sensors

MDPI

Article

DSF-BRNet: Dual-Gated Semantic Fusion and Boundary Refinement for Efficient Endoscopic Polyp Segmentation

Botao Liu [1], Changqi Shi [1] and Ming Zhao [2,*]

[1] School of Computer Science, Yangtze University, Jingzhou 434023, China; 500327@yangtzeu.edu.cn (B.L.); shichangqi17@gmail.com (C.S.)
[2] School of Cyber Science and Engineering, Wuxi University, Wuxi 214105, China
* Correspondence: mingzhao@cwxu.edu.cn

Abstract

Early detection and accurate segmentation of colorectal polyps during colonoscopy are crucial for the prevention of colorectal cancer. However, automated polyp segmentation remains challenging because of high inter-class variance, complex intestinal backgrounds, and blurred boundaries. To address these issues while maintaining computational efficiency, DSF-BRNet was developed for endoscopic polyp segmentation. In this framework, a Dual-Gated Semantic Fusion (DSF) module is introduced to reduce spatial misalignment between cross-level features and to provide a more reliable semantic basis for lesion localization. To further alleviate boundary ambiguity, a High-Frequency Boundary Refinement (HBR) module is used to sharpen segmentation contours under aligned semantic guidance. Together, these components form an Align-then-Refine framework in which semantic localization is strengthened before boundary refinement is performed. Experiments on four public benchmark datasets—Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-LaribPolypDB—showed competitive performance with favorable computational efficiency. Mean Dice scores of 0.943 on CVC-ClinicDB and 0.818 on ETIS-LaribPolypDB were achieved, with 25.55 M parameters and an inference speed of 80.08 FPS. These results indicate that accurate semantic localization and fine boundary preservation can be achieved simultaneously, suggesting that the method may be promising for real-time computer-aided diagnosis (CAD).

Keywords: polyp segmentation; efficient network; real-time segmentation; semantic fusion; boundary refinement

Check for updates

Academic Editor: Jan Cornelis

Received: 16 March 2026
Revised: 15 April 2026
Accepted: 26 April 2026
Published: 28 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Colorectal cancer (CRC) is one of the most common and lethal malignancies globally, posing a severe threat to public health [1,2]. Clinical studies have extensively demonstrated that the early detection and resection of precancerous colorectal polyps during colonoscopy can significantly reduce the mortality rate associated with CRC [3,4]. However, traditional manual screening highly depends on the experience and concentration of the endoscopists. Due to the prolonged operation time and visual fatigue, the polyp miss rate remains non-negligible, particularly for small or flat polyps that are camouflaged within the intestinal mucosa [5,6]. Consequently, developing an automated, robust, and accurate computer-aided diagnosis system for endoscopic polyp segmentation is of paramount clinical significance.

In recent years, deep learning models have significantly advanced automatic polyp segmentation [7]. Early architectures like U-Net [8] and UNet++ [9] rely on encoder–decoder structures with skip connections, but they struggle with the extreme scale variations and diverse morphologies of polyps. To address this, task-specific models have been proposed; for instance, PraNet [10] utilizes reverse attention for boundary mining, while Polyp-PVT [11] employs a transformer backbone to capture long-range dependencies.

Despite the promising performance of these task-specific networks, automated polyp segmentation remains highly challenging due to two primary obstacles: (1) Semantic Gap and Spatial Misalignment: High-level features contain rich semantics but lack spatial resolution, whereas low-level features possess fine details but are easily contaminated by background noise [12]. Directly fusing these cross-level features often causes spatial misalignment and false positives in complex mucosal backgrounds. (2) Boundary Ambiguity: Due to high intra-class variance and low contrast between polyps and surrounding tissues, precise contour delineation remains a bottleneck [13,14]. To explicitly tackle these intertwined challenges, various boundary-enhancement techniques have been explored [15]. However, rigorous experimental investigation reveals a critical underlying mechanism: boundary refinement is largely ineffective, or even counterproductive, when fundamental semantic features are spatially misaligned. Aggressively injecting high-frequency edge cues without precise semantic localization tends to amplify background noise and generate misplaced boundaries. Therefore, the main issue addressed in this paper is the mismatch between reliable semantic localization and accurate boundary recovery in challenging endoscopic scenes.

To overcome these accuracy bottlenecks, one might look toward the recent emergence of vision foundation models, which have demonstrated remarkable zero-shot generalization capabilities. However, a critical clinical dilemma persists: their staggering parameter counts (often exceeding hundreds of millions) and massive memory consumption render high-framerate real-time inference practically unattainable on the resource-constrained edge computing platforms of current clinical endoscopy towers. Consequently, achieving a pragmatic balance between lightweight architecture and robust accuracy remains a pressing necessity for real-world clinical deployment.

To address this issue, DSF-BRNet, a Dual-Gated Semantic Fusion and Boundary Refinement Network, was developed for efficient endoscopic polyp segmentation. The framework follows a sequential Align-then-Refine strategy, in which semantic misalignment is reduced before contour sharpening is performed. A Dual-Gated Semantic Fusion (DSF) module is used to model channel-wise and spatial relationships for cross-level alignment, thereby providing a more reliable semantic basis. On this basis, a High-Frequency Boundary Refinement (HBR) module employs a fixed-weight Sobel operator to extract structural cues, which are fused with aligned semantics through learnable weights to refine polyp contours without introducing a notable computational burden. In summary, the main contributions of this paper are four-fold:

1. A novel and efficient deep learning network, termed DSF-BRNet, is presented for polyp segmentation, achieving a favorable balance between efficiency and accuracy while addressing the challenges of complex intestinal backgrounds and blurred boundaries.
2. The Dual-Gated Semantic Fusion (DSF) module is designed to effectively correct spatial misalignments between hierarchical features, establishing a precise semantic localization foundation.
3. A High-Frequency Boundary Refinement (HBR) module is introduced. Guided by aligned semantics, it utilizes high-frequency structural cues to actively refine and sharpen segmentation contours.

4. An Align-then-Refine strategy is comprehensively conceptualized and validated. Extensive experiments on four public benchmark datasets demonstrate that the proposed DSF-BRNet achieves competitive accuracy with superior real-time efficiency.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the DSF-BRNet method. Section 4 presents the experiments and discussion. Section 5 concludes the paper.

2. Related Work

2.1. Deep Learning for Medical Image Segmentation

Deep learning architectures have fundamentally driven the progress of medical image analysis [16]. Early frameworks, such as Fully Convolutional Networks (FCNs) [17], formulated segmentation as a pixel-level classification task. Subsequently, U-Net established the standard encoder–decoder paradigm, utilizing skip connections to transfer low-level spatial details directly to high-level semantic maps. To further mitigate the semantic gap between encoder and decoder stages, UNet++ introduced nested and dense skip pathways. Recently, Vision Transformers [18] have been adapted for medical image segmentation to capture long-range dependencies and global contextual representations. While these general-purpose models demonstrate robust performance across various medical imaging modalities, applying them directly to colonoscopy images often yields suboptimal results due to the severe scale variations and low contrast typical of endoscopic environments.

2.2. Automated Polyp Segmentation

To address the unique characteristics of colorectal polyps, task-specific segmentation networks have been extensively investigated. PraNet introduced a reverse attention mechanism to systematically mine boundary cues and refine the localization of polyps. SANet [19] proposed a shallow attention module combined with a color exchange strategy to filter out background noise from low-level features. Furthermore, transformer-based architectures, such as Polyp-PVT, and adaptive multi-scale perception frameworks [20] have been utilized to extract robust global contexts, effectively handling multi-scale variations. However, when segmenting camouflaged polyps that share similar textures with the surrounding intestinal mucosa, these methods frequently suffer from false positives and ambiguous contours. This limitation primarily stems from the lack of explicit spatial alignment constraints during the aggregation of cross-level features.

2.3. Feature Fusion and Boundary-Aware Learning

Effective cross-layer feature fusion and precise boundary delineation are critical for accurate polyp segmentation. Conventional fusion strategies predominantly rely on direct channel concatenation or element-wise addition [21]. These operations, however, often lead to spatial misalignment and semantic dilution, given the inherent resolution and representation disparities between hierarchical levels. Furthermore, existing attention mechanisms, such as SE [22] and CBAM [23], often rely on fully connected layers, which entails severe dimensionality reduction, inevitably destroying the fine-grained spatial details crucial for medical images. To alleviate boundary ambiguity, explicit boundary-aware learning paradigms have been developed. These methods typically employ edge operators or auxiliary edge-supervision branches to explicitly constrain the network. Although beneficial for highlighting structural contours, empirical evidence indicates that blindly injecting high-frequency boundary cues without a stable semantic foundation can inadvertently amplify background noise and generate misplaced boundaries [24]. Similarly, prevalent boundary-aware methods like PraNet utilize reverse attention to implicitly erase foregrounds for edge mining, which can be fragile when dealing with severely blurred

polyp contours. Related ideas have also been explored in other image segmentation tasks. For instance, a recent study on road connectivity enhancement in remote sensing imagery employed a multi-modal attention fusion strategy to improve structural integrity [25]. Although it focuses on a different application, it still suggests that feature fusion and structure-aware refinement can be beneficial in challenging segmentation scenarios. To overcome these inherent limitations, a decoupled learning paradigm is presented in this work. The proposed DSF module avoids dimensionality reduction via 1D convolutions for efficient cross-channel interaction, while the HBR module shifts from implicit erasure to explicitly injecting deterministic structural gradients, fundamentally differing from previous paradigms.

3. Methodology

3.1. Overall Architecture

The overall architecture of the proposed DSF-BRNet is illustrated in Figure 1a. Adopting a standard encoder–decoder paradigm, the network primarily comprises three integral components: a multi-scale feature extraction backbone, a top-down Dual-Gated Semantic Fusion (DSF) cascade, and a High-Frequency Boundary Refinement (HBR) module.

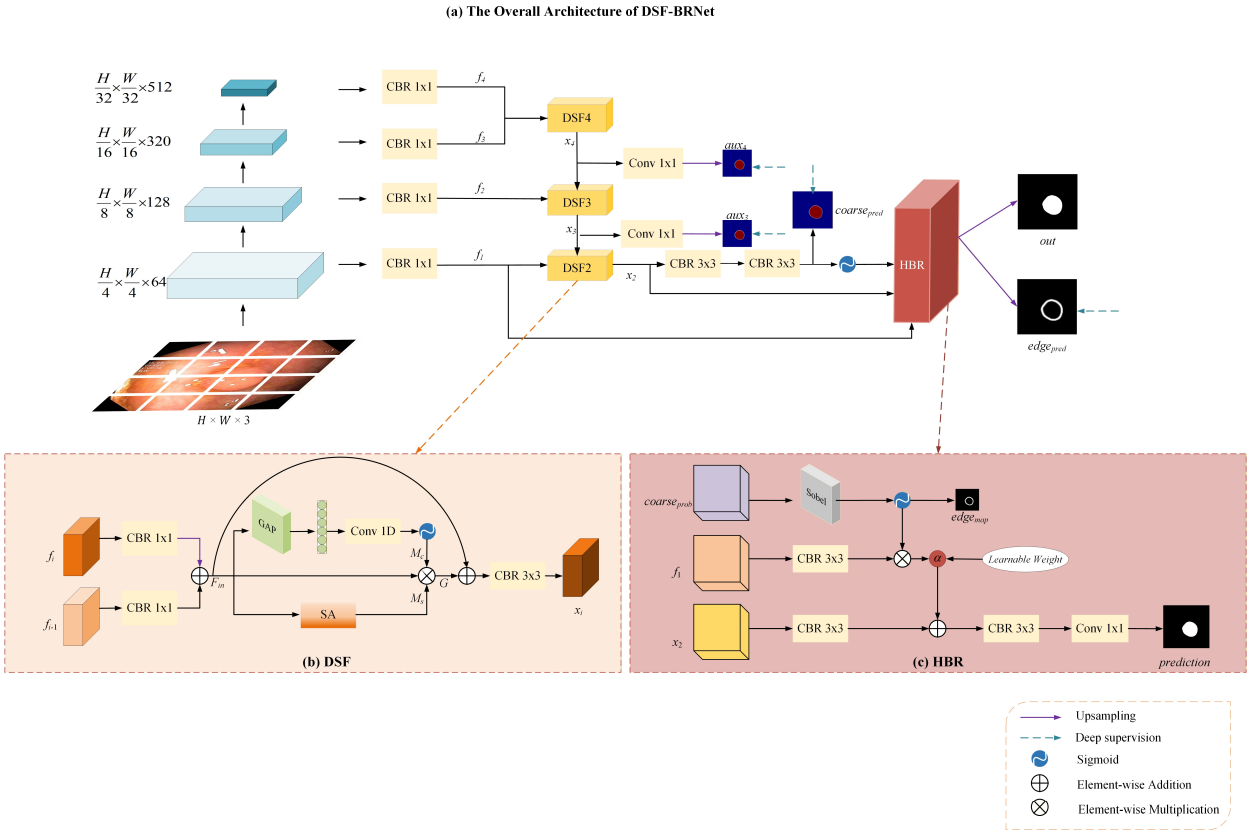

Figure 1. (**a**) Overall architecture of the proposed DSF-BRNet. (**b**) Architecture of DSF; (**c**) Architecture of HBR. CBR denotes Convolution+BatchNorm+ReLU.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the Pyramid Vision Transformer v2 (PVTv2) [26] is initially employed as the hierarchical backbone to extract multi-scale features at four distinct stages. Its pyramid structure naturally produces hierarchical representations at different resolutions, which fits well with the proposed top-down DSF cascade for cross-level feature fusion. At the same time, it provides strong contextual modeling while maintaining relatively high efficiency, making it suitable for polyp segmentation where both multi-scale representation and real-time inference are important. To streamline the

computational cost and unify the channel dimensions for subsequent feature aggregation, a 1×1 Convolution-BatchNorm-ReLU (CBR) block is applied to each encoder stage. This operation yields a set of channel-reduced features denoted $\{f_i\}_{i=1}^{4}$, with spatial resolutions ranging from $\frac{H}{4} \times \frac{W}{4}$ to $\frac{H}{32} \times \frac{W}{32}$.

