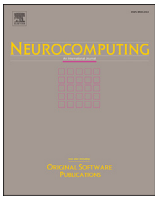
Neurocomputing 667 (2026) 132339

Contents lists available at ScienceDirect

Neurocomputing

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



Efficient transformer-based semantic segmentation of colonic polyps using SegFormer

Gul E Arzu [a], Muhammad Fayaz [a], Usman Ali [a], L. Minh Dang [b, c, d], Hyeonjoon Moon [a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
[b] The Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam
[c] Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam
[d] Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, South Korea

ARTICLE INFO

Communicated by W. Wang

Keywords:
Automated polyp segmentation
Colorectal cancer
SegFormer
Semantic segmentation
Test time augmentation
Medical image analysis

ABSTRACT

This paper introduces a lightweight and efficient polyp segmentation framework, which is built on the SegFormer-B3 model and trained on the MMSegmentation platform. The proposed technique was tested and trained on four popular colonoscopy datasets: CVC-ClinicDB, Kvasir-SEG, Kvasir-Capsule-SEG, and ETIS-Larib. It was also assessed on a mixed dataset to determine generalization in a variety of clinical situations. By exploiting transformer-based encoding, the framework encodes the global contextual information, and fine-grained local details, which greatly boosts recognition of polyp morphology and boundaries greatly. The hybrid of Dice and Focal loss functions was used to trade off region-level and pixel-level accuracy, and test-time augmentations, including horizontal and vertical flips, were used to enhance inference stability. It achieved a Dice score of 0.969, a mean IoU of 0.942, and is computationally efficient with an inference speed of 94 FPS, proving to be highly accurate in segmentation and computationally efficient. These findings demonstrate that the framework can be applied in clinical practice and can be regarded as a reliable and strong solution to endoscopic polyp segmentation. This approach provides a valuable resource for enhancing clinical decision support systems and facilitating further automated analysis in gastrointestinal imaging, as it integrates the latest transformer architecture with practical efficiency.

1. Introduction

Colorectal cancer (CRC) is one of the most frequently diagnosed malignancies worldwide and the cause of cancer-related death among men and women in the first part of the world [1]. As per some of the recent statistics in the United States, colorectal cancer is the third leading cause of cancer-related deaths, with an estimated 52,550 cancer-related deaths and 53,020 new cancer cases per annum. Moreover, in the stomach, esophagus, and small intestines, cancers of the digestive system are major contributors to the global burden of cancer in the world today [2]. Other gastrointestinal (GI) diseases such as polyps, ulcers and internal bleeding pose a significant challenge to clinicians in detecting and locating, especially in the small intestine. The colonic polyps are among the most common gastrointestinal conditions that are difficult to diagnose because of their size, irregular geometry, and the complex surface architecture. Regardless of such complications, the current diagnostic methods have significantly enhanced the accuracy and reliability of polyp detection.

Colonoscopy remains the primary and most reliable approach for colorectal CRC screening and prevention. A flexible tube equipped with a miniature camera is inserted through the rectum during this procedure to visually examine the inside of the colon for anomalies [3]. During the process, clinicians identify and categorize polyps into two main types: neoplastic and non-neoplastic. Non-neoplastic polyps are usually non-cancerous, whereas neoplastic polyps are capable of developing into malignant lesions [1]. Based on size, colorectal polyps are typically classified as diminutive (≤ 5 mm), small (6–9 mm), and advanced (≥ 10 mm) [4]. Although colonoscopy facilitates both detection and removal, it has drawbacks: it is invasive, can cause discomfort, often requires sedation, and may overlook smaller or anatomically concealed polyps [5].

To overcome some of these limitations, wireless capsule endoscopy (WCE) has been established as a minimally invasive diagnostic method.

* Corresponding author.
Email addresses: arzurabani@sju.ac.kr (G.E. Arzu), muhammadfayaz@sju.ac.kr (M. Fayaz), usman.ali@sejong.ac.kr (U. Ali), minhdl@sejong.ac.kr (L.M. Dang), hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.neucom.2025.132339
Received 9 September 2025; Received in revised form 27 November 2025; Accepted 3 December 2025
Available online 9 December 2025

The method involves the use of a swallowable capsule, sized at about 11 mm, that records over 55,000 high-resolution images each time that the small intestine, 6 meters long, is examined, a process that offers a detailed view of the inside of this tube [6]. The WCE enables patients to be on the move throughout the process and it does not require much preparation [7,8]. Nonetheless, there is a long and time-consuming task of studying the huge amount of generated images, and this aspect highlights the necessity of sophisticated computing support. The latest developments in computer-aided detection (CAD) systems that operate on artificial intelligence (AI) have significantly helped the analysis of GI medical imaging substantially. Deep learning (DL) and machine learning (ML) methods have taken the centre stage in the diagnosis of GI malignancies and have provided better diagnostic accuracy and operational efficiency. Such artificial intelligence-based systems can assist the healthcare professionals in making a more informed clinical decisions, which will eventually lead to better patient outcomes.

