



DEGF-Net: Dual-Encoder Global–Local Joint Feature Aggregation Network for colorectal polyp segmentation

He Yu ^{a,c}, Jinming Guo ^b, Xiaorui Cao ^b, Xuan Li ^b, Yulong Li ^a, Jianying Hao ^a,
Haoming Yang ^a, Xi Zhou ^{d,e}, Tao Huang ^{d,e}*

^a College of Computer Science and Technology, Changchun University, Changchun, 130022, PR China

^b College of Electronic and Information Engineering, Changchun University, Changchun, 130022, PR China

^c Key Laboratory of Intelligent Rehabilitation and Barrier-free for the Disabled Ministry of Education, Changchun University, Changchun, 130022, PR China

^d College of Science and Engineering, James Cook University, Cairns, 4878, Queensland, Australia

^e Centre for AI and Data Science Innovation, James Cook University, Cairns, 4878, Queensland, Australia

ARTICLE INFO

Keywords:

Polyp segmentation

Feature fusion

Medical image processing

Swin transformer

Convolutional neural network

ABSTRACT

Accurate segmentation of colorectal polyps in colonoscopy images is crucial for early prevention and computer-aided diagnosis of colorectal cancer, yet large variations in polyp appearance, low polyp-mucosa contrast, and device-related imaging discrepancies still hinder robust performance, especially for small and flat lesions and cross-dataset generalization. To address these challenges, we propose a Dual-Encoder Global–Local Joint Feature Aggregation Network (DEGF-Net) that enhances feature fusion and improves generalization. DEGF-Net adopts a dual-encoder architecture that separately models long-range global context and fine-grained local textures. A Global Joint Feature Fusion Module (GFFM) employs global attention to align and aggregate high-level features from both branches into a unified representation, while an Upper-Lower Level Feature Fusion Module (UL-FM) performs residual multi-scale cross-layer fusion in the decoder to narrow the semantic gap between high-level semantics and low-level details and refine polyp boundaries. In addition, a multi-output hybrid loss is applied to the final and intermediate predictions to leverage deep supervision, accelerate convergence, and improve robustness. Experiments on two benchmark colonoscopy datasets, Kvasir-SEG and CVC-ClinicDB, show that under a unified setting, DEGF-Net achieves mean Dice scores of 0.933 and 0.958, respectively, surpassing recent CNN-based, Transformer-based, and hybrid architectures and exhibiting strong cross-dataset generalization. These results indicate that DEGF-Net can substantially improve automatic polyp segmentation and provide a promising technical basis for computer-aided colorectal cancer screening.

1. Introduction

Colorectal cancer (CRC) remains a major public health burden [1], with considerable mortality driven by delayed diagnosis owing to its often asymptomatic early course. Most CRCs originate from colorectal polyps, prevalent yet treatable precancerous lesions, whose timely detection and removal can interrupt malignant progression. Population-wide screening can reduce morbidity and mortality [2], and colonoscopy remains a highly effective preventive modality. In this context, accurate segmentation of polyps in colonoscopic images is crucial for reliable computer-assisted detection and improved screening outcomes. Nevertheless, this task is complicated by intrinsic data characteristics, including variability in polyp shape and color, the difficulty

of detecting small or flat lesions, and the low color and texture contrast between polyps and surrounding mucosa, which often results in poorly defined boundaries.

Traditional polyp segmentation methods often struggle to achieve high segmentation accuracy since they primarily rely on shallow features, such as basic linear iterative clustering [3], texture [4], and shape [5]. By contrast, deep learning has substantially improved feature representation and driven significant progress in colorectal polyp segmentation. For example, U-Net [6] established the canonical encoder–decoder paradigm based on Convolutional Neural Networks (CNNs) [7]. The encoder extracts hierarchical features, while the decoder restores spatial resolution via transposed convolution (deconvolution),

* Correspondence to: Room B1002, Zhi Xing Building, No.6543 Weixing Road, Changchun University, Changchun City, 130022, China.

** Correspondence to: A2.217, Cairns Campus, James Cook University, 14-88 McGregor Road, Smithfield, QLD 4878, Australia.

E-mail addresses: yuh82@ccu.edu.cn (H. Yu), 230402200@mails.ccu.edu.cn (J. Guo), 210401133@mails.ccu.edu.cn (X. Cao), 230402207@mails.ccu.edu.cn (X. Li), 230702310@mails.ccu.edu.cn (Y. Li), 241502532@mails.ccu.edu.cn (J. Hao), 241501516@mails.ccu.edu.cn (H. Yang), xi.zhou@jcu.edu.au (X. Zhou), tao.huang1@jcu.edu.au (T. Huang).

<https://doi.org/10.1016/j.bspc.2026.110023>

Received 21 August 2025; Received in revised form 29 January 2026; Accepted 1 March 2026

Available online 2 March 2026

1746-8094/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

aligning output and input dimensions for precise, pixel-level predictions. Numerous U-Net derivatives, such as UNet++ [8], nnU-Net [6], and Res-UNet [9], have achieved further performance gains in polyp segmentation. Despite the success of CNN-based architectures, their reliance on localized convolutional operations [10,11] inherently limits their ability to capture global contextual relationships and model long-range dependencies, which are critical for accurately segmenting complex polyp structures.

In contrast to convolutional operations, Transformers can model global dependencies, thereby capturing long-range contextual relationships [12] and achieving state-of-the-art performance in image segmentation tasks [13,14]. They offer notable advantages, including enhanced scalability [15] and resilience to performance degradation [10]. In medical imaging, Transformer-based architectures have demonstrated impressive success across a wide range of tasks [16,17]. Nonetheless, despite their strength in capturing global context, Transformers tend to lose fine spatial detail when transforming image data into one-dimensional token sequences [18]. Moreover, their strong performance often depends on the availability of extensive, annotated datasets, which is challenging to achieve in many medical imaging applications, where data scarcity remains a significant constraint. Against this background, hybrid CNN-Transformer architectures have gradually become an important development trend in current polyp segmentation. TransFuse [19] employs parallel CNN and Transformer branches and uses BiFusion modules for feature fusion at multiple scales, thereby preserving shallow spatial details while fully incorporating global context, and has achieved excellent performance on medical segmentation tasks such as polyp and skin lesion segmentation. CSwinDoubleU-Net [20] combines the CSwin Transformer with the U-Net architecture, enhancing segmentation accuracy in complex scenarios through a dual-encoder-dual-decoder design and cross-scale attention. These hybrid architectures clearly demonstrate that combining CNNs and Transformers in parallel or cascaded fashions is a key direction for improving performance in current polyp segmentation.

From the perspective of feature fusion, existing methods still exhibit the following common limitations: Most methods perform CNN and Transformer feature concatenation or BiFusion independently at each scale, lacking global-level joint modeling and alignment over the “entire high-level semantic space”, which can easily lead to discrepancies between the two branches in their understanding of complex background regions. In multi-scale fusion strategies, the decoding stage is typically based on simple top-down aggregation or skip connections, without an explicit design for cross-layer residual interaction to bridge the semantic gap between high-level semantics and low-level details. As a result, boundary delineation for small, flat, and low-contrast polyps still has substantial room for improvement. To tackle these issues, this paper aims to build a polyp segmentation network that simultaneously accounts for global semantics and local details and exhibits strong generalization capability. We propose a novel framework for colorectal polyp segmentation, termed Dual Encoder Global Joint Feature Network (DEGF-Net). Its core idea is to synergistically fuse the complementary strengths of CNNs and Transformers and to systematically bridge discrepancies between global and local representations, as well as between multi-scale semantics, through explicit cross-branch global fusion and cross-scale residual fusion. The main contributions of this work can be summarized as follows:

- A global-local dual-encoder segmentation framework (DEGF-Net) is proposed, which integrates a CNN encoder and a Swin Transformer encoder in parallel to jointly capture fine-grained local textures and long-range global context, addressing the inherent limitations of single-encoder CNN or Transformer-based architectures.
- A Global Joint Feature Fusion Module (GFFM) is introduced to explicitly align and aggregate high-level features from the CNN and Transformer branches via shared attention weighting, enabling global semantic consistency and reducing cross-branch feature discrepancies under complex and low-contrast clinical backgrounds.

- An Upper-Lower Level Feature Fusion Module (UL-FM) is designed to perform residual multi-scale cross-layer fusion in the decoder, effectively bridging the semantic gap between high-level semantics and low-level spatial details, leading to improved boundary delineation for small, flat, and irregular polyps.
- Comprehensive experiments on Kvasir-SEG and CVC-ClinicDB demonstrate Dice scores of 0.933 and 0.958, respectively. Additional evaluations on the DRIVE retinal dataset and a public cell segmentation dataset further verify the strong cross-domain generalization of the proposed method.

The remainder of this paper is organized as follows. Section 2 reviews the current research progress on colorectal polyp segmentation from the perspective of different network structures. Section 3 details the overall architecture of DEGF-Net, including the dual-encoder design and the structures of the two feature-fusion modules. Section 4 presents the datasets, experimental settings, evaluation metrics, and comparative results, and analyzes the performance and generalization of the proposed model. Finally, Section 6 concludes the paper, discusses future research directions, and emphasizes the clinical significance of deep-learning-based colorectal polyp segmentation for promoting early cancer diagnosis and reducing misdiagnosis and missed diagnosis.

2. Related work

For medical image segmentation tasks involving complex structures such as colorectal polyps, it is highly challenging to simultaneously account for local detail modeling and global context modeling, and to fuse multi-scale features effectively. This section focuses on the specific task of colorectal polyp segmentation and reviews existing deep learning-based methods. Based on their network architectures, related work can be roughly categorized into three groups: CNN-based methods, Transformer-based methods, and CNN-Transformer hybrid architectures.

2.1. CNN-based methods for colorectal polyp segmentation

Early automatic segmentation methods for colorectal polyps are primarily based on classical fully convolutional networks and encoder-decoder architectures. The U-Net [6] model proposed by Ronneberger et al. employs a symmetric encoder-decoder convolutional architecture, where a contracting path extracts contextual information and an expansive path enables precise localization. By introducing skip connections and strong data augmentation, U-Net has achieved milestone results in medical image segmentation and has been widely adopted as a baseline network for polyp segmentation tasks. Building on the U-Net framework, UNet++ [8] introduces densely nested skip connections and multi-scale deep supervision, significantly reducing the semantic gap between the encoder and decoder and improving segmentation performance on complex structures.

Badrinarayanan et al. proposed SegNet [21], a typical convolutional encoder-decoder network for semantic segmentation. By using pooling indices for upsampling during the decoding stage, SegNet preserves object boundaries and fine-grained structures while maintaining computational and memory efficiency. This architecture has been widely adopted by subsequent medical image segmentation methods. PolypSegNet [22] builds on the SegNet encoder-decoder framework and introduces sequential deep dilated Inception modules and deeply fused skip modules to aggregate contextual information at multiple scales and narrow the semantic gap. On several colonoscopy datasets, it achieves higher Dice/IoU scores than U-Net and SegNet, and shows stronger robustness in scenarios involving small and low-contrast polyps, thereby providing an important structural benchmark for subsequent work. This class of methods generally relies on convolutions with fixed receptive fields and thus remains limited in modeling long-range contextual relationships and complexly morphed polyps. When

dealing with small or flat polyps, multiple downsampling operations can easily overwhelm or miss the target features. Moreover, in scenarios with complex backgrounds and blurred boundaries, relying solely on local convolutions and simple skip connections often fails to reconstruct complete and sharp polyp contours.

2.2. Transformer-based methods for colorectal polyp segmentation

Given the success of Vision Transformers (ViT) [23] and their variants in natural image tasks, an increasing number of studies have sought to incorporate Transformer-based backbones into colorectal polyp segmentation to better model long-range dependencies and global context. Compared with traditional networks that rely purely on convolutions, these methods typically use a Transformer as the primary feature encoder and add dedicated multi-scale feature-fusion and boundary-refinement modules on top to address challenges such as diverse polyp morphologies and blurred boundaries.

Polyp-PVT [24], proposed by Dong et al. is one of the earlier works that systematically introduced the Pyramid Vision Transformer (PVT) into polyp segmentation. This method uses PVT as a single encoder backbone, exploits its hierarchical Transformer structure to extract global semantic features at multiple scales, and designs a Cascaded Fusion Module (CFM), a Camouflage Identification Module (CIM), and a Similarity Aggregation Module (SAM) to aggregate high-level semantics and positional information, enhance responses in low-level camouflaged polyp regions, and propagate similarity features across layers, respectively. Polyp-PVT outperforms many traditional CNN-based methods on five public datasets, validating the effectiveness of Transformer backbones in complex polyp scenarios. ColonFormer [25], proposed by Nguyen et al. further explores the use of efficient Transformers for polyp segmentation. The model uses a lightweight Transformer encoder that explicitly captures long-range semantic dependencies across multiple scales, and a hierarchical decoder to learn multi-level feature representations. It also incorporates refinement modules and improved skip-connection strategies to refine boundaries. Experiments on several datasets demonstrate that ColonFormer outperforms representative contemporaneous methods on multiple metrics, indicating that a pure Transformer backbone, without an additional CNN encoder branch, can also achieve high accuracy in polyp segmentation. In their study on early detection of colorectal cancer, Ayana et al. [26] proposed an overall framework based on Vision Transformers and designed a polyp segmentation subnetwork, PUTS, within it. PUTS is built upon Swin-Unet, a pure Transformer-based encoder-decoder architecture, and introduces a spatial Transformer localization module that explicitly relocates input images to highlight potential lesion regions, after which Swin-Unet performs fine-grained segmentation. On the Kvasir-SEG [27] and CVC-ClinicDB [28] datasets, PUTS achieves mIoU scores of 0.8673 and 0.9092, respectively, which are significantly higher than those of many classical CNN and hybrid models. This indicates that a Transformer framework combined with spatial relocation can effectively enhance the saliency and segmentation quality of polyp regions under complex views and imaging conditions.

RSAFormer [29], proposed by Yin et al. is another polyp segmentation method that employs a Transformer backbone as the main encoder. The network uses a Transformer encoder to capture global semantic information and introduces a Region Self-Attention (RSA) module to enhance features specifically in regions with blurred boundaries and highly similar tissue structures. It further employs a dual-decoder structure to generate diverse feature maps and coarse segmentation maps, which are then fed back into the RSA module to reinforce region-level attention. Across multiple polyp datasets, RSAFormer outperforms various existing methods in Dice and mIoU, indicating that combining region-level self-attention with a Transformer backbone improves modeling of complex boundaries.

Overall, this category of polyp segmentation methods, which uses Transformers as the sole backbone encoder, mainly improves segmentation performance by explicitly modeling long-range dependencies and global semantic relationships at multiple scales, combined with multi-scale feature fusion and region-level attention mechanisms. These designs enhance the characterization of small, camouflaged, and boundary-blurred polyps. However, these methods typically rely on a single Transformer encoder pathway and still have certain limitations in explicitly aligning “global Transformer semantics” with “local spatial details”. For example, although Polyp-PVT, ColonFormer, and PVT-MA incorporate specialized multi-scale fusion designs, they lack explicit complementary mechanisms with an independent CNN branch, and local texture and fine boundary information may still be insufficient under complex backgrounds and extreme morphological variations.

2.3. CNN-Transformer hybrid architectures for colorectal polyp segmentation

To jointly leverage the strengths of CNNs for modeling local, fine-grained features and Transformers for capturing global contextual dependencies, numerous CNN-Transformer hybrid networks have been proposed for polyp segmentation in recent years, leading to a variety of dual-branch or dual-encoder architectures with multi-scale feature fusion.

TransEFusionNet [19] employs Transformer and CNN branches as parallel encoding paths and introduces a Spatial Feature Fusion Module (SFM) and an Edge Feature Fusion Module (EFM) to collaboratively fuse deep and shallow, edge and region information in the decoding stage. It outperforms single-CNN or single-Transformer architectures on several polyp datasets. CSwinDoubleU-Net [20] combines convolutions with the CSwin Transformer into a dual U-shaped structure and enhances modeling capacity for complex polyp morphologies through multi-level, multi-branch feature interaction, achieving high Dice and mIoU scores in experiments. The Ens2 model [30] integrates multiple CNN and Transformer segmenters and achieves stable, robust overall performance across several datasets, indicating that multi-model ensembles can improve generalization to some extent under complex noise and varying imaging conditions. Building on this line of work, Cao et al. proposed DEMF-Net [31], which combines Swin Transformer and CNN into a dual-branch encoder structure through a dual-encoder multi-scale feature fusion design, and introduces multi-scale fusion modules to alleviate cross-layer semantic discrepancies. DEMF-Net achieves better segmentation performance than many CNN and Transformer baselines on several polyp datasets.

Collectively, these studies show that dual-encoder or dual-branch structures, combined with feature fusion modules, have become an important avenue of development in current polyp segmentation research. Designing more effective cross-branch and multi-scale feature fusion mechanisms that fully leverage the complementary strengths of CNNs and Transformers while maintaining model efficiency remains a key factor shaping both performance and generalization. Despite these efforts, several research gaps remain. Most methods emphasize either the global context modeled by the Transformer branch or the local textures captured by the CNN branch, but lack a global feature fusion mechanism tailored to polyp segmentation that explicitly aligns and jointly encodes semantic and spatial positional information. Meanwhile, the semantic gap between high- and low-level multi-scale features remains unbridged. Current cross-layer fusion strategies mostly rely on simple addition or concatenation, ignoring differences in semantic abstraction levels and noise distributions across layers, which can easily introduce information conflicts or redundancy when deep semantics are propagated to shallow layers.

To address these shortcomings, the proposed DEGF-Net builds upon a dual-encoder architecture consisting of Swin Transformer and CNN, and introduces a Global Joint Feature Fusion Module (GFFM) and an Upper-Lower Level Feature Fusion Module (UL-FM) to explicitly realize

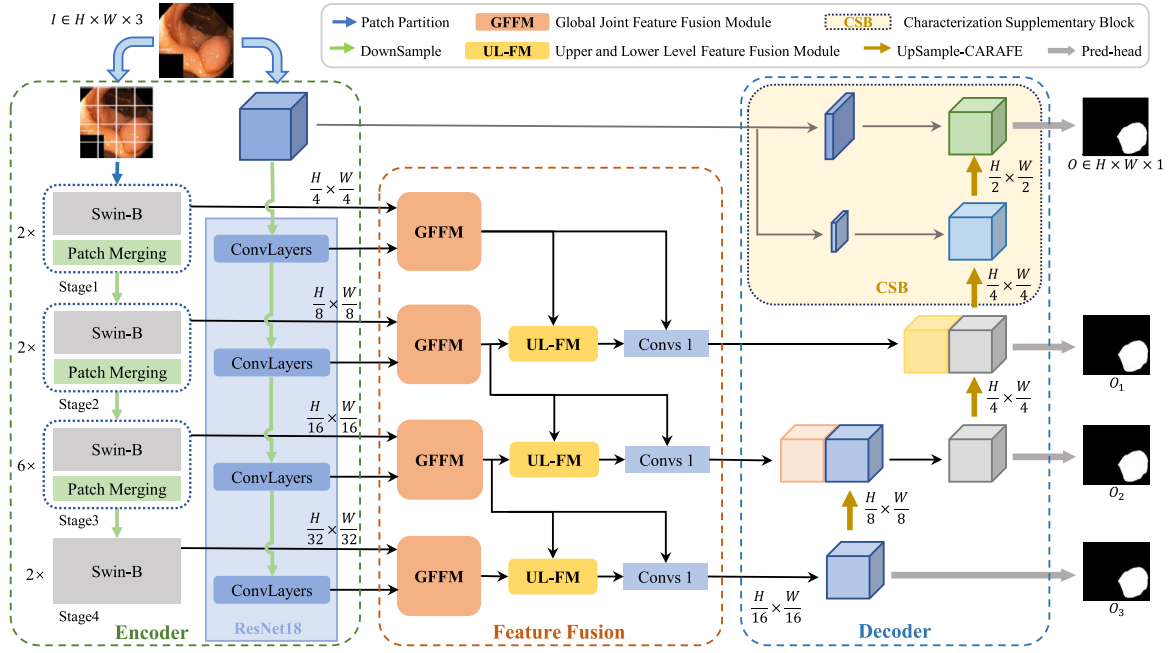


Fig. 1. DEGF-Net overall architecture.

joint modeling and interaction between global and local information, as well as between multi-level semantics and boundary cues. GFFM shares attention weights across branches to align and fuse high-level features from the two encoders, thereby strengthening unified representations that integrate global semantics and local structures. UL-FM employs a multi-scale cross-layer selection mechanism to effectively mitigate the semantic gap between high- and low-level features, suppress noise interference, and preserve critical boundary information. Experimental results show that this design achieves better segmentation performance and stronger generalization than existing methods on multiple polyp segmentation datasets and other medical imaging datasets. In this way, it partially remedies the limitations of current pure Transformer-backbone methods in local detail modeling and cross-layer information alignment, and effectively addresses the aforementioned key gaps in current research (see Table 1).

3. Methods

3.1. Network architecture

This study proposes a dual-encoder-based global feature fusion method for colorectal polyp segmentation, as depicted in Fig. 1. This architecture consists of four core components: the encoder, the decoder, the Global Joint Feature Fusion Module (GFFM), and the Upper and Lower Level Feature Fusion Module (UL-FM). The encoder capitalizes on a dual-path design that merges Swin Transformer and CNN branches to jointly encode global contextual cues and fine-grained spatial details. In the decoder, features at both high and low levels are jointly reconstructed to preserve essential semantic information and restore the output segmentation to the original image dimensions. The GFFM merges features extracted by the dual encoder to enhance semantic and positional understanding. The UL-FM fuses multi-scale features from various layers to better capture structural details. Finally, a feature-supplementation block is introduced to recover edge information that may be lost during up-sampling, thereby improving the model's representativeness.

Given a colorectal polyp image $I \in \mathbb{R}^{H \times W \times 3}$, the input is processed simultaneously through both the Swin Transformer and CNN channels. As shown in Fig. 3, the image is processed through four layers of dual encoders, each of which is a dual encoder, to obtain global and local

information to generate a feature map with dimensions $\mathbb{R}^{\frac{H}{2^{(i+1)}} \times \frac{W}{2^{(i+1)}}}$, where $i \in \{1, 2, 3, 4\}$. The encoded features are subsequently passed into the Global Joint Feature Fusion Module (GFFM), where features are fused while maintaining dimensional consistency with the input at each hierarchical level. The fused feature maps from both lower and upper levels are subsequently passed to the UL-FM, facilitating multi-scale integration across hierarchical layers. These integrated features are then consolidated via a convolutional layer before being forwarded to the decoder. Within the decoder, layer-wise concatenation is performed, followed by a feature-complementation block, which ultimately produces the final polyp segmentation output. It is particularly worth noting that the CSB (Characterization Supplementary Blocks) in the decoder is a characterization supplementary block, as part of the authors' earlier work. Its structure is described in DEMF-Net [31].

3.2. Dual encoding mechanism

3.2.1. Swin transformer encoder

The Swin Transformer encoder in the proposed method adopts the Swin-B model, a large-scale variant of the Swin Transformer architecture [32]. Compared to Swin-T, Swin-B offers a deeper network architecture and more parameters, resulting in enhanced representational capacity and improved segmentation performance. The specific parameter configurations are outlined in Table 2.

3.2.2. CNN encoder

To balance network depth, ResNet-18 is employed as the CNN encoder to model the local spatial characteristics of polyp images. This branch comprises five components: the pre-downsampling module, ConvLayers1, ConvLayers2, ConvLayers3, and ConvLayers4.

The input $I \in \mathbb{R}^{H \times W \times 3}$ is first processed by ResNet-18 and passed through a pre-downsampling module. In this module, two consecutive convolutional blocks, each with a 3×3 kernel and a stride of 2, are applied to progressively reduce the spatial resolution. This step not only reduces the spatial resolution but also prepares the network for residual concatenation during the output stage. The operation is expressed using the following function, which $I' \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 64}$:

$$I' = F(I). \quad (1)$$

Table 1

Summary of colorectal polyp segmentation methods reviewed in Section 2, including year, main evaluation datasets, a representative accuracy, and remaining research gaps or future directions.

Method	Year	Main evaluation datasets	Evaluation metrics	Main research gaps/future directions
CNN-based methods				
U-Net [6] (CNN encoder–decoder)	2015	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, CVC-T, ETIS	Dice 0.812/mIoU 0.722 (Kvasir-SEG)	Classical FCN-style baseline; limited receptive field and weak long-range context modeling, performance degrades on very small/flat and low-contrast polyps, motivating stronger global modeling and multi-scale fusion.
UNet++ [8] (CNN nested U-Net)	2018	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS	Dice 0.724/mIoU 0.633 (Kvasir-SEG)	Dense skip connections narrow the encoder–decoder gap but still rely on local convolutions; boundary sharpness and cross-dataset robustness remain limited in complex scenes.
PolypSegNet [22] (SegNet variant)	2021	CVC-ClinicDB, CVC-ColonDB, ETIS, Kvasir-SEG	Dice 0.881/mIoU 0.809 (Kvasir-SEG)	Dilated Inception and deep skip fusion enhance context, but fixed convolutional receptive fields still limit modeling of very long-range dependencies and extreme morphology; generalization across centers is moderate.
Transformer-based methods				
Polyp-PVT [24] (pure PVT Transformer encoder)	2023	CVC-ClinicDB, Kvasir-SEG, CVC-ColonDB, CVC-T, ETIS	Dice 0.937/mIoU 0.889 (Kvasir-SEG)	Hierarchical PVT backbone captures strong global semantics, while local textures and boundary details remain suboptimal in heavily camouflaged and low-contrast regions due to the absence of an explicit CNN branch.
ColonFormer [25] (pure Transformer, MiT encoder)	2022	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, CVC-T, ETIS	Dice 0.927/mIoU 0.877 (Kvasir-SEG)	Efficient ViT backbone and refined decoder improve performance, but the model remains a single Transformer pathway; complementary CNN-style edge and texture modeling is not explicitly exploited, leaving room for finer boundary control.
PUTS [26] (Swin-Unet based Transformer with spatial localization)	2024	Kvasir-SEG, CVC-ClinicDB	mIoU 0.8673 (Kvasir-SEG), 0.9092 (CVC-ClinicDB)	Combines spatial relocation with a Swin-Unet backbone to enhance saliency, yet mainly focuses on global cues; further work is needed on high-precision boundary refinement and broader cross-dataset generalization.
RSAFormer [29] (region self-attention Transformer)	2024	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS	Dice 0.922 (Kvasir-SEG)	Region self-attention improves blurred-boundary modeling, but there is still no explicit cross-branch alignment between global Transformer semantics and local spatial details, which can limit robustness to severe appearance variations.
CNN-Transformer hybrid methods				
TranSEFusionNet [19] (dual-branch CNN + Transformer)	2023	CVC-ClinicDB, Kvasir-SEG, CVC-ColonDB, ETIS	Dice 0.865/mIoU 0.791 (Kvasir-SEG)	Parallel CNN/Transformer encoders with spatial and edge fusion improve performance, but cross-branch interaction is still relatively shallow; feature fusion is not globally attentive and may introduce redundancy.
CSwinDoubleU-Net [20] (double U-Net with CSwin Transformer)	2024	CVC-ClinicDB, Kvasir-SEG, CVC-ColonDB, ETIS	Dice 0.931/mIoU 0.883 (Kvasir-SEG)	Double U-shaped architecture boosts capacity but increases complexity; multi-branch feature interaction mainly occurs locally, and explicit global–local alignment is still insufficient.
Ens2 [30] (ensemble of CNN and Transformer segmenters)	2023	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB	Dice 0.935/mIoU 0.893 (Kvasir-SEG)	Multi-model ensemble significantly improves robustness and average performance, but at the cost of heavy inference and memory; lacks a unified mechanism for efficient feature-level fusion across models.
DEMF-Net [31] (dual encoder Swin Transformer + CNN)	2024	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS, EndoScene	Dice 0.953/mIoU 0.919 (Kvasir-SEG)	Dual encoders plus DAF/AFFM/CSB enhance multi-scale fusion and edge extraction, yet cross-branch interaction remains relatively coarse; more explicit joint modeling of global semantics and local boundaries is needed, which motivates our DEGF-Net design.

Note: Evaluation Metrics are mainly reported as mean Dice (and mIoU where available) on the Kvasir-SEG and/or CVC-ClinicDB datasets as given in the original papers or in unified comparison tables; readers are referred to the cited works for complete metric tables and experimental settings.

The remaining 16 residual convolutional layers are evenly divided across the four convolutional layers: ConvLayers1 through ConvLayers4. Each block consists of four convolutional layers and an additional residual connection. The 2-times downsampling operation is applied at each stage to ensure that the output dimensions at each level match those produced by the corresponding layers in the Swin Transformer encoder. This alignment guarantees that the features from both encoding paths are compatible for subsequent fusion. The process is

formalized as:

$$R_1 = F(I') + \text{Conv}_{3 \times 3}(I'), \quad (2)$$

$$R_{i+1} = R_i + 4\text{Conv}_{3 \times 3}(R_{i+1}), i \in \{1, 2, 3\}, \quad (3)$$

where $R_1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$, $R_2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$, $R_3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$, $R_4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$.

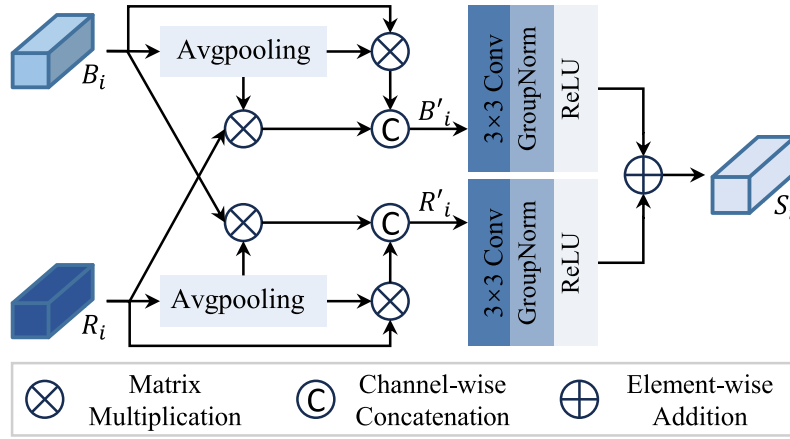


Fig. 2. Modular diagram of Global Joint Feature Fusion Module (GFFM).

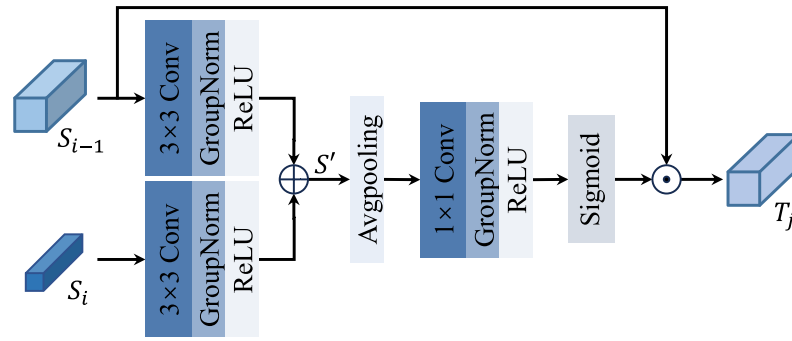


Fig. 3. Detailed schematic of the Upper and Lower Level Feature Fusion Module (UL-FM).

Table 2

Network parameters.

Component	Parameter	
SwinTransformer	Patch_size	4
	Embed_dims	128
	num heads	[4,8,16,32]
	Window size	7
	depths	[2,2,18,2]
	Drop_path_rate	0.5
CNN	Pre-downsampling	[3,64,3,2]
	ConvLayer1	[64,128,3,2]; [128,128,3,1]
	ConvLayer2	[128,256,3,2]; [256,256,3,1]
	ConvLayer3	[256,512,3,2]; [512,512,3,1]
	ConvLayer4	[512,512,3,2]; [512,512,3,1]
GFFM	Layer1	[128,128,3,1]
	Layer2	[256,256,3,1]
	Layer3	[512,512,3,1]
	Layer4	[1024,1024,3,1]
UL-FM	Layer1	[384,128,3,1]
	Layer2	[786,256,3,1]
	Layer3	[1536,512,3,1]
	Convs 1	$3 \times 3, s = 1$

To further leverage temporal information and enhance robustness against feature corruption, we aggregate the cooperative features from consecutive timesteps into a feature sequence:

3.3. Global joint feature fusion module

Although many CNN-Transformer hybrid segmenters combine the same-scale features via simple concatenation/addition (or apply attention independently within each branch), such local and per-scale fusion does not explicitly align the two encoder branches in the high-level

semantic space. Consequently, the CNN and Transformer branches may exhibit inconsistent responses to complex mucosal backgrounds, specular highlights, and low-contrast regions, and naive fusion can propagate these discrepancies to the decoder, undermining boundary accuracy and cross-dataset robustness. To address this gap beyond basic feature fusion, we propose the Global Joint Feature Fusion Module (GFFM), as illustrated in Fig. 2. GFFM introduces an attention-guided cross-branch enhancement mechanism: attention-derived channel weights from one branch are shared to reweight the features of the other branch, enabling explicit cross-branch alignment and complementary aggregation of salient semantic cues while suppressing irrelevant background responses. This design provides a measurable performance gain. As shown in the ablation study (Table 4), incorporating GFFM yields Dice/mIoU/Precision/Recall of 0.917/0.862/0.917/0.931 on Kvasir-SEG, demonstrating that attention-guided cross-branch alignment contributes positively to the overall segmentation performance.