As shown in Figure 1a, the extracted backbone features are not fused by direct skip concatenation. Instead, they are progressively integrated through the top-down DSF cascade, where higher-level semantic features interact with lower-level features stage by stage. In this way, multi-scale features from different backbone stages are used consistently during decoding. The final coarse prediction, together with the shallow feature and the aligned semantic feature, is then sent to HBR for boundary refinement.

In the decoding phase, directly concatenating features from different levels often leads to semantic dilution. To address this, the DSF modules are deployed in a cascaded top-down manner to progressively aggregate the high-level semantic features and the corresponding low-level spatial features. This cascaded decoding process generates a series of semantically aligned intermediate features, denoted as x_4, x_3, and x_2. To facilitate the training process and enhance gradient propagation, deep supervision is introduced by generating auxiliary predictions (aux_4, aux_3) from x_4 and x_3 via 1×1 convolutions.

Subsequently, the fused feature x_2 is passed through consecutive 3×3 CBR blocks and a Sigmoid function to produce a coarse segmentation probability map $\text{coarse}_{\text{pred}}$. Finally, to explicitly tackle the boundary ambiguity prevalent in endoscopic images, $\text{coarse}_{\text{pred}}$, the shallow spatial feature f_1, and the aligned semantic feature x_2 are fed into the HBR module. The HBR module dynamically extracts edge representations $\text{edge}_{\text{pred}}$ and sharpens the segmentation contours, yielding the final accurate prediction out.

3.2. Dual-Gated Semantic Fusion (DSF) Module

In conventional U-shaped networks, cross-layer features are typically fused via simple element-wise addition or channel concatenation. However, due to the inherent semantic gap—where high-level features possess strong semantics but low spatial resolution, while low-level features contain rich detail but abundant background noise—direct fusion frequently causes spatial misalignment. To mitigate this issue, the Dual-Gated Semantic Fusion (DSF) module is designed to recalibrate the fused features along both channel and spatial dimensions before passing them to the next stage.

As depicted in Figure 1b, let f_i represent the high-level input feature and f_{i-1} denote the low-level input feature. Initially, both features undergo a 1×1 CBR block to align their representations, followed by an element-wise addition to generate the initial fused feature F_{in}:

$$F_{in} = \text{CBR}_{1\times1}(f_i) \oplus \text{CBR}_{1\times1}(f_{i-1}) \tag{1}$$

where $\oplus$ denotes element-wise addition. Note that $\text{CBR}_{1\times1}(f_i)$ is first upsampled to match the spatial resolution of f_{i-1} via bilinear interpolation before the addition operation.

To selectively emphasize informative regions and suppress background noise, the fused feature is subsequently routed into two parallel recalibration branches. For the channel dimension, the Efficient Channel Attention (ECA) mechanism [27] is employed instead of the fully connected design used in standard attention modules, which often introduces destructive dimensionality reduction. A Global Average Pooling (GAP) operation is first applied to aggregate the spatial information into a channel descriptor. A 1D convolution is then used to model channel-wise dependencies without compressing the channel dimension, followed by a Sigmoid activation to generate the channel attention map: F_{in} F_{in} M_c

$$M_c = \sigma(\text{Conv1D}(\text{GAP}(F_{in}))) \tag{2}$$

where σ represents the Sigmoid function.

Concurrently, along the spatial dimension, a spatial attention mechanism is applied to F_{in} to highlight the positional correspondences of the polyp regions. Average-pooling and max-pooling operations are first applied independently along the channel axis of F_{in} to generate two 2D spatial maps. These maps are then concatenated and convolved by a standard convolution layer to produce the spatial attention map M_s:

$$M_s = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}(F_{in}); \text{MaxPool}(F_{in})])) \quad (3)$$

where $[\cdot;\cdot]$ denotes the concatenation operation.

Finally, the attention maps M_c and M_s are multiplied with the initial fused feature F_{in} to perform dual-gated recalibration. To preserve the integrity of the original feature flow and prevent gradient vanishing, a residual connection is incorporated. The recalibrated feature is then smoothed by a 3×3 CBR block to output the final aligned feature x_i:

$$x_i = \text{CBR}_{3\times3}(F_{in} \oplus (F_{in} \otimes M_c \otimes M_s)) \quad (4)$$

where $\otimes$ denotes element-wise multiplication. Through this dual-gated mechanism, the DSF module effectively corrects spatial misalignments, providing a robust and clean semantic foundation for the subsequent boundary refinement process.

3.3. High-Frequency Boundary Refinement (HBR) Module

While the cascaded DSF modules successfully establish a reliable semantic foundation by correcting spatial misalignments, the inherently low contrast between polyps and normal intestinal mucosa often leads to blurred segmentation boundaries. To explicitly address this degradation, a High-Frequency Boundary Refinement (HBR) module is introduced. This module performs the final contour enhancement stage, leveraging high-frequency structural cues to sharpen contours under the strict guidance of aligned semantics. As explicitly illustrated in Figure 1c, the HBR module processes three distinct inputs: the coarse probability map $coarse_{prob}$, the shallow spatial feature f_1, and the semantically aligned deep feature x_2. Initially, a Sobel operator, implemented as fixed-weight 3×3 convolutional kernels to allow end-to-end training, is applied to $coarse_{prob}$ to deterministically extract spatial gradients. The Sobel operator is used here as a lightweight and deterministic structural prior for boundary refinement. Because the HBR module operates on the coarse segmentation map rather than the raw image, stable gradient cues can be obtained with minimal computational overhead and without introducing additional trainable parameters. The extracted gradients are then followed by a Sigmoid activation σ to generate a normalized edge mask, denoted as M_{edge}:

$$M_{edge} = \sigma\Big(\text{Sobel}(coarse_{prob})\Big) \quad (5)$$

This mask serves a dual purpose—it acts as a spatial attention map for subsequent feature filtering, and it is explicitly supervised by the ground truth edge contours $edge_{map}$ during training to enforce structural fidelity. Subsequently, the shallow feature f_1 undergoes a 3×3 CBR block for feature adaptation. The resulting feature is multiplied element-wise $\otimes$ by M_{edge} to suppress background noise and highlight boundary-relevant details. A learnable weight parameter α is then applied to dynamically scale these isolated high-frequency features, formulated as:

$$f_{edge_scaled} = \alpha \cdot (\text{CBR}_{3\times3}(f_1) \otimes M_{edge}) \quad (6)$$