Earlier approaches to polyp detection relied on manually designed feature extraction techniques, including Scale-Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) [9–11]. These features were then classified using conventional machine learning models, such as Random Forests (RF), Support Vector Machines (SVM), Decision Trees (DT), and K-Nearest Neighbours (KNN) [12–14]. However, these handcrafted techniques face limitations in terms of scalability and generalization, as they are often task-specific and dataset-dependent. Additionally, the manual process of selecting and engineering features for WCE and colonoscopy images is both time-intensive and laborious [15–17]. Deep learning, particularly CNNs [18], addressed this with automatic hierarchical feature extraction, and encoder–decoder models like U-Net [19] and SegNet [20] achieved strong results in polyp segmentation. Nonetheless, CNNs struggle with modelling global context and require large, labelled datasets. Currently Vision Transformers (ViTs) have surpassed CNNs in performance by applying self-attention modules to collect global context and long-range interdependence [20]. Their integration into segmentation tasks opens new opportunities for detecting small and morphologically complex polyps with improved accuracy. Polyp segmentation, particularly for small or flat lesions, continues to be a difficult endeavor for a number of reasons: reduced contrast, varying illumination, irregular shapes and textures, and the presence of complex backgrounds. As illustrated in Fig. 1, experienced experts often face difficulties in correctly detecting polyps due to their visual resemblance to the surrounding healthy tissues. These complications highlight the crucial necessity for automated, resilient, and intelligent decision support systems to assist clinicians in interpreting GI images.

To address these issues, this study proposes a lightweight yet powerful deep learning (DL) framework based on the SF (SF) architecture for the semantic segmentation of colonic polyps. SF capably incorporates the Mix Vision Transformer (MiT) backbone for effective global characteristic acquisition with a lightweight all-MLP decoder, ensuring both accuracy and computational efficiency. The segmentation head generates accurate masks without incurring additional computational load, enabling its effective application in real-time workflows and resource-constrained environments. At the same time, the transformer-based encoder utilizes self-attention mechanisms to gather contextual information across various scales. This research provides the following key contributions:

- We propose a lightweight deep learning model that is entirely implemented on the SF architecture to accurately and semantically segment small colonic polyps, leveraging vision transformers to effectively capture both global and local features.
- The SF encoder utilizes a hierarchical MiT backbone, which enables effective multi-level spatial feature extraction of WCE and colonoscopy images, thereby overcoming polyp variability and complexity.
- The decoder is developed as a lightweight, all-MLP-based module utilising multi-scale feature representations. This design does not rely on complicated attention patterns and is computationally efficient, achieving high segmentation accuracy.

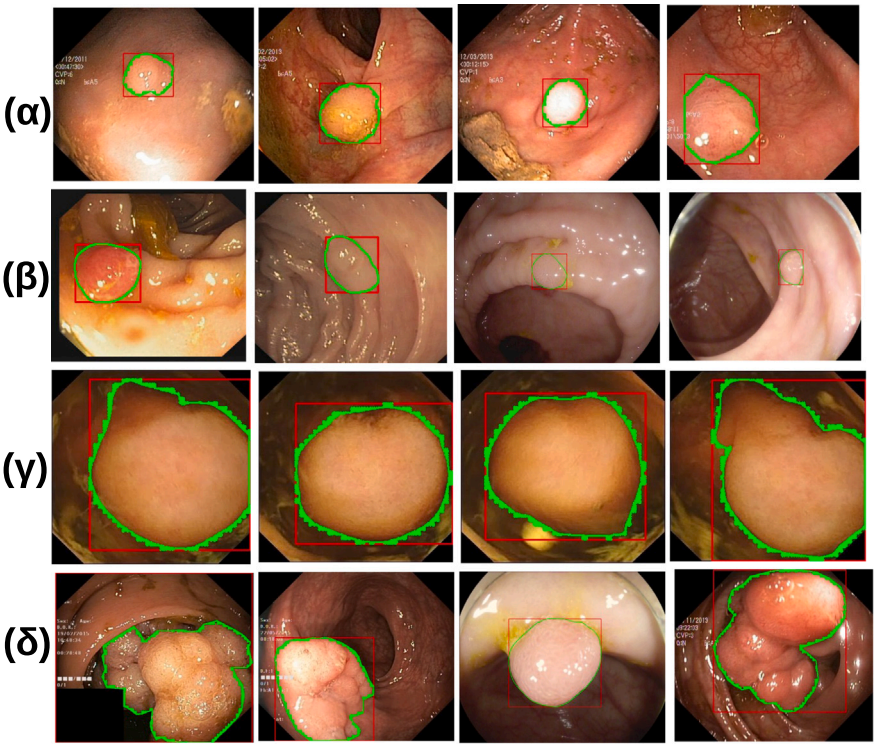


Fig. 1. The visual examples demonstrate variation in polyp sizes: (α) and (β) correspond to small polyps, whereas (γ) and (δ) illustrate large polyp formations with clear boundaries.

- We present a cross-dataset generalization policy to test the stability of the model on a variety of public polyp segmentation benchmarks and show that it can be used to fit different clinical settings without the need for retraining.
- The model features a lightweight design (46.3M parameters, 42.8G FLOPs) and can be inferred in real-time (94 FPS), which verifies it as a clinically deployable model.
- The gains of the proposed SF-based architecture are consistent with those of current state-of-the-art approaches in segmentation accuracy, boundary preservation, and small-polyp detection, indicating its potential use in the real world of the medical field.

The remaining section of this paper will be organized as follows: Section 2 provides a literature review of work on related fields, such as colon polyp detection and segmentation, with particular attention to recent deep learning tools, transformer-based architectures, and benchmark medical imaging datasets. Section 3, Methodology, contains the description of the proposed methodology, its datasets (CVC-ClinicDB, Kvasir-SEG, Kvasir-Capsule-SEG, ETIS-Larib, and combined one), data preprocessing, and the annotation approach. Section 4 setup presents the experimental setup, including hardware and software set-up, the SegFormer-B3 model architecture, the training procedure, and the ablation study design. A detailed discussion of the findings, including the limitations, failure cases, their applicability to other medical segmentation tