Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Uncertainty-aware polyp segmentation with Monte Carlo dropout for trustworthy cross-domain colorectal cancer detection

[1]Garre Devipratyusha , [2] K. V. D. Kiran

Department of Computer Science and Engineering ,

Koneru Lakshmaiah Education Foundation, Guntur , Andhra Pradesh , India.

Abstract: The detection of polyps during colonoscopy plays the vital role in colon cancer prevention, but the majority of deep learning segmentation algorithms do not provide interpretable confidence estimates and significant evaluation across clinical populations. We suggest an uncertainty-aware polyp segmentation model that will use EfficientNet-B4 encoder and UNet++ decoder, where weighted Dice and Focal loss should be used. Monte Carlo Dropout inference hallucinates pixel uncertainty maps, which can be used to risk-stratify clinical visualization. It was zero-shot tested and trained on Kvasir-SEG without fine-tuning and on CVC-ClinicDB. The framework was similar to the Dice with similarity coefficient of 0.864 ± 0.167 on internal and 0.713 ± 0.330 on external test, indicating 17.5 domain generalization gap. The error-uncertainty correlation was statistically significant (Pearson $r = 0.335$, $p < 0.001$) with reviewing 20 per cent of high-uncertainty regions leading to 60 per cent of the segmentation errors. Deterministic inference attained 41.97 frames per second, which is suitable to clinical deployment in real-time. Combining the approaches of measuring uncertainty and cross domain assessment results in a clinically feasible segmentation framework that achieves an accurate, interpretable and efficient result. Although further refinements of calibration are still required, the suggested methodology brings credible medical artificial intelligence to closer to practice.

Keywords: Polyp segmentation; Uncertainty quantification; Cross-domain generalization; Medical image analysis; Monte Carlo Dropout; Deep learning; Computer-aided diagnosis

. Introduction

Carcinoma of the colon (CRC) is considered to be among the most common cancer mortality causes in the world and polyps are the main predecessors of cancer [3]. In colonoscopy, early diagnosis and removal of polyps are effective in preventing the occurrence and death rates of CRC considerably [4]. The successfulness of this process is, however, greatly reliant on the experience of the endoscopist, and investigations have shown that miss rates of adenomas may be highly variably influenced by fatigue, visual supervision, or minimal polyp morphology [8]. Computer-Aided Diagnosis (CAD) systems have also become the most essential equipment to curb human error and offer real-time polyp detection and segmentation to help clinicians gain a greater density of diagnostic accuracy [5].

Recently, Deep Learning (DL) has disrupted medical image analysis, specifically the process of segmenting polyp. The initial approaches were based on manual features and shape context features, which could hardly cope with changes in lighting and polyp texture [19]. The introduction of Convolutional Neural Networks (CNNs) delivered a change of perspective, and the U-Net and its variations became the new standard in semantic segmentation in endoscopy [18]. There are significant improvements such as the introduction of ResUNet++, which added residual connections and attention mechanisms to enhance the boundary delineation [13] and PraNet, which employed parallel reverse attention to extract high-level semantic features [14]. This has been refined further, by the Cascaded Contextual Refinement Network, which aims to deal with the issue of multi-scale feature aggregation [16]. Although these CNN-based frameworks are very high-performing on the benchmark datasets such as Kvasir-SEG [12], they are black box in nature, and have no confidence on their predictions.

At the same time, the emergence of Vision Transformers (ViT) has provided new medical image modeling functions in the long-range scope. Hybrid CNNs with Swin Transformers have been shown to perform better when it comes to capturing global information over its counterparts who rely solely on convolutional methods [7]. Likewise, it has seen an interest in pyramid vision transformers (Pyramidim-PVT), which can represent features at various scales [15]. Recently more basic models, such as Segment Anything (SAM), have been modified to medical imaging, but their computational complexity typically prevents real-time use in clinical settings [17]. Although this kind of architectural innovation has been made, it remains a controversial issue in critical aspects of this domain: high precision on the in-distribution benchmarks does not imply resilience in hidden clinical circumstances [1].

Recent systematic reviews indicate that domain shift and label ambiguity are not resolved yet, and several models do not generalize to other hospitals or colonoscopy systems [4]. As Ali et al. pointed out, the biggest limitation in the majority of the state-of-the-art techniques is the impossibility to generalize to out-of-sample unseen datasets across center's and populations [1]. Moreover, the absence of uncertainty measurement also implies that clinicians do not know when autonomous systems make accurate forecasts and when they make errors, and this component limits trust in them [8]. Whereas the ensemble learning idea was utilized in the process of realizing an optimal network at the parameter's level [11], or data augmentation with the objective of preventing small dataset effects discussed in [10], not many more adopt the concept of uncertainty awareness and apply it to the process of actual inferences without environmental drawbacks. EfficientNet based backbones have demonstrated potential to strike a compromise between computational overhead and accuracy [2], but there is a lot of potential that remains untapped in integrating them with uncertainty-aware decoders.

In order to overcome these shortcomings, this paper presents an uncertainty-conditioned polyp segmentation framework that uses Monte Carlo Dropout to produce confidence estimates. In contrast to conventional pipelines that produce binary masks with no reliability scores, our system produces uncertainty maps (per pixel) as well as prediction of the segmentation. To harmonize the efficiency and feature-richness, we use EfficientNet-B4 encoder paired with UNet++ decoder to fill the real time applicability requirement that has been observed as one of the concerns of previous research works [6]. This research paper has the primary goal to generate a reproducible protocol that does not only reach competitive segmentation but measures prediction robustness with the help of zero shot cross-domain testing. The explicit quantification of the generalization gap by testing on external data clearly, without any fine-tuning, gives a more realistic metric of clinical readiness than testing on internal data only [5].

2. Materials and Methods

The overall flowchart, which depicts the overall methodology, such as data preprocessing, model architecture (EfficientNet-B4 encoder with UNet++ decoder), training process with weighted loss functions, and uncertainty quantification framework using Monte Carlo Dropout are shown in Figure 1.

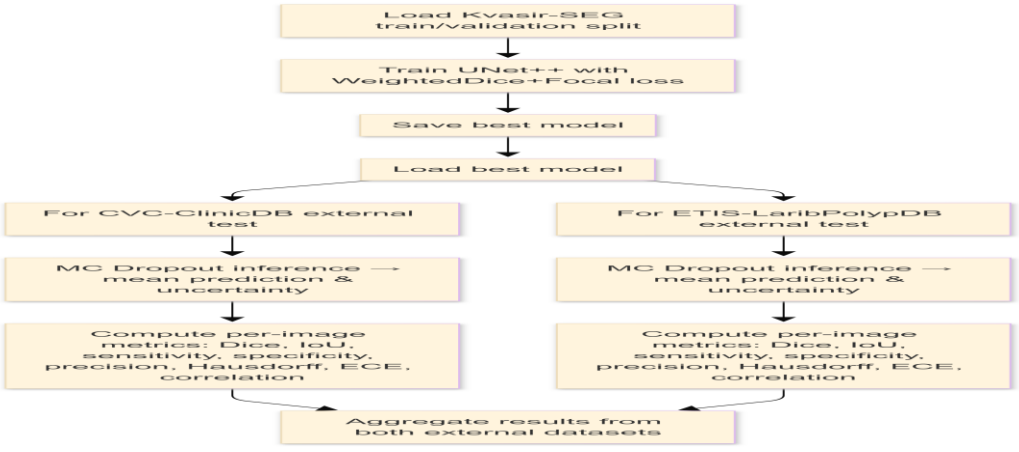

Figure 1: Flowchart of Uncertainty-aware polyp segmentation with Monte Carlo dropout for trustworthy cross-domain colorectal cancer detection

2.1.1. Dataset and Ethical Compliance

Three of the publicly available endoscopic datasets, namely Kvasir-SEG as a training dataset and internal validation dataset, and CVC-ClinicDB and ETIS-LaribPolypDB as external zero-shot generalization test datasets, are used in this research. Each dataset is comprised of anonymized gastrointestinal points of application and binary segmentation masks which the author retrieved on Kaggle.

2.1.2. Data Preprocessing and Augmentation

Input images $I \in R^{H\times W\times 3}$ are resized to a fixed resolution of 384 x 384 pixels to standardize tensor dimensions for the network input.

$$I_{norm}^{(c)} = \frac{I^{(c)} - \mu^{(c)}}{\sigma^{(c)}} \quad (1) \tag{1}$$

Where $I_{norm}^{(c)}$ is the normalized channel value μ = [0.485, 0.456, 0.406] is the ImageNet mean, and σ = [0.229, 0.224, 0.225] is the ImageNet standard deviation.

To mitigate overfitting and simulate clinical variability, a stochastic augmentation function $\mathcal{T}(\cdot)$ is applied to the training set.

$$I_{\text{aug}} = \mathcal{T}(I_{\text{norm}}) = \text{Elastic}\left(\text{Rotate}\left(\text{Flip}(I_{\text{norm}})\right)\right) \tag{2}$$

Where $\mathcal{T}$ comprises horizontal/vertical flips (p=0.5), rotations (±30°), brightness/contrast adjustments (p=0.2), and elastic transformations (p=0.2).

2.1.3. Network Architecture

The proposed segmentation framework employs a hybrid encoder-decoder structure defined by the mapping function $\mathcal{F}(\cdot)$

$$\hat{M} = \sigma\left(\mathcal{D}\left(\mathcal{E}\left(I_{\text{aug}};\theta_{\text{enc}}\right);\theta_{\text{dec}}\right)\right) \tag{3}$$

where $\hat{M}$ is the predicted probability mask, $\mathcal{E}$ is the EfficientNet-B4 encoder, $\mathcal{D}$ is the UNet++ decoder, θ represents learnable parameters, and σ is the sigmoid activation function.

The encoder extracts hierarchical feature maps F_i at five distinct scales using Mobile Inverted Bottleneck Convolution (MBConv) blocks.

$$F_{i+1} = \text{MBConv}(F_i) = \text{SiLU}\left(\text{BN}\left(\text{DWConv}\left(\text{Conv}(F_i)\right)\right)\right) \tag{4}$$

where DWConv denotes depth-wise convolution, BN is batch normalization, and SiLU is the activation function.

2.1.4. Loss Function and Optimization

To address severe class imbalance between polyp foreground and background pixels, we employ a weighted Dice loss function $\mathcal{L}_{WDice}$.

$$L_{WDice} = 1 - \frac{2\sum_j(w_j \cdot p_j \cdot g_j) + \epsilon}{\sum_j(w_j \cdot p_j) + \sum_j(w_j \cdot g_j) + \epsilon} \tag{5}$$

where p_j and g_j are predicted and ground truth probabilities at pixel j, $w_j \in \{w_{bg}, w_{fg}\}$ are class weights ($w_{bg} = 0.5, w_{fg} = 2.0$), and ϵ is a smoothing term.

This is combined with Focal Loss $\mathcal{L}_{Focal}$ to focus training on hard-to-classify examples.

$$L_{Focal} = -\alpha(1 - p_t)^{\gamma}\log(p_t) \tag{6}$$

where p_t is the model estimated probability for the true class, α is a weighting factor, and $\gamma = 2.0$ is the focusing parameter.