Concurrently, the deep semantic feature x_2 is processed through a 3×3 CBR block. The scaled edge feature f_{edge_scaled} is injected into this semantic representation through element-wise addition $\oplus$. Ultimately, the fused feature passes through a final 3×3 CBR block and a 1×1 convolution to generate the final segmentation prediction:

$$prediction = \mathrm{Conv}_{1\times1}\Big(\mathrm{CBR}_{3\times3}(\mathrm{CBR}_{3\times3}(x_2) \oplus f_{edge_scaled})\Big) \tag{7}$$

Through this explicit structural guidance pathway, the HBR module effectively sharpens polyp boundaries without compromising the semantic accuracy established by the preceding cascaded fusion.

3.4. Loss Function

To comprehensively optimize the proposed network, a hybrid loss function is formulated to simultaneously supervise the regional semantic segmentation and explicit boundary delineation. For the semantic segmentation predictions, a structure loss $\mathcal{L}_{struct}$, originally introduced by PraNet [10] and widely adopted in recent architectures like DCANet [28], is employed. It combines a weighted Binary Cross-Entropy (wBCE) loss and an Intersection over Union (IoU) loss [10]

$$\mathcal{L}_{struct}(P, G) = \mathcal{L}_{wBCE}(P, G) + \mathcal{L}_{IoU}(P, G) \tag{8}$$

where P represents the predicted segmentation map and G denotes the corresponding ground truth mask. The wBCE loss provides pixel-level supervision while assigning higher weights to hard pixels. The IoU loss, in turn, focuses on the global structure of the segmented target and helps alleviate the foreground/background imbalance commonly encountered in endoscopic images. The structure loss is defined as the direct sum of these two terms, without introducing an additional balancing coefficient. In this way, the semantic prediction branch is jointly constrained at both the pixel level and the region level. For the boundary branch, Dice loss is adopted because the target edge map is relatively sparse, making Dice-based supervision more suitable in this setting.

Furthermore, to guide the High-Frequency Boundary Refinement (HBR) module, an explicit edge supervision mechanism is incorporated. Given that target edge maps are inherently highly sparse, a Dice loss [29] $\mathcal{L}_{\mathrm{dice}}$ is utilized to supervise the edge prediction $edge_{pred}$:

$$\mathcal{L}_{dice}(E, G_{edge}) = 1 - \frac{2\sum(E \times G_{edge}) + \epsilon}{\sum E + \sum G_{edge} + \epsilon} \tag{9}$$

where E is the predicted edge map, $\mathrm{G_{edge}}$ is the ground truth boundary contour, and ϵ is a smoothing factor set to 10^{-5} to ensure numerical stability.

Following standard practices in deep supervision architectures, let $\mathcal{P} = \left\{prediction, coarse_{prob}, aux_3, aux_4\right\}$ denote the set of multi-level semantic segmentation outputs generated by the network. The total loss function $\mathcal{L}_{total}$ is elegantly formulated as a weighted summation of the structure losses from all side outputs and the explicit boundary loss:

$$\mathcal{L}_{total} = \sum_{P_i \in \mathcal{P}} \mathcal{L}_{struct}(P_i, G) + \mathcal{L}_{dice}(edge_{pred}, G_{edge}) \tag{10}$$

This design allows the network to optimize region-level segmentation and sparse boundary supervision at the same time, which matches the roles of the semantic prediction branch and the boundary refinement branch in our framework.

4. Experiments

4.1. Evaluation Metrics

To quantitatively assess the performance of the proposed DSF-BRNet and other comparison methods, two widely adopted semantic segmentation metrics were primarily utilized: the mean Dice Similarity Coefficient (mDice) and the mean Intersection over Union (mIoU).

The Dice Similarity Coefficient calculates the spatial overlap index between the predicted segmentation mask P and the corresponding ground truth G. It places a strong penalty on false positives and false negatives, making it highly suitable for medical image segmentation tasks. mDice is formulated as:

$$\text{mDice} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times |P_i \cap G_i|}{|P_i| + |G_i|} \tag{11}$$

where N denotes the total number of images in the testing dataset, and $|\cdot|$ calculates the number of pixels within the specified region.

Similarly, the Intersection over Union computes the ratio of the intersection area to the union area between the prediction and the ground truth. mIoU is defined as:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{|P_i \cup G_i|} \tag{12}$$

In addition to mDice and mIoU, which directly measure region-level similarity, other comprehensive evaluation metrics including *F-measure* (F_β^w) [30], *S-measure* (S_α) [31] and *E-measure* (E_ϕ) [32] are also reported to evaluate boundary quality, structural similarity, and both pixel-level and image-level alignment, providing a holistic assessment of the segmentation performance.

4.2. Datasets

The proposed framework was evaluated on four publicly available colorectal polyp segmentation benchmark datasets to examine its robustness and generalization ability. Since these datasets differ in image quality, polyp appearance, and background complexity, they provide a suitable test bed for a more comprehensive performance assessment.

Kvasir-SEG [33]: This dataset contains 1000 high-resolution endoscopic images acquired using standard endoscopic equipment, accompanied by pixel-level ground truth masks manually annotated by expert gastroenterologists.

CVC-ClinicDB [34]: This dataset comprises 612 images extracted from 31 distinct colonoscopy video sequences. It features various polyp sizes and distinct lighting conditions, posing challenges for consistent feature extraction.

CVC-ColonDB [35]: A highly challenging dataset consisting of 380 images collected from 15 different colonoscopy sequences. It is characterized by severe variations in polyp appearance and heterogeneous intestinal backgrounds.

ETIS-LaribPolypDB [36]: This dataset contains 196 images predominantly featuring small, flat, and camouflaged polyps. It is extensively utilized as a rigorous testbed to evaluate the out-of-distribution generalization capability of models against complex mucosal structures.

Following the commonly used evaluation protocol in PraNet, 900 images from Kvasir-SEG and 550 images from CVC-ClinicDB were used for training, while the remaining 100 Kvasir-SEG images and 62 CVC-ClinicDB images were used for in-domain testing. In addition, the entire CVC-ColonDB and ETIS-LaribPolypDB datasets were used only for

cross-dataset evaluation, in order to assess the generalization ability of the proposed model on unseen test sets.

4.3. Implementation Details

The proposed DSF-BRNet was implemented in PyTorch 2.0.0 and trained on a single NVIDIA RTX 3090 GPU. All input images were uniformly resized to 352 × 352 pixels. All images were further normalized under the same settings across all datasets. To prevent overfitting, standard data augmentations including random flipping, multi-scale rotation, and random cutout were applied. The PVTv2 backbone was initialized with ImageNet pre-trained weights, while the newly added modules were optimized via AdamW [37] with a weight decay set to 1×10^{-5}. The network was trained for 150 epochs with a batch size of 16. An initial learning rate of 1×10^{-4} was employed, which decayed according to a cosine annealing schedule following a 5-epoch linear warmup. Early stopping was adopted during training. The training process was terminated when the monitored mDice did not improve for 30 consecutive epochs. All experiments were conducted with a fixed random seed.

4.4. Quantitative Analysis

To comprehensively evaluate the effectiveness of the proposed DSF-BRNet, it was compared against seven representative medical image segmentation models. The comparison baseline included classic encoder–decoder architectures such as U-Net [8] and UNet++ [9] alongside specialized polyp segmentation networks including PraNet [10], EU-Net [38], SANet [19], Polyp-PVT [11], and PraNetV2 [39]. For comparison, the results of several representative methods, including PraNet, EU-Net, SANet, and Polyp-PVT, were taken from the literature under the same dataset split protocol adopted in this study. The remaining methods were re-implemented and evaluated under the same experimental setting. The quantitative evaluation results across the four benchmark datasets are detailed comprehensively in Tables 1–4. Tables 1 and 2 show that DSF-BRNet achieved mDice values of 0.933 and 0.943 on Kvasir-SEG and CVC-ClinicDB, respectively. This performance indicates that the proposed framework can effectively capture the main semantic and structural characteristics of polyp regions, leading to accurate segmentation on these two benchmark datasets.

Table 1. Comparative experimental results of the models on Kvasir-SEG.

Model	mDice	mIoU	F_β^w	S_α	E_ϕ
U-Net [8]	0.818	0.746	0.794	0.858	0.881
UNet++ [9]	0.821	0.743	0.808	0.862	0.886
PraNet [10]	0.898	0.840	0.885	0.915	0.944
EU-Net [38]	0.908	0.854	0.893	0.917	0.951
SANet [19]	0.904	0.847	0.892	0.915	0.949
Polyp-PVT [11]	0.917	0.864	0.911	0.925	0.956
PraNet-V2 [39]	0.914	0.861	0.917	0.925	0.966
DSF-BRNet (Ours)	**0.933**	**0.885**	**0.938**	**0.935**	**0.983**

In addition to the in-domain results, the generalization ability of the model was further examined on the unseen and more challenging CVC-ColonDB and ETIS-LaribPolypDB datasets, with detailed results reported in Tables 3 and 4. These datasets provide a more demanding test setting and offer a clearer view of the robustness of the proposed method under cross-dataset evaluation.

As shown in Figures 2 and 3, strong performance was maintained on both CVC-ColonDB and ETIS-LaribPolypDB. On CVC-ColonDB, an mDice of 0.812 was obtained, as

shown in Figure 2. On ETIS-LaribPolypDB, which contains a larger proportion of small and camouflaged polyps, the mDice reached 0.818, exceeding the values reported for Polyp-PVT (0.787) and PraNetV2 (0.774), as shown in Figure 3. Because segmentation accuracy alone is insufficient for practical deployment, computational efficiency is further analyzed in the following subsection.

Table 2. Comparative experimental results of the models on CVC-ClinicDB.

Model	mDice	mIoU	F_β^w	S_α	E_ϕ
U-Net [8]	0.823	0.755	0.811	0.889	0.913
UNet++ [9]	0.794	0.729	0.785	0.873	0.891
PraNet [10]	0.899	0.849	0.896	0.936	0.963
EU-Net [38]	0.902	0.846	0.891	0.936	0.959
SANet [19]	0.916	0.859	0.909	0.939	0.971
Polyp-PVT [11]	0.937	0.889	0.936	0.949	0.985
PraNet-V2 [39]	0.931	0.881	0.928	0.945	0.982
DSF-BRNet (Ours)	**0.943**	**0.896**	**0.939**	**0.950**	**0.993**

Table 3. Comparative experimental results of the models on CVC-ColonDB.

Model	mDice	mIoU	F_β^w	S_α	E_ϕ
U-Net [8]	0.512	0.444	0.498	0.712	0.696
UNet++ [9]	0.483	0.410	0.467	0.691	0.681
PraNet [10]	0.712	0.641	0.699	0.821	0.847
EU-Net [38]	0.756	0.681	0.732	0.831	0.863
SANet [19]	0.753	0.671	0.726	0.837	0.869
Polyp-PVT [11]	0.808	0.727	0.795	**0.865**	0.913
PraNet-V2 [39]	0.766	0.684	0.779	0.841	0.964
DSF-BRNet (Ours)	**0.812**	**0.728**	**0.819**	0.862	**0.967**

Table 4. Comparative experimental results of the models on ETIS-LaribPolypDB.

Model	mDice	mIoU	F_β^w	S_α	E_ϕ
U-Net [8]	0.398	0.335	0.366	0.684	0.643
UNet++ [9]	0.401	0.344	0.391	0.683	0.629
PraNet [10]	0.628	0.567	0.601	0.794	0.808
EU-Net [38]	0.687	0.609	0.636	0.793	0.807
SANet [19]	0.750	0.654	0.685	0.849	0.881
Polyp-PVT [11]	0.787	0.706	0.751	0.871	0.906
PraNet-V2 [39]	0.774	0.697	0.757	0.865	0.903
DSF-BRNet (Ours)	**0.818**	**0.744**	**0.803**	**0.886**	**0.988**

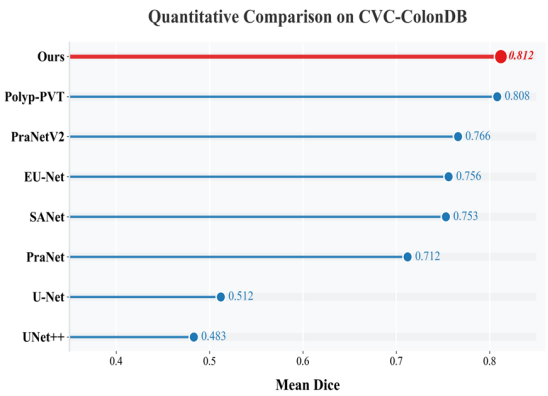

Figure 2. Horizontal bar comparison chart for ColonDB.

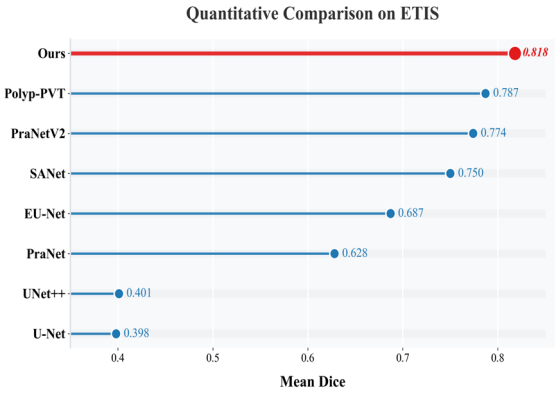

Figure 3. Horizontal bar comparison chart for ETIS-LaribPolypDB.

4.5. Computational Efficiency Analysis

To evaluate practical applicability, Table 5 and Figure 4 present the efficiency–accuracy trade-off of DSF-BRNet and several representative baselines. In clinical endoscopy, real-time processing generally requires inference speeds above 30 FPS. Although heavily parameterized architectures such as UNet++ and EU-Net incur higher computational costs, DSF-BRNet achieves 80.08 FPS with 25.55 M parameters. Compared with PraNet-V2 and Polyp-PVT, similar or lower model complexity was maintained while competitive or better segmentation accuracy was obtained. As shown in Figure 4, the lightweight 1D convolution in the DSF module and the fixed-weight Sobel operator in the HBR module added little latency, resulting in a favorable balance between efficiency and accuracy.

Table 5. Computational efficiency comparison of different models.

Metric	UNet	UNet++	PraNet	SANet	EU-Net	Polyp-PVT	PraNet-V2	DSF-BRNet
Params (M)	32.52	48.99	30.5	23.9	31.36	25.11	30.8	25.55
FLOPs (G)	20.24	108.83	13.15	11.34	23.15	10.02	11.23	11.37
FPS	180.17	48.99	71.67	110.81	11.40	79.26	63.74	80.08

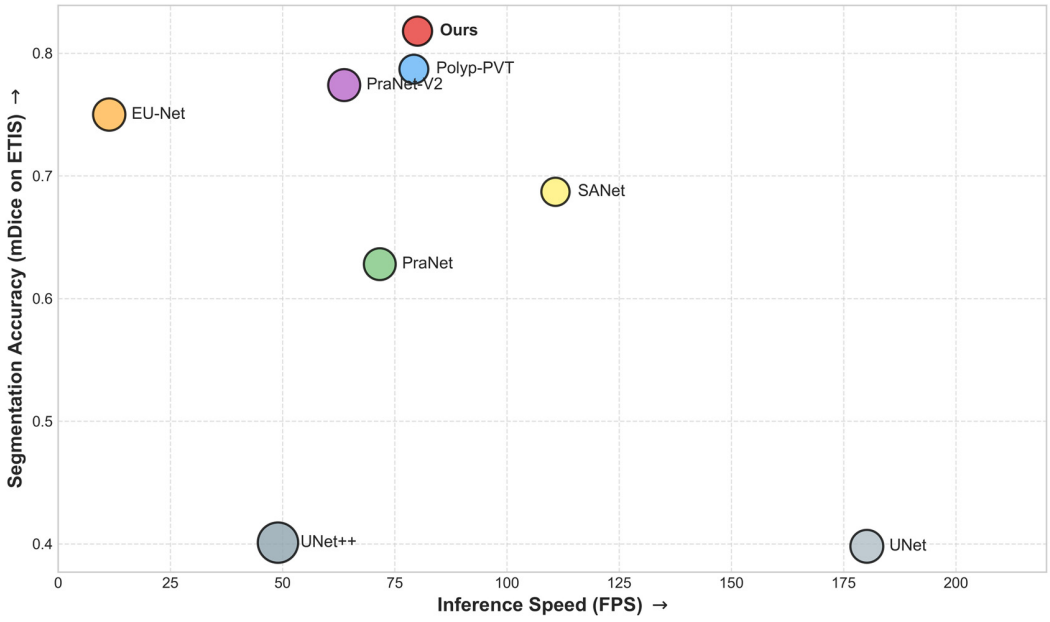

Figure 4. Trade-off between segmentation accuracy (mDice) and inference speed (FPS) on the ETIS dataset.

4.6. Qualitative Analysis

To provide an intuitive comparison, the visual segmentation results of different models under challenging colonoscopy scenarios are presented in Figure 5. The selected cases encompass massive polyps with irregular shapes in the first and second rows, small polyps in the fourth and fifth rows, and flat polyps sharing similar textures with the surrounding mucosa in the third row.

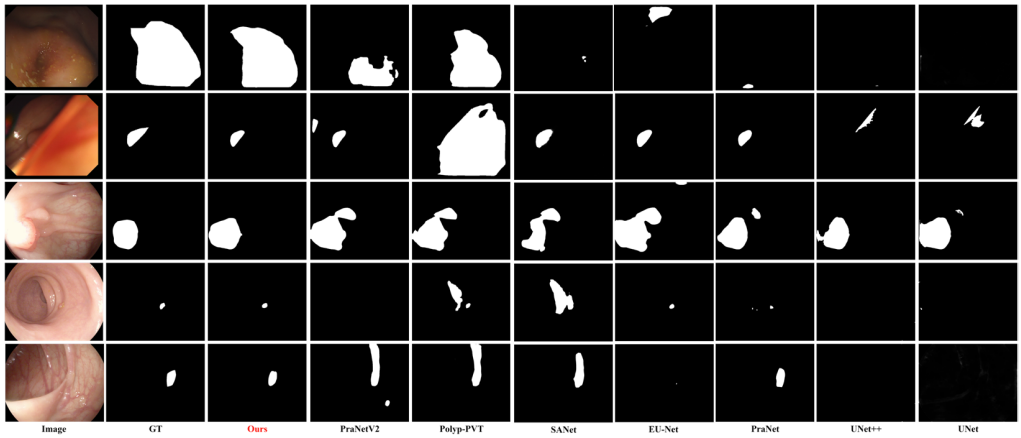

Figure 5. Comparison of experimental results.

As observed in the visual comparisons in Figure 5, early CNN-based models, including U-Net and UNet++, occasionally fail to completely locate the lesions, resulting in under-segmentation. Advanced models like PraNet and Polyp-PVT exhibit improved localization but can still be affected by complex background interference; for example, in the second row, Polyp-PVT and SANet segment areas of normal intestinal folds as foreground targets, resulting in false positives. In comparison, the proposed DSF-BRNet generates segmentation masks that more closely align with the ground truth. Guided by the HBR module, the contours generated by DSF-BRNet remain continuous and clear, mitigating the boundary ambiguity observed in other methods.

4.7. Ablation Experiments

To validate the effectiveness of the core components in the proposed DSF-BRNet, an ablation study was conducted on the Kvasir-SEG and ETIS-LaribPolypDB datasets, with quantitative results summarized in Table 6.

