Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PolyDetect: A Quality-Aware Hybrid Deep Learning Framework for Robust Polyp Segmentation in Degraded Colonoscopy Images

Azizur Rahman

azizurusa22@gmail.com

Indiana Wesleyan University https://orcid.org/0009-0007-6945-6953

Nakib Uddin Ahmed

nakibuddin33@gmail.com

Shahjalal University of Science and Technology https://orcid.org/0009-0006-6515-5750

Kadirur Rahman Chowdhury

Birmingham City University

Mehjabin Ferdous

St. Francis College

Research Article

Keywords: Polyp segmentation, Colonoscopy, Image quality assessment, Hybrid deep learning, PVT-v2, Degradation robustness, Boundary refinement, Colorectal cancer

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10731156/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

PolyDetect: A Quality-Aware Hybrid Deep Learning Framework for Robust Polyp Segmentation in Degraded Colonoscopy Images

Abstract

Colorectal cancer (CRC) is one of the most common cancer-related deaths, and colonoscopy is the preferred method of screening for CRC. In practice, the images captured by endoscopes are often corrupted by motion blur, noise, compression artifacts, low contrast and poor illumination, adversely affecting the existing segmentation models and only a few algorithms explicitly use the image quality signal within the segmentation pipeline. We suggest a hybrid scheme, PolyDetect, which combines five components: a five-type degradation simulator for robustness training, a no-reference image-quality assessment (IQA) module combining BRISQUE statistics with a lightweight CNN (~16K parameters) to form a per-image quality score, a quality-score-gated three-tier adaptive enhancement pipeline (Retinex, CLAHE, DnCNN), a PVT-v2 encoder with a CNN decoder and a dedicated boundary-refinement head, and a quality-weighted composite loss to down-weight degraded training samples. PolyDetect is trained on 900 Kvasir-SEG images and evaluated zero-shot on CVC-ColonDB (n=380) and ETIS (n=196) datasets, and achieves the highest Dice (0.8066) and IoU (0.7262) on all three tests, improvements statistically significant after multiple-comparison correction over CNN-based baselines. However, the ablation of the boundary-refinement head was found as the most critical component to remove (−6.47 pp Dice), while the ablation of the quality-gated enhancement pipeline results in an improvement of mean Dice by 0.94 pp in 91.8% of the frames with quality score $q \geq 0.5$. The segmentation network has an average FPS (frames per second) of 53.62 on a Tesla T4 GPU (27.35M parameters, 12.47 GFLOPs), which is in the real time range for clinical application. PolyDetect is a feasible way to achieve a comprehensive quality assessment, quality-gated enhancement, and quality-weighted training in the same pipeline for polyp segmentation in degraded, real-world colonoscopy images.

Keywords: Polyp segmentation · Colonoscopy · Image quality assessment · Hybrid deep learning · PVT-v2 · Degradation robustness · Boundary refinement · Colorectal cancer

1 Introduction

In the world, colorectal cancer (CRC) represents the third most common cancer and the second causes of cancer related death [1]. The gold-standard screening test for early detection of polyps is colonoscopy, and early removal of polyps is likely to save about 53% of CRC deaths [2] while also being recommended for colon cancer screening by principals [3]. In recent years, automatic polyp segmentation has made significant progress, particularly with the rise of deep convolutional neural networks (CNNs) and, more recently, transformer-based architectures, with the best models reporting Dice similarity index (DSI) scores over 0.90, when tested on the standard benchmark datasets [4,5,6]. There remains a long-standing gap between performance of the controlled benchmark and its deployment in the real world, however, which has not been fully uncovered.

The five types of degradation that occur commonly in colonoscopy images in practice are: movement blur due to movement of the scope, additive-gaussian noise due to using low gain settings, lack of contrast in flat or sessile polyps, JPEG compression artifacts that occur in video encoding pipelines, and the poor illumination of images in the distal and angled portions of the colon [7,8]. Together, these degradation conditions lead to a 10–30 point drop in segmentation Dice for CNN-based methods for the same segmentation results, and are poorly addressed in the literature. No current state-of-the-art method (such as PraNet [4], SANet [5], Polyp-PVT [6], SSFormer [9], ColonFormer [10] and the recent VMDU-Net [11]) includes any kind of IQR, quality-conditioned IQR, QCPP, or QALC.

This study outlines three gaps: (i) the lack of no-reference image quality assessment (IQA) in the polyp segmentation training and inference pipeline, (ii) the lack of adaptively scoring with an IQA score that allows for the avoidance of artifact-introducing over-processing of already clean frames, and (iii) the uniform weighting of samples in loss functions where degraded images, with artifacts and noise, are considered the same as clean high-quality images. In this study, a unified framework, PolyDetect, is proposed which directly addresses all three gaps with five integrated components to enhance the quality of the multimedia system.

The principal contributions of this work are:

(1) IQA Module with BRISQUE–CNN Fusion: The first polyp segmentation framework, to our knowledge, to embed a live, differentiable image-quality signal as a training-time loss modulator, fusing classical BRISQUE statistics with a CNN quality-regressor branch (∼16 K parameters) to produce a per-image quality score $q \in [0, 1]$.

(2) Quality-Score-Gated Adaptive Enhancement: A three-tier enhancement pipeline (Retinex illumination correction → CLAHE contrast enhancement → DnCNN denoising) that activates selectively based on q, avoiding artifact-introducing over-processing of high-quality images — novel in medical image segmentation.

(3) Quality-Weighted Composite Loss: A per-sample loss weighting scheme $L = \text{mean}[(L_{Dice} + L_{BCE} + \lambda_b \times L_{Boundary}) \times q]$ that proportionally reduces the gradient contribution of degraded training samples, preventing noise-pattern memorization.

(4) Hybrid PVT-v2 + CNN Decoder with Morphological Boundary Refinement: A segmentation backbone combining hierarchical transformer (PVT-v2-B2) global context with CNN U-Net decoder local spatial detail, augmented by a boundary refinement head trained on morphologically derived 1-pixel edge maps.

(5) Systematic Degradation Robustness Benchmark: The first per-degradation robustness evaluation protocol in polyp segmentation, benchmarking Dice drop across five controlled degradation types (blur, noise, low contrast, JPEG compression, poor illumination) for all four competing models.

PolyDetect outperforms CNN baselines on the Kvasir-SEG, CVC-ColonDB (zero-shot), ETIS-LaribPolypDB (zero-shot) datasets with the highest average Dice (0.8066) and IoU (0.7262) metrics, which

is statistically significant with respect to these CNN models ($p < 10^{-10}$) across all three datasets and is statistically significant with respect to SANet on ETIS zero-shot ($p = 0.037$). The model has 27.35M parameters, 12.47 GFLOPs and runs at 53.62 FPS to meet clinical real-time deployment requirements. The rest of this paper is organized as follows: Section 2 reviews related work, Section 3 discusses the PolyDetect methodology, Section 4 writes about experimental setup and Section 5 outlines the results and Section 7 concludes the study.

2 Related Work

2.1 CNN-Based Medical Image Segmentation

The study [12] proposed a U-Net architecture that laid the groundwork for the encoder–decoder paradigm with skip connections as the basis for medical image segmentation. U-Net++ [13] was a netting of U-Net which added dense inter-scale skip connections, which enhanced multi-scale feature integration with an increased number of parameters. Although used extensively, both architectures are lacking of support for real-world images degradation and suffer a significant drop in performance, when trained on Kvasir-SEG only, on challenging colonoscopy datasets like CVC-ColonDB and ETIS-LaribPolypDB.

2.2 Transformer-Based and Hybrid Polyp Segmentation

To tackle this challenge, PraNet [4] proposed a parallel reverse attention mechanism with feature aggregation based on ResNet-50 to obtain competitive Dice scores in the PraNet benchmark protocol. SANet [5] presented shallower variant of its attention model, which can capture small and flat polyps. In Polyp-PVT [6] a Pyramid Vision Transformer (PVT-v2) is used as the encoder, which obtains the multi-scale attention features for superior global context modeling. In the field of progressively structured feature fusion, SSFormer [9] and ColonFormer [10] boosted transformer-encoding to state-of-the-art (Dice > 0.917 on Kvasir-SEG following training protocol of PraNet). Recently, VMDU-Net [11] adopted Visual Mamba using cross-shape transformer features which achieved a Dice of 0.934 on the task of Kvasir-SEG.

2.3 Image Quality Assessment and Enhancement

No-reference image quality assessment (NR-IQA) refers to a set of image quality assessment metrics that do not need a clean reference image. BRISQUE [14] estimates the naturalness by analyzing the spatial relations within the scene; NIQE [15] and PIQUE calculate the naturalness index based on the statistics of natural scenes. The study of image quality has been carried out in the field of medical image in histopathology screening [16] and in fundus photography [17]. This line of work has a long history within endoscopy, and informative-frame selection and characterization of artifacts such as the Endoscopy Artifact Detection (EAD) challenge series [18] have existed, but this paper attempts to fill the void of using a quality signal to inform segmentation loss or at training time, using gated preprocessing, of artifacts. Some of the image enhancement techniques that are commonly used as a fixed preprocessing step in medical imaging applications are adaptive image enhancement methods like Retinex-based illumination correction [19],

Contrast Limited Adaptive Histogram Equalization (CLAHE) [20,21] and other denoising methods are residual denoising networks like DnCNN [22].

2.4 Quality-Aware and Curriculum Learning in Deep Networks

Curriculum learning [23] and hard-example mining methods like Focal Loss [24] and its online extension Online Hard Example Mining (OHEM) [25] have been proposed to enhance convergence when doing general detection tasks. These methods however are geared towards difficulty in prediction (loss magnitude), and not towards quality of the input image. Another conceptually very different and complementary technique is called "quality-aware sample weighting" in which each sample's gradient contribution is modulated directly by the continuous IQA score. CFFormer [26] is the most related work that analyzes dataset quality statistics using the PIQUE scores, but doesn't involve IQA in the model architecture and loss function. To our knowledge, PolyDetect is the first medical image segmentation method that employs a live IQA signal as training modulator.

3 Methodology

The PolyDetect framework is made of five tightly coupled modules that together create a full quality-aware segmentation pipeline. The overall system architecture is shown in Figure 1. During inference, every input colonoscopy image is sent into the IQA module, which outputs a scalar quality score q; then on the basis of q, the adaptive enhancement pipeline applies correction stages for enhanced image; and then the enhanced image is inputted into the hybrid segmentation backbone to get the final binary polyp mask. The quality score also adjusts the loss weights per sample during training.

3.1 Degradation Simulation Module

To address the challenge of robustness to real-world artifacts in colonoscopy, we simulate 5 kinds of image degradations during training. To obtain images of different quality, each image is independently exposed to one or two randomly selected degradation types, with application probability $p = 0.7$, such that the training distribution covers the range of image quality but there are clean images in the training distribution.

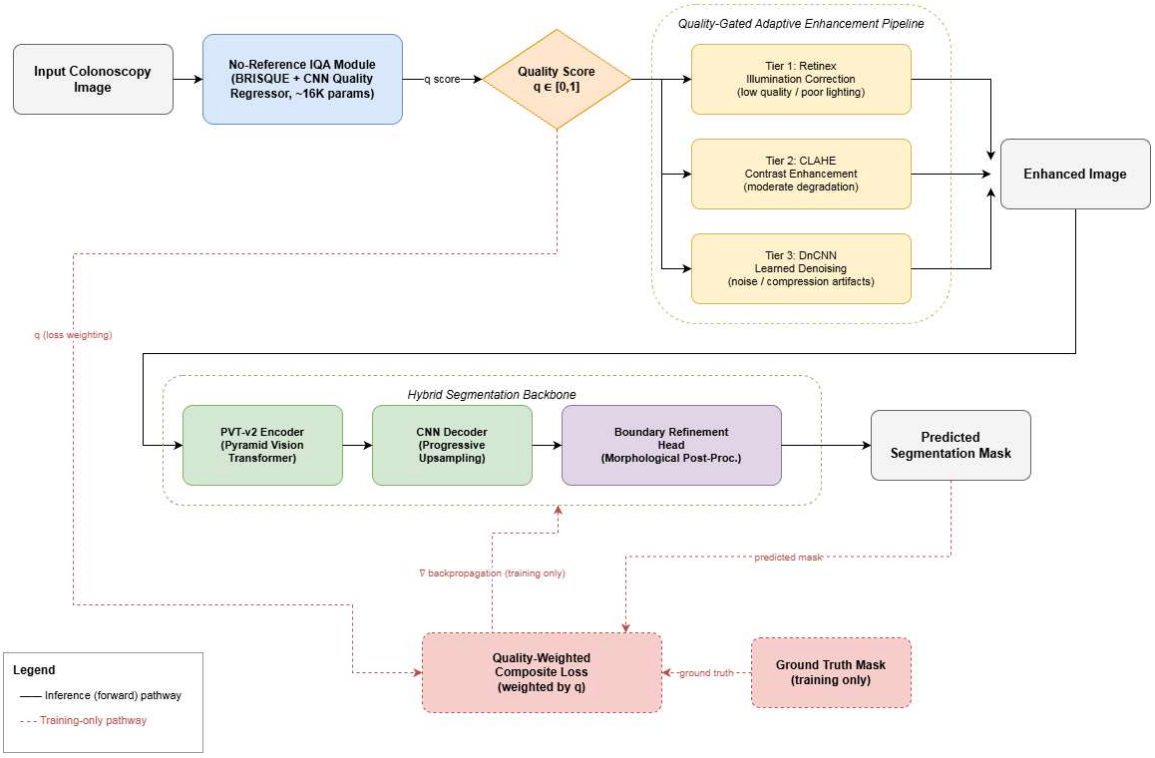


Figure 1 Overview of the PolyDetect Framework.

Table 1 Five degradation types simulated during training.