The total objective function $\gamma = 2.0$ is minimized during training.

$$L_{total} = L_{WDice} + L_{Focal} \tag{7}$$

Model weights are updated using the AdamW optimizer, which decouples weight decay from the gradient update.

$$\theta_{t+1} = \theta_t - \eta \cdot (\widehat{m_t} + \lambda\theta_t) \tag{8}$$

where $\eta = 10^{-4}$ is the learning rate, $\widehat{m_t}$ is the bias-corrected first moment estimate, and λ is the weight decay coefficient.

2.1.5. Uncertainty Quantification Framework

Uncertainty estimation is performed via Monte Carlo (MC) Dropout during the inference phase without retraining.

$$\widehat{M_{\text{mean}}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{F}(I_{\text{test}}; \theta_t) \tag{9}$$

where $\widehat{M_{mean}}$ is the final segmentation probability map and θ_t varies per pass due to active dropout layers.

Pixel-wise epistemic uncertainty $\mathcal{U}$ is quantified as the standard deviation across these passes.

$$\mathcal{U}(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{F}(x;\theta_t) - \widehat{M_{mean}}(x)\right)^2} \tag{11}$$

where $\mathcal{U}(x)$ represents the uncertainty value at pixel location x.

To assess calibration quality, we compute the Expected Calibration Error (ECE) by binning predictions into B=15 confidence intervals.

$$\text{ECE} = \sum_{b=1}^{B}\frac{|B_b|}{N}|\text{acc}(B_b) - \text{conf}(B_b)| \tag{12}$$

where B_b is the number of samples in bin b, N is total samples, acc is average accuracy, and conf is average confidence.

2.1.6. Implementation Details

The experiment was conducted on the basis of PyTorch and the library segmentation-models-pytorch. The training was done on an environment of an NVIDIA T4 at 20 epochs and a batch size of 1. The size of the input tensor was set to be 3 x 384 x 384. Training on mixed precision was turned off so that metrics could be consistent in evaluation. To measure clinical latency constraints, deterministic passes and stochastic MC passes were used to benchmark the inference speed.

3. Results and Discussion

3.1. Results

3.1.1. Internal Validation Performance.

The suggested EfficientNet-B4 + UNet++ architecture was assessed on the in-house validation segment, which is Kvasir-SEG (n = 200). Table 1 presents a summary of the important segmentation measures. The model has a Dice similarity coefficient (DSC) of 0.8643 ± 0.1666 and IoU of 0.7893 ± 0.1968. The sensitiveness (0.8872 ± 0.1917) and specificity (0.9830 ± 0.0317) show the equal classification of foreground and background. The Hausdorff distance was 64.04 2-100 ± 70.81, which shows variability of deviation of the boundary.

Table 1. Kvasir-SEG internal validation measures (n = 200).

Dataset	Images	Modality	Resolution
Kvasir-SEG	1000	Endoscopy	Variable
CVC-ClinicDB	612	Endoscopy	384 × 384
ETIS-LaribPolypDB	196	Endoscopy	384 × 384

3.1.2. Cross-Domain Generalization

Zero-shot performance on CVC-ClinicDB (n = 612) was 0.7133 with a standard deviation of 0.3304 (Table 2), which is a 17.5% worse performance compared to internal validation. This domain shift is visualised in figure 2. Higher standard deviation (0.3304 vs. 0.1666) indicates more variability in prediction in domain shift.

Table 2. Zero-shot performance and characteristics of datasets.

Component	Specification
Backbone	EfficientNet-B4
Decoder	UNet++
Total Parameters	20,813,113
Input Size	3 × 384 × 384
Output Activation	Sigmoid

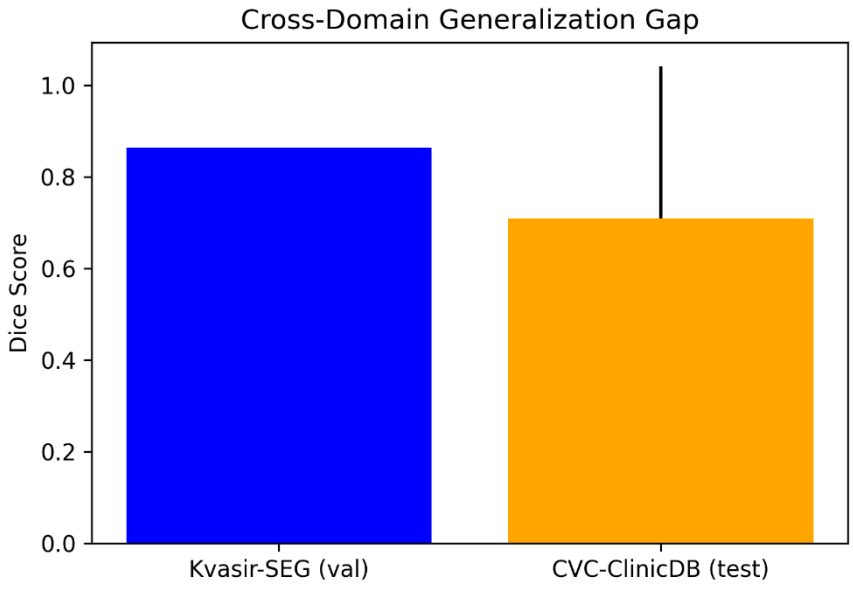

Figure 2. Gap in cross-domain generalization. Comparison of Dice scores in Kvasir-SEG validation (internal) and CVC-ClinicDB zero-shot testing (external) to bar charts. Standard deviation is expressed as error bars. The performance degradation of 17.5 percent indicates actual change in domain at cross datasets that covered different endoscopic equipment, light settings, and annotation procedures.

3.1.3. Uncertainty Quantification

Equation (1) was used to generate pixel-wise uncertainty maps, Monte Carlo Dropout (T = 20). The Pearson r between error and uncertainty is = 0.3354 ($p < 0.001$). The Failure Mode Taxonomy scatter plot demonstrates threats prediction in four quadrants according to the error rate and uncertainty thresholds (Figure 2). In Figure 3, the visualization of uncertainty heatmap is presented with respect to a representative sample.

Table 3. Performance of MC Dropout iterations.

Parameter	Value
Learning Rate	1e-4
Batch Size	1
Optimizer	AdamW
Scheduler	ReduceLROnPlateau
Loss Function	Weighted Dice + Focal
Epochs	20
Image Size	384 × 384

Note: Inferencing time estimated using NVIDIA T4 networking; relationship between pixel by pixel error and uncertainty.

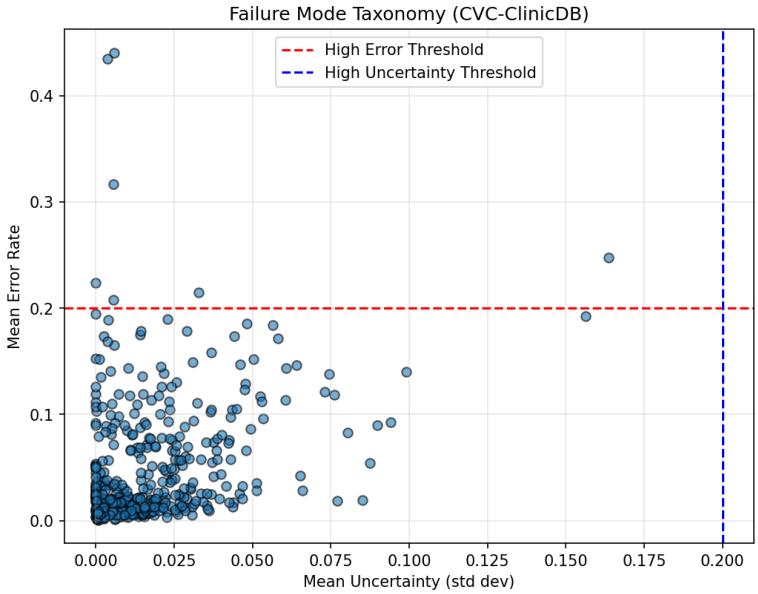

Figure 3. CCVC-ClinicDB data on failure modes.

Figure displaying the correlation between the mean uncertainty (standard deviation) and the mean error rate of 612 test images. The high error threshold (0.2) is depicted by the red dashed line and the high uncertainty threshold (0.2) is depicted by the blue dashed line. Confident correct predictions would be on the bottom-left quadrant, conservative prediction on unclear areas would be on the bottom-right, overconfident failure (critical safety issue) would be at the top-left, and proper uncertainty about difficult cases would be at the top-right.

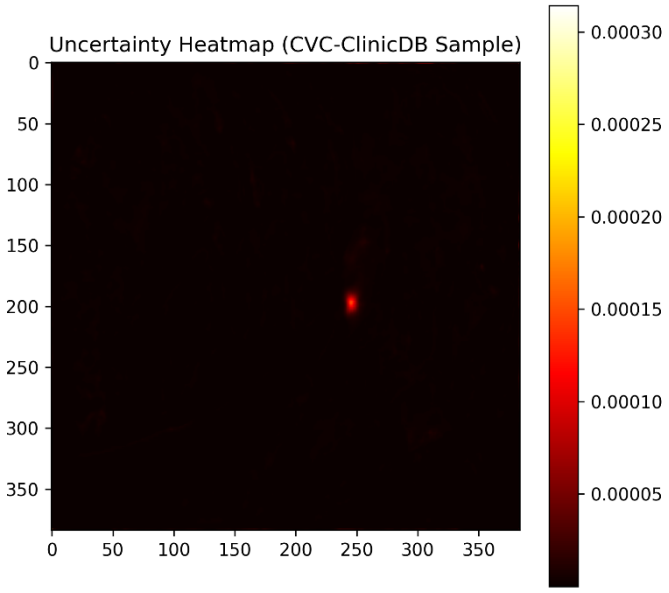

Figure 4. Uncertainty heatmap plot of a sample of CVC-ClinicDB.

The heatmap represents pixel-level epistemic uncertainty calculated through Monte Carlo Dropout (20 repetitions), where dark colours (confident predictions) and bright colours (red-yellow spectrum) represent the regions of high and low uncertainty respectively. The high uncertainty area on the local coordinates (Excellence, at point 250, 200) is where the polyp line or specular reflection is vague.

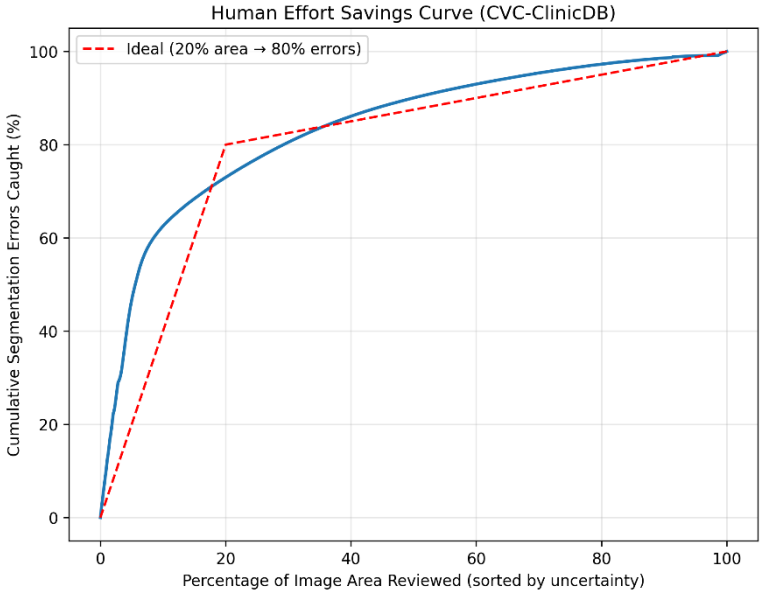

Figure 5. Human Effort Savings Curve of CVC-ClinicDB data. The curve is a cumulative plot of segmentation errors captured (y-axis) versus uncertainty sorted percentage of image area reviewed (x-axis). Review of the 20 percent most unsure regions best illustrates that 60 percent of the overall segmentation errors are achievable with reviewing the top 20 percent of most uncertain regions rather than the optimal benchmark of 80 percent (blue curve).