Table 6. Ablation results on the Kvasir-SEG and ETIS-LaribPolypDB datasets.

Method	Baseline	HBR	DSF	Kvasir-SEG		ETIS-LaribPolypDB	
				mDice	mIoU	mDice	mIoU
1	√			0.910	0.856	0.759	0.668
2	√	√		0.916	0.865	0.762	0.687
3	√		√	0.921	0.871	0.781	0.702
4	√	√	√	**0.933**	**0.885**	**0.818**	**0.744**

The baseline model achieved an mDice of 0.759 on the challenging ETIS-LaribPolypDB dataset. Adding only the HBR module yielded a marginal improvement to 0.762, suggesting that boundary refinement alone is less effective without reliable semantic alignment, since high-frequency cues may also respond to surrounding mucosal structures. In contrast, incorporating only the DSF module improved the mDice to 0.781, indicating that cross-level semantic alignment plays a more critical role in establishing discriminative feature

representations. When DSF and HBR were jointly employed, the model achieved the best performance with an mDice of 0.818. These results support the effectiveness of the proposed align-then-refine framework, where DSF provides a reliable semantic basis for subsequent boundary refinement.

To further investigate the internal design of the proposed DSF module, three representative variants were additionally evaluated on the ETIS-LaribPolypDB dataset: channel-only gating, spatial-only gating, and the full dual-gated DSF.

As shown in Table 7, both the channel-only and spatial-only variants improved performance, which indicates that channel recalibration and spatial filtering are both helpful for cross-level feature fusion. The full DSF achieved the best result, showing that these two gating operations are complementary rather than redundant.

Table 7. Additional ablation analysis of the DSF gating design on ETIS-LaribPolypDB.

Variant	ETIS-LaribPolypDB	
	mDice	**mIoU**
DSF-Channel	0.776	0.695
DSF-Spatial	0.775	0.690
DSF-Full	0.781	0.702

4.8. Visualization Analysis of the DSF Module

To provide more intuitive evidence for the role of the proposed DSF module, feature representations before and after DSF were visualized on several representative samples.

As shown in Figure 6, before DSF refinement, the fused features still contained noticeable background responses and structural interference from surrounding mucosal patterns. After DSF processing, the activations became more concentrated on the lesion regions, while irrelevant responses were reduced. These observations suggest that DSF helps improve cross-level semantic alignment and enhances the discriminability of fused feature representations.

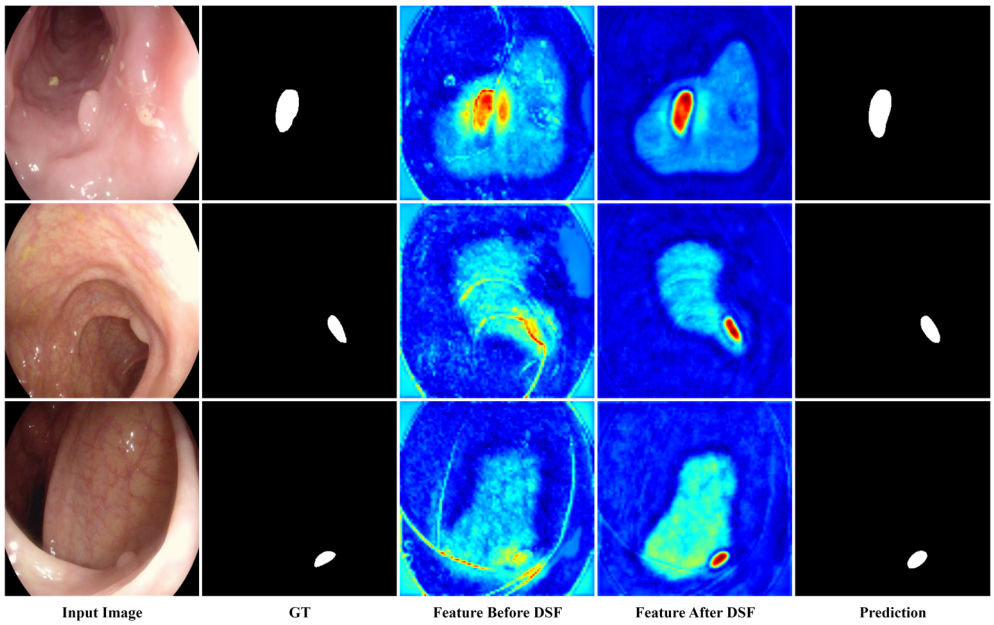

Figure 6. Visualization of feature representations before and after DSF.

4.9. Failure Cases and Limitations

Although favorable overall performance is achieved, several challenging cases remain difficult to handle. As shown in Figure 7, these failure cases are often associated with specular highlights, bubbles or debris, and visual adhesion to surrounding mucosal structures. Under such conditions, false-positive responses, boundary deviation, or over-segmentation may still occur. These observations indicate that highly ambiguous local appearance and strong structural interference remain challenging for the proposed framework.

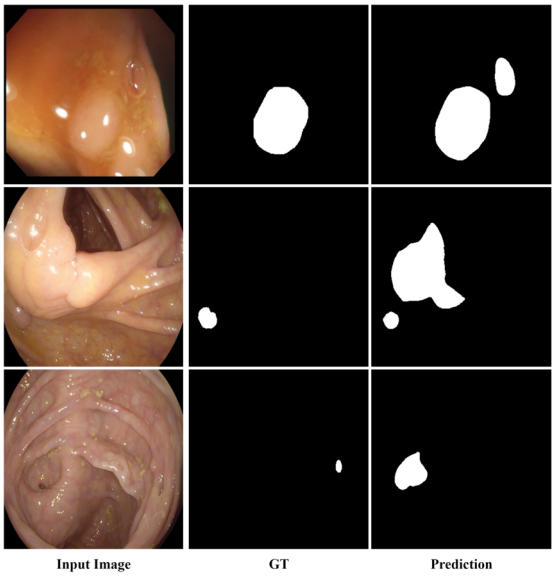

Figure 7. Representative failure cases of the proposed method.

5. Conclusions

A novel deep learning framework, termed DSF-BRNet, is presented for endoscopic polyp segmentation. To address complex intestinal backgrounds and boundary ambiguity, an Align-then-Refine mechanism is introduced so that semantic alignment and boundary refinement can be handled in a progressive manner.

Under this strategy, robust semantic localization is performed before precise structural delineation. The Dual-Gated Semantic Fusion module reduces information loss by avoiding dimensionality reduction and by progressively correcting spatial misalignment between cross-level features, thereby establishing a cleaner semantic foundation. The High-Frequency Boundary Refinement module then uses explicit structural gradients to refine polyp contours without relying on implicit foreground erasure.

Extensive experiments on four public benchmark datasets, including Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-LaribPolypDB, demonstrate competitive accuracy with favorable computational efficiency. With a real-time inference speed of 80.08 FPS and 25.55 M parameters, a favorable balance between segmentation accuracy and computational cost is achieved. These findings indicate that the Align-then-Refine paradigm may be useful for efficient polyp segmentation.

Given the favorable inference speed of DSF-BRNet, extension of the framework to video polyp segmentation will be investigated in future work. Evaluation on more diverse clinical data and further exploration of practical applicability in real-world endoscopic settings will also be pursued.

Author Contributions: Conceptualization, B.L. and C.S.; formal analysis, M.Z.; methodology, C.S.; resources, B.L.; software, C.S.; supervision, B.L. and M.Z.; validation, M.Z. and C.S.; writing–original

draft, B.L. and C.S.; writing–review and editing, B.L. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Fund of Hubei Province, China, grant numbers 2026AFB713 and 2026AFB686.

Data Availability Statement: The original contributions presented in this study are included in the article. The code used in this study will be made publicly available upon acceptance. Further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

DSF-BRNet	Dual-Gated Semantic Fusion and Boundary Refinement Network
DSF	Dual-Gated Semantic Fusion
GT	Ground Truth
HBR	High-Frequency Boundary Refinement
mDice	Mean Dice Similarity Coefficient
mIoU	Mean Intersection over Union

References

1. Siegel, R.L.; Miller, K.D.; Wagle, N.S.; Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **2023**, *73*, 17–48. [CrossRef] [PubMed]
2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
3. Zauber, A.G.; Winawer, S.J.; O'Brien, M.J.; Lansdorp-Vogelaar, I.; van Ballegooijen, M.; Hankey, B.F.; Shi, W.; Bond, J.H.; Schapiro, M.; Panish, J.F.; et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *N. Engl. J. Med.* **2012**, *366*, 687–696. [CrossRef] [PubMed]
4. Winawer, S.J.; Zauber, A.G.; Ho, M.N.; O'Brien, M.J.; Gottlieb, L.S.; Sternberg, S.S.; Waye, J.D.; Schapiro, M.; Bond, J.H.; Panish, J.F.; et al. Prevention of colorectal cancer by colonoscopic polypectomy. *N. Engl. J. Med.* **1993**, *329*, 1977–1981. [CrossRef] [PubMed]
5. Ahn, S.B.; Han, D.S.; Bae, J.H.; Byun, T.J.; Kim, J.P.; Eun, C.S. The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. *Gut Liver* **2012**, *6*, 64–70. [CrossRef] [PubMed]
6. Le Clercq, C.M.; Bouwens, M.W.; Rondagh, E.N.; Bakker, C.M.; Keulen, E.T.; Koek, G.H.; Sanduleanu, S. Postcolonoscopy colorectal cancers are preventable: A population-based study. *Gut* **2014**, *63*, 957–963. [CrossRef] [PubMed]
7. Nanni, L.; Fantozzi, C.; Loreggia, A.; Lumini, A. Ensembles of convolutional neural networks and transformers for polyp segmentation. *Sensors* **2023**, *23*, 4688. [CrossRef] [PubMed]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
9. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11.
10. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel reverse attention network for polyp segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Lima, Peru, 4–8 October 2020; pp. 263–273.
11. Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp segmentation with pyramid vision transformers. *CAAI Artif. Intell. Res.* **2023**, *2*, 9150015. [CrossRef]
12. Ji, G.P.; Chou, Y.C.; Fan, D.P.; Chen, G.; Fu, H.; Jha, D.; Shao, L. Progressively normalized self-attention network for video polyp segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 27 September–1 October 2021; pp. 142–152.
13. Zhang, R.; Lai, G.; Huang, X.; Li, S. Boundary-guided context-aware network for polyp segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 27 September–1 October 2021; pp. 455–464.

14. Qing, W.; Ouyang, Y.; Yin, P. Research on polyp segmentation via dynamic multi-scale feature fusion and global–local semantic enhancement. *Sensors* **2025**, *25*, 6495. [CrossRef] [PubMed]
15. Yun, H.; Du, Q.; Han, Z.; Li, M.; Yang, L.; Liu, X.; Wang, C.; Ma, W. SGNet: A structure-guided network with dual-domain boundary enhancement and semantic fusion for skin lesion segmentation. *Sensors* **2025**, *25*, 4652. [CrossRef] [PubMed]
16. Litjens, G.; Kooi, T.; Bejnordi, E.B.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
19. Wei, J.; Hu, Y.; Zhang, R.; Li, Z.; Zhou, S.K.; Cui, S. Shallow attention network for polyp segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 27 September–1 October 2021; pp. 699–708.
20. Qiu, W.; Yang, X.; Liu, Z.; Qiu, C. Advancing real-time polyp detection in colonoscopy imaging: An anchor-free deep learning framework with adaptive multi-scale perception. *Sensors* **2025**, *25*, 7524. [CrossRef] [PubMed]
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
24. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U-2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
25. Yuan, Y.; Cheng, Y.; Pan, B.; Jin, G.; Yu, D.; Ye, M.; Zhang, Q. A Multi-Modal Attention Fusion Framework for Road Connectivity Enhancement in Remote Sensing Imagery. *Mathematics* **2025**, *13*, 3266. [CrossRef]
26. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVTv2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
28. Li, X.; Huo, H.; Zhang, C. DCANet: Disentanglement and category-aware aggregation for medical image segmentation. *Sensors* **2026**, *26*, 1300. [CrossRef] [PubMed]
29. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
30. Margolin, R.; Zelnik-Manor, L.; Tal, A. How to evaluate foreground maps? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 248–255.
31. Fan, D.P.; Cheng, M.M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4548–4557.
32. Fan, D.P.; Ji, G.P.; Cheng, M.M.; Shao, L. Enhanced-alignment measure for binary foreground map evaluation. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 698–704.
33. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-SEG: A segmented polyp dataset. In Proceedings of the International Conference on Multimedia Modeling (MMM), Daejeon, Republic of Korea, 5–7 January 2020; pp. 451–462.
34. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [CrossRef] [PubMed]
35. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **2015**, *35*, 130–144. [CrossRef] [PubMed]
36. Silva, J.; 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*, 283–293. [CrossRef] [PubMed]
37. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.

38. Patel, K.; Bur, A.M.; Wang, G. Enhanced U-Net: A Feature Enhancement Network for Polyp Segmentation. In Proceedings of the 2021 18th Conference on Robots and Vision (CRV), Burnaby, BC, Canada, 26–28 May 2021; pp. 181–188.
39. Hu, B.C.; Ji, G.P.; Shao, D.; Fan, D.P. PraNet-V2: Dual-supervised reverse attention for medical image segmentation. *arXiv* **2025**, arXiv:2504.10986. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.