Degradation Type	Implementation	Parameter Range
Gaussian Blur	cv2.GaussianBlur(img, (k,k), 0)	Kernel k ∈ {3, 5, 7, 9}
Gaussian Noise	np.clip(img + N(0,σ), 0, 255)	σ ∈ Uniform(5, 30)
Low Contrast	α × img + β	α ∈ (0.4, 0.7), β ∈ (−10, 10)
JPEG Compression	cv2.imencode / cv2.imdecode at quality q	q ∈ Randint(10, 40)
Poor Illumination	Gamma correction: ((i/255)^(1/γ))×255	γ ∈ Uniform(1.8, 3.5)

The five types of degradation that were simulated at training and testing were displayed on the representative Kvasir-SEG colonoscopy frame in figure 2. In the original image (a), the image quality under the endoscope is normal and the texture of the tissues is clear and sharp with the polyp area easy to see. Gaussian blur (b) reproduces blurring due to movement that occurs during scope movements, while maintaining the overall outline of the tissue. Gaussian noise (c) adds sensor level noise, providing a mimic for low-light or high ISO exposure. Low contrast (d) affects the range by reducing the difference between the contours of mucosal tissue and those of the polyps, one of the more challenging degradations to address, as it directly reduces the distinction that segmentation models use. JPEG compression (e) includes blocking artifacts commonly observed in pipelines that record endoscopy processes. Poor illumination (f) dims the frame and makes it harder to see in the shadowed areas, as is the case in the distal colon segments or at an angle. These six panels illustrate that the training and evaluation data of PolyDetect not only spans the idealized benchmark images but also covers a representative range of visual degradation that occurs during

standard colonoscopy procedures, directly motivating the adaptive enhancement pipeline, described in Section 3.3, that leverages the concept of IQA.

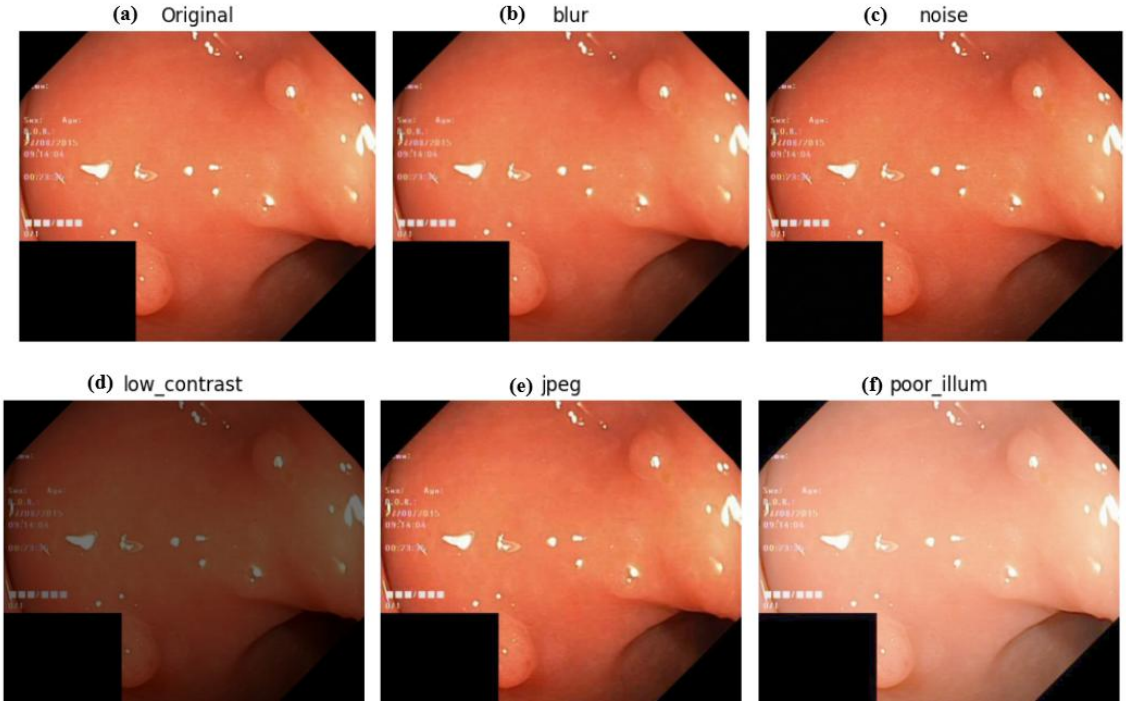


Figure 2 Visual examples of the five degradation types applied to a representative Kvasir-SEG colonoscopy image. From left to right: (a) clean, (b) Gaussian blur, (c) Gaussian noise, (d) low contrast, (e) JPEG compression, (f) poor illumination.

3.2 Image Quality Assessment (IQA) Module

The IQA module computes a scalar quality score $q \in [0, 1]$ (1 = highest quality) for each image by fusing two complementary signals.

BRISQUE Branch: BRISQUE (Blind/Reference less Image Spatial Quality Evaluator) [14] quantifies perceptual quality via locally normalized luminance coefficients. The raw score is normalized to $b_{norm} = 1 - \text{clamp}(\text{BRISQUE} / 100, 0, 1)$.

Lightweight CNN Branch: The four layer convolutional network Conv → ReLU × 3 → AdaptiveAvgPool → FC → Sigmoid is a small network (has about 16 K parameters) that takes an image as an input and outputs a quality feature $c \in [0, 1]$. Here this branch is used with random initialization (as output computed under no-gradient inference and parameters not used in the segmentation optimizer), for this reason no adaptation is performed on c with respect to the statistics of the colonoscopy domain, in this case. Joint/or auxiliary supervision of this branch is pointed as a condition of its follow-up, prior to which the fused signal cannot be considered learned - see Section 6.3.

Fusion: $q = 0.5 \times b_{norm} + 0.5 \times c$. IQA fusion ablation (Section 5.7) shows that cnn_only, brisque_only, and fusion_50_50 differ by less than 0.9 pp Dice, with fusion_50_50 achieving the highest recall (0.8492) and F2 (0.7992). As the CNN branch is not trained in the current results (see above), this cnn_only condition

in this comparison does not represent a learned quality signal, but a random projection of the input, hence not very robust to the choice of the per-image weighting, rather in favor of the CNN branch in its own right, if the joint-training experiment described in Section 6.3 is confirmed. The fused formulation is kept because of its recall/F2 benefit.

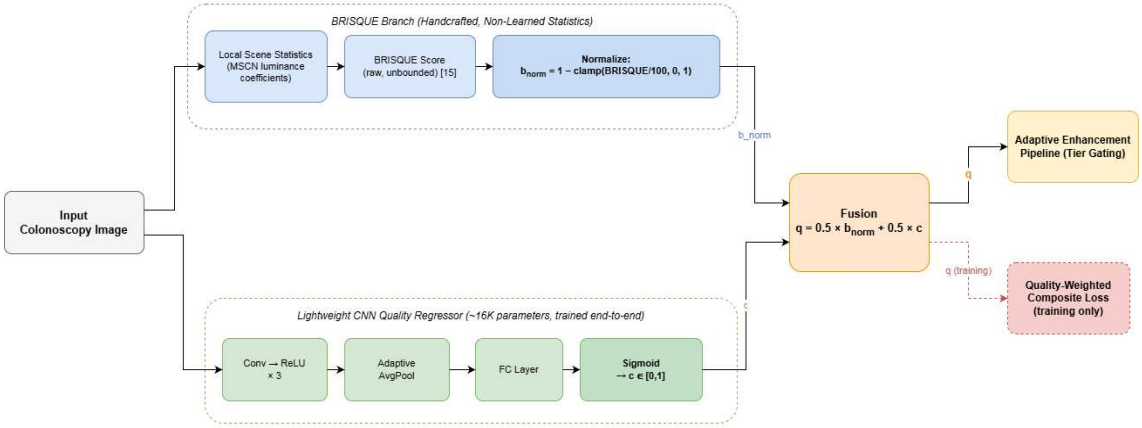


Figure 3 IQA Module Architecture: Dual-Branch Quality Estimation with BRISQUE-CNN Fusion.