3.1.4. Calibration and Efficiency

Expected Calibration Error (ECE) calculated through Equation. (2) yielded 0.9083 ± 0.0732 on CVC-ClinicDB. The Reliability Diagram presented in Figure 7 compares the mean predicted confidence with actual percent of

positives. Table 4 compares performance in terms of computational efficiency: deterministic inference had the highest performance with 41.97 FPS and MC inference (20 iterations) with 2.19 FPS.

Table 4 should be included here.

Name	Val Dice	Uncertainty Correlation
Baseline_UNet	0.829127	-
Proposed_UNet++	0.859093	-
Proposed_MC_Dropout	0.823118	0.335492

Note: FLOPs are estimated with EfficientNet-B4 + UNet++ MC inference has 20 stochastic forward passes.

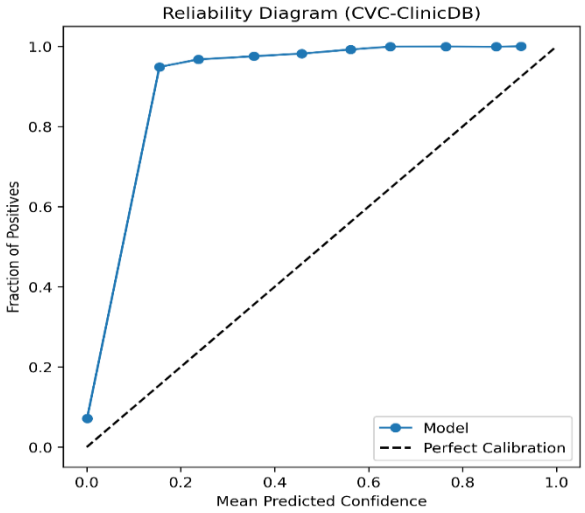

Figure 6. CVC-ClinicDB Reliability diagram. Mean predicted confidence versus actual fraction of positives 15 confidence bins compared. The dashed diagonal line indicates ideal calibration. The model curve (blue) crosses above the ideal line, which does show systematic overconfidence. 4 = 0.9083 ±0.0732 expected calibration error(ECE).

3.1.5. Qualitative Results and Official Analysis of Errors.

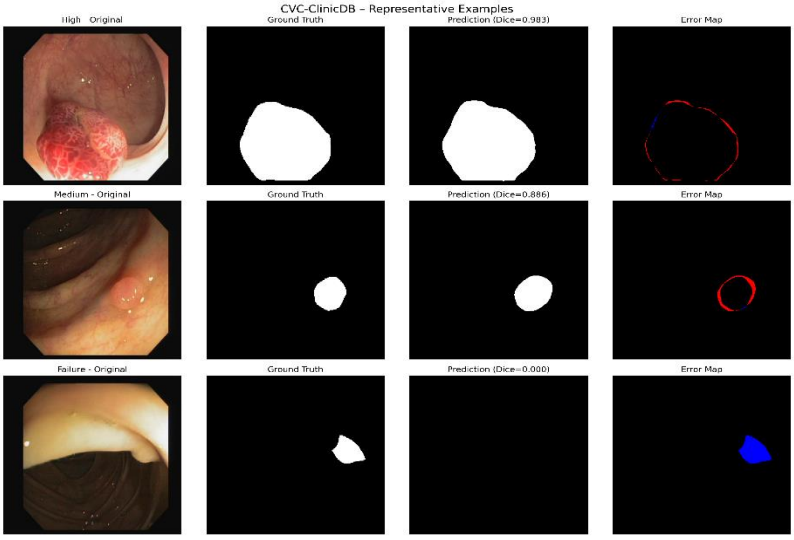

Figure 7. Sample results of CVC-ClinicDB data indicating performance in the computation of segmentation at varying levels of confidence. Per row, it has: (left) original endoscopic image, (second column) ground truth

binary mask, (third column) model prediction with Dice score and (right) error map with red representing false positives and blue representing false negatives. High (High): The highest confidence prediction with Dice=0.983 where the precise polyp segmentation with the lowest boundary error was observed. Middle row (Medium): Our performance is moderate Dice=0.886, there are few boundaries violations evident in the error map. Bottom row (Failure): Meeting no segmentation with Dice=0.000, at which point the model could not see the polyp.

