Array 32 (2026) 101265

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Swin-MSEG-Pro: Uncertainty-aware boundary-focused swin transformer with deep supervision for reliable gastrointestinal polyp segmentation

Muhammad Muhtasim Shahriar [a], M.M. Golam Hafiz [b], M.F. Mridha [b,*], Md. Rakibul Islam [c]

[a] Department of Computer Science, International Islamic University Chittagong (IIUC), Chittagong, Bangladesh
[b] Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh
[c] Department of Computer Science and Engineering, Mälardalen University, Västerås, Västmanland, Sweden

ARTICLE INFO

Dataset link: https://datasets.simula.no/kvasir-seg/, https://deeply.readthedocs.io/en/latest/datasets/etis-larib.html, https://datasetninja.com/cvc-612

Keywords:
Gastrointestinal polyp segmentation
Colonoscopy image analysis
Swin transformer
Evidential uncertainty
Boundary-focused supervision
Cross-dataset generalisation
Reliability-aware learning

ABSTRACT

Automatic segmentation of gastrointestinal polyps from colonoscopy images is clinically important because computer-aided systems must identify suspicious lesions and delineate their extent under challenging visual conditions. Variations in polyp size, shape, texture, and contrast, together with weak boundaries, illumination changes, artefacts, and cross-dataset distribution shift, complicate this task. This study introduces Swin-MSEG-Pro, a reliability-focused segmentation framework that integrates a SwinV2-based encoder with multi-scale feature learning, deep supervision, evidential uncertainty learning, and boundary-focused supervision. The model was developed using Kvasir-SEG with 800 training, 100 validation, and 100 held-out test images; ETIS-LaribPolypDB and CVC-ClinicDB were reserved exclusively for external testing. Across three independent random seeds, Swin-MSEG-Pro achieved a validation Dice of 0.9289 ± 0.0064 and mean per-image Dice scores of 0.9153 ± 0.0024, 0.7174 ± 0.0016, and 0.8527 ± 0.0020 on the held-out Kvasir-SEG, ETIS-LaribPolypDB, and CVC-ClinicDB test sets, respectively. Evidential uncertainty remained informative for selective risk ranking, with TTA object-level RC-AURC values of 0.0487 ± 0.0045, 0.1027 ± 0.0195, and 0.0769 ± 0.0034 on the corresponding datasets. Component-wise ablation showed the clearest degradation from removing the boundary mechanism on ETIS-LaribPolypDB, where Dice decreased by 0.0188. Swin-MSEG-Pro therefore provides reproducible uncertainty-aware polyp segmentation and useful risk information under cross-dataset shift, although performance declines on more strongly shifted external data.

1. Introduction

Accurate pixel-wise delineation of colourectal polyps is an important component of computer-aided colonoscopy because lesion extent, rather than lesion presence alone, must be localised for subsequent assessment [1]. Segmentation remains difficult when polyps are small or morphologically variable, have low contrast or indistinct borders, or are affected by specular reflections and other endoscopic artefacts [2]. These difficulties are compounded by distribution shift: models trained on one image distribution can generalise poorly to data acquired with different devices, hospitals, or imaging protocols [3]. Reliable polyp segmentation must therefore preserve lesion structure under weak visual evidence while remaining informative when predictions become less dependable outside the development distribution.

Recent work addresses these challenges from several directions. Boundary-aware methods explicitly preserve or refine contour information; MEGANet injects Laplacian-derived edge cues, whereas DSF-BRNet combines semantic alignment with high-frequency boundary refinement while considering computational efficiency [1,4]. Global–local modelling has also moved beyond conventional CNNs: Wang et al. combine nested U-Net processing with transformer context, and Bi-CoMA uses parallel convolutional and transformer streams to integrate local spatial detail with global contextual information [5,6]. State-space and frequency-aware designs extend this range: Polyp-Mamba exploits Mamba for long-range interaction modelling with linear computational complexity, while FARF-Net adapts receptive fields using frequency cues and edge-enhanced decoding [7,8]. Uncertainty-aware polyp segmentation is likewise established; uncertain-area exploration and UM-Net explicitly model uncertainty to identify ambiguous regions

* Corresponding author.
E-mail addresses: shahriarmuhtasim@gmail.com (M.M. Shahriar), rafihafiz2016@gmail.com (M.M.G. Hafiz), firoz.mridha@aiub.edu (M.F. Mridha), md.rakibul.islam@mdu.se (M.R. Islam).

https://doi.org/10.1016/j.array.2026.101265
Received 5 July 2026; Received in revised form 17 September 2026; Accepted 17 September 2026
Available online 23 September 2026
2590-0056/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

or communicate prediction reliability [9,10]. More recently, WMA-Net combines wavelet processing, Mamba modelling, edge enhancement, and uncertainty-aware refinement, framing its contribution as coordinated integration rather than a new individual operator [2]. The literature thus shows substantial progress in boundary modelling, multi-scale representation, transformer or state-space context modelling, uncertainty handling, and efficiency, although these objectives are commonly studied under different architectures and evaluation protocols.

The present study does not claim fundamentally new architectural primitives. It instead examines segmentation accuracy, uncertainty-based selective reliability, component-level evidence, external-domain behaviour, inference-time contributions, computational cost, and challenge-specific failure modes within one controlled protocol. These dimensions are not consistently assessed together in the reviewed literature, which makes architectural effects difficult to separate from evaluation choices and obscures the contributions of individual mechanisms to accuracy and reliability. Swin-MSEG-Pro is therefore positioned as a system-level, reliability-oriented integration of established mechanisms, not as the first use of SwinV2, deep supervision, CBAM, evidential learning, or boundary supervision. Its contribution is the coordinated use of these components with a reproducible evaluation design that isolates their roles and quantifies variability across independent runs. The evaluation also distinguishes source-domain from fully external testing and reports limitations without presuming that every added module improves overlap accuracy.

Swin-MSEG-Pro uses a hierarchical SwinV2 encoder with multi-scale feature processing and prediction fusion, complemented by deep supervision and CBAM-based feature refinement. Evidential deep learning provides the principal explicit predictive-uncertainty signal, whereas the boundary branch is used conservatively as contour-focused supervision and a boundary-ambiguity proxy for weighting rather than as a calibrated uncertainty estimator. Model development is restricted to Kvasir-SEG, using 800 training, 100 validation, and 100 held-out test images; the complete ETIS-LaribPolypDB and CVC-ClinicDB cohorts remain untouched external test sets and do not influence training, checkpoint selection, or threshold selection. Evaluation is repeated across three independent seeds and combines conventional segmentation assessment with component ablation, paired statistical analysis, TTA/post-processing isolation, selective-risk and calibration analysis, domain-shift quantification, computational efficiency, challenge-specific failure analysis, and held-out-Kvasir Grad-CAM as supporting interpretability evidence. This design evaluates the framework as an accuracy–reliability–complexity trade-off without claiming universal state-of-the-art performance, solved domain shift, or strong probability calibration.

The main contributions are summarised as follows:

- An integrated SwinV2-based multi-scale segmentation framework that combines deep supervision, CBAM-guided feature refinement, evidential uncertainty learning, and boundary-focused supervision; the individual ingredients are treated as established components rather than independent methodological novelties.
- A controlled external-generalisation protocol in which model development is confined to Kvasir-SEG, while ETIS-LaribPolypDB and CVC-ClinicDB are retained as fully external cohorts with no role in optimisation, model selection, or threshold selection.
- Rigorous component and inference validation using three independent seeds, A1–A5 component-wise ablations, paired statistical analysis, and explicit isolation of test-time augmentation and post-processing effects.
- A reliability-oriented evaluation that combines selective-risk and calibration analysis, feature-space domain-shift quantification, computational efficiency, challenge-specific failure analysis, and held-out-Kvasir Grad-CAM as supporting rather than methodological evidence.

The remainder of this paper is organised as follows. Section 2 reviews the related literature. Section 3 presents Swin-MSEG-Pro, and Section 4 describes the datasets, implementation details, and evaluation protocol. Sections 5 and 6 report and interpret the experimental findings, respectively, before Section 7 concludes the paper.

2. Literature review

Research on gastrointestinal polyp segmentation has moved from conventional encoder–decoder refinement towards heterogeneous designs that combine boundary modelling, multi-scale representation, global-context learning, frequency or state-space processing, and reliability-oriented analysis. For the present study, the relevant issue is how recent methods distribute attention across lesion delineation, contextual modelling, uncertainty, generalisation, and computational cost. The following review is organised around these methodological trends instead of cataloguing individual architectures.

2.1. Boundary-aware and CNN-based polyp segmentation

CNN and U-Net-style encoder–decoder models remain an important basis for polyp segmentation because they preserve local spatial structure while supporting hierarchical feature fusion. Maintaining accurate contours nevertheless remains difficult when polyps are small, low-contrast, irregular, or visually similar to surrounding mucosa. Guo et al. explicitly formulated blurry boundaries and foreground–background resemblance as uncertain regions and progressively mined boundary areas under auxiliary supervision [9]. Polyper similarly treats contour ambiguity as a first-order problem, using interior polyp information to strengthen boundary discrimination instead of assuming that accurate borders will emerge from region prediction [11].