3.3 Quality-Score-Gated Adaptive Enhancement Pipeline

The enhancement pipeline has three tiers of processing and uses the q parameter to select between them so as to avoid the possibility of over-processing already clean frames that will contribute to artifacts.

Table 2 Quality-gated three-tier adaptive enhancement pipeline. Processing intensity is determined by the IQA score q. Note: these tier boundaries ($q \geq 0.7$, $0.4 \leq q < 0.7$, $q < 0.4$) differ from the $q \geq 0.5$ / $q < 0.5$ split used for the quality-stratified analysis in Table 14 and Section 5.6; occupancy of the three tiers, measured on the CVC-ColonDB test set (n = 380), is 35.8% good ($q \geq 0.7$, no enhancement), 64.2% moderate ($0.4 \leq q < 0.7$, CLAHE + DnCNN), and 0.0% poor ($q < 0.4$, full Retinex→CLAHE→DnCNN pipeline); see Section 5.6 for how this reconciles with the Table 14 stratification.

Quality Score	Enhancement Applied	Rationale
$q \geq 0.7$ (Good)	No enhancement	Image is clean; any processing risks introducing artifacts
$0.4 \leq q < 0.7$ (Moderate)	CLAHE + DnCNN denoising	Moderate contrast/noise correction; skip Retinex to avoid halo artifacts
$q < 0.4$ (Poor)	Retinex → CLAHE → DnCNN	Full pipeline: illumination normalization, contrast boost, denoising

Retinex: Single-Scale Retinex [19] corrects uneven illumination via $\log(I) - \log(I \otimes G\sigma)$ where $G\sigma$ is a Gaussian with $\sigma = 50$, normalized to [0, 255].

CLAHE: Contrast Limited Adaptive Histogram Equalization [20,21] is applied in LAB color space on the L channel (clipLimit = 2.0, tileGridSize = (8, 8)), preserving color fidelity while enhancing local contrast.

DnCNN: A residual denoising CNN [22] with depth = 10, 64 channels per layer. Residual learning: output = input − noise_estimate. Architecture: [Conv→ReLU, (Conv→BN→ReLU) ×8, Conv]. Note: In the experiments reported here the DnCNN is employed with random initialization without fine tuning on colonoscopy data; this is the same as for the CNN quality branch (Section 3.2) in that the parameters of the DnCNN are not passed to the training optimizer: dncnn.parameters() does not exist. The denoising stage is

now a fixed random nonlinear filter instead of learned denoising, and the reasons for this can be explained in the following ways: (1) pre-training or jointly training DnCNN on the synthetic colonoscopy noise pairs is a needed correction, to which learned denoising will be attributed in Section 6.3; (2) training the denoising network on the denoising pairs of the imaging scenario is not feasible, since the imaging process is not simply noise reduction but demands many other capabilities as well.

3.4 Hybrid Segmentation Backbone

This segmentation backbone includes the skill of global context modeling of a Pyramid Vision Transformer (PVT-v2-B2) [27] encoder and the local spatial detail of a CNN U-Net-style decoder, and a dedicated Boundary Refinement Head.

Table 3 Hybrid PVT-v2 + CNN decoder architecture. The encoder provides global context via self-attention; the CNN decoder restores local spatial detail through skip connections; the boundary head refines edge predictions.

Component	Details	Design Rationale
Encoder	PVT-v2-B2 (timm); features_only=True; out_indices=(0,1,2,3)	ImageNet-1K pre-training; hierarchical features at 4 scales
Decoder Blocks	3 × DecoderBlock: ConvBNReLU + bilinear upsample + skip concatenation	CNN U-Net-style skip connections preserve fine local detail
Segmentation Head	Conv2d(64, 1, kernel=1) → logit map	Single-channel binary mask prediction
Boundary Refine Head	2 × Conv(3×3) + Conv(1×1); edge_logits added to seg_logits	Sharpens polyp edges; critical for HD95 metric
Final Output	Bilinear interpolation to input H×W	Mask aligned with original image resolution

3.5 Quality-Weighted Composite Loss Function

The loss function combines three per-sample terms, weighted by the IQA quality score q:

$$\mathbf{L = mean[\ (L_{Dice} + L_{BCE} + \lambda_b \times L_{Boundary})\ \times q\ ]} \qquad \mathbf{(1)}$$

where L_{Dice} is per-sample Dice loss; L_{BCE} is per-sample binary cross-entropy; $L_{Boundary}$ is morphological edge BCE loss computed from the dilated minus eroded mask (3 × 3 structuring element); q is the IQA quality score ∈ [0, 1]; and $\lambda_b = 0.5$ is the boundary loss weight. This weighting has the effect of making it harder for an image to be "memorized" by performing correctly, so that degraded images that are more difficult to learn from contribute less to the updates of the gradient. Within the ablation protocol used throughout the paper (Table 13), further fine-tuning results in a reduction in Dice by 0.75 pp and increase in HD95 by 2.76 on CVC-ColonDB.

4 Experimental Setup

4.1 Datasets

The main in-distribution benchmark is Kvasir-SEG [28] which consists of 1000 polyp images with masks at the level of a single pixel provided by the Simula Research Laboratory. The zero-shot cross-dataset tests

are CVC-ColonDB (380 images, high visual variability) and ETIS-LaribPolypDB (196 high resolution images, most challenging). Experiments are reproducible by keeping the same random seed (42).

Table 4 Benchmark datasets. The model is trained exclusively on 900 Kvasir-SEG images; CVC-ColonDB and ETIS are used for zero-shot cross-domain generalization testing.

Dataset	Source	Images	Resolution	Split
Kvasir-SEG	Simula Research Lab	1,000	Variable (332–1920 px)	900 total: 800 train / 100 val
CVC-ColonDB	UAB Computer Vision	380	574 × 500 px	Zero-shot evaluation (all 380)
ETIS-LaribPolypDB	ETIS Challenge	196	1225 × 966 px	Zero-shot evaluation (all 196)

The studied protocol differs from widely-used PraNet protocol (1450 images combined for training) and is called training-on-Kvasir, zero-shot-on-CVC-ColonDB-and-ETIS. Therefore, the direct numerical comparisons with the methods of PraNet protocol is an approximation to the closest extent, Kvasir-SEG Dice is the most similar single-dataset metric.

4.2 Baseline Models

Four well-known models are used: (1) U-Net [12] standard encoder–decoder combined with four down/up-sampling blocks (64 channels vs. 1024 channels); (2) U-Net++ [13] with nested skip connections, base channels = 32; (3) PraNet [4] with ResNet-50, Parallel Reverse Attention and RFB blocks; (4) SANet [5] with ResNet-50, shallow attention and structure loss. All baselines use the same dataloaders, augmentation code, optimizer and evaluation pipeline used by PolyDetect. Two comparability caveats: first, epoch budget is different across the models (20 for U-Net/U-Net++, 25 for PraNet/SANet, 30 for the proposed model), but not epoch-matched; second, the architecture of the two models (PraNet/SANet) differs from their published loss and epoch training schemes, with the proposed model using the same loss and epoch training recipe as PolyDetect, which may lead to understating their published performance. The baseline comparison should be epoch-matched and be an original recipe re-run before it would be considered to be fully controlled (Section 6.3).

4.3 Training Configuration

The training settings are summarized in Table 5. The same configurations, AdamW optimization, cosine learning-rate scheduling, AMP, geometric augmentation and early stopping were used in all models. The total training time for all models was 4.05 hours and the models converged at epoch 13 in the process of training.