As shown in Fig. 2, the feature B_i , which is output from Swin-B, is subjected to channel attention selection by average pooling, and $H_{\text{avgpool}}(\cdot)$ represents the operation of channel attention selection. The obtained weight matrix $H_{\text{avgpool}}(B_i)$ is multiplied by the features B_i and R_i respectively, and then spliced to obtain the feature B'_i . Similarly, the CNN output feature R_i is averaged-pooled for channel attention selection. The obtained weight matrix $H_{\text{avgpool}}(R_i)$ is multiplied by the features B_i and R_i respectively, and also spliced to obtain the feature R'_i . The features B'_i and R'_i are both passed through a convolution kernel with sizes of 1 and 3×3 , respectively.

As illustrated in Fig. 2, the feature B_i , produced by the Swin-B module, undergoes channel attention via average pooling, denoted by the function $H_{\text{avgpool}}(\cdot)$. The resultant attention weight matrix $H_{\text{avgpool}}(B_i)$ is independently multiplied with features B_i and R_i , respectively. These weighted features are subsequently concatenated to generate the refined feature B'_i . Analogously, the CNN-generated feature R_i is processed through the average pooling-based channel attention, producing

the attention weight matrix $\mathcal{H}_{\text{avgpool}}(R_i)$. This weight matrix is separately applied to features B_i and R_i , and the results are concatenated to obtain the enhanced feature R'_i . Finally, both B'_i and R'_i are further refined by two convolutional operations employing a kernel of size 3×3 . At last, the processed features are merged to produce the output S_i :

$$B'_i = [B_i \times \mathcal{H}_{\text{avgpool}}(B_i); R_i \times \mathcal{H}_{\text{avgpool}}(B_i)], \quad (4)$$

$$R'_i = [R_i \times \mathcal{H}_{\text{avgpool}}(B_i); R_i \times \mathcal{H}_{\text{avgpool}}(R_i)], \quad (5)$$

$$S_i = \text{Conv}_{3 \times 3}(B'_i) + \text{Conv}_{3 \times 3}(R'_i), \quad (6)$$

where $i \in \{1, 2, 3, 4\}$, and $[\cdot]$ denotes channel-wise concatenation.

3.4. Upper and lower level feature fusion module

The complete architecture of the UL-FM is depicted in Fig. 3. As depicted, UL-FM receives two inputs representing features at varying scales. Leveraging multi-scale information is crucial for reducing the semantic discrepancy between the encoder and decoder. Conventional approaches, such as directly adding or multiplying feature maps with different spatial resolutions, often fail to achieve effective fusion. Therefore, UL-FM adopts a multi-scale feature selection strategy to facilitate more meaningful integration, ensuring that features from different levels are effectively aligned and semantically enriched before decoding. Compared with conventional multi-scale feature fusion strategies that simply upsample and aggregate features from different resolutions, UL-FM performs adaptive cross-level selection and re-weighting driven by the joint representation, which explicitly aligns upper- and lower-level semantics and suppresses redundant background responses. This design enables the decoder to exploit more complementary semantic and spatial cues, leading to more accurate reconstruction of small, flat, and low-contrast polyps.

After the GFFM produces S_i , $i \in \{1, 2, 3, 4\}$, the features from the upper and lower levels, which differ in semantic and spatial information, are first processed by a 3×3 convolution with stride one and then summed to form an intermediate representation S' :

$$S' = \text{Conv}_{3 \times 3}(S_i) + \text{Conv}_{3 \times 3}(S_{i-1}). \quad (7)$$

A selector module then derives attention weights from S' via average pooling, $\mathcal{H}_{\text{Avgpool}}(S')$, to preserve informative cross-level semantic cues while suppressing redundant features. To obtain the final fused output T_j (with $j \in \{1, 2, 3\}$), the selected features are first transformed by a 1×1 convolution, passed through a sigmoid activation, and are then multiplied element-wise with the up-level features S_{i-1} to facilitate improved fusion of cross-level features and obtain the final fused output T_j :

$$T_j = S_{i-1} \odot \text{Sigmoid}(\text{Conv}_{1 \times 1}(\mathcal{H}_{\text{Avgpool}}(S'))), \quad (8)$$

where $\odot$ denotes element-wise multiplication.

3.5. Decoder

The decoder section of this paper references the traditional U-Net decoder workflow, which proceeds as follows: abstract features from the bottom-most layer of the encoder undergo upsampling, concatenation with features from corresponding encoder layers, refinement through convolutional blocks, and repetition of this process until the original resolution is restored, as shown in Fig. 1. The CSB block diagram in this paper is simplified. The flat cubes represent supplementary edge detail features within the CSB module. For detailed descriptions, please refer to DEMF-Net [31]. Among the remaining cubes in the decoder section, the bottom blue cube and the orange cube represent the feature-fusion results for each encoder layer. The fusion results from each layer are passed upward, upsampled, and aligned with those from the previous layer (indicated by yellow arrows in Fig. 1). This process yields the feature fusion results for the current layer, represented by the gray cube. The results from the topmost layer will pass through the CSB module to produce the final predicted polyp segmentation result, O .

3.6. Loss function

The loss function employed for training the model is specified as follows:

$$\mathcal{L} = \mathcal{L}_{\text{IoU}}^W + \mathcal{L}_{\text{BCE}}^W, \quad (9)$$

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}(G, O_1) + \beta \mathcal{L}(G, O_2) + \delta \mathcal{L}(G, O_3) + \gamma \mathcal{L}(G, O). \quad (10)$$

This loss function integrates two components, $\mathcal{L}_{\text{IoU}}^W$ and $\mathcal{L}_{\text{BCE}}^W$. Polyp segmentation typically suffers from severe foreground-background class imbalance and large variations in polyp size. In this setting, using a single BCE loss is often dominated by abundant background pixels and may lead to under-segmentation of small or flat polyps, whereas using only an IoU loss primarily constrains region-level overlap and often produces overly smooth boundaries. Therefore, we adopt a hybrid loss to combine the complementary advantages of both terms. The weighted IoU loss $\mathcal{L}_{\text{IoU}}^W$ encourages high region-level agreement between prediction and ground truth and alleviates the effect of class imbalance, while the weighted BCE loss $\mathcal{L}_{\text{BCE}}^W$ provides pixel-wise supervision that preserves fine structures and sharp boundaries, especially around small and low-contrast polyps. As a result, the hybrid loss leads to more accurate and robust polyp segmentation than using either IoU or BCE alone. To optimize the model and enhance its predictive accuracy, a weighted loss function is employed. The hyperparameters α , β , δ , and γ are set to values of 0.1, 0.1, 0.1, and 0.7, respectively. As stated in Eq. (10), O_1 corresponds to the results of the encoder, O_2 refers to the results of the second-to-last layer, O_3 denotes the results of the third-to-last encoder layer, and O represents the final prediction map.

4. Experiments and results

4.1. Colorectal polyp segmentation dataset

To train and evaluate the proposed methodological model, its effectiveness in diagnosing colorectal polyps is assessed to ensure reliable diagnostic support for physicians. Five polyp datasets that are publicly accessible: Kvasir-SEG [27], CVC-ClinicDB [28], CVC-ColonDB [33], EndoScene [34], and ETIS [35] are selected. These datasets typically contain polyps of varying types and sizes, thereby simulating real clinical scenarios. Fig. 4 shows examples of images from the five publicly available polyp datasets described above. The details of each dataset are as follows:

Kvasir-SEG: the dataset represents the largest publicly accessible collection for polyp segmentation. It includes 1000 images of polyps, each accompanied by its corresponding annotation. The resolutions of the images span from 332×487 pixels to 1920×1072 pixels, and the polyp regions exhibit significant variation in size and shape.

CVC-ClinicDB, which was officially utilized during the training phase of the MICCAI 2015 Colonoscopy Video Automated Polyp Detection Sub-Challenge, contains 612 polyp images with annotations. The dataset comprises frames extracted from 29 distinct endoscopic video sequences, each at a spatial resolution of 384×288 pixels.

CVC-ColonDB: the dataset is designed for tasks related to the detection and classification of colorectal polyps. It comprises 380 annotated polyp images, derived from 15 distinct colonoscopy videos. The images feature both normal tissue and polyps, with a resolution of 574×500 pixels each.

EndoScene: the dataset comprises 912 annotated images and is a fusion of CVC-ClinicDB and CVC300. This dataset is frequently used for training and evaluating automated algorithms for polyp detection in colonoscopy images.

ETIS: the dataset consists of 192 polyp images, each annotated, and sourced from 29 different colonoscopy sequences. Every image in the collection has a resolution of 1225×996 pixels.

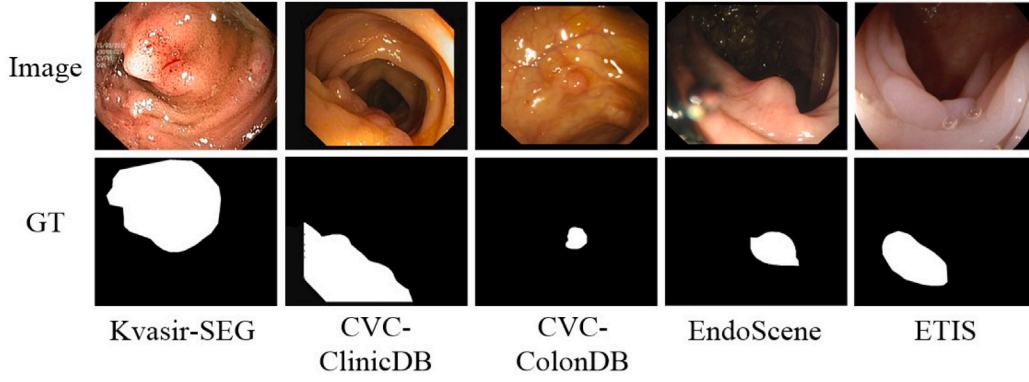


Fig. 4. Five public polyp dataset images.

4.2. Experiment setup

The experiments are implemented in PyTorch, utilizing GPU acceleration provided by an NVIDIA RTX 3090 graphics card. The Swin Transformer is initialized using pre-trained Swin-B weights for the encoder. Instead of conventional data augmentation methods, a multi-scale training strategy [36] is employed. The model parameters are optimized using the Stochastic Gradient Descent (SGD) optimizer. To further enhance training efficiency and performance, an early-stopping mechanism and a cosine annealing learning rate schedule are integrated. The initial learning rate is set at 1×10^{-4} and progressively increases to a peak value of 0.01. The training procedure is run for 60 epochs with a batch size of 4. No additional geometric or color augmentation is used. Multi-scale training serves as the only augmentation strategy, introducing scale variation across iterations to reduce overfitting and improve generalization. For optimization, we employ SGD with momentum and cosine annealing. This choice follows several recent CNN-Transformer hybrid segmentation studies in which SGD has demonstrated stable convergence during multi-scale training. Although adaptive optimizers such as Adam may offer additional improvements, exploring such alternatives is left for future work.

4.3. Evaluation metrics

This study employs four widely used evaluation metrics to assess the performance of medical image segmentation methods: the Dice similarity coefficient (Dice), mean Intersection over Union (mIoU), Precision, and Recall. These metrics are extensively recognized in the literature for their effectiveness in quantifying the accuracy and reliability of segmentation outcomes within medical imaging applications.

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN}, \quad (11)$$

$$\text{mIoU} = \frac{1}{N} \sum_{n=1}^N \frac{TP_n}{TP_n + FP_n + FN_n}, \quad (12)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (13)$$

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (14)$$

Here, Dice for each test image is computed according to Eq. (11), and the final Dice score reported in our experiments is obtained by averaging the Dice values over all test images. TP_n , FP_n , and FN_n denote the numbers of true positives, false positives, and false negatives for the n th test image, respectively, and N is the total number of test images.

The Dice similarity coefficient measures the overlap between two sets and is widely used in image segmentation tasks to evaluate the correspondence between predicted segmentations and their ground truths. The mIoU metric measures the overlap between predicted segmentation

masks and their corresponding ground-truth annotations, providing a comprehensive indication of segmentation quality. Precision quantifies the proportion of positively predicted samples that are correctly identified, thereby reflecting the reliability of the segmentation model's positive predictions. Conversely, recall measures the model's ability to accurately detect true positives, serving as an indicator of sensitivity, particularly relevant to polyp detection accuracy.

4.4. Analysis of learning ability

All experiments are conducted on the public Kvasir-SEG and CVC-ClinicDB datasets. For both datasets, all images and their corresponding masks are uniformly resized to 512×512 pixels. For each dataset, we first randomly select 90% of the images as the development data and reserve the remaining 10% as a held-out test set. In the development data, 10% of the images are reserved as a validation set for hyperparameter tuning and early stopping. In total, this yields 1305 training images (81%), 145 validation images (9%), and 162 test images (10%) across the two datasets. The validation and test sets are strictly disjoint, and all quantitative results are reported on the test set.

The performance of the proposed method for polyp segmentation is evaluated by comparing it with several widely used open-source models using four evaluation metrics. These models include U-Net [6], UNet++ [8], HardNet-MSEG [37], SANet [38], TransUNet [39], and DS-TransUNet [40]. Additional polyp segmentation methods are also included for comparison. The performance of the proposed model is evaluated by comparing it with state-of-the-art approaches on two colorectal polyp image collections: the CVC-ClinicDB and Kvasir-SEG datasets. To enhance clarity, the top-performing results for each metric are highlighted in bold in Table 3. As illustrated in Table 3, the metrics achieved by the DEGF-Net surpass those of classical polyp segmentation approaches, including both traditional CNN-based models and hybrid CNN-Transformer architectures. Specifically, on the CVC-ClinicDB dataset, DEGF-Net outperforms HardNet-MSEG by 2.2%, 3.2%, 0.2%, and 3.2% across the respective metrics. For the Kvasir-SEG dataset, the improvements over DS-TransUNet are 2.0%, 2.7%, 1.7%, and 0.3%, respectively. The results indicate that the proposed method outperforms existing approaches in segmenting colorectal polyps.

Note that the results presented in Table 3 were partially obtained using the original author's publicly available code repository and pre-trained weights. Retraining and testing were conducted under identical software and hardware conditions (e.g., PyTorch and CUDA versions), with some results directly sourced from the original paper. However, our ablation studies during experimentation demonstrate that the performance of our proposed method remains robust across a reasonable range of variations in its core configuration.

Our comparative analysis employs standard segmentation metrics (Dice, mIoU, Precision, Recall) as the primary basis for evaluation, consistent with prior works (e.g., HardNet-MSEG [46], DS-TransUNet [40],

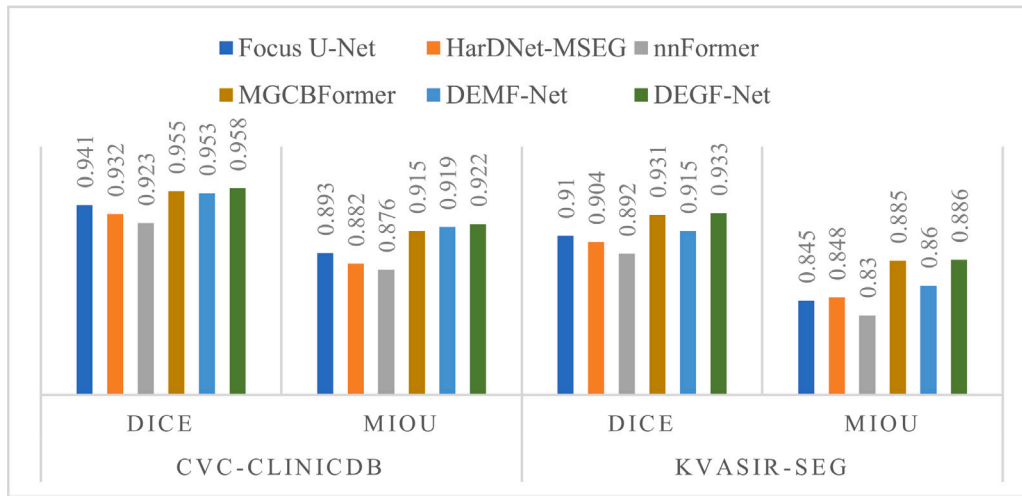


Fig. 5. Performance comparison using the evaluation metrics of Dice and mIoU on CVC-ClinicDB and Kvasir-SEG datasets.

Table 3
Quantitative results of learning ability.

Encoder type	Model	CVC-ClinicDB				Kvasir-SEG			
		Dice	mIoU	Precision	Recall	Dice	mIoU	Precision	Recall
CNN	U-Net [6]	0.832	0.750	0.887	0.851	0.812	0.722	0.855	0.835
	UNet++ [8]	0.715	0.602	0.814	0.773	0.724	0.633	0.836	0.769
	ResUNet++ [41]	0.795	0.796	0.878	0.702	0.813	0.793	0.877	0.706
	Focus U-Net ^a [42]	0.941	0.893	0.956	0.930	0.910	0.845	0.917	0.916
	PraNet [43]	0.899	0.849	0.961	0.912	0.898	0.840	0.903	0.910
	DDANet [44]	0.941	0.893	0.951	0.936	0.857	0.780	0.864	0.888
	PolypSegNet ^a [21]	0.915	0.862	0.961	0.911	0.887	0.826	0.917	0.925
	ACSNet [45]	0.882	0.826	0.923	0.911	0.898	0.838	0.902	0.930
	HardNet-MSEG [46]	0.932	0.882	0.960	0.923	0.904	0.848	0.923	0.907
	SANet [47]	0.899	0.859	0.925	0.914	0.904	0.847	0.916	0.923
Transformer	nnFormer ^a [48]	0.923	0.876	0.924	0.913	0.892	0.830	0.959	0.912
	Polyp-PVT ^a [24]	0.937	0.889	N/A	N/A	0.917	0.864	N/A	N/A
	TransUnet ^a [39]	0.935	0.887	N/A	N/A	0.913	0.856	N/A	N/A
Hybrid	TranSEFusionNet ^a [19]	0.865	0.791	N/A	N/A	0.845	0.781	N/A	N/A
	CSwinDoubleU-Net ^a [20]	0.931	0.883	0.965	0.872	0.910	0.844	0.903	0.854
	DS-TransUnet ^a [40]	0.942	0.894	0.937	0.950	0.913	0.859	0.914	0.936
	MGCBFormer ^a [49]	0.955	0.915	0.963	0.949	0.931	0.885	0.955	0.919
	Ens2 ^a [30]	0.935	0.893	N/A	N/A	0.927	0.883	N/A	N/A
	DEMF-Net [31]	0.953	0.919	0.962	0.953	0.915	0.860	0.913	0.937
	DEGF-Net (Ours)	0.958	0.922	0.962	0.955	0.933	0.886	0.931	0.939

^a The labeling results were directly adopted from the data in the original paper.

Table 4
Ablation study on Kvasir-SEG evaluating the contribution of the proposed modules.

Index	Setting			Kvasir-SEG			
	Backbone	GFFM	UL-FM	Dice	mIoU	Precision	Recall
Med.1	✓	×	×	0.907	0.847	0.897	0.913
Med.2	✓	✓	×	0.917	0.862	0.917	0.931
Med.3	✓	×	✓	0.918	0.865	0.913	0.933
DEGF-Net	✓	✓	✓	0.933	0.886	0.931	0.939

MGCBFormer [49], Ens2 [30], DEMF-Net [31]). Although AUC-ROC curves can provide complementary insights, their calculation relies on model confidence scores, which are not consistently obtainable from all benchmark methods. Therefore, AUC-ROC curves are not provided in this paper.

Fig. 5 presents a comparative analysis of the DEGF-Net method against several state-of-the-art techniques, evaluated using the Dice similarity coefficient and mIoU metrics. The evaluation is performed on the CVC-ClinicDB and Kvasir-SEG datasets.

Fig. 6 presents a comparative analysis of several leading medical image segmentation models on the CVC-ClinicDB and Kvasir-SEG datasets, with an emphasis on their precision and recall performance.

From the results, the DEGF-Net (green columns) proposed in this study demonstrates the best Recall performance in both cases: 0.955 on CVC-ClinicDB and 0.939 on Kvasir-SEG, outperforming the other models compared. In medical image analysis, Recall is a key indicator of the model's ability to identify lesion regions, as it measures the proportion of actual lesions that are covered, which directly reflects the risk of missed detection. In clinical diagnosis, omitting a single lesion can have serious consequences, so a high recall rate is particularly important for practical applications. Meanwhile, although DEGF-Net performs well in Recall, its precision remains high (0.962 and 0.931, respectively), which does not lead to many false alarms due to an overemphasis on detection rate and reflects a good balance in prediction.

