
Trainable Parameters	239.36M (100%)	**48.49M (19.82%)**
Frozen Parameters	0	196.14M
Parameter Reduction	–	**80.18% reduction**
Memory Requirements		
Model Size (Disk)	**913.15 MB**	933.62 MB
Runtime Memory	**913.15 MB**	933.24 MB
Training Efficiency		
Total Training Time	7h 11min	**2h 23min**
Time per Epoch	∼8.6 min	**∼2.9 min**
Training Speedup	–	**3.0× faster**
Inference Performance (Single Image, 512×512, Batch Size 4)		
Mean Inference Time	**167.5 ms**	198.5 ms
Standard Deviation	±3.25 ms	±3.75 ms
Throughput (FPS)	**5.97**	5.04
Min-Max Range	161–174 ms	191–206 ms
Stability (CV%)	1.94%	**1.89**%

Experimental Setup for Fair Comparison Both full fine-tuning and LoRA-based models were trained under identical settings, including hardware, optimizer, loss function, and training schedule. A fair comparison was conducted at 50 epochs (Table 4), the practical limit for full fine-tuning, where LoRA achieved comparable segmentation performance despite using only 19.82% trainable parameters. Although LoRA introduces a increase in model size, this overhead is negligible relative to the substantial gains in efficiency. LoRA reduced training time by approximately 3×, enabling extended training to 100 epochs within the same time budget required for full fine-tuning. This additional training led to superior final performance while remaining computationally efficient. These results confirm LoRA as an efficient and practical alternative to full fine-tuning for medical image segmentation

Practical Implications for Clinical Deployment The comparison reveals that LoRA offers a compelling value proposition for medical image segmentation:

1. **Accessibility**: With only 48.49M trainable parameters, LoRA enables fine-tuning on consumer-grade GPUs with 8-12 GB VRAM, democratizing access to state-of-the-art models. Full fine-tuning typically requires high-end GPUs with 16+ GB VRAM.
2. **Rapid Prototyping**: The 3× training speedup enables faster experimentation with hyperparameters, architectures, and loss functions, accelerating research and development cycles.
3. **Cost Efficiency**: Reduced training time and hardware requirements translate to lower computational costs-particularly relevant for resource-constrained clinical institutions in developing regions.

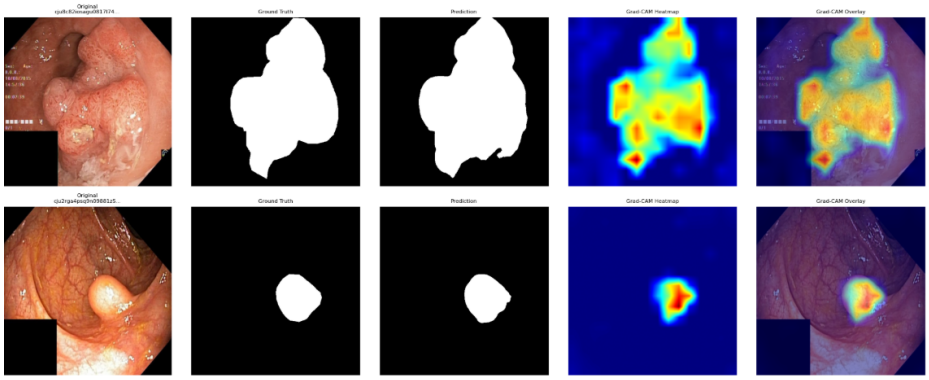

Fig. 3. Gradcam analysis with ground truth, prediction and original image

5 Explainable AI

Model interpretability was examined using Gradient-weighted Class Activation Mapping (Grad-CAM) to identify image regions that most strongly influenced the segmentation predictions. As illustrated in Figure 3, the Grad-CAM heatmaps consistently highlight polyp regions, with high activations aligning closely with both the predicted masks and ground truth annotations. Large and irregular polyps exhibit broad, contiguous activation patterns, while smaller lesions show localized responses, indicating sensitivity to fine-grained visual cues. Minimal activation is observed in background regions. This indicates that the model bases its predictions on clinically meaningful structures.