Table 5 Training hyperparameters.

Hyperparameter	Value	Justification
Input Resolution	352 × 352 px	Standard for polyp benchmarks; balances detail and GPU memory
Batch Size	8	Single-GPU fit; larger batch degrades Dice convergence in experiments

Epoch Schedule	UNet/UNet++ 20; PraNet/SANet 25; Proposed 30	Longer schedule for the more complex proposed model; early stopping monitors val Dice
Early Stopping Patience	7 epochs	Prevents overfitting; Proposed converged at epoch 13
Optimizer	AdamW (lr = 1×10^{-4}, wd = 1×10^{-4})	Weight decay regularization; stable for transformer encoders
LR Scheduler	CosineAnnealingLR (T_max = epochs)	Smooth decay without abrupt drops
Mixed Precision	torch.cuda.amp (GradScaler + autocast)	Faster training; reduces GPU memory
Augmentation	H-Flip (p=0.5), V-Flip (p=0.5), RandomRotate90 (p=0.5)	Standard geometric; avoids color jitter that conflicts with IQA
Normalization	ImageNet μ=[0.485,0.456,0.406], σ=[0.229,0.224,0.225]	Required for PVT-v2 ImageNet pretrained weights

4.4 Evaluation Metrics

There are two main metrics for segmentation performance: the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU / Jaccard Index), with the latter being the secondary metric. Both are computed per image. Assessment of statistical significance is made using one-sided paired Wilcoxon signed-rank test (H_1: Proposed > Baseline; at the 5% level). The speed of the system is evaluated by the number of parameters (M), number of floating-point operations (GFLOPs for 352 x 352 input), and inference speed (FPS, batch = 1, Tesla T4 GPU) with the thop library.

5 Results

5.1 Main Segmentation Performance

5.1.1 Kvasir-SEG (In-Distribution)

Table 6 shows the results on the test set Kvasir. The highest Dice (0.9106), IoU (0.8511), precision, recall, and F2 scores were obtained with PolyDetect, which demonstrated best and most stable performance. It also gave the lowest HD95 (24.8) which meant that the accuracy of the boundary delineation was better than for the models compared.

Table 6 Main segmentation results on Kvasir-SEG test set (n = 100, in-distribution). ↑ higher is better; ↓ lower is better. Highlighted row = proposed model.

Model	Dice ↑	±Std	IoU ↑	Prec ↑	Recall ↑	F2 ↑	HD95 ↓
U-Net	0.8055	±0.196	0.7107	0.8237	0.8555	0.8489	58.0
U-Net++	0.7986	±0.181	0.6964	0.7946	0.8743	0.8571	64.1
PraNet	0.9027	±0.143	0.8448	0.9182	0.9111	0.9125	29.0
SANet	0.9022	±0.140	0.8435	0.9117	0.9247	0.9221	27.3
Proposed	**0.9106**	**±0.112**	**0.8511**	**0.9227**	**0.9238**	**0.9236**	**24.8**

5.1.2 CVC-ColonDB (Zero-Shot)

The zero-shot results for CVC-ColonDB are reported in Table 7, where all models were only trained on Kvasir-SEG. PolyDetect scored the best Dice (0.7543) and IoU (0.6605), and its predictions remain fairly consistent. CNN based U-Net models are poorly performing, which reflects the lack of cross-dataset

generalization. While the overall results showed improved robustness to unseen colonoscopy data, the PolyDetect had a higher HD95 than PraNet.

Table 7 Zero-shot results on CVC-ColonDB (n = 380). Proposed achieves the highest Dice (+0.87 pp over SANet) and IoU. CNN-only baselines collapse (Dice < 0.52).

Model	Dice ↑	±Std	IoU ↑	Prec ↑	Recall ↑	F2 ↑	HD95 ↓
U-Net	0.5189	±0.358	0.4281	0.6055	0.5531	0.5628	58.7
U-Net++	0.4894	±0.346	0.3938	0.5628	0.5544	0.5561	68.6
PraNet	0.7328	±0.292	0.6453	0.7809	0.7455	0.7523	38.1
SANet	0.7456	±0.281	0.6568	0.7634	0.7929	0.7868	35.2
Proposed	**0.7543**	**±0.260**	**0.6605**	**0.7777**	**0.7880**	**0.7859**	**45.0**

5.1.3 ETIS-LaribPolypDB (Zero-Shot)

The ETIS-LaribPolypDB (n = 196) is the most challenging dataset in this study due to the high-resolution images, and the significant domain shift between the dataset containing images used for training Kvasir-SEG and the dataset containing images used for the testing phase. Under these conditions, CNN-only baselines are essentially not competitive (the Dice is lower than 0.41), while U-Net and U-Net++ are underperforming when applied to unseen data with higher resolution in colonoscopy.

Table 8 Zero-shot results on ETIS-LaribPolypDB (n = 196). Proposed achieves highest Dice (0.7548) and Recall (0.8373). CNN-only baselines essentially fail (Dice < 0.41).

Model	Dice ↑	±Std	IoU ↑	Prec ↑	Recall ↑	F2 ↑	HD95 ↓
U-Net	0.4105	±0.382	0.3390	0.4373	0.4800	0.4708	70.6
U-Net++	0.3728	±0.357	0.2977	0.3846	0.4900	0.4645	98.0
PraNet	0.7035	±0.326	0.6234	0.6940	0.7862	0.7659	55.6
SANet	0.6965	±0.337	0.6177	0.6584	0.8162	0.7789	41.2
Proposed	**0.7548**	**±0.279**	**0.6669**	**0.7169**	**0.8373**	**0.8101**	**43.5**

The scores for PolyDetect are a clinically meaningful increase over those of SANet with Dice (+5.83 pp) and the highest score for Recall (+5.13 pp) and the highest score for F2 (+5.81 pp) (with Recall being more important to weigh heavily to minimize missed polyps in screening). Also included is the HD95 (43.5) of PolyDetect, which is the lowest of the top three models, and which means that PolyDetect has shown its boundary localization in the difficult dataset. The results demonstrate the generalization capabilities of the IQA based enhancement and final hybrid PVT-v2 + CNN decoder is especially strong in the hardest zero-shot setting, which provides the robustness motivation, central to PolyDetect.

5.1.4 Cross-Dataset Summary

Table 9 provides an overview of all three benchmarks. PolyDetect also demonstrates strong overall robustness and generalization, with the highest average Dice (0.8066) and IoU (0.7262) across all of the test sets, Kvasir-SEG, CVC-ColonDB and ETIS, and despite being trained on 900 images across the three sets.

Table 9 Cross-dataset summary.

Model	Kvasir Dice	ColonDB Dice	ETIS Dice	Avg Dice ↑	Avg IoU ↑
U-Net	0.8055	0.5189	0.4105	0.5783	0.4926
U-Net++	0.7986	0.4894	0.3728	0.5536	0.4626
PraNet	0.9027	0.7328	0.7035	0.7797	0.7045

SANet	0.9022	0.7456	0.6965	0.7814	0.7060
Proposed	**0.9106**	**0.7543**	**0.7548**	**0.8066**	**0.7262**

Qualitative segmentation results of PolyDetect on the representative test images from three datasets are presented in Figure 4 (a) CVC-ColonDB, (b) ETIS-LaribPolypDB and (c) Kvasir-SEG. For each row, the image on the left is the ground-truth mask, the middle is the model prediction, and the right is an overlay (red) between the prediction and the ground-truth mask (green) showing spatial agreement. In all three datasets, the predicted masks are in close agreement with the ground truth in both shape and boundary contour; the two masks are visually strongly aligned in red and green sides and have a very small amount of residual error, as can be seen in the overlay graphics and measurements (Table 6–9).