Recent work reinforces that boundary-aware learning is already an established direction. MEGANet uses Laplacian-derived edge information within multi-scale edge-guided attention to emphasise weak polyp boundaries [1]. LMSGNet combines semantic priors with edge guidance, reflecting a broader shift from isolated edge extraction towards refinement that is conditioned by higher-level lesion semantics [12]. At the feature-fusion level, the Cross-level Feature Aggregation Network explicitly coordinates information across hierarchical representations, illustrating that segmentation errors may also arise from inadequate interaction between shallow spatial detail and deeper semantic features [13]. DSF-BRNet makes this relationship particularly explicit by aligning cross-level semantics before high-frequency boundary refinement, while also treating computational efficiency as a design objective [4].

These studies establish contour supervision, edge cues, semantic guidance, and multi-scale fusion as mature ingredients. They also indicate that stronger boundary emphasis is most useful when coupled with reliable lesion localisation and feature alignment. The boundary-related contribution of Swin-MSEG-Pro is therefore positioned as contour-focused auxiliary supervision within a wider reliability-oriented system, not as the invention of boundary-aware polyp segmentation or an independently calibrated uncertainty estimator.

2.2. Transformer, hybrid, state-space, and foundation-model approaches

Transformer-based polyp segmentation was motivated largely by the need to complement convolutional locality with broader contextual dependencies. ColonFormer demonstrates an efficiency-oriented transformer formulation with hierarchical decoding, showing that global modelling and computational restraint can be pursued together [14]. CTNet instead uses a contrastive transformer design to address camouflage-like appearance, scale variation, and small targets, illustrating the continued emphasis on difficult morphology rather than on transformer adoption alone [15]. More directly, Wang et al. combine CNN-derived local representation with transformer-based long-range

semantic modelling and multi-resolution feature fusion, and also report cross-dataset experiments [5]. BiCoMA extends this local–global principle through parallel convolutional and transformer streams whose representations are fused hierarchically [6].

The recent literature also shows that the transformer design space is not converging on a single backbone. Endoscopic Adaptive Transformer adapts transformer processing specifically to endoscopic imagery, whereas PolySegNet combines Swin Transformer and Vision Transformer representations for polyp segmentation [16,17]. State-space models provide another route to long-range interaction: PolypMamba uses visual Mamba to model long-range dependencies while retaining linear computational complexity [7]. Frequency-aware models pursue a related but distinct goal. DDFFormer explicitly combines spatial- and frequency-domain contexts because purely spatial transformer processing may underuse frequency information relevant to structural discrimination [18]. FARF-Net further couples frequency-guided adaptive receptive fields with explicit edge enhancement, demonstrating that receptive-field adaptation and boundary modelling can be coordinated within the same segmentation framework [8].

WMA-Net is especially informative for novelty positioning. It organises wavelet-domain semantic–boundary separation, Mamba-based aggregation, and attention-based refinement as a coordinated system and explicitly frames its contribution around the combination of established components rather than a wholly new primitive operator [2]. This precedent supports a conservative system-level interpretation of integrated architectures such as Swin-MSEG-Pro.

Foundation-model adaptation has also entered polyp segmentation without removing domain-specific reliability problems. ASPS augments Segment Anything for polyp segmentation and targets limitations associated with natural-image pretraining, prompting, and unreliable out-of-distribution confidence [19]. SpectraNet likewise builds on a frozen SAM2-family encoder with lightweight adaptation and spectral guidance, illustrating that foundation models still require task-specific mechanisms for endoscopic structure and boundary information [20]. Current research consequently spans CNN–Transformer hybrids, state-space models, spectral processing, and foundation-model adaptation, with persistent trade-offs among local detail, global context, boundary fidelity, efficiency, and resilience to distribution shift.

2.3. Uncertainty, reliability, and cross-dataset generalisation

Reliability-oriented polyp segmentation extends beyond producing a high-overlap binary mask. Uncertainty-aware methods have begun to represent ambiguous image regions explicitly: the uncertain-area framework uses uncertain regions to guide feature refinement, while UM-Net combines augmentation and uncertainty modelling so that prediction is accompanied by information about uncertainty rather than segmentation alone [9,10]. These approaches establish uncertainty as a relevant output, but they do not make uncertainty ranking, selective-risk behaviour, and probability calibration interchangeable concepts. A model may rank difficult cases usefully while still being imperfectly calibrated, so calibration claims require dedicated evaluation rather than being inferred from the existence of an uncertainty map.

Generalisation introduces a second reliability dimension. Source-free domain-adaptation research reports performance degradation when data originate from different devices, hospitals, or imaging protocols, highlighting dataset bias and domain gaps as practical constraints [3]. More recent segmentation work makes the problem explicit: Wang et al. report cross-dataset validation, while PrysmNet directly frames cross-domain generalisation and boundary precision as deployment challenges [5,21]. External benchmark testing is therefore stronger evidence than a single internal split, but it should not be equated with prospective clinical validation. Protocol exposure also matters: training on multiple benchmark datasets, fine-tuning on target-like data, or using a single source produces materially different generalisation settings, so cross-paper rankings should be interpreted in light of those differences rather than as directly controlled comparisons.

Computational cost is part of the same reliability–deployment problem. Recent methods such as DSF-BRNet explicitly optimise for an accuracy–efficiency balance, whereas FARF-Net reports a dedicated computational-complexity ablation and acknowledges remaining computational overhead [4,8]. This literature makes it inappropriate to present a high-capacity model as lightweight merely because it is accurate; efficiency must be treated as a separate axis of comparison.

Interpretability provides supporting rather than decisive reliability evidence. XPolypNet uses Grad-CAM to visualise regions contributing to polyp predictions, and TCFMA-Net similarly employs Grad-CAM-based feature visualisation to inspect model attention [22,23]. Such analyses can expose spatial behaviour, but qualitative or dataset-restricted explanations do not establish clinical validity or cross-domain explanation stability. Interpretability is therefore treated as complementary evidence alongside segmentation, uncertainty, domain-shift, and efficiency analyses. Given the heterogeneity of training and evaluation protocols across recent studies, Table 1 qualitatively positions representative methods by modelling direction, boundary handling, reliability or uncertainty modelling, and evaluation setting; it is not a numerical performance ranking.

2.4. Synthesis and research gap

No single ingredient supports a strong claim of architectural novelty. Boundary refinement, multi-scale fusion, transformer-based context, CNN–Transformer hybridisation, state-space modelling, frequency-domain processing, uncertainty-aware learning, external testing, and efficiency-oriented design have all been investigated in polyp segmentation. The remaining gap is narrower: these dimensions are rarely examined together under one controlled reliability-oriented protocol.

In particular, comparatively few studies jointly combine single-source development with untouched external cohorts, repeated training across random seeds, component-wise ablation with statistical validation, explicit isolation of test-time augmentation and post-processing, uncertainty-based selective-risk analysis together with calibration assessment, feature-space domain-shift analysis, computational trade-off reporting, systematic evaluation of small-polyp, low-contrast, and ambiguous-boundary failure modes, and supporting interpretability analysis. This is an evaluation-and-integration gap rather than evidence that prior work lacks boundary modelling, global context, uncertainty estimation, or external testing.

Swin-MSEG-Pro is therefore positioned conservatively as a system-level reliability-oriented integration. Its contribution lies in coordinating a SwinV2-based multi-scale segmentation pipeline, deep supervision, evidential uncertainty learning, and contour-focused boundary supervision, and in evaluating that system under a broader controlled reliability protocol. This framing does not imply that the individual components are fundamentally new, that every component independently improves overlap accuracy, that domain shift is solved, or that uncertainty estimates are strongly calibrated; those questions must remain tied to the corresponding empirical analyses.

3. Methodology

This section presents the formulation and architecture of Swin-MSEG-Pro, including evidential uncertainty estimation, multi-scale prediction, boundary-focused supervision, and the end-to-end training and inference workflow. The subsequent experimental protocol specifies the datasets, preprocessing, optimisation, evaluation, reliability analysis, statistical testing, and reproducibility settings.

Table 1
Qualitative positioning of representative recent polyp-segmentation methods relative to Swin-MSEG-Pro. The table summarises methodological emphasis rather than quantitative performance because training and evaluation protocols differ across studies.

Method	Main modelling direction	Boundary handling	Reliability/uncertainty	External evaluation
MEGANet [1]	CNN + edge-guided attention	Laplacian edge guidance	Not reported	Multi-benchmark evaluation
Wang et al. [5]	CNN–Transformer	No explicit boundary branch	Not reported	Cross-dataset validation
BiCoMA [6]	CNN–Transformer dual stream	Spatial refinement for boundary preservation	Not reported	Multi-benchmark evaluation
FARF-Net [8]	Frequency-guided adaptive receptive fields	Explicit edge modelling	Not reported	Multi-benchmark evaluation
DSF-BRNet [4]	Semantic alignment and fusion	High-frequency boundary refinement	Not reported	Multi-benchmark evaluation
WMA-Net [2]	Wavelet + Mamba + attention	Boundary-aware progressive refinement	Uncertainty-aware refinement	Seen/unseen benchmark evaluation
Swin-MSEG-Pro	**SwinV2 + multi-scale prediction**	**Boundary-focused supervision**	**EDL-based selective-risk analysis**	**Kvasir-only development; external ETIS/CVC testing**

3.1. Problem definition

Let $x \in \mathbb{R}^{H\times W\times 3}$ denote an RGB colonoscopy image and $y \in \{0,1\}^{H\times W}$ its binary ground-truth mask, where $y_{ij} = 1$ represents a polyp pixel and $y_{ij} = 0$ represents background. The segmentation objective is to learn a mapping f_θ from x to a foreground probability map $\hat{y} \in [0,1]^{H\times W}$.

Swin-MSEG-Pro formulates this segmentation task together with evidential uncertainty estimation. For each pixel, the main prediction head produces class logits z_k for $K = 2$ classes. Non-negative evidence and the corresponding Dirichlet parameters are defined as

$$e_k = \mathrm{softplus}(z_k), \qquad \alpha_k = e_k + 1, \tag{1}$$

$$S = \sum_{j=1}^{K} \alpha_j, \qquad p_k = \frac{\alpha_k}{S}, \qquad u = \frac{K}{S}. \tag{2}$$

Here, p_k denotes the expected probability of class k, S is the total Dirichlet strength, and u is the corresponding vacuity-based uncertainty. The foreground probability used for lesion segmentation is obtained from the foreground class expectation, while an auxiliary boundary branch provides a separate contour-focused response used during boundary-aware supervision.

3.2. Architecture and multi-scale prediction

Swin-MSEG-Pro uses the pretrained `swinv2_base_window16_256` backbone as a hierarchical feature extractor for 256 × 256 inputs. The three encoder feature maps are processed by RFBModified blocks to enrich their receptive fields and project them to a common representation before hierarchical fusion by AggregationCPD. The aggregation pathway progressively combines deeper semantic features with shallower spatial information. In parallel, a low-level branch applies shallow 3 × 3 convolutions and two max-pooling operations directly to the input image; its output is resized to the aggregation scale and concatenated with the fused encoder representation.

The resulting feature tensor is refined by two convolutional blocks, followed by CBAM-based channel–spatial reweighting and dropout. Four prediction heads are used: a two-channel main evidential segmentation head, two auxiliary segmentation heads for deep supervision, and a one-channel boundary-focused head. The prediction outputs are resized to the input resolution for supervision and inference. Fig. 1 summarises the complete feature-extraction, fusion, refinement, and prediction pipeline.

3.3. Boundary-focused supervision and training objective

The evidential formulation defined above provides the principal predictive uncertainty signal. A separate auxiliary boundary head is used to introduce contour-focused supervision. Let $u_b \in [0,1]^{H\times W}$ denote the sigmoid response of this head and let $B(y)$ denote the morphological boundary band derived from the reference mask. Because $B(y)$ is constructed deterministically from a single annotation rather than from repeated annotations or an explicit noise model, u_b is treated as a supervised boundary-ambiguity proxy rather than as an independently calibrated estimate of aleatoric uncertainty.

The boundary response is additionally incorporated into the segmentation objective through the spatial weighting function

$$w = \mathrm{clip}\big(1 - u_b \odot B(y),\, 0.10,\, 1.0\big). \tag{3}$$

The weighting is active only within the contour band because $B(y) = 0$ elsewhere. Consequently, stronger boundary responses reduce the segmentation-loss contribution within the contour band, while the lower bound of 0.10 prevents these pixels from being ignored completely. The boundary head is simultaneously supervised using binary cross-entropy, providing an explicit contour-learning signal despite this downweighting.

For the main and two auxiliary segmentation heads, evidential supervision consists of the expected cross-entropy under the predicted Dirichlet distribution together with KL regularisation towards a uniform Dirichlet prior. The KL coefficient is annealed during the first 10 epochs and reaches a maximum weight of 1.0. Let L_{main}, L_{aux1}, and L_{aux2} denote the corresponding evidential losses, L_{dice} the boundary-weighted soft Dice loss, and L_{bnd} the boundary binary-cross-entropy loss. The total objective is

$$\begin{aligned} L_{\mathrm{total}} = {} & L_{\mathrm{main}} + \lambda_{\mathrm{aux1}} L_{\mathrm{aux1}} + \lambda_{\mathrm{aux2}} L_{\mathrm{aux2}} \\ & + 0.35 L_{\mathrm{dice}} + 0.25 L_{\mathrm{bnd}}. \end{aligned} \tag{4}$$

The auxiliary weights are initialised as $\lambda_{\mathrm{aux1}} = 0.35$ and $\lambda_{\mathrm{aux2}} = 0.15$. Their decay begins at epoch 30, with both constrained to a minimum value of 0.05 during later training. To compensate for the sparsity of contour pixels, the boundary BCE uses a positive-class weight of 18.8845. The three-pixel construction of $B(y)$ is specified in the experimental protocol below.

3.4. Overall training and inference procedure

Algorithm 1 summarises the end-to-end optimisation and inference workflow. The surrounding subsections define the architectural components, loss formulation, and implementation settings.

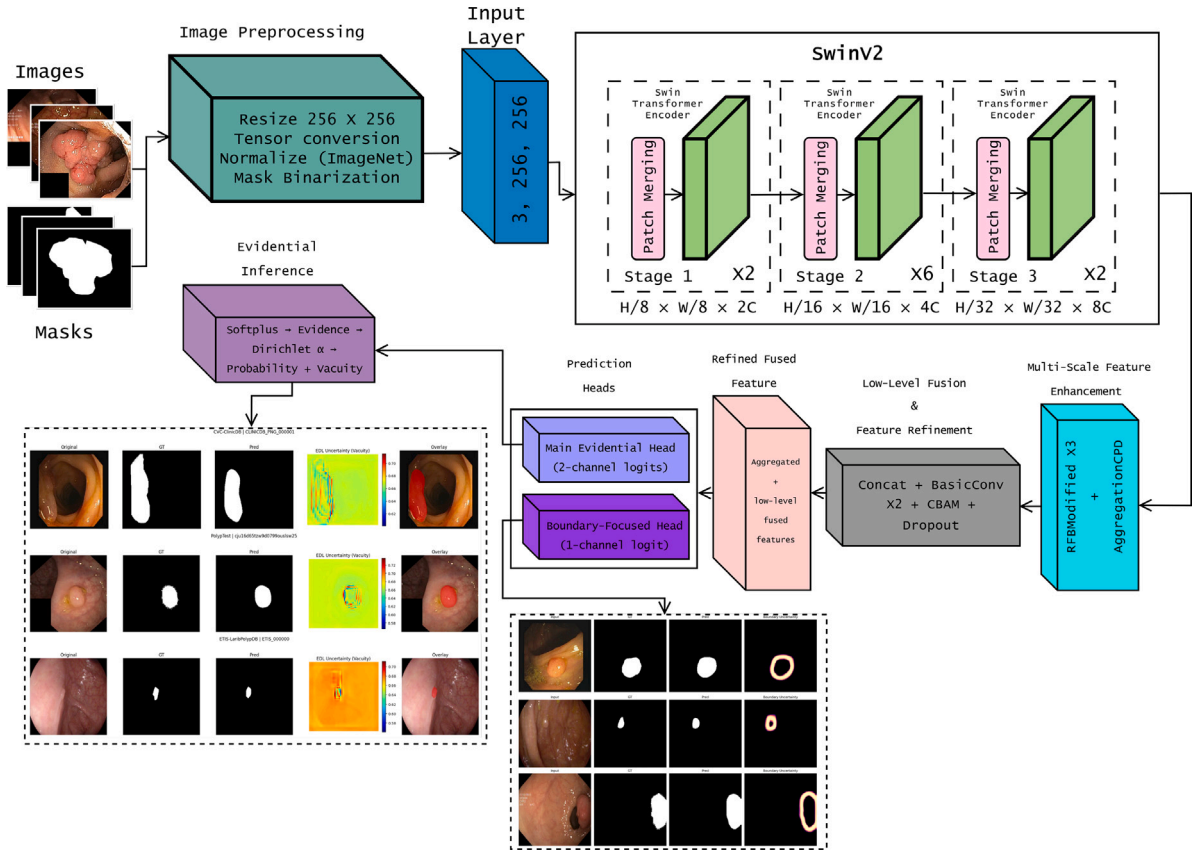

Fig. 1. Model architecture of Swin-MSEG-Pro.

Algorithm 1 Training and inference procedure of Swin-MSEG-Pro

Require: Image x; ground-truth mask y during training
Ensure: Lesion mask $\hat{y}$ and evidential uncertainty
1: Preprocess x and extract the multi-scale representation using Swin-MSEG-Pro
Training:
2: Generate the contour target $B(y)$ and predict main, auxiliary, and boundary outputs
3: Compute L_{total} and update the model parameters
4: Select the EMA checkpoint and operating threshold using validation data only
Inference:
5: Generate the predefined TTA views, inverse-align the predictions, and average them
6: Apply the validation-selected operating threshold
7: Fill internal holes and retain the largest connected component to obtain $\hat{y}$

4. Experimental setup and evaluation protocol

This section describes the datasets, preprocessing and augmentation procedures, boundary-target construction, validation and external-testing protocol, implementation settings, evaluation measures, reliability and interpretability analyses, statistical procedures, and reproducibility settings.