6 Conclusion & Future work

This work presented a LoRA-enhanced ConvNeXt U-Net framework for efficient and robust polyp segmentation in colonoscopy images. The proposed method

achieves competitive or superior segmentation performance while reducing trainable parameters by over 80% compared to full fine-tuning. Extensive evaluations on benchmark datasets demonstrate strong in-domain accuracy and cross-dataset generalization under distribution shifts. The comparative analysis confirms that LoRA enables substantially faster training with minimal trade-offs. Grad-CAM visualizations further enhance model transparency by validating that predictions are driven by relevant anatomical structures.

While the proposed LoRA-based framework achieves competitive segmentation performance with significantly reduced training cost, several limitations remain. The current study evaluates only a single LoRA configuration and does not investigate sensitivity to rank selection or other hyperparameters. In addition, performance comparisons are based on single training runs, and further experiments with multiple random seeds are needed to establish statistical significance. Broader ablation studies, including alternative PEFT methods and loss-function settings, were beyond the scope of this work. Future research will address these aspects while extending the framework to multi-class and video-based colonoscopy segmentation and evaluating its suitability for real-time clinical deployment.

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015). https://doi.org/10.1016/j.compmedimag.2015.02.007
2. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data **7**(1), 1–14 (2020)
3. Feng, Z., Zhang, Y., Chen, Y., Shi, Y., Liu, Y., Sun, W., Du, L., Chen, D.: Swinsam: Fine-grained polyp segmentation in colonoscopy images via segment anything model integrated with a swin transformer decoder. Biomedical Signal Processing and Control **100**, 107055 (2025)
4. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
6. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
7. Lafraxo, S., El Ansari, M., Koutti, L., Kerkaou, Z., Souaidi, M.: Attdenseunet: Segmentation of polyps from colonoscopic images based on attention-densenet-unet architecture. In: 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM). pp. 1–6. IEEE (2024)

8. Leufkens, A.M., van Oijen, M.G., Vleggaar, F.P., Siersema, P.D.: Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy **44**(05), 470–475 (2012)
9. Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z.R., Inamura, K., Kim, S.A., Kuchiba, A., Yamauchi, M., et al.: Long-term colorectal-cancer incidence and mortality after lower endoscopy. New England Journal of Medicine **369**(12), 1095–1105 (2013)
10. Oukdach, Y., Garbaz, A., Kerkaou, Z., El Ansari, M., Koutti, L., El Ouafdi, A.F., Salihoun, M.: Uvit-seg: an efficient vit and u-net-based framework for accurate colorectal polyp segmentation in colonoscopy and wce images. Journal of Imaging Informatics in Medicine **37**(5), 2354–2374 (2024)
11. Pham, T.H., Li, X., Nguyen, K.D.: Seunet-trans: A simple yet effective unet-transformer model for medical image segmentation. IEEE Access (2024)
12. Rex, D.K., Schoenfeld, P.S., Cohen, J., Pike, I.M., Adler, D.G., Fennerty, M.B., Lieb, J.G., Park, W.G., Rizk, M.K., Sawhney, M.S., et al.: Quality indicators for colonoscopy. Gastrointestinal Endoscopy **81**(1), 31–53 (2015)
13. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: a cancer journal for clinicians **73**(1), 17–48 (2023)
14. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery **9**(2), 283–293 (2014). https://doi.org/10.1007/s11548-013-0926-3
15. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
16. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging **35**(2), 630–644 (2015)
17. Wan, Y., Zhou, D., Wang, C.: Umf-net: A unet-based multi-branch feature fusion network for colon polyp segmentation. Biomedical Signal Processing and Control **99**, 106851 (2025)
18. Wang, Z., Fu, S., Zhang, H., Wang, C., Xia, C., Hou, P., Shun, C., Shun, G.: Dual-branch dynamic hierarchical u-net with multi-layer space fusion attention for medical image segmentation. Scientific Reports **15**(1), 8194 (2025)
19. Zauber, A.G., Winawer, S.J., O'Brien, M.J., Lansdorp-Vogelaar, I., Van Ballegooijen, M., Hankey, B.F., Shi, W., Bond, J.H., Schapiro, M., Panish, J.F., et al.: Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. New England Journal of Medicine **366**(8), 687–696 (2012)
20. Zhang, M., Zhang, Y., Liu, S., Han, Y., Cao, H., Qiao, B.: Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. Scientific Reports **14**(1), 25704 (2024)
21. Zhong, X., Lu, G., Li, H.: Vision mamba and xlstm-unet for medical image segmentation. Scientific Reports **15**(1), 8163 (2025)