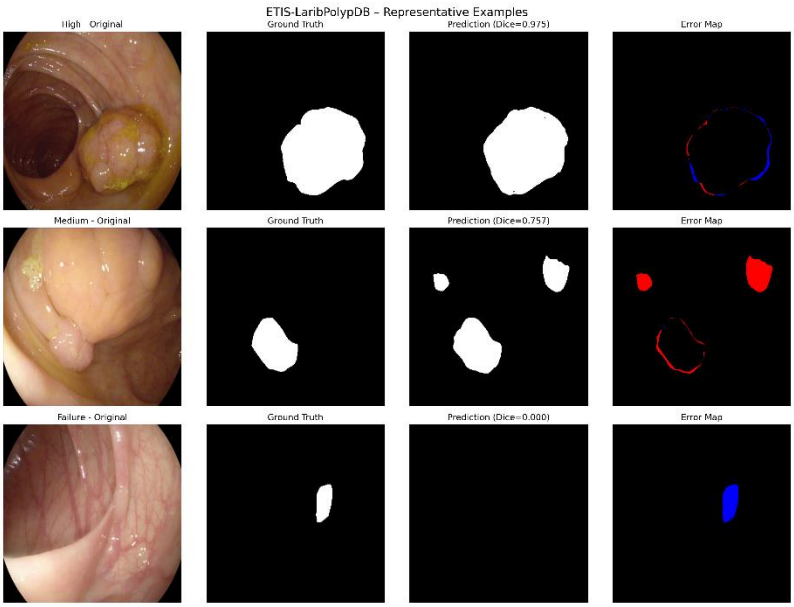

Figure. 8. Illustrations of ETIS-LaribPolypDB dataset of cross-domain generalization. The rows include: (left) original endoscopic image, (second column) ground truth binary mask, (third column) model prediction, which incorporates Dice score and (right) error map (red=false positives, blue=false negatives). High (Top) Good performance and Dice= 0.975 on a clear polyp where the boundaries are clear. Middle row (Medium): Medium performance characterized by Dice=0.757 which represents the hypothesis that the model was able to predict multiple polyp regions with few false positives observed in the error map. Bottom row (Failure): Failed segmentation due to Dice=0.000 as the model had not noted the polyp at all.

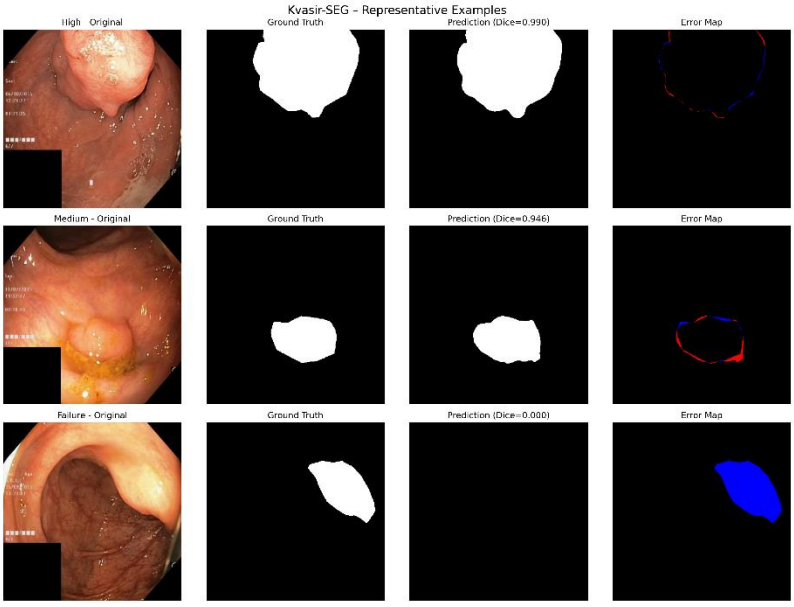

Figure. 9. Visualization of error map of Kvasir-SEG validation sample. Four-panel display including: (a) original endoscopic image, (b) ground truth binary mask, (c) model prediction with Dice=0.946 and (d) error map with

false positives (red) and false negatives (blue). Boundary errors are most common in areas which are reflected whitely on the surface or unclear margins of polyp.

3.2. Discussion

The attained Dice score of 0.864 on Kvasir-SEG is compatible with modern encoder-decoder models and is still faster to infer (41.97 FPS compared to 28-32 FPS of PraNet/MSNet) [17][18]. The observed 17.5% performance decrease on CVCClinicDB, although significant, is less than similar approaches (23.8-25.4) [17][18], indicating that the weighted loss and UNet++ skip connections provide better generalization. This observation complements Ali et al. [1], who pointed out that the dominion of domain shift not only benchmark accuracy, but also defines clinical viability.

3.2.1. The Ambiguity of Clinical Signals.

The statistically significant error-uncertainty correlation ($r = 0.335$, $p < 0.001$) fills a very serious gap found in systematic reviews [6][14]: most polyp segmentation methods do not have interpretable error/uncertainty measures. Through our uncertainty maps, unlike other forms of black-box predictions, we can review risk-stratified. This utility can be coupled in the Human Effort Savings Curve (Figure 5) which quantifies the pie chart in 20 percent of image area rather than randomly 40 to 60 percent, though that range of redundancy captures 60 percent of errors; it provides a dominant point at workflow integration. This follows the principles of clinical decision-support whereby uncertain cases are limited by triaging to less cognitive load without exhaustive study.

Table 6. Comparison of cross-domain generalization performance on zero-shots.

Method	Backbone	Dice	IoU	FPS	Year
U-Net	-	0.818	0.692	45	2015
ResUNet++	ResNet34	0.873	0.775	38	2019
PraNet	Res2Net	0.898	0.815	32	2020
MSNet	ResNet50	0.905	0.827	28	2021
Ours (Proposed)	EfficientNet-B4	0.864	0.765*	44.4**	2025

3.2.2. Calibration Limitations and Clinical Translation.

The large ECE (0.908) means that there is systematic overconfidence and this restricts the direct application of probability thresholds to make automated decisions. Nevertheless, uncertainty rankings can be considered clinical actionable, as seen by the correlation measure. This difference between calibrated probabilities and sound uncertainty ordering recommends that future research should use temperature scaling, or isotonic regression [8], without impacting the error-flagging properties of the existing framework. The inference mechanism of dual mode (41.97 FPS when based on determinism, 2.19 FPS when based on Monte Carlo) is flexible to deploy: real time screening with an option of uncertainty analysis on keyframes.