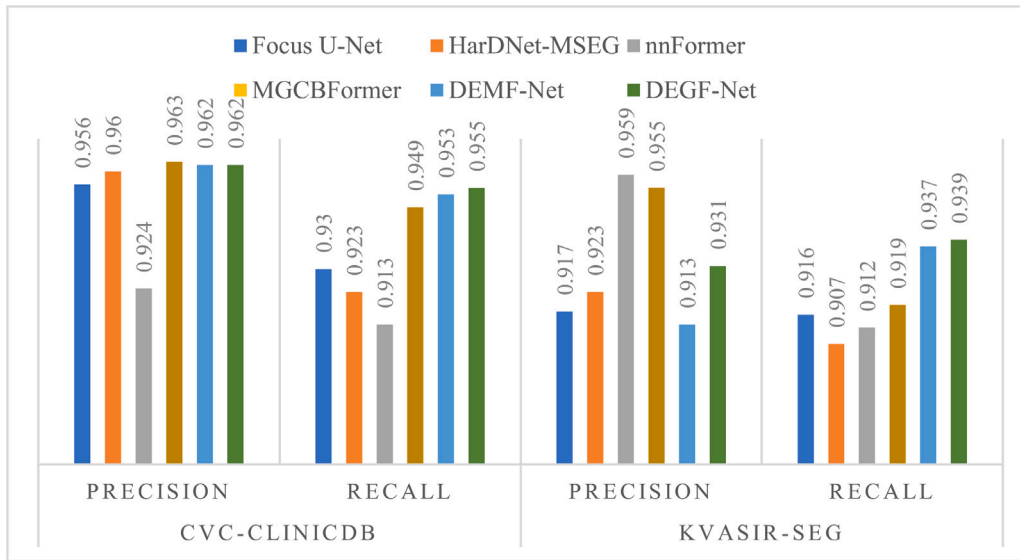


Fig. 6. Performance comparison using the evaluation metrics of Precision and Recall on CVC-ClinicDB and Kvasir-SEG datasets.

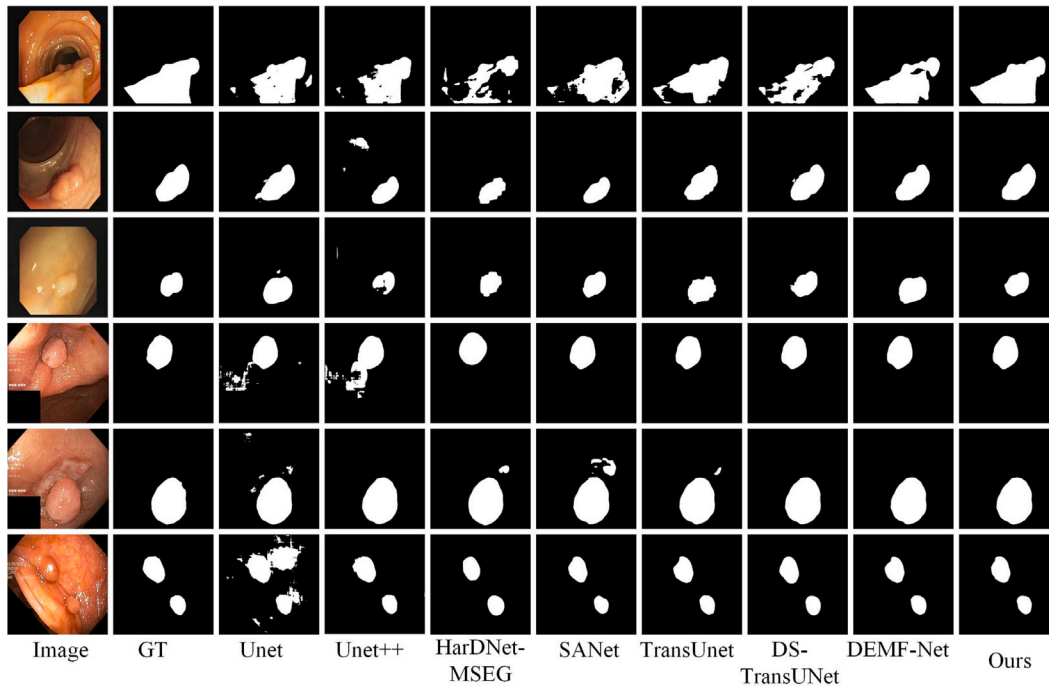


Fig. 7. Visualization results of polyp segmentation.

Fig. 7 depicts the segmentation outputs of DEGF-Net on both the CVC-ClinicDB and Kvasir-SEG datasets, with the first three examples drawn from CVC-ClinicDB and the subsequent three from Kvasir-SEG. From the analyses presented in Table 3 and Fig. 7, it is evident that the DEGF-Net excels in polyp segmentation. The advantages of the model can be summarized below:

1. Adaptability for clinical variability: The segmentation performance is effectively enhanced to address real-world issues, including irregular polyp shapes, size variability, and the coexistence of multiple polyps. In particular, the model achieves more accurate edge segmentation for irregularly shaped polyps.
2. Hybrid encoder architecture: The integration of Swin Transformer and CNN to a dual encoder enables the model to leverage

the complementary strengths of both architectures. Moreover, pruning branches in the Swin Transformer reduces computational cost while maintaining significant improvements in polyp detection rate and segmentation accuracy.

3. Advanced feature fusion: Two newly proposed feature fusion modules capture both structural details and positional information by integrating global features and multi-level feature representations, enabling effective segmentation of polyp regions, even in complex clinical scenarios.

4.5. Ablation experiments

An ablation study was conducted on model variants using the Kvasir-SEG dataset to assess the segmentation performance of the

Table 5
Quantitative results for generalizability, the best results are highlighted.

Model	CVC-ColonDB		ETIS		EndoScene	
	Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net [6]	0.471	0.390	0.401	0.340	0.627	0.535
UNet++ [8]	0.318	0.253	0.297	0.247	0.428	0.357
ResUNet++ [41]	0.300	0.222	0.121	0.081	0.834	0.777
PraNet [43]	0.709	0.640	0.628	0.567	0.871	0.797
DDANet [44]	0.520	0.439	0.247	0.195	0.834	0.777
ACSNNet [45]	0.716	0.649	0.578	0.509	0.863	0.787
HarDNet-MSEG [46]	0.731	0.660	0.677	0.613	0.887	0.821
SANet [47]	0.745	0.665	0.750	0.654	0.888	0.815
DEMF-Net [31]	0.751	0.668	0.736	0.635	0.908	0.882
DEGF-Net (Ours)	0.792	0.712	0.750	0.675	0.924	0.886

DEGF-Net for colorectal polyp images. The study also aimed to highlight the contribution of each component to the model's overall effectiveness.

The baseline configuration, referred to as 'Backbone', consists solely of a two-branch encoder and decoder. Variants of the model were constructed by including or omitting the Global Federated Feature Fusion Module (GFFM) and the Upper and Lower Level Feature Fusion Module (UL-FM). The standard configuration, incorporating both modules, is denoted as 'Backbone+GFFM+UL-FM' (DEGF-Net). The validity of each configuration was assessed by comparing performance metrics on the Kvasir-SEG dataset.

4.5.1. Effectiveness of the global federated feature fusion module (GFFM)

As shown in Table 4, the model variant Med.2, which includes GFFM, achieves improvements of 1.0% in Dice, 1.5% in mIoU, 2.0% in Precision, and 1.8% in Recall, over the baseline 'Backbone'. These results suggest that the GFFM effectively integrates features from the dual encoder and captures both hidden semantic and positional information in polyp images, enhancing global contextual understanding and improving output accuracy.

4.5.2. Effectiveness of the upper and lower level feature fusion module (UL-FM)

According to Table 4, the model Med.3, which incorporates UL-FM, shows improvements of 1.1% in Dice, 1.8% in mIoU, 1.6% in Precision, and 2.0% in Recall over the baseline. This indicates that UL-FM effectively reduces the semantic gap across various feature levels, enabling the propagation of high-level semantic and positional information throughout the polyp region. Consequently, it facilitates the successful fusion of cross-level features and enhances segmentation performance.

4.5.3. Effectiveness of the full configuration (DEGF-Net)

The complete DEGF-Net model integrates a dual-encoder architecture comprising a Swin Transformer and a CNN, along with the GFFM and UL-FM modules. By employing a dual-encoder architecture, the model addresses CNNs' inability to capture long-range dependencies and overcomes the shortcomings of the Swin Transformer in extracting fine-grained local details, thereby harnessing the complementary strengths of both. Moreover, the inclusion of GFFM and UL-FM allows the module to capture fine-grained structural and positional details from both global and hierarchical feature levels, resulting in robust and accurate polyp segmentation, even in complex clinical scenarios.

4.6. Generalization capability

4.6.1. Generalization ability on polyp datasets

To assess the generalization capability of the DEGF-Net, which achieved outstanding results on the CVC-ClinicDB and Kvasir-SEG datasets, it was further tested on the EndoScene, CVC-ColonDB, and ETIS datasets. The segmentation models that performed best on the CVC-ClinicDB and Kvasir-SEG polyp datasets were selected for testing

on the CVC-ColonDB, EndoScene, and ETIS datasets. None of these three datasets was used in the model training.

As shown in Table 5, the DEGF-Net demonstrates strong generalization performance when compared to existing models. It effectively adapts to varying characteristics of different polyp datasets and achieves superior segmentation results on all three previously unseen datasets. On the CVC-ColonDB dataset, particularly, the DEGF-Net achieves improvements of 2.3% in Dice and 2.1% in mIoU over the traditional segmentation algorithm, TransUnet. On the ETIS dataset, the method outperforms DEMF-Net, achieving gains of 1.7% in Dice and 3.8% in mIoU. For the EndoScene dataset, the DEGF-Net outperforms the second-best model, TransUnet, achieving a 0.8% higher Dice score and a 0.2% improvement in mIoU. The results suggest that the DEGF-Net demonstrates strong generalization across various clinical scenarios. It maintains high segmentation performance across datasets with varying imaging environments and polyp characteristics, demonstrating its potential for widespread use in real-world clinical settings.

4.6.2. Generalization ability on other medical datasets

To assess the adaptability and performance of the DEGF-Net across different medical imaging domains, experiments were conducted on the DRIVE retina dataset and a publicly available cell segmentation dataset. This evaluation assesses the algorithm's potential beyond colorectal polyp segmentation, highlighting its broader applicability to diverse medical image segmentation tasks.

The DRIVE dataset, provided by the Image Processing Group [50], consists of 40 color retinal images that can be used for training and an additional 20 for testing, all provided at a resolution of 584×565 pixels. The cell segmentation dataset, which also includes 30 cell images for training and an additional 20 for testing, all with a resolution of 512×512 pixels, is commonly utilized for evaluating medical image segmentation models.

During the experiments, all images from both datasets were uniformly resized to 512×512 pixels. These experiments help to validate the generalization capability of the DEGF-Net across various medical imaging contexts.

Table 6 illustrates that the DEGF-Net outperforms existing models in terms of segmentation accuracy on both the retina and cell segmentation datasets. Fig. 8 provides a visual comparison of the DEGF-Net model with other existing models for retinal segmentation. These results indicate that DEGF-Net can be effectively adapted to a range of medical image segmentation tasks beyond colorectal polyp segmentation. The demonstrated generalization capability provides strong evidence supporting the applicability of the DEGF-Net across a broader range of medical imaging domains.

5. Discussion

While the colorectal polyp segmentation method presented in this paper has yielded encouraging results, several aspects still require further refinement in future investigations. Specifically, these areas include:

Table 6
Generalization ability of other datasets.

Model	DRIVE				Cell			
	Dice	mIoU	Precision	Recall	Dice	mIoU	Precision	Recall
U-Net [6]	0.770	0.627	0.775	0.769	0.950	0.905	0.956	0.944
UNet++ [8]	0.721	0.616	0.637	0.628	0.932	0.873	0.956	0.910
DEMF-Net [31]	0.886	0.795	0.869	0.903	0.982	0.965	0.984	0.980
DEGF-Net (Ours)	0.896	0.823	0.893	0.908	0.994	0.985	0.993	0.990

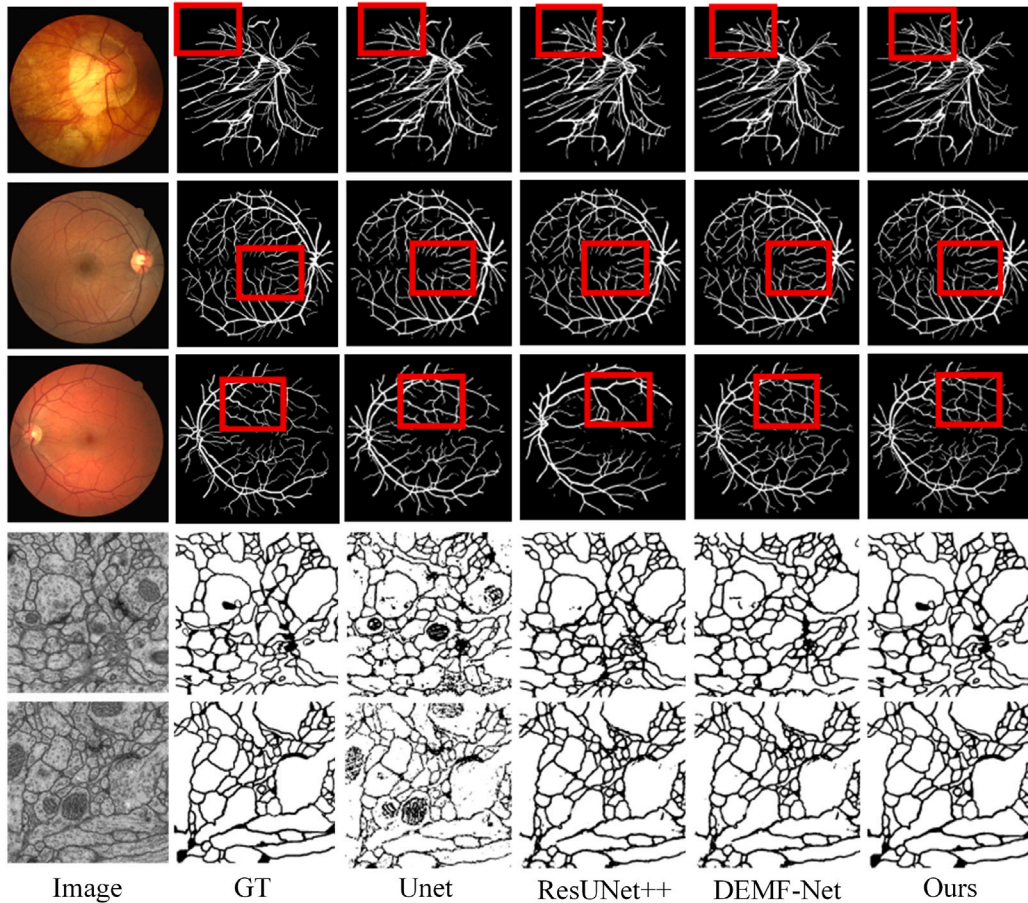


Fig. 8. Visualization results of retinal segmentation and cell segmentation.

1. The values of the hyperparameters α , β , δ , γ , in the loss function are set based on empirical experiments as well as the relative importance assigned to the different loss components throughout the training process. The weights ($\gamma = 0.7$) of the final prediction graph O emphasize that the final output should be the main focus of the training process, which is typical for many tasks, as the final prediction is the most important aspect, and it is crucial to ensure that the model performs well in this regard. By assigning higher weights, the model is encouraged to learn to generate high-quality final predictions. While the final output should be the main priority, O_1 , O_2 , and O_3 , the intermediate outputs of each encoder layer, capture different levels of feature abstraction, and the intermediate outputs can provide valuable supervisory signals for fine-grained tuning at different levels of abstraction, so the remaining weights are equally distributed among the three hyperparameters. These features can help the model converge faster, prevent overfitting, and improve generalization by using multiple feature levels in the model representation. For now, the selection of loss-function hyperparameters remains largely empirical, with room for further improvement.
2. Recent studies predominantly concentrate on the segmentation of colorectal polyps in images, with most publicly available datasets being extracted from colonoscopy videos. This presents an opportunity to adapt existing colorectal segmentation methods for video data, enabling their application to colonoscopy videos. Furthermore, these methods can be integrated into hardware devices to delineate the areas and locations of colorectal polyps in real time and efficiently during colonoscopy, making them an effective adjunct to colonoscopy.
3. Although the DEGF-Net shows strong segmentation performance on the retina and cellular datasets. Moving forward, comparative experiments will be conducted on additional medical image segmentation datasets, focusing on developing feasible, reliable, and accurate segmentation methods based on existing techniques. Further optimization of the model is essential to reduce training time and minimize its high hardware memory usage.
4. Despite the strong overall performance of DEGF-Net on Kvasir-SEG and CVC-ClinicDB, several limitations remain. The method can still under-segment or miss extremely small or flat polyps with very low contrast to the surrounding mucosa, and it is

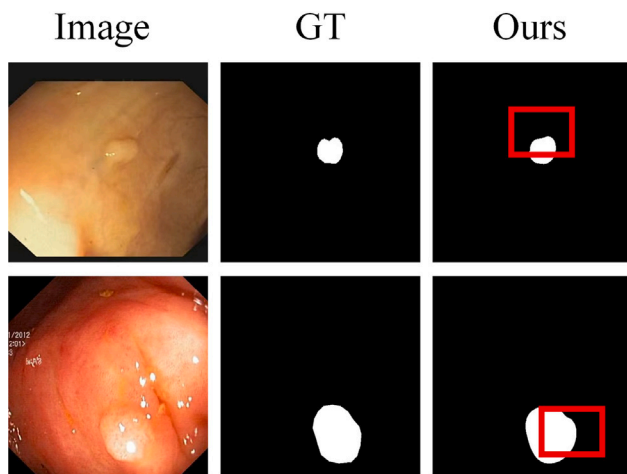


Fig. 9. Poor segmentation results for small or flat samples.

susceptible to severe colonoscopy artefacts such as specular highlights, motion blur, and bowel-preparation residues, which may induce false positives on mucosal folds or background tissue. As shown in Fig. 9, we present two sets of poor segmentation results for small, flat polyps. In the first set, the segmentation of the upper portion of the polyp is suboptimal, exhibiting significant shape deviation. In the second group, the original image reveals that the polyps on the right side of this sample are flatter with less distinct boundaries. Consequently, the segmentation results on the right side show a noticeable discrepancy compared to the ground truth (GT). We have not yet fully identified the reasons for this outcome, which will also be a challenge for future research. In addition, the current evaluation is limited to two public datasets acquired under relatively controlled conditions, so the robustness of DEGF-Net to images from other endoscope systems and clinical centers remains to be established. These observations indicate that further work on small-lesion enhancement, artefact-robust representations, and broader multi-center validation is still required.

5. Analyzing the model's performance across different polyp sizes is crucial for a comprehensive assessment of its robustness. However, the datasets employed in this study (e.g., Kvasir-SEG, CVC-ClinicDB) are publicly available benchmark datasets that do not provide official standardized annotations grouping polyps by size (small, medium, large). Manually defining classification criteria, alongside verifying compliance with medical standards and conducting rigorous re-evaluation. Furthermore, we shall include a performance across polyps of varying sizes (small, medium, large), which represents a crucial direction for future research in quantitative polyp segmentation analysis.
6. Whilst the model demonstrates consistent performance across multiple datasets, confirming its sound reliability and robustness, we acknowledge that more rigorous statistical analysis, such as paired t-tests or other statistical significance tests applied to all test images, would furnish stronger quantitative evidence. Conducting such analyses requires access to single-image predictions from all comparison methods, which is not possible under the current experimental setup because not all benchmark models can be reproduced. Achieving this objective will be a key focus of our subsequent research, with substantive exploration and implementation of this goal in future work.

6. Conclusions

This paper proposes a dual-encoder framework, DEGF-Net, for colorectal polyp segmentation. DEGF-Net combines Swin Transformer and

CNN branches to exploit the Transformer's strength in modeling long-range dependencies together with the CNN's ability to capture fine-grained local features.

By integrating Swin Transformer and CNN in the encoder, the network effectively leverages complementary global and local representations, while a layer-by-layer CNN-based decoder enables efficient feature fusion. To further enhance segmentation performance in complex scenarios, two dedicated modules, the Global Joint Feature Fusion Module (GJFFM) and the Upper-Lower Level Feature Fusion Module (UL-FM), are introduced. These modules jointly encode semantic and positional information and help bridge the gap between hierarchical features. Experimental results on the CVC-ClinicDB and Kvasir-SEG datasets confirm the effectiveness of DEGF-Net, with Dice scores of 0.958 and 0.933, respectively. Additional evaluations on the DRIVE retina dataset and a public cell segmentation dataset further demonstrate the model's generalization ability across different medical imaging domains.

Although the proposed method achieves promising results, several directions remain for future work. We plan to extend DEGF-Net to video-based polyp segmentation to enable real-time detection during colonoscopy procedures. In addition, we intend to conduct more extensive experiments on diverse medical datasets and explore model optimization strategies to reduce computational cost and memory usage, thereby facilitating deployment in resource-constrained clinical environments.

CRedit authorship contribution statement

He Yu: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Jinming Guo:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Xiaorui Cao:** Writing – original draft, Methodology, Investigation. **Xuan Li:** Writing – original draft, Visualization, Validation, Methodology, Investigation. **Yulong Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Jianying Hao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis. **Haoming Yang:** Writing – original draft, Visualization, Validation, Software, Conceptualization. **Xi Zhou:** Writing – review & editing, Writing – original draft, Visualization, Methodology. **Tao Huang:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- [1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: Cancer J. Clin.* 71 (3) (2021) 209–249, <http://dx.doi.org/10.3322/caac.21660>.
- [2] L. Hol, M.E. Van Leerdam, M. Van Ballegooijen, A.J. Van Vuuren, H. Van Dekken, J.C. Reijerink, A.C. Van der Togt, J. Habbema, E.J. Kuipers, Screening for colorectal cancer: randomised trial comparing guaiac-based and immunochemical faecal occult blood testing and flexible sigmoidoscopy, *Gut* 59 (01) (2010) 62–68, <http://dx.doi.org/10.1136/gut.2009.177089>.