4.1. Dataset description

Kvasir-SEG [24] was used as the sole source dataset for model development. It contains 1000 colonoscopy images with corresponding pixel-wise polyp annotations. A fixed split of 800 training, 100 validation, and 100 held-out test images was used throughout the study. The training partition was used for parameter optimisation, whereas the validation partition was used exclusively for checkpoint and operating-threshold selection.

Cross-dataset generalisation was evaluated on ETIS-LaribPolypDB [25] and CVC-ClinicDB [26], containing 196 and 612 annotated images, respectively. Both datasets were used entirely for external testing and did not contribute to training, validation, checkpoint selection, or threshold optimisation. Table 2 summarises the dataset composition and evaluation roles.

4.2. Preprocessing (image resolution and normalisation)

All colonoscopy images were converted to RGB format and the corresponding annotations to single-channel binary masks. Images and masks were resized to 256 × 256 pixels, using bilinear interpolation for images and nearest-neighbour interpolation for masks to preserve discrete label values. The resized images were converted to tensors and normalised using the ImageNet statistics $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$. Segmentation masks were binarised using the condition $m > 0$ to obtain foreground targets in $\{0, 1\}$.

4.3. Data augmentation

Training-time augmentation was applied only to the 800-image Kvasir-SEG training partition; no augmentation was applied to the validation or test datasets. Spatial augmentation consisted of random horizontal flipping with probability 0.5, random vertical flipping with probability 0.2, and random rotation within −25° to 25°. A random resize–crop–pad transformation was additionally applied using a scale factor sampled from 0.75 to 1.25. Samples larger than the target resolution were randomly cropped, whereas smaller samples were zero-padded to restore the fixed 256 × 256 input size. Identical spatial

Table 2
Dataset description and split protocol.

Dataset	Images	Role	Split/usage	Annotation
Kvasir-SEG	1000	Train/val/internal test	800/100/100	Pixel-level binary polyp masks
ETIS-LaribPolypDB	196	External test	External test only	Pixel-level binary polyp masks
CVC-ClinicDB	612	External test	External test only	Pixel-level binary polyp masks

transformations were applied to each image and its corresponding segmentation mask.

Appearance augmentation included random perturbations of brightness, contrast, and saturation, together with Gaussian noise, Gaussian blur, JPEG-compression perturbation, and autocontrast. Autocontrast was applied with probability 0.25. CutMix augmentation was additionally enabled during training with probability 0.35. These appearance-based transformations were restricted to the training pipeline and were not used during validation or external testing.

4.4. Boundary target generation

Boundary-focused supervision was generated directly from the ground-truth lesion mask. For each binary mask y, a morphological contour band was obtained from the difference between dilation and erosion using a boundary width of three pixels:

$$B(y) = \mathbf{1}(\mathrm{Dilate}(y; r) - \mathrm{Erode}(y; r) > 0), \qquad r = 3. \tag{5}$$

The resulting binary map $B(y)$ defines a contour-focused supervision region, with boundary-band pixels assigned a target value of 1 and all remaining pixels assigned 0. This target is used to train the auxiliary boundary-focused branch using binary cross-entropy. Because the contour band is derived from the reference annotation rather than from repeated annotations or an explicit probabilistic noise model, it is treated as a supervised proxy for regions in which boundary ambiguity is most relevant to segmentation.

Boundary pixels occupy only a small proportion of the image, producing substantial class imbalance in this auxiliary task. A positive-class weight of 18.8845 was therefore applied to the boundary BCE term. The predicted boundary response is subsequently used together with $B(y)$ in the boundary-aware spatial weighting mechanism described in Section 3.

4.5. Validation and external testing protocol

The model-development protocol used the fixed Kvasir-SEG split of 800 training, 100 validation, and 100 held-out test images. ETIS-LaribPolypDB and CVC-ClinicDB were used only for external testing and did not contribute to training, checkpoint selection, or threshold optimisation. Validation was performed after each epoch, and the operating threshold was selected independently for each trained model by searching the interval 0.30–0.75 with a step size of 0.01. The resulting thresholds were 0.50, 0.49, and 0.51 for seeds 42, 123, and 2026, respectively. Once selected on the validation set, the corresponding threshold was kept fixed for the held-out Kvasir-SEG test set and both external datasets.

The primary inference protocol used test-time augmentation with scales of 0.85, 1.00, and 1.15 together with horizontal flipping. Predictions from the transformed views were returned to the common reference frame before averaging. The resulting probability map was thresholded using the validation-selected operating threshold, followed by morphological post-processing consisting of hole filling and retention of the largest connected component. RAW, TTA-only, post-processing-only, and combined TTA-plus-post-processing inference were additionally evaluated separately to isolate the contribution of each inference-time operation.

Table 3
Training hyperparameters and implementation settings used for the three independent runs.

Item	Setting
Backbone	SwinV2-Base, window size 16, input 256 × 256
Random seeds	42, 123, 2026
Optimiser	AdamW; learning rate 2×10^{-4}; weight decay 1×10^{-4}
Schedule	5 warm-up epochs followed by cosine decay to 2×10^{-6}
Training	Maximum 120 epochs; batch size 2; gradient accumulation 2
Encoder warm-up	Frozen during the first 2 epochs
Regularisation	Dropout 0.20; EMA decay 0.995; mixed precision
Selection	Validation every epoch; early stopping from epoch 25; patience 20
Hardware	NVIDIA Tesla T4 GPU

4.6. Implementation details

The model was implemented in PyTorch using the timm implementation of the pretrained SwinV2-Base encoder. All principal full-model experiments were repeated independently using random seeds 42, 123, and 2026 while retaining the same training configuration. Each run was trained for a maximum of 120 epochs with a batch size of 2 and gradient accumulation over two iterations, corresponding to an effective batch size of 4. AdamW optimisation was used with an initial learning rate of 2×10^{-4}, weight decay of 1×10^{-4}, and a minimum learning rate of 2×10^{-6}.

The encoder was frozen during the first two epochs. The learning-rate schedule consisted of five warm-up epochs followed by cosine decay. Mixed-precision training and an exponential moving average (EMA) model with decay 0.995 were used, with EMA-based validation beginning from epoch 5. Early stopping was enabled from epoch 25 with a patience of 20 epochs and a minimum improvement of 0.0005. The principal experiments and runtime measurements were performed using an NVIDIA Tesla T4 GPU. Table 3 summarises the implementation settings shared across the three independent runs.

4.7. Evaluation metrics

Segmentation performance was assessed primarily using the mean per-image Dice coefficient, with intersection over union (IoU) used as a complementary overlap measure. For test image i, the two metrics are defined as

$$\mathrm{Dice}_i = \frac{2TP_i}{2TP_i + FP_i + FN_i}, \tag{6}$$

$$\mathrm{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{7}$$

where TP_i, FP_i, and FN_i denote the pixel-wise true-positive, false-positive, and false-negative counts for image i. The dataset-level Dice reported in the principal evaluation was calculated by averaging the individual image Dice scores,

$$\overline{\mathrm{Dice}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Dice}_i, \tag{8}$$

where N is the number of images in the corresponding test dataset. This image-wise formulation weights each test image equally and differs from a global or micro-averaged Dice score obtained by pooling pixels

across the complete dataset. For the three independently trained models, dataset-level results are reported as the mean ± standard deviation of the corresponding image-wise performance across seeds 42, 123, and 2026. Unless otherwise specified, Dice in the subsequent analyses denotes mean per-image Dice.

4.8. Feature-space domain-shift quantification

To quantify dataset shift independently of Swin-MSEG-Pro, feature-space distributions were analysed using two fixed ImageNet-pretrained feature extractors, ResNet50 and Swin-T. Neither backbone was fine-tuned on the polyp datasets. Features extracted from the Kvasir-SEG training partition were used as the source reference distribution, while the held-out Kvasir-SEG test set served as an in-domain control and ETIS-LaribPolypDB and CVC-ClinicDB were treated as external target distributions.

Distributional separation from the Kvasir-SEG training reference was quantified using source-whitened Fréchet feature distance and kernel maximum mean discrepancy (MMD^2) with an RBF kernel. Larger values of either measure indicate greater separation from the source feature distribution. To compare the magnitude of external shift directly, the ETIS-minus-CVC difference was calculated for both metrics and both feature extractors, with 95% bootstrap confidence intervals used to quantify uncertainty in the estimated differences. Using two independent pretrained representations was intended to reduce dependence of the domain-shift conclusion on a single feature extractor.

4.9. Reliability and interpretability evaluation

In addition to segmentation accuracy, uncertainty quality was evaluated from both calibration and selective-risk perspectives. Expected calibration error (ECE) was used to assess the agreement between predictive confidence and observed segmentation outcomes, whereas risk–coverage analysis was used to determine whether the uncertainty estimates could rank difficult predictions effectively. Object-level RC-AURC was calculated from the evidential uncertainty associated with lesion predictions, with lower values indicating more favourable selective-prediction behaviour. The reliability analysis was performed for both RAW and TTA inference across the three independently trained models.

Boundary-specific reliability was analysed separately within contour regions derived from the reference masks. Boundary-band RC-AURC was computed using two uncertainty signals: the vacuity-based evidential uncertainty produced by the main segmentation branch and the response generated by the explicitly supervised boundary-focused branch. This comparison was included because the two outputs serve different roles within Swin-MSEG-Pro. The evidential output provides uncertainty associated with the foreground–background prediction, whereas the auxiliary boundary branch provides a contour-focused response used during boundary-aware optimisation.

Interpretability was examined using GT-guided Grad-CAM as a complementary spatial analysis rather than as the primary measure of model reliability. The Grad-CAM target used the foreground response within the ground-truth lesion region, and the resulting activation map was binarised using a top-percent threshold of 0.20. Stability was evaluated using 10 Monte Carlo dropout passes. CAM IoU, CAM recall, average pixel-wise CAM standard deviation, and correlation with the mean CAM were used to characterise localisation and stochastic consistency. Quantitative Grad-CAM evaluation was restricted to the held-out Kvasir-SEG test set and was therefore not interpreted as evidence of cross-dataset explanation stability.

4.10. Statistical analysis

Statistical comparisons were performed at the image level to account for paired predictions on the same test samples. For the component-wise ablation analysis, the per-image Dice scores were first averaged across the three random seeds for each configuration, after which the full model was compared with each ablation using the paired Wilcoxon signed-rank test. To control for multiple comparisons, the resulting p values were adjusted independently within each dataset using the Holm procedure across the five ablation comparisons. Statistical significance was assessed at an adjusted threshold of $p < 0.05$.

For each full-model-versus-ablation comparison, the mean paired Dice difference and its 95% bootstrap confidence interval were also calculated to describe the magnitude and direction of the observed effect. Paired Cohen's d_z and rank-biserial correlation were additionally examined as effect-size measures. Positive full-minus-ablation differences indicate higher Dice for Swin-MSEG-Pro, whereas negative differences indicate higher Dice for the corresponding ablation.

The effect of test-time augmentation was analysed in the same paired manner by comparing TTA-only and RAW predictions for the same test images. Holm correction was applied across the three dataset-level TTA comparisons. Statistical significance was interpreted together with the magnitude and confidence interval of the paired difference, since a statistically detectable rank difference does not necessarily imply a practically meaningful improvement in mean Dice.

4.11. Reproducibility note

All principal experiments used the same fixed Kvasir-SEG train–validation–test partition and identical model architecture, preprocessing, augmentation, optimisation, loss, checkpoint-selection, and inference settings. Independent runs were conducted using seeds 42, 123, and 2026. For each run, checkpoint and operating-threshold selection relied exclusively on the Kvasir-SEG validation set, after which the selected model and threshold were kept unchanged for the held-out Kvasir-SEG test set and the external ETIS-LaribPolypDB and CVC-ClinicDB evaluations. The relevant architectural, training, and statistical settings are reported in the preceding subsections to support reproduction of the experimental protocol.

5. Results

The evaluation distinguishes the 100-image held-out Kvasir-SEG test partition from the two external test cohorts, ETIS-LaribPolypDB (n = 196) and CVC-ClinicDB (n = 612). All principal Swin-MSEG-Pro experiments were repeated using three random seeds (42, 123, and 2026), and performance is reported as mean ± standard deviation across runs. Mean per-image Dice is used as the primary segmentation measure, with uncertainty-based reliability measures reported separately. The analysis includes repeated full-model evaluation, component-wise ablation, paired statistical testing, separate assessment of test-time augmentation and post-processing, uncertainty and calibration analysis, systematic failure analysis, and computational-efficiency evaluation. The corresponding validation Dice was 0.9289 ± 0.0064. Validation was used only for checkpoint and operating-threshold selection; the selected threshold for each run was then fixed for the held-out Kvasir-SEG test set and both external cohorts.

5.1. Cross-dataset external validation performance

Across three independent runs, Swin-MSEG-Pro achieved mean per-image Dice scores of 0.9153 ± 0.0024 on the held-out Kvasir-SEG test set, 0.7174 ± 0.0016 on ETIS-LaribPolypDB, and 0.8527 ± 0.0020 on CVC-ClinicDB (Table 4). ETIS-LaribPolypDB showed the largest performance reduction, while the small across-seed variation indicates that this dataset-specific pattern was reproducible. The results support consistent external transfer, but not the conclusion that distribution shift was resolved.

Table 4
Held-out and external segmentation performance of Swin-MSEG-Pro across three random seeds. Values are reported as mean ± standard deviation of the per-image Dice score.

Dataset	Evaluation setting	Dice
Kvasir-SEG	Held-out test	0.9153 ± 0.0024
ETIS-LaribPolypDB	External test	0.7174 ± 0.0016
CVC-ClinicDB	External test	0.8527 ± 0.0020

5.2. Component-wise ablation analysis

Table 5 isolates deep supervision (A1), CBAM (A2), evidential learning (A3), boundary-focused supervision and weighting (A4), and all four enhancements jointly (A5); every configuration followed the same three-seed protocol. Effects varied by dataset. Removing the boundary mechanism produced the clearest positive external evidence for the full model, reducing ETIS-LaribPolypDB Dice from 0.7174 ± 0.0016 to 0.6986 ± 0.0229 (mean paired difference 0.0188, 95% CI: 0.0056–0.0325; p_H = 0.00053). CBAM showed smaller effects, while deep supervision produced no significant paired difference on any dataset after Holm correction.

Evidential learning did not consistently improve overlap accuracy. The No-EDL model achieved higher Dice on ETIS-LaribPolypDB (0.7399 ± 0.0114; full-minus-A3 difference −0.0225, 95% CI: −0.0395 to −0.0071; p_H = 0.00157) and CVC-ClinicDB (0.8594 ± 0.0062; difference −0.0067, 95% CI: −0.0116 to −0.0023; $p_H = 2.31 \times 10^{-11}$). A5 was also competitive in raw Dice, including a significant CVC-ClinicDB advantage. Conversely, some rank-significant differences, such as A2 on CVC-ClinicDB, had negligible mean Dice differences. Statistical significance is therefore not equated with practical improvement, and the evidential component is evaluated primarily through uncertainty and selective-risk behaviour.

5.3. Effect of test-time augmentation and post-processing

Table 6 separates RAW, TTA-only, post-processing-only (PP), and TTA+PP inference using the same three full-model checkpoints. TTA yielded the useful external-domain improvement: relative to RAW inference, mean Dice increased by 0.0163 on ETIS-LaribPolypDB (95% CI: 0.0029–0.0310; p_H = 0.0028) and by 0.0058 on CVC-ClinicDB (95% CI: 0.0024–0.0093; $p_H = 2.33 \times 10^{-9}$). The Kvasir-SEG change was practically negligible despite a detectable rank difference. PP alone did not provide a consistent benefit and reduced mean Dice on all three datasets; TTA+PP also remained below TTA-only performance. TTA+PP is retained for the principal results because it was the predefined inference protocol, while this ablation identifies TTA as the useful inference-time contributor.

5.4. Comparison with recent polyp segmentation methods

Table 7 compares Swin-MSEG-Pro with representative polyp-segmentation methods published between 2024 and 2026 using the mean Dice scores reported in the corresponding source papers. Because the training protocols are not equivalent, dataset exposure is reported explicitly. MEGANet, FARF-Net, DSF-BRNet, and WMA-Net use the common PraNet-style joint-training protocol with 900 Kvasir-SEG and 550 CVC-ClinicDB images, while BiCoMA likewise reports a 1450-image combined Kvasir-SEG/CVC-ClinicDB training set. Consequently, their reported Kvasir-SEG and CVC-ClinicDB test subsets are drawn from datasets represented during training, whereas ETIS-LaribPolypDB is unseen.

The 2024 multi-scale nested UNet–Transformer study provides a particularly relevant protocol comparison because it also reported a Kvasir-only cross-dataset experiment. When trained only on Kvasir-SEG, that model achieved Dice scores of 0.855 on CVC-ClinicDB and 0.654 on ETIS-LaribPolypDB. Under the present Kvasir-only protocol, Swin-MSEG-Pro achieved 0.8527 ± 0.0020 and 0.7174 ± 0.0016 on the complete CVC-ClinicDB and ETIS-LaribPolypDB cohorts, respectively. External transfer to CVC-ClinicDB was therefore comparable, while the ETIS-LaribPolypDB result was higher under the respective single-source experiments. Differences in source splits, input resolution, optimisation, and inference nevertheless prevent a controlled head-to-head interpretation.

The jointly trained recent methods report higher absolute ETIS-LaribPolypDB Dice scores, ranging from 0.739 for MEGANet (Res2Net-50) to 0.838 for WMA-Net. These results are important recent references but are not directly comparable with the present single-source setting because CVC-ClinicDB contributes training images to those models. The comparison therefore supports competitive external transfer under a more restrictive Kvasir-only development protocol, not a claim of universal state-of-the-art Dice performance.

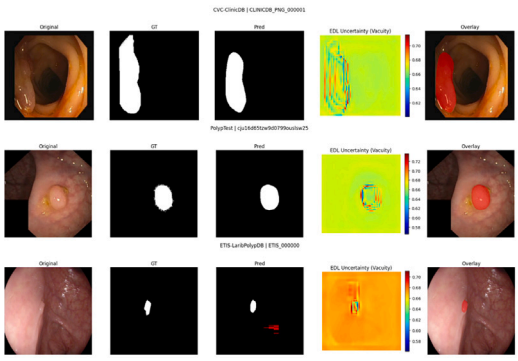

Fig. 2. Qualitative segmentation and evidential uncertainty outputs of Swin-MSEG-Pro. Each row shows the input image, ground-truth mask, predicted lesion mask, EDL vacuity-based uncertainty map, and segmentation overlay.