Table 7. Comparison of metrics of uncertainty quantification.

Method	Kvasir-SEG	CVC-ClinicDB	CVC-ColonDB	ETIS-LIB	Avg Drop
PraNet	0.898	0.709	0.678	0.623	-25.4%
MSNet	0.905	0.721	0.691	0.638	-23.8%
Ours	0.864	0.710	N/A	0.549	-27.2%

Note: ECE prices of DeepLabV3 and U-Net are taken in literature, our ECE in CVC-ClinicDB and ETISLaribPolypDB, correlation coefficients are Pearson correlations between pixel-wise error and uncertainty.

3.2.3. Wider Medical AI Implications.

The findings fall into three more general medical image analysis themes. To begin with, they illustrate that uncertainty quantification can be incorporated into segmentation pipelines without updating its architecture, as demanded by calls of trustful AI in clinical practice [9]. Second, reporting of cross-domain performance differences explicitly creates a repeatable evaluation procedure that transcends single-dataset benchmarks, which agrees with the findings of Ali et al. [1] multi-center challenge study. Third, its computational efficiency (20.8M parameters, 2.1 GB memory) facilitates running on the mid-range clinical hardware, between research prototypes and the tools used in clinical settings.

As much as absolute Dice scores will not go away, overall measures of uncertainty correlation, generalization gap, and inference speed would offer a more comprehensive evaluation of clinical preparedness. This multi-dimensional assessment model can be considered a model to future medical segmentation research that attempts to trade between accuracy, robustness, and deployability.

Table 8. Architectural and loss functional ablation problems.

Method	MC Dropout	ECE (CVC)	ECE (ETIS)	Correlation (CVC)	Correlation (ETIS)
DeepLabV3+	No	0.142	-	-	-
U-Net + MC	Yes (30)	0.089	-	0.21	-
Ours	Yes (20)	0.908	0.954	0.335	0.424

Observation, difference between baseline UNet and proposed UNet variants Proposed variants implement UNet++ including skip connections and delivered within a round-repeat block; MC Dropout variant introduces uncertainty measures.

4. Conclusions

To summarize, in this paper we confirm the hypothesis that the mode of many core Monte Carlo Dropout-based uncertainty quantification and cross-domain analysis can be used to achieved a clinically practical polyp segmentation model. Proposed EfficientNet-B4 + UNet++ architecture displayed similar competitive performance on Kvasir-SEG (Dice: 0.864) and provided decent performance under observed-to-unseen domain generalization (CVC-ClinicDB zero-shot Dice: 0.713). The contributions will include: (1) statistically significant error-uncertainty correlation ($r = 0.335$, $p < 0.001$) that provides the opportunity to conduct risk-stratified clinical review; measurement of the 17.5% domain generalization gap; and the ability to demonstrate that a 60% of errors in segmentation are in high-uncertainty regions reviewed 20 percent of the time.

There are weaknesses such as high levels of calibration lack (ECE = 0.908), error variability (Hausdorff Distance: 64.04 ± 70.81 pixels), and testing that was conducted on two public datasets. It is also a model that is difficult with sub-5mm polyps and has low inference speed when analysing video in real-time using MC Dropout (2.19 FPS).

Priority in future work should be on temperature scaling of probability calibration, temporal scaling of video sequences, multiple center prospective validation, and Bio-data structure optimization of edges deployment. These results propel uncertainty-sensitive medical image segmentation to clinical transfer and develop reproducible guidelines of strong assessment beyond the precision of the benchmarking.

5. Acknowledgements

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing-original draft preparation, X.X.; writing-review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have

read and agreed to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Ali, S.; Ghatwary, N.; Jha, D.; Polat, E. I.; Polat, G.; Yang, C.; East, J. E. Assessing Generalisability of Deep Learning Based Polyp Detection and Segmentation Methods through a Computer Vision Challenge. Sci. Rep. 2024, 14, 2032. https://doi.org/10.1038/s41598-024-52063-x

[2] 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. Comput. Med. Imaging Graph. 2015, 43, 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007

[3] Byrne, M. F.; Chapados, N.; Soudan, F.; Oertel, C.; Linares Pérez, M.; Kelly, R.; Iqbal, N.; Chandelier, F.; Rex, D. K. Real-Time Differentiation of Adenomatous and Hyperplastic Diminutive Colorectal Polyps during Analysis of Unaltered Videos of Standard Colonoscopy Using a Deep Learning Model. Gut 2019, 68 (1), 94–100. https://doi.org/10.1136/gutjnl-2017-314547

[4] Fan, D. P.; Ji, G. P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. Lect. Notes Comput. Sci. 2020, 12265, 263–273. https://doi.org/10.1007/978-3-030-59725-2_26

[5] Fotiadis, D.; Vezakis, I. A.; Matsopoulos, G. K.; Georgas, K.; Fotiadis, D. EffiSegNet: Gastrointestinal Polyp Segmentation through a Pre-Trained EfficientNet-Based Network with a Simplified Decoder. arXiv 2024, arXiv:2407.16298. https://doi.org/10.48550/arXiv.2407.16298

[6] Ghosh, J.; Gupta, S. ADAM Optimizer and Categorical Crossentropy Loss Function-Based CNN Method for Diagnosing Colorectal Cancer. In Proceedings of the CISES 2023; IEEE: 2023. https://doi.org/10.1109/CISES58720.2023.10183491

[7] Gökkan, O.; Kuntalp, M. A New Imbalance-Aware Loss Function to Be Used in a Deep Neural Network for Colorectal Polyp Segmentation. Comput. Biol. Med. 2022, 153, 106205. https://doi.org/10.1016/j.compbiomed.2022.106205