- [3] O.H. Maghsoudi, Superpixel-based segmentation and classification of polyps in wireless capsule endoscopy, in: 2017 IEEE Signal Processing in Medicine and Biology Symposium, SPMB, IEEE, Philadelphia, PA, USA, 2017, pp. 1–4, <http://dx.doi.org/10.1109/SPMB.2017.8257027>.
- [4] M. Fiori, P. Musé, G. Sapiro, A complete system for candidate polyps detection in virtual colonoscopy, *Int. J. Pattern Recognit. Artif. Intell.* 28 (07) (2014) 1460014, <http://dx.doi.org/10.1142/S0218001414600143>.
- [5] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automatic polyp detection using global geometric constraints and local intensity variation patterns, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cambridge, MA, USA, 2014, pp. 179–187, http://dx.doi.org/10.1007/978-3-319-10470-6_23.
- [6] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol. 9351, Springer, Munich, Germany, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.
- [7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *nature*, 521, (7553) 2015, pp. 436–444, <http://dx.doi.org/10.1038/nature14539>.
- [8] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: International Workshop on Deep Learning in Medical Image Analysis, Springer, Granada, Spain, 2018, pp. 3–11, http://dx.doi.org/10.1007/978-3-030-00889-5_1.
- [9] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th International Conference on Information Technology in Medicine and Education, ITME, IEEE, Hangzhou, China, 2018, pp. 327–331, <http://dx.doi.org/10.1109/ITME.2018.00080>.
- [10] M.M. Naseer, K. Ranasinghe, S.H. Khan, M. Hayat, F. Shahbaz Khan, M.-H. Yang, Intriguing properties of vision transformers, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 23296–23308, URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/c404a5adbf90e09631678b13b05d9d7a-Abstract.html.
- [11] H.-Y. Zhou, C. Lu, S. Yang, Y. Yu, Convnets vs. transformers: Whose visual representations are more transferable? in: 2021 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW, Montreal, BC, Canada, 2021, pp. 2230–2238, <http://dx.doi.org/10.1109/ICCVW54120.2021.00252>.
- [12] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A.K. Nandi, Medical image segmentation using deep learning: A survey, *IET Image Process.* 16 (5) (2022) 1243–1267, <http://dx.doi.org/10.1049/ipr2.12419>.
- [13] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: A general vision transformer backbone with cross-shaped windows, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, LA, USA, 2022, pp. 12114–12124, <http://dx.doi.org/10.1109/CVPR52688.2022.01181>.
- [14] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, in: Advances in neural information processing systems, Vol. 34, 2021, pp. 9355–9366, URL <https://proceedings.neurips.cc/paper/2021/hash/4e0928de075538c593fbdabb0c5ef2c3-Abstract.html>.
- [15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, LA, USA, 2022, pp. 11966–11976, <http://dx.doi.org/10.1109/CVPR52688.2022.01167>.
- [16] J. Chen, E.C. Frey, Y. He, W.P. Segars, Y. Li, Y. Du, Transmorph: Transformer for unsupervised medical image registration, *Med. Image Anal.* 82 (2022) 102615, <http://dx.doi.org/10.1016/j.media.2024.103280>.
- [17] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, Waikoloa, HI, USA, 2022, pp. 1748–1758, <http://dx.doi.org/10.1109/WACV51458.2022.00181>.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, Vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.
- [19] Y. Zhang, L. Liu, Z. Han, F. Meng, Y. Zhang, Y. Zhao, Transefusionnet: Deep fusion network for colorectal polyp segmentation, *Biomed. Signal Process. Control.* 86 (2023) 105133, <http://dx.doi.org/10.1016/j.bspc.2023.105133>, URL <https://www.sciencedirect.com/science/article/pii/S1746809423005669>.
- [20] Y. Lin, X. Han, K. Chen, W. Zhang, Q. Liu, Cswindoubleu-net: A double u-shaped network combined with convolution and swin transformer for colorectal polyp segmentation, *Biomed. Signal Process. Control.* 89 (2024) 105749, <http://dx.doi.org/10.1016/j.bspc.2023.105749>, URL <https://www.sciencedirect.com/science/article/pii/S1746809423011825>.
- [21] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (12) (2017) 2481–2495, <http://dx.doi.org/10.1109/TPAMI.2016.2644615>.
- [22] T. Mahmud, B. Paul, S.A. Fattah, Polypsegnet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images, *Comput. Biol. Med.* 128 (2021) 104119, <http://dx.doi.org/10.1016/j.compbiomed.2020.104119>.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of the 9th International Conference on Learning Representations, ICLR, 2021, URL <https://openreview.net/forum?id=YicbFdNTTy>.
- [24] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, *CAAI Artif. Intell. Res.* 2 (2023) 9150015, <http://dx.doi.org/10.26599/AIR.2023.9150015>, URL <https://www.sciopen.com/article/10.26599/AIR.2023.9150015>.
- [25] T.D. Nguyen, T.O. Nguyen, T.T. Nguyen, M.T. Tran, V.S. Dinh, Colonformer: An efficient transformer-based method for colon polyp segmentation, *IEEE Access* 10 (2022) 110879–110889, <http://dx.doi.org/10.1109/ACCESS.2022.3206819>, URL <https://arxiv.org/abs/2205.08473>.
- [26] G. Ayana, H. Barki, S.-W. Choe, Pathological insights: Enhanced vision transformers for the early detection of colorectal cancer, *Cancers* 16 (7) (2024) 1441, <http://dx.doi.org/10.3390/cancers16071441>, URL <https://www.mdpi.com/2072-6694/16/7/1441>.
- [27] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, H.D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International conference on multimedia modeling, Vol. 11962, Springer, 2019, pp. 451–462, http://dx.doi.org/10.1007/9783030377342_37.
- [28] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer, *Int. J. Comput. Assist. Radiol. Surg.* 9 (2) (2014) 283–293, <http://dx.doi.org/10.1007/s11548-013-0926-3>.
- [29] X. Yin, J. Zeng, T. Hou, R. Zhang, Y. Qian, Rsaformer: A method of polyp segmentation with region self-attention transformer, *Comput. Biol. Med.* 172 (2024) 108268, <http://dx.doi.org/10.1016/j.compbiomed.2024.108268>, URL <https://www.sciencedirect.com/science/article/pii/S0010482524003524>.
- [30] L. Nanni, A. Lumini, C. Fantozzi, Exploring the potential of ensembles of deep learning networks for image segmentation, *Information* 14 (12) (2023) <http://dx.doi.org/10.3390/info14120657>, URL <https://www.mdpi.com/2078-2489/14/12/657>.
- [31] X. Cao, H. Yu, K. Yan, R. Cui, J. Guo, X. Li, X. Xing, T. Huang, Demf-net: A dual encoder multi-scale feature fusion network for polyp segmentation, *Biomed. Signal Process. Control.* 96 (2024) 106487, <http://dx.doi.org/10.1016/j.bspc.2024.106487>, URL <https://www.sciencedirect.com/science/article/pii/S1746809424005457>.
- [32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, Montreal, QC, Canada, 2021, pp. 9992–10002, <http://dx.doi.org/10.1109/ICCV48922.2021.00986>.
- [33] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, *Comput. Med. Imaging Graph.* 43 (2015) 99–111.
- [34] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, *IEEE Trans. Med. Imaging* 35 (2) (2015) 630–644, <http://dx.doi.org/10.1109/TMI.2015.2487997>.
- [35] D. Vázquez, J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, A.M. López, A. Romero, M. Drozdal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, *J. Heal. Eng.* 2017 (1) (2017) 4037190, <http://dx.doi.org/10.1155/2017/4037190>.
- [36] C. Sun, Y. Liu, Y. Huang, B. Li, W. Rang, Colorectal cancer incidence and mortality trends and analysis of risk factors in China from 2005 to 2015, *Int. J. Gen. Med.* (2021) 9965–9976, <http://dx.doi.org/10.2147/IJGM.S344448>.
- [37] C.-H. Huang, H.-Y. Wu, Y.-L. Lin, Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, 2021, arXiv preprint [arXiv:2101.07172](https://arxiv.org/abs/2101.07172).
- [38] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, H. Zhang, Sanet: A slice-aware network for pulmonary nodule detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (8) (2021) 4374–4387, <http://dx.doi.org/10.1109/TPAMI.2021.3065086>.
- [39] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al., Transunet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers, *Med. Image Anal.* 97 (2024) 103280, <http://dx.doi.org/10.1016/j.media.2024.103280>.
- [40] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, *IEEE Trans. Instrum. Meas.* 71 (2022) 1–15, <http://dx.doi.org/10.1109/TIM.2022.3178991>.
- [41] D. Jha, P.H. Smedsrud, D. Johansen, T. de Lange, H.D. Johansen, P. Halvorsen, M.A. Riegler, A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation, *IEEE J. Biomed. Health Inform.* 25 (6) (2021) 2029–2040, <http://dx.doi.org/10.1109/JBHI.2021.3049304>.
- [42] M. Yeung, E. Sala, C.-B. Schönlieb, L. Rundo, Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy, *Comput. Biol. Med.* 137 (2021) 104815, <http://dx.doi.org/10.1016/j.compbiomed.2021.104815>, URL <https://www.sciencedirect.com/science/article/pii/S0010482521006090>.

- [43] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: *Medical Image Computing and Computer Assisted Intervention, MICCAI 2020*, Springer International Publishing, Cham, 2020, pp. 263–273, http://dx.doi.org/10.1007/978-3-030-59725-2_26.
- [44] N.K. Tomar, D. Jha, S. Ali, H.D. Johansen, D. Johansen, M.A. Riegler, P. Halvorsen, Ddanet: Dual decoder attention network for automatic polyp segmentation, in: *Pattern Recognition. ICPR International Workshops and Challenges*, Springer International Publishing, Cham, 2021, pp. 307–314.
- [45] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, Y. Yu, Adaptive context selection for polyp segmentation, in: *Medical Image Computing and Computer Assisted Intervention, MICCAI 2020*, Springer International Publishing, Cham, 2020, pp. 253–262.
- [46] C.-H. Huang, H.-Y. Wu, Y.-L. Lin, Hardnet-mseg: A simple encoder–decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, 2021, [arXiv:2101.07172](https://arxiv.org/abs/2101.07172) URL <https://arxiv.org/abs/2101.07172>.
- [47] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, H. Zhang, Sanet: A slice-aware network for pulmonary nodule detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (8) (2022) 4374–4387, <http://dx.doi.org/10.1109/TPAMI.2021.3065086>.
- [48] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, Y. Yu, Nnformer: Volumetric medical image segmentation via a 3d transformer, *IEEE Trans. Image Process.* 32 (2023) 4036–4045, <http://dx.doi.org/10.1109/TIP.2023.3293771>.
- [49] Y. Xia, H. Yun, Y. Liu, J. Luan, M. Li, Mgcformer: The multiscale grid-prior and class-inter boundary-aware transformer for polyp segmentation, *Comput. Biol. Med.* 167 (2023) 107600, <http://dx.doi.org/10.1016/j.compbiomed.2023.107600>, URL <https://www.sciencedirect.com/science/article/pii/S001048252301065X>.
- [50] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, *IEEE Trans. Med. Imaging* 23 (4) (2004) 501–509, <http://dx.doi.org/10.1109/TMI.2004.825627>.