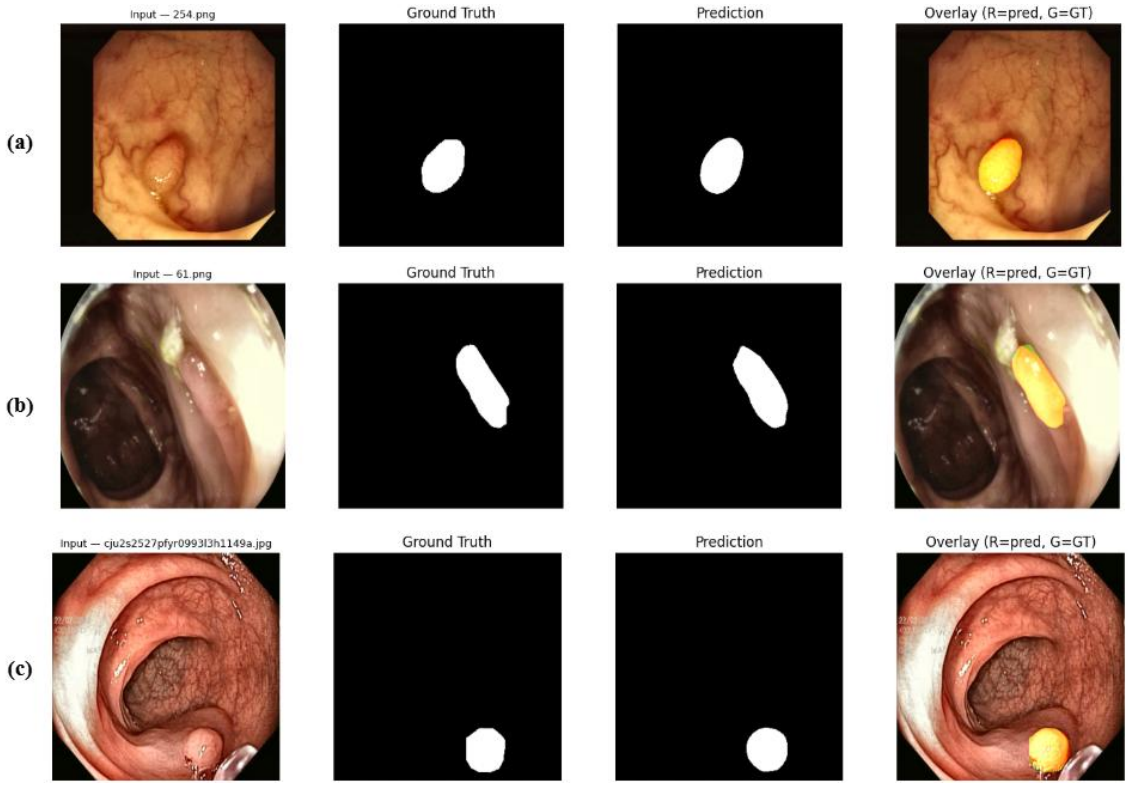


Figure 4 Prediction on Test Set of Proposed Model for each Dataset; (a) ColonDB; (b) ETIS; (c) Kvasir.

5.2 Statistical Significance

Table 10 shows 12 paired one-sided Wilcoxon tests and all p-values are nominal – this means that the p-values for each individual test cannot be considered significant unless corrected for multiplicity. To assess the robustness and consistency of this advantage over non-pretrained CNN architectures, the six comparisons against the CNN-only baselines U-Net, U-Net++ are made across all three datasets, with the results all remaining significant after correction (corrected $p < 10^{-7}$ at all comparisons). None of the six comparisons survived Holm correction, including the one previously mentioned ETIS-vs-SANet (nominal $p = 0.037$), which is the 7th smallest of a total of 12 p-values; the value is therefore below the adjusted p value at its rank (which is 0.0083). The honest reading of Table 10 is that at this sample size under single-

seed, the advantage of the CNN baseline is solid, whereas the advantage of the transformer/attention baseline is not established, and should not be reported as such.

Table 10 One-sided paired Wilcoxon signed-rank test (H_1: Proposed > Baseline) on per-image Dice scores.

Comparison	Proposed Dice	Baseline Dice	Δ Dice	p-value	Significant?
vs U-Net (Kvasir)	0.9106	0.8055	+0.1051	5.43×10^{-10}	Yes ✓
vs U-Net++ (Kvasir)	0.9106	0.7986	+0.1120	7.78×10^{-12}	Yes ✓
vs PraNet (Kvasir)	0.9106	0.9027	+0.0079	3.63×10^{-1}	No ✗
vs SANet (Kvasir)	0.9106	0.9022	+0.0085	4.86×10^{-1}	No ✗
vs U-Net (ColonDB)	0.7543	0.5189	+0.2354	5.63×10^{-44}	Yes ✓
vs U-Net++ (ColonDB)	0.7543	0.4894	+0.2649	7.02×10^{-46}	Yes ✓
vs PraNet (ColonDB)	0.7543	0.7328	+0.0215	6.05×10^{-1}	No ✗
vs SANet (ColonDB)	0.7543	0.7456	+0.0087	6.72×10^{-1}	No ✗
vs U-Net (ETIS)	0.7548	0.4105	+0.3443	1.43×10^{-27}	Yes ✓
vs U-Net++ (ETIS)	0.7548	0.3728	+0.3820	5.86×10^{-30}	Yes ✓
vs PraNet (ETIS)	0.7548	0.7035	+0.0514	5.17×10^{-2}	Marginal
vs SANet (ETIS)	0.7548	0.6965	+0.0584	3.67×10^{-2}	Yes ✓

5.3 Degradation Robustness

Table 11 compares the robustness of the five types of degradation on CVC-ColonDB. PolyDetect shows the Best ABSOLUTE Dice for ALL conditions. The score for Poor Illumination (0.772) is worse than clean (0.754), and we are presenting this as a hypothesis, not a proven mechanism, because if gamma correction is applied then the distribution of intensities in ColonDB is shifted towards the distribution of intensities used in the training set of Kvasir-SEG (a domain-shifting artifact rather than a true enhancement capability). To distinguish these two, the enhancement pipeline is applied to otherwise-clean ColonDB images and the repeated check is to see if Dice still rises (Section 6.3), which has not yet been performed. However, it is not the case of PraNet: it does see a smaller numerical reduction (0.015), but it has an extremely low baseline to begin with, meaning it's not even a real reduction, it's just a weaker curve that's flatter. For all models' JPEG compression has been found to be the most harmful degradation.

Table 11 Degradation robustness on CVC-ColonDB zero-shot (n = 380).

Model	Clean	Blur	Noise	Low Contr.	JPEG	Poor Illum.	Mean Deg.	Drop ↓
U-Net	0.519	0.490	0.490	0.453	0.461	0.501	0.479	0.040
U-Net++	0.489	0.463	0.443	0.383	0.379	0.430	0.420	0.070
PraNet	0.733	0.737	0.698	0.727	0.696	0.733	0.718	0.015
SANet	0.746	0.736	0.680	0.721	0.681	0.730	0.710	0.036
Proposed	**0.754**	**0.735**	**0.693**	**0.743**	**0.697**	**0.772**	**0.728**	**0.026**

5.4 Computational Efficiency

On the Tesla T4, it is seen to exceed the real-time constraint of 25 FPS with 27.35M parameters and 12.47 GFLOPs as reported in Table 12. This is because the costs of OpenCV enhancement, BRISQUE, and DnCNN are not included in these estimates. Thus, the end-to-end/worst case wall clock latency should be measured for efficiency in deployment.