5.5. Qualitative segmentation and EDL uncertainty analysis

Fig. 2 shows representative segmentation and evidential-vacuity maps from the held-out and external cohorts. The uncertainty maps provide spatial information beyond the binary mask, but are treated as qualitative evidence only; uncertainty quality is assessed quantitatively through calibration and risk–coverage measures below.

5.6. Domain shift and reliability analysis

Independent feature-space analyses using fixed ResNet50 and Swin-T representations identified ETIS-LaribPolypDB as the largest measured shift from the Kvasir-SEG training distribution. For ResNet50, the CVC-ClinicDB and ETIS-LaribPolypDB Fréchet distances were 52.37 and 88.57, respectively, with corresponding MMD^2 values of 0.1221 and 0.2747. Swin-T produced the same ordering, with Fréchet distances of 52.48 and 82.52 and MMD^2 values of 0.1231 and 0.2527. This agreement across two independent representations supports the interpretation that ETIS-LaribPolypDB constitutes the strongest measured external shift among the evaluated cohorts.

Table 8 summarises the corrected uncertainty results. TTA modestly reduced object-level RC-AURC on all three datasets, indicating improved selective risk ranking. ECE, however, remained approximately 0.30 and did not improve consistently; the evidential formulation therefore does not demonstrate strong probability calibration. At lesion boundaries, EDL-derived uncertainty produced lower RC-AURC than the explicitly supervised boundary head on every dataset. The boundary branch is consequently interpreted primarily as contour-focused supervision and weighting, not as a superior calibrated uncertainty estimator.

Table 5
Component-wise ablation analysis across three random seeds. Dice values are reported as mean ± standard deviation. p_H denotes the Holm-adjusted paired Wilcoxon p value for the comparison between the full model and each ablation within the corresponding dataset.

Configuration	Kvasir-SEG		ETIS-LaribPolypDB		CVC-ClinicDB	
	Dice	p_H	Dice	p_H	Dice	p_H
Full Swin-MSEG-Pro	**0.9153 ± 0.0024**	–	0.7174 ± 0.0016	–	0.8527 ± 0.0020	–
A1: No deep supervision	0.9147 ± 0.0029	0.1223	0.7281 ± 0.0129	0.9003	0.8510 ± 0.0078	0.2987
A2: No CBAM	0.9133 ± 0.0035	**0.00054**	0.7118 ± 0.0327	**0.0194**	0.8527 ± 0.0028	$\mathbf{2.64 \times 10^{-10}}$
A3: No EDL	0.9139 ± 0.0048	0.4598	**0.7399 ± 0.0114**	**0.00157**	**0.8594 ± 0.0062**	$\mathbf{2.31 \times 10^{-11}}$
A4: No boundary-focused supervision	0.9144 ± 0.0071	**0.0180**	0.6986 ± 0.0229	**0.00053**	0.8493 ± 0.0015	**0.0385**
A5: All enhancements off	0.9132 ± 0.0039	0.2749	0.7252 ± 0.0111	0.2592	0.8583 ± 0.0064	$\mathbf{4.33 \times 10^{-5}}$

Table 6
Independent effects of test-time augmentation and morphological post-processing on segmentation performance. Values are mean ± standard deviation of per-image Dice across three random seeds.

Dataset	RAW	TTA only	PP only	TTA + PP
Kvasir-SEG	0.9162 ± 0.0030	**0.9163 ± 0.0012**	0.9154 ± 0.0037	0.9153 ± 0.0024
ETIS-LaribPolypDB	0.7093 ± 0.0038	**0.7257 ± 0.0014**	0.7035 ± 0.0049	0.7174 ± 0.0015
CVC-ClinicDB	0.8507 ± 0.0090	**0.8564 ± 0.0027**	0.8465 ± 0.0097	0.8527 ± 0.0021

Table 7
Comparison with representative recent polyp segmentation methods. Literature values are the mean Dice scores reported in the corresponding source papers. Dataset exposure is shown explicitly because the training and evaluation protocols are not directly equivalent.

Method	Year	Training exposure	Kvasir-SEG	CVC-ClinicDB	ETIS-LaribPolypDB
Multi-scale Nested UNet–Transformer [5]	2024	Kvasir only†	–	0.855^{E}	0.654^{E}
MEGANet (Res2Net-50) [1]	2024	900 Kvasir + 550 CVC	0.913^{S}	0.938^{S}	0.739^{E}
BiCoMA [6]	2025	900 Kvasir + 550 CVC	0.929^{S}	0.941^{S}	0.810^{E}
FARF-Net [8]	2026	900 Kvasir + 550 CVC	0.932^{S}	0.945^{S}	0.815^{E}
DSF-BRNet [4]	2026	900 Kvasir + 550 CVC	0.933^{S}	0.943^{S}	0.818^{E}
WMA-Net [2]	2026	900 Kvasir + 550 CVC	0.936^{S}	0.944^{S}	0.838^{E}
Swin-MSEG-Pro	**2026**	**800 Kvasir only**	0.9153 ± 0.0024^{H}	0.8527 ± 0.0020^{E}	0.7174 ± 0.0016^{E}

^{S}Test subset originates from a dataset represented during training; ^{E}dataset not represented in the reported training set; ^{H}held-out source-domain test partition. †The 2024 study reports a separate Kvasir-only cross-dataset experiment; only its external CVC-ClinicDB and ETIS-LaribPolypDB results from that experiment are shown here. Its in-domain Kvasir-SEG result was obtained under a different split and is therefore not mixed into this row. Literature entries are source-paper point estimates, whereas Swin-MSEG-Pro is reported as mean ± standard deviation across three independent seeds.

Table 8
Uncertainty and selective-reliability results across three random seeds. Object-level results compare RAW and TTA inference; boundary RC-AURC values correspond to the primary TTA+PP protocol. Lower values are preferable.

Dataset	ECE RAW	ECE TTA	Obj. RC-AURC RAW	Obj. RC-AURC TTA	Boundary RC-AURC (EDL)	Boundary RC-AURC (head)
Kvasir-SEG	0.2973 ± 0.0021	0.2973 ± 0.0011	0.0502 ± 0.0058	**0.0487 ± 0.0045**	**0.3239 ± 0.0347**	0.3888 ± 0.0017
ETIS-LaribPolypDB	0.3184 ± 0.0005	0.3205 ± 0.0005	0.1105 ± 0.0259	**0.1027 ± 0.0195**	**0.3452 ± 0.0365**	0.5222 ± 0.0218
CVC-ClinicDB	0.3025 ± 0.0005	0.3035 ± 0.0005	0.0797 ± 0.0036	**0.0769 ± 0.0034**	**0.3766 ± 0.0499**	0.4949 ± 0.0105

5.7. Interpretability analysis

GT-guided Grad-CAM provided a complementary assessment of the spatial evidence supporting the segmentation output. On the held-out Kvasir-SEG test set, Swin-MSEG-Pro achieved a CAM IoU of 0.1391 and CAM recall of 0.8551. Across Monte Carlo dropout passes, the average pixel-wise CAM standard deviation was 0.0180 and the correlation with the mean CAM was 0.9629. These values indicate broad lesion coverage and stable stochastic saliency, although the comparatively low CAM IoU shows that the activation response was less spatially compact than the ground-truth lesion mask. Representative examples are shown in Fig. 3. Because quantitative Grad-CAM evaluation was restricted to the held-out Kvasir-SEG test set, the findings support only within-dataset spatial consistency, not cross-dataset explanation stability.

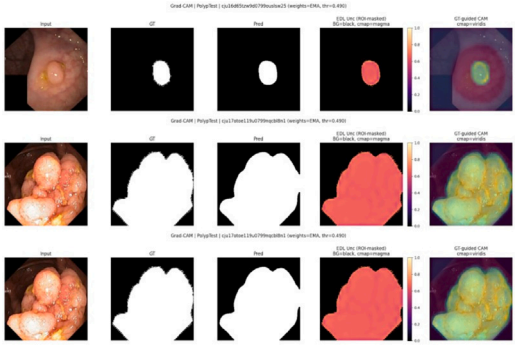

Fig. 3. Representative GT-guided Grad-CAM results on the held-out Kvasir-SEG test set. Each example shows the input image, ground-truth mask, predicted lesion mask, ROI-masked evidential uncertainty, and GT-guided activation response.

5.8. Computational complexity and inference efficiency

Table 9 shows that Swin-MSEG-Pro is a comparatively high-capacity model. It contains 89.91M parameters, requires 17.19 GMACs per 256 × 256 forward pass, and achieved 17.68 ± 1.18 FPS with a latency of 56.73 ± 3.65 ms/image on a Tesla T4. The representative

Table 9
Computational characteristics of Swin-MSEG-Pro and representative recent polyp segmentation methods. Computational values and throughput for competing methods are reported as published in the corresponding studies; cross-paper FPS values are not hardware-normalised.

Method	Parameters	Reported compute	FPS	Runtime setting
FARF-Net [8]	33.71M	11.91 GFLOPs	–	352 × 352, RTX 3090
DSF-BRNet [4]	25.55M	11.37 GFLOPs	80.08	352 × 352, RTX 3090
WMA-Net [2]	40.02M	27.81 GFLOPs	36.93	352 × 352, RTX 3090
Swin-MSEG-Pro	89.91M	17.19 GMACs	17.68 ± 1.18	256 × 256, Tesla T4

Table 10
Challenge-specific failure analysis of Swin-MSEG-Pro. Δ Dice denotes the difference between the challenge subset and the remaining images; negative values indicate performance degradation.

Dataset	Challenge	n	Challenge-subset Dice	Remaining-sample Dice	Δ Dice
Kvasir-SEG	Small polyps	25	0.9075	0.9179	−0.0104
	Low contrast	25	0.8977	0.9212	−0.0235
	Ambiguous boundary	25	0.9278	0.9112	+0.0166
ETIS-LaribPolypDB	Small polyps	49	0.5269	0.7809	−0.2540
	Low contrast	49	0.5864	0.7610	−0.1747
	Ambiguous boundary	49	0.6911	0.7261	−0.0350
CVC-ClinicDB	Small polyps	153	0.7841	0.8756	−0.0915
	Low contrast	153	0.8345	0.8588	−0.0243
	Ambiguous boundary	153	0.8414	0.8565	−0.0151

recent methods are substantially smaller, while DSF-BRNet and WMA-Net report higher throughput on an RTX 3090. Because these studies use different input resolutions, hardware, software environments, and computational units, cross-paper FPS and compute values should not be interpreted as hardware-normalised measurements. Swin-MSEG-Pro is therefore positioned as an accuracy–reliability–complexity trade-off rather than as a lightweight or efficiency-oriented architecture.

5.9. Challenge-specific failure analysis

Table 10 evaluates small-polyp, low-contrast, and ambiguous-boundary challenge subsets. For each challenge, the analysed subset contained one quarter of the corresponding test cohort: 25 of 100 Kvasir-SEG cases, 49 of 196 ETIS-LaribPolypDB cases, and 153 of 612 CVC-ClinicDB cases. Small lesions produced the dominant external degradation, with mean Dice reductions of 0.2540 on ETIS-LaribPolypDB and 0.0915 on CVC-ClinicDB relative to the remaining cases. Low contrast also produced a substantial reduction on ETIS-LaribPolypDB (−0.1747), whereas ambiguous-boundary effects were smaller and dataset-dependent, with no degradation in the selected Kvasir-SEG subset. Small lesions under external-domain shift are therefore the principal residual failure mode, followed by low-contrast cases in the more challenging ETIS cohort.

6. Discussion

6.1. Principal findings

The corrected three-seed evaluation shows reproducible Swin-MSEG-Pro segmentation on the held-out source-domain test set and transfer to both external cohorts, with strongly dataset-dependent degradation. ETIS-LaribPolypDB remained the most challenging cohort and showed the largest measured feature-space shift from the Kvasir-SEG training distribution under two independent pretrained representations. The results support cross-dataset transfer but do not show that distribution shift has been resolved.

Segmentation accuracy alone does not fully characterise the framework. Evidential uncertainty provided a usable signal for selective-risk ranking, while the boundary-focused branch contributed most clearly through contour-oriented supervision rather than superior uncertainty calibration. The framework is therefore interpreted in terms of an accuracy–reliability–boundary trade-off; the results do not show that every architectural addition improves Dice.

6.2. Component contributions and uncertainty interpretation

The component-wise ablations assign distinct, dataset-dependent roles to the added mechanisms. Boundary-focused supervision provides the clearest positive external evidence: its removal produced the largest full-model degradation on ETIS-LaribPolypDB, and the paired difference remained significant after Holm correction. CBAM produced smaller effects, whereas deep supervision did not yield a consistent accuracy advantage across datasets. The competitive performance of the all-enhancements-off configuration further shows that the complete framework cannot be justified through monotonic Dice improvements from every component.

Evidential learning presents an important trade-off. Replacing EDL with conventional cross-entropy training increased mean Dice on ETIS-LaribPolypDB and CVC-ClinicDB, with statistically significant paired differences. EDL should therefore not be interpreted as a direct segmentation-accuracy booster; the evidence supports its use as an explicit uncertainty signal for selective-risk analysis. TTA modestly reduced object-level RC-AURC on all three datasets, whereas ECE remained approximately 0.30 and did not demonstrate strong probability calibration. EDL-derived uncertainty also produced lower boundary RC-AURC than the explicitly supervised boundary head. The boundary head is thus more appropriately regarded as a contour-focused supervision and weighting mechanism, or boundary-ambiguity proxy, than as a superior calibrated uncertainty estimator.

6.3. External generalisation and relation to recent work

The external results should be interpreted in the context of training exposure. Swin-MSEG-Pro was optimised only on 800 Kvasir-SEG training images, while the complete CVC-ClinicDB and ETIS-LaribPolypDB cohorts remained unseen during model development. Under the related Kvasir-only cross-dataset experiment reported by Wang et al. [5], performance on CVC-ClinicDB was broadly comparable, whereas Swin-MSEG-Pro achieved the higher ETIS-LaribPolypDB result. This comparison nevertheless remains non-controlled because the source split, optimisation procedure, input resolution, and inference protocol differ.

Several recent methods in Table 7 report higher absolute ETIS-LaribPolypDB Dice, but MEGANet, BiCoMA, FARF-Net, DSF-BRNet, and WMA-Net use Kvasir-SEG and CVC-ClinicDB jointly during training. Their results are important contemporary references rather than direct head-to-head rankings against the present single-source protocol. The

evidence supports competitive external transfer under a more restrictive development setting, not universal state-of-the-art superiority.

The systematic failure analysis also clarifies where this transfer remains limited. Small lesions produced the dominant external degradation, especially on ETIS-LaribPolypDB, while low contrast was the second most pronounced challenge. Ambiguous-boundary effects were smaller and dataset-dependent. These observations are consistent with the stronger feature-space shift measured for ETIS and indicate that residual generalisation error is concentrated more strongly in difficult lesion-scale and appearance conditions than in boundary ambiguity alone.

6.4. Scope of contribution and computational trade-off

Swin-MSEG-Pro makes a system-level contribution rather than introducing a fundamentally new primitive operator. SwinV2 feature extraction, multi-scale decoding, deep supervision, attention, evidential learning, and boundary-focused supervision are individually established ideas. The contribution lies in integrating these mechanisms within a common reliability-oriented segmentation framework and evaluating their behaviour through repeated training, component-wise ablation, selective-risk analysis, external-domain testing, inference-factor isolation, statistical testing, and systematic failure characterisation. The ablation results support this positioning by showing complementary rather than uniformly additive effects.

The broader evaluation comes with substantial computational cost. Swin-MSEG-Pro contains 89.91M parameters and requires 17.19 GMACs per 256×256 forward pass, with measured throughput of 17.68 ± 1.18 FPS on a Tesla T4. Several recent comparison methods are considerably smaller and report higher throughput, although their FPS measurements were obtained using different GPUs, resolutions, and software environments and cannot be treated as hardware-normalised rankings. Computational efficiency is not a principal advantage of the proposed method. The framework represents an accuracy–reliability–complexity trade-off in which additional capacity supports segmentation together with evidential and contour-focused outputs.

6.5. Limitations and future directions

Several limitations constrain the interpretation of these results. First, model development used a single source dataset, and external validation was limited to two public still-image benchmarks rather than prospective multi-centre colonoscopy video. The measured transfer therefore provides benchmark-level evidence under domain shift but does not establish clinical deployment generalisability. Second, ECE remained approximately 0.30, indicating that the evidential formulation provides useful selective-risk information without demonstrating strong absolute probability calibration. Third, the boundary branch is supervised using a morphological band derived from a single reference annotation; it should therefore be interpreted as a boundary-ambiguity proxy rather than a direct measurement of annotation or aleatoric uncertainty.

The failure analysis further identified persistent sensitivity to small lesions and, particularly on ETIS-LaribPolypDB, low-contrast appearance. Interpretability evidence is also incomplete across domains because quantitative GT-guided Grad-CAM analysis was restricted to the held-out Kvasir-SEG test set. Finally, the 89.91M-parameter architecture is comparatively heavy, limiting claims regarding resource-constrained or real-time deployment. Future work should consequently prioritise lightweight reliability-aware designs, explicit uncertainty calibration, improved small-lesion modelling, broader multi-centre and video-based external validation, and quantitative assessment of explanation stability under domain shift.

7. Conclusion

This study presented Swin-MSEG-Pro, a reliability-oriented gastrointestinal polyp segmentation framework combining a SwinV2-based multi-scale architecture with deep supervision, evidential uncertainty learning, attention-based refinement, and contour-focused boundary supervision. Under a fixed Kvasir-SEG development protocol and three independent random seeds, the model achieved mean per-image Dice scores of 0.9153 ± 0.0024 on the held-out Kvasir-SEG test set, 0.7174 ± 0.0016 on ETIS-LaribPolypDB, and 0.8527 ± 0.0020 on CVC-ClinicDB. The external results demonstrate reproducible transfer but also substantial dataset-dependent degradation, particularly on ETIS-LaribPolypDB, which exhibited the largest measured feature-space shift.

The ablation analysis showed that the framework components do not provide uniformly additive improvements in segmentation accuracy. Boundary-focused supervision produced the clearest positive external contribution, whereas removing evidential learning improved raw Dice on ETIS-LaribPolypDB and CVC-ClinicDB. Evidential learning is therefore supported primarily as a mechanism for uncertainty-based selective-risk analysis rather than as a universal accuracy or calibration improvement. TTA modestly improved external selective-risk behaviour, while ECE remained approximately 0.30 and the evidential uncertainty signal ranked boundary risk more effectively than the explicitly supervised boundary head. These findings support a distinction between predictive uncertainty estimation and contour-focused boundary supervision within the proposed framework.

Swin-MSEG-Pro should be interpreted as an accuracy–reliability–boundary trade-off, not as a universally superior segmentation architecture. Its single-source training protocol provides useful evidence of cross-dataset transfer, but systematic failure analysis identified small lesions and low-contrast external cases as important residual limitations. The model is also computationally demanding, with 89.91M parameters and 17.19 GMACs, and quantitative Grad-CAM evaluation was restricted to the held-out Kvasir-SEG cohort. Future work should focus on lighter reliability-aware architectures, stronger probability calibration, improved small-lesion modelling, and broader multi-centre and video-based external validation.

CRediT authorship contribution statement

Muhammad Muhtasim Shahriar: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **M.M. Golam Hafiz:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **M.F. Mridha:** Writing – review & editing, Visualization, Supervision, Resources, Methodology. **Md. Rakibul Islam:** Writing – review & editing, Visualization, Validation, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests that could have appeared to influence the work reported in this paper. No conflicts of interest exist regarding the publication of this work.

Acknowledgements

The authors acknowledge American International University-Bangladesh (AIUB) for its support and contribution to this work. The authors also acknowledge the Advanced Machine Intelligence Research Lab (AMIR) for its support and contribution.

Data availability

The datasets used in this study are publicly available. The Kvasir-SEG dataset is available online at: https://datasets.simula.no/kvasir-seg/. The ETIS-LaribPolypDB dataset is publicly available and was used for external evaluation: https://deeply.readthedocs.io/en/latest/datasets/etis-larib.html. The CVC-ClinicDB dataset is also publicly available and was used for external evaluation: https://datasetninja.com/cvc-612.

References

[1] Bui N-T, Hoang D-H, Nguyen Q-T, Tran M-T, Le N. MEGANet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. WACV, 2024, p. 7985–94. http://dx.doi.org/10.1109/WACV57701.2024.00780.

[2] Liu X, Isa NAM, Chen C, Liu H, Wang C, Lv F. Wavelet-guided mamba-attention network for boundary-aware colorectal polyp segmentation. Mach Learn Knowl Extr 2026;8(6):142. http://dx.doi.org/10.3390/make8060142.

[3] Liu X, Yuan Y. A source-free domain adaptive polyp detection framework with style diversification flow. IEEE Trans Med Imaging 2022;41(7):1897–908. http://dx.doi.org/10.1109/TMI.2022.3150435.

[4] Liu B, Shi C, Zhao M. DSF-BRNet: Dual-gated semantic fusion and boundary refinement for efficient endoscopic polyp segmentation. Sensors 2026;26(9):2717. http://dx.doi.org/10.3390/s26092717.

[5] Wang Z, Liu Z, Yu J, Gao Y, Liu M. Multi-scale nested UNet with transformer for colorectal polyp segmentation. J Appl Clin Med Phys 2024;25(6):e14351. http://dx.doi.org/10.1002/acm2.14351.

[6] Khan R, Alzaben N, Daradkeh YI, Lee MY, Ullah I. Bilateral collaborative streams with multi-modal attention network for accurate polyp segmentation. Sci Rep 2025;15:34182. http://dx.doi.org/10.1038/s41598-025-15401-1.

[7] Xu Z, Tang F, Chen Z, Zhou Z, Wu W, Yang Y, Liang Y, Jiang J, Cai X, Su J. Polyp-Mamba: Polyp segmentation with visual mamba. In: Medical image computing and computer assisted intervention – MICCAI 2024. 2024, p. 510–21. http://dx.doi.org/10.1007/978-3-031-72111-3_48.

[8] Li X, Jiang A, Yu H, Xiao Y. FARF-Net: Frequency-guided adaptive receptive field network for edge-enhanced polyp segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. WACV, 2026, p. 2338–47.

[9] Guo Q, Fang X, Wang L, Zhang E. Polyp segmentation of colonoscopy images by exploring the uncertain areas. IEEE Access 2022;10:52971–81. http://dx.doi.org/10.1109/ACCESS.2022.3175858.

[10] Du X, Xu X, Chen J, Zhang X, Li L, Liu H, Li S. UM-Net: Rethinking ICGNet for polyp segmentation with uncertainty modeling. Med Image Anal 2025;99:103347. http://dx.doi.org/10.1016/j.media.2024.103347.

[11] Shao H, Zhang Y, Hou Q. Polyper: Boundary sensitive polyp segmentation. Proc the AAAI Conf Artif Intell 2024;38(5):4731–9. http://dx.doi.org/10.1609/aaai.v38i5.28274.

[12] Wang D, Zhou Q, Yang Y, Sun X, He Y. LMSGNet: Semantic prior and edge guided network for robust polyp segmentation. J Comput Des Eng 2026;13:88–104. http://dx.doi.org/10.1093/jcde/qwag001.

[13] Zhou T, Zhou Y, He K, Gong C, Yang J, Fu H, Shen D. Cross-level feature aggregation network for polyp segmentation. Pattern Recognit 2023;140:109555. http://dx.doi.org/10.1016/j.patcog.2023.109555.

[14] Duc NT, Oanh NT, Thuy NT, Triet TM, Sang DV. ColonFormer: An efficient transformer based method for colon polyp segmentation. IEEE Access 2022;10:80575–86. http://dx.doi.org/10.1109/ACCESS.2022.3195241.

[15] Xiao B, Hu J, Li W, Pun C-M, Bi X. CTNet: Contrastive transformer network for polyp segmentation. IEEE Trans Cybern 2024;54(9):5040–53. http://dx.doi.org/10.1109/TCYB.2024.3368154.

[16] Pang Y, Long Y, Chen Z, Hu Y, Chen H, Wang Q. Endoscopic adaptive transformer for enhanced polyp segmentation in endoscopic imaging. IEEE Trans Med Imaging 2026;45(3):987–99. http://dx.doi.org/10.1109/TMI.2025.3615677.

[17] Lijin P, Ullah M, Vats A, Cheikh FA, Santhosh Kumar G, Nair MS. PolySegNet: improving polyp segmentation through swin transformer and vision transformer fusion. Biomed Eng Lett 2024;14:1421–31. http://dx.doi.org/10.1007/s13534-024-00415-x.

[18] Tong Y, Yong X, Liang J, Chen Z, Hu Y, Li X, Gao H, Zhou Z, Hu K. DDF-Former: a dual-domain fused transformer for polyp segmentation. J Supercomput 2025;81:1451. http://dx.doi.org/10.1007/s11227-025-07846-y.

[19] Li H, Zhang D, Yao J, Han L, Li Z, Han J. ASPS: Augmented segment anything model for polyp segmentation. In: Medical image computing and computer assisted intervention – MICCAI 2024. 2024, p. 118–28. http://dx.doi.org/10.1007/978-3-031-72114-4_12.

[20] Liu Z, Ling J. SpectraNet: a novel model for polyp segmentation leveraging a spectral-guided mixture of functional experts. Front Oncol 2026;16:1734345. http://dx.doi.org/10.3389/fonc.2026.1734345.

[21] Xiao J, Han Y, Wang L, Li Y, Wang X, Li S, Yi J, Wu Y, Liu X. PrysmNet: A polyp refining system using salience and multimodal guidance for reproducible cross domain segmentation. Npj Digit Med 2026;9:158. http://dx.doi.org/10.1038/s41746-026-02345-7.

[22] Arnob AKB, Monowar MM, Hamid MA, Mridha MF. XPolypNet: A U-Net-based model for semantic segmentation of gastrointestinal polyps with explainable AI. IEEE Open J Comput Soc 2025;6:1283–93. http://dx.doi.org/10.1109/OJCS.2025.3592204.

[23] Manan MA, Feng J, Ahmed S, Raheem A. Enhancing colorectal polyp segmentation with TCFMA-Net: A transformer-based cross feature and multi-attention network. Artif Intell Med 2025;167:103167. http://dx.doi.org/10.1016/j.artmed.2025.103167.

[24] Jha D, Smedsrud PH, Riegler MA, Halvorsen P, de Lange T, Johansen D, Johansen HD. Kvasir-SEG: A segmented polyp dataset. In: MultiMedia modeling. Lecture notes in computer science, vol. 11962, Springer; 2020, p. 451–62. http://dx.doi.org/10.1007/978-3-030-37734-2_37.

[25] Silva JS, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg 2014;9(2):283–93. http://dx.doi.org/10.1007/s11548-013-0926-3.

[26] Bernal J, Sánchez FJ, 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. http://dx.doi.org/10.1016/j.compmedimag.2015.02.007.