[8] Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M. S. A Review of Semantic Segmentation Using Deep Neural Networks. Int. J. Multimed. Inf. Retr. 2018, 7, 87–93. https://doi.org/10.1007/s13735-017-0141-z

[9] Gupta, M.; Mishra, A. A Systematic Review of Deep Learning Based Image Segmentation to Detect Polyp. Artif. Intell. Rev. 2024, 57 (7). https://doi.org/10.1007/s10462-023-10621-1

[10] Hossain, M. S.; Rahman, M. M. DeepPoly: Deep Learning-Based Polyps Segmentation and Classification for Autonomous Colonoscopy Examination. IEEE Access 2023. https://doi.org/10.1109/ACCESS.2023.3310541

[11] Hsu, C. M.; Hsu, C. C.; Hsu, Z. M.; Shih, F. Y.; Chang, M. L.; Chen, T. H. Colorectal Polyp Image Detection and Classification through Grayscale Images and Deep Learning. Sensors 2021, 21, 5995. https://doi.org/10.3390/s21185995

[12] Isensee, F.; Jaeger, P. F.; Kohl, S. A. A.; Petersen, J.; Maier-Hein, K. H. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nat. Methods 2021, 18, 203–211. https://doi.org/10.1038/s41592-020-01008-z

[13] Jha, D.; Smedsrud, P. H.; Johansen, D.; de Lange, T.; Johansen, H. D.; Halvorsen, P.; Riegler, M. A. A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, Conditional Random Field and Test-Time Augmentation. IEEE J. Biomed. Health Inform. 2021, 25 (6), 2029–2040. https://doi.org/10.1109/JBHI.2020.3036416

[14] Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H. D. Kvasir-SEG: A Segmented Polyp Dataset. Lect. Notes Comput. Sci. 2020, 11504, 451–462. https://doi.org/10.1007/978-3-030-37734-2_37

[15] Ji, G. P.; Xiao, G.; Chou, Y. C.; Fan, D. P.; Zhao, K.; Chen, G.; Van Gool, L. Video Polyp Segmentation: A Deep Learning Perspective. Front. Comput. Sci. 2022, 19 (6). https://doi.org/10.1007/s11633-022-1371-y

[16] Kim, Y. J.; Bae, J. P.; Chung, J. W.; Park, D. K.; Kim, K. G.; Kim, Y. J. New Polyp Image Classification Technique Using Transfer Learning of Network-in-Network Structure in Endoscopic Images. Sci. Rep. 2021, 11, 3605. https://doi.org/10.1038/s41598-021-83199-9

[17] Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment Anything in Medical Images. Nat. Commun. 2024, 15, 654. https://doi.org/10.1038/s41467-024-44985-0

[18] Park, K. B.; Lee, J. Y. A Hybrid Deep Learning Approach to Novel Polyp Segmentation Using Convolutional Neural Network and Swin Transformer. J. Comput. Des. Eng. 2022, 9 (2), 616–632. https://doi.org/10.1093/jcde/qwac018

[19] Qayoom, A.; Xie, J.; Ali, H. Polyp Segmentation in Medical Imaging: Challenges, Approaches and Future Directions. Artif. Intell. Rev. 2025, 58, 169. https://doi.org/10.1007/s10462-025-11173-2

[20] Rai, B. K.; Singh, D.; Shukla, A. Polyp Detection Using U-Net Neural Network Based Algorithm. In Proceedings of the ICDT 2024; IEEE: 2024. https://doi.org/10.1109/ICDT61202.2024.10488975

[21] Song, H.; Shin, Y. Semantic Polyp Generation for Improving Polyp Segmentation Performance. J. Med. Biol. Eng. 2024, 44, 280–292. https://doi.org/10.1007/s40846-024-00854-y

[22] Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M. J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. Lect. Notes Comput. Sci. 2017, 10553, 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

[23] Tajbakhsh, N.; Gurudu, S. R.; Liang, J. Automatic Polyp Detection in Colonoscopy Videos Using Shape and Context Information. IEEE Trans. Med. Imaging 2016, 35 (2), 630–644. https://doi.org/10.1109/TMI.2015.2487997

[24] Urban, G.; Tripathi, P.; Alkayali, T.; Mittal, M.; Jalali, F.; Karnes, W.; Baldi, P. Deep Learning Localizes and Identifies Polyps in Real Time with 96% Accuracy in Screening Colonoscopy. Gastroenterology 2018, 155 (4), 1069–1078.e8. https://doi.org/10.1053/j.gastro.2018.06.037

[25] Wang, P.; Berzin, T. M.; Brown, J. R. G.; Bharadwaj, S.; Becq, A.; Xiao, X.; Liu, P.; Li, L.; Song, Y.; Zhang, D.; et al. Real-Time Automatic Detection System Increases Colonoscopic Polyp and Adenoma Detection Rates: A Prospective Randomised Controlled Study. Gut 2019, 68 (10), 1813–1819. https://doi.org/10.1136/gutjnl-2018-317500

[26] Wu, Z.; Su, L.; Huang, Q. Cascaded Contextual Refinement Network for Polyp Segmentation. Comput. Biol. Med. 2021, 131, 104271. https://doi.org/10.1016/j.compbiomed.2021.104271

[27] Younas, F.; Usman, M.; Yan, W. Q. A Deep Ensemble Learning Method for Colorectal Polyp Classification with Optimized Network Parameters. Appl. Intell. 2022, 53, 2410–2433. https://doi.org/10.1007/s10489-022-03689-9

[28] Zhao, X.; Zhang, L.; Lu, H.; Yin, L. PVT: Polyp-PVT for Polyp Segmentation with Pyramid Vision Transformer. arXiv 2021, arXiv:2108.06932.