Table 12 Computational efficiency (Tesla T4 GPU, batch = 1, input 352 × 352).

Model	Params (M)	FLOPs (G)	FPS ↑	Real-Time?	Notes
U-Net	31.04	103.50	12.32	✗ No	High FLOPs; lowest accuracy
U-Net++	9.16	65.99	25.83	✓ Barely	Fewest parameters; borderline real-time
PraNet	26.55	11.91	85.59	✓ Yes	Fastest inference; lightweight RFB heads
SANet	38.28	16.51	70.12	✓ Yes	Most parameters; efficient attention
Proposed	**27.35**	**12.47**	**53.62**	**✓ Yes (×2.14)**	**53.62 FPS exceeds 25 FPS clinical threshold by ×2.14**

5.5 Component Ablation Study

The result shown in Table 13 uses continued fine-tuning based on the main checkpoint for 15 epochs to assess the contribution of each component relatively. As such, they are not directly comparable to the results shown in Table 7 and can only be interpreted in the context of the ablation table. The largest Dice reduction after removing the boundary-refinement head was −6.47 pp (the Dice reduction is from 0.64 before removing the boundary-refinement head to 0.55 afterward), showing the boundary-refinement head's importance in accurate segmentation for the HD95. The shift in precision–recall curve, however, could be due to the change in the decision threshold, and not just the quality of the boundary in general. In this ablation the degradation-simulation module has not been evaluated separately.

Table 13 Ablation study on CVC-ColonDB zero-shot (n = 380).

Variant	Dice ↑	IoU ↑	Precision ↑	Recall ↑	F2 ↑	HD95 ↓	Δ Dice
Full Proposed (all components)	**0.7817**	**0.6968**	**0.8188**	**0.7931**	**0.7816**	**34.97**	—
(−) Quality-weighted loss	0.7742	0.6813	0.8057	0.7986	0.7797	37.73	−0.0075
(−) Boundary refinement head	0.7170	0.6170	0.6724	0.8776	0.7750	55.50	−0.0647
(−) Adaptive enhancement	0.7578	0.6684	0.7992	0.7755	0.7596	40.50	−0.0239
(−) IQA & enhancement	0.7455	0.6555	0.7957	0.7590	0.7449	37.41	−0.0362

The removal of the IQA and enhancement module together accounts for the second largest decrease (−3.62 pp) and the removal of the enhancement module alone accounts for a smaller but still significant decrease (−2.39 pp). However, a quality-weighted loss is very small in each of its components (−0.75 pp) but effective, suggesting that there are no redundant components; all five components help the performance.

5.6 Quality-Stratified Enhancement Analysis

Table 14 separates CVC-ColonDB images at (q=0.5) while the enhancement pipeline separates at 0.7 and 0.4. For this reason, the CLEAN subset includes 136 pristine images and 213 "lightly" enhanced images, and all 31 DEGRADED images are subjected to CLAHE+DnCNN, but none to the full Retinex pipeline. Enhancement decreases Dice by 0.94 pp on the CLEAN subset and increases it by 3.76 pp on the small DEGRADED subset. This decrease might be due to the unexperienced DnCNN or enhancement artifacts, which needs more validation.

Table 14 Quality-stratified enhancement analysis on CVC-ColonDB (n = 380).

Subset	n	Quality Range	With Enhancement	Without Enhancement	Δ Dice

CLEAN ($q \geq 0.5$)	349 (91.8%)	$q \geq 0.5$	0.7591	0.7497	+0.0094
DEGRADED ($q < 0.5$)	31 (8.2%)	$q < 0.5$	0.6997	0.7374	−0.0376

5.7 IQA Fusion Ablation

The fusion and CNN only results cannot be compared directly with the BRISQUE only one due to the particular fine-tuning protocol used. Their Dice scores differ by at most 0.81 pp, but the CNN branch is not trained and randomly initialized; thus, the performance of the branch should not be considered competitive with BRISQUE.

Table 15 IQA fusion ablation on CVC-ColonDB zero-shot.

IQA Mode	Dice ↑	IoU ↑	Precision ↑	Recall ↑	F2 ↑	HD95 ↓
BRISQUE-only	0.7695	0.6811	0.7927	0.8117	0.7810	40.71
CNN-only	0.7776	0.6877	0.8217	0.7876	0.7783	39.16
Fusion 50/50 (Proposed)	**0.7714**	**0.6781**	**0.7690**	**0.8492**	**0.7992**	**43.91**

The fusion approach is kept as the default value not for raw Dice but because it gives the highest Recall (0.8492) and F2-score (0.7992), which focuses on putting in as many missed polyps in the correct classification while not putting in too many polyps in the wrong classification, a clinically crucial trade-off for polyp screening; missing a polyp is a larger risk than a false positive.

5.8 Comparison with State-of-the-Art

Table 16 compares the performance of PolyDetect with published state-of-the-art methods, but caution needs to be taken when comparing directly because of the differences in the training and testing protocols. PolyDetect uses an in-domain protocol with 900 Kvasir-SEG images for training while ColonDB and ETIS are zero-shot test sets following the PraNet protocol (1,450 training images combined). Although PolyDetect requires approximately 38% less training data than the other strong methods in the set, e.g., PraNet protocols SANet (0.904) and Polyp-PVT (0.917) on the other two datasets, it performs competitively on the directly comparable Kvasir-SEG Dice metric. This indicates that, in addition to being in-distribution accurate, PolyDetect also shows cross-dataset robustness, which is something that was not directly evaluated by the combined training method of the PraNet protocol.

Table 16 Quantitative comparison with state-of-the-art.

Method	Year	Encoder	Kvasir Dice	ColonDB Dice	ETIS Dice	Protocol
U-Net [12]	2015	VGG/CNN	0.818	0.512	0.398	PraNet
PraNet [4]	2020	ResNet-50	0.898	0.712	0.628	PraNet
SANet [5]	2021	ResNet-50	0.904	0.753	0.750	PraNet
Polyp-PVT [6]	2021	PVT-v2	0.917	0.808	0.787	PraNet
SSFormer-L [9]	2022	PVT+Trans.	0.917	0.802	0.796	PraNet
ColonFormer-L [10]	2022	MiT-B3	0.918	0.809	0.793	PraNet
CFA-Net [31]	2023	ResNet-50	0.917	0.810	0.793	PraNet
VMDU-Net [11]	2025	ViM+Trans.	0.934	—	—	PraNet
PolyDetect	**2025**	**PVT-v2+CNN**	**0.9106**	**0.7543**	**0.7548**	**In-domain***

6 Discussion

6.1 Key Findings

With only 900 training images, PolyDetect had the highest average Dice (0.8066) and IoU (0.7262) in all of three datasets. It demonstrated good robustness under degraded images and better performance than CNN baselines. The gains on PraNet and SANet were not statistically significant, however, after correction, thus a multi-seed validation is required.

6.2 Architectural Insights from Ablation

The findings of Ablation Study (Table 13) clearly indicate the importance of each component in CVC-ColonDB under continued-fine-tuning ablation: (1) Boundary Refinement Head (−6.47 pp Dice, HD95: 34.97 → 55.50) is by far the most critical single component; (2) IQA & Enhancement (−3.62 pp); (3) Adaptive Enhancement alone (−2.39 pp); (4) Quality-Weighted Loss (−0.75 pp). This paper does not have an ablation reported on ETIS or Kvasir-SEG, so the importance of the boundary head has not been quantified yet on this dataset, and it will not be possible to claim any cross-dataset importance of this boundary head until there is an ablation reported on this dataset. The enhancement paradox, where adaptive enhancement causes Dice on Kvasir-SEG clean images to increase slightly but decreases HD95, suggests that most of the benefit of enhancement is in improving boundary precision, rather than overlaps in regions, as expected because the illumination and contrast artifacts are a primary influence on hyperplastic areas around the edges of polyps.

6.3 Limitations and Future Work

The study is based on non-standard evaluation protocol and shared validation/test set; single-seed results and different evaluation protocols restrict direct comparison and statistical confidence. Future research efforts should employ standard protocols, make use of multiple data splits, multiple seeds, full-length ablations, baseline training parameters from the original experiments, and end-to-end latency measurements. Recent methods, the PolypGen data set and prospective testing using private clinical data should also be evaluated.

6.4 Clinical Deployment Potential

PolyDetect is capable of running at 53.62 fps for 27.35 M parameters, suggesting that it can be used for real-time colonoscopy support. The end-to-end latency, however, which includes the quality assessment and enhancement must be validated prior to clinical deployment. It is quality-gated, which could potentially lower the processing costs and have good prospects for use in resource-limited environments.

7 Conclusion

PolyDetect provides a novel framework that combines aspects of both approaches and introduces a quality-aware strategy for confidently segmenting polyps in degraded colonoscopy images. The framework attained the best overall scores of Dice and IoU on Kvasir-SEG, CVC-ColonDB, and ETIS by incorporating image-

quality assessment, gated enhancement, a PVT-v2 encoder, boundary refinement, a CNN decoder and a quality-weighted loss. It was also significantly better than CNN-based baselines on all three datasets. Ablation studies show that the architecture, in particular the boundary-refinement head, is the prime source of robustness, not degradation augmentation. Another key advantage of the quality-gated enhancement pipeline is the reduction of processing of higher quality frames, which also enhances segmentation. The segmentation time of 53.62 FPS represents good potential to use in clinical applications in real time. Multi-seed experiments, end to end latency testing, extensive benchmark evaluation and prospective validation using real clinical data should verify these results in the future.

References

[1] Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. 71, 209–249 (2021).

[2] Zauber, A.G. et al. Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths. N. Engl. J. Med. 366, 687–696 (2012).

[3] Rex, D.K. et al. Colorectal Cancer Screening: Recommendations for Physicians and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer. Gastroenterology 153, 307–323 (2017).

[4] Fan, D.-P. et al. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In Proc. MICCAI 2020, LNCS 12266, 263–273 (2020).

[5] Wei, J. et al. Shallow Attention Network for Polyp Segmentation. In Proc. MICCAI 2021, LNCS 12901, 699–708 (2021).

[6] Dong, B. et al. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. arXiv:2108.06932 (2021).

[7] Ali, S. et al. A Comprehensive Study on Colorectal Polyp Segmentation with ResUNet++, ConditionalRandom Field and Test-Time Augmentation. IEEE J. Biomed. Health Inform. 25, 714–725 (2021).

[8] Tajbakhsh, N. et al. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. IEEE Trans. Med. Imaging 35, 630–644 (2016).

[9] Wang, J. et al. Stepwise Feature Fusion: Local Guides Global. In Proc. MICCAI 2022, LNCS 13433, 110–120 (2022).

[10] Duc, N.T. et al. ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation. IEEE Access 10, 80575–80586 (2022).

[11] Li, X. et al. VMDU-Net: Visual Mamba with Dual-Branch Transformer for Polyp Segmentation. Front. Artif. Intell. (2025). doi:10.3389/frai.2025.1557508.

[12] Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proc. MICCAI 2015, LNCS 9351, 234–241 (2015).

[13] Zhou, Z. et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Deep Learning in Medical Image Analysis 2018, LNCS 11045, 3–11 (2018).

[14] Mittal, A., Moorthy, A.K. & Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans. Image Process. 21, 4695–4708 (2012).

[15] Mittal, A., Soundararajan, R. & Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. IEEE Signal Process. Lett. 20, 209–212 (2013).

[16] Kohlberger, T. et al. Whole-Slide Image Focus Quality: Automatic Assessment and Impact on AI Cancer Detection. J. Pathol. Inform. 10, 39 (2019).

[17] Zago, G.T. et al. Retinal Image Quality Assessment Using Deep Learning. Comput. Biol. Med. 103, 64–70 (2018).

[18] Ali, S. et al. Endoscopy Artifact Detection (EAD 2019) Challenge Dataset. arXiv:1905.03209 (2019).

[19] Land, E.H. The Retinex Theory of Color Vision. Sci. Am. 237, 108–128 (1977).

[20] Pizer, S.M. et al. Adaptive Histogram Equalization and its Variations. Comput. Vis. Graph. Image Process. 39, 355–368 (1987).

[21] Zuiderveld, K. Contrast Limited Adaptive Histogram Equalization. In Graphics Gems IV, 474–485 (Academic Press, 1994).

[22] Zhang, K. et al. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. IEEE Trans. Image Process. 26, 3142–3155 (2017).

[23] Bengio, Y. et al. Curriculum Learning. In Proc. ICML 2009, 41–48 (2009).

[24] Lin, T.-Y. et al. Focal Loss for Dense Object Detection. In Proc. ICCV 2017, 2980–2988 (2017).

[25] Shrivastava, A. et al. Training Region-based Object Detectors with Online Hard Example Mining. In Proc. CVPR 2016, 761–769 (2016).

[26] Li, J., Xu, Q., He, X., Liu, Z., Zhang, D., Wang, R., Qu, R. & Qiu, G. CFFormer: Cross CNN-Transformer Channel Attention and Spatial Feature Fusion for Improved Segmentation of Heterogeneous Medical Images. arXiv:2501.03629 (2025).

[27] Wang, W. et al. PVT v2: Improved Baselines with Pyramid Vision Transformer. Comput. Vis. Media 8, 415–424 (2022).

[28] Jha, D. et al. Kvasir-SEG: A Segmented Polyp Dataset. In Proc. MMM 2020, LNCS 11962, 451–462 (2020).

[29] Bernal, J., Sánchez, J. & Vilariño, F. Towards Automatic Polyp Detection with a Polyp Appearance Model. Pattern Recognit. 45, 3166–3182 (2012).

[30] Silva, J. et al. Toward Embedded Detection of Polyps in WCE Images for Early Diagnosis of Colorectal Cancer. Int. J. Comput. Assist. Radiol. Surg. 9, 283–293 (2014).

[31] Zhou, T. et al. Cross-Level Feature Aggregation Network for Polyp Segmentation. Pattern Recognit. 140, 109555 (2023).

[32] Ali, S., Jha, D., Ghatwary, N. et al. A Multi-Centre Polyp Detection and Segmentation Dataset for Generalisability Assessment (PolypGen). Sci. Data 10, 75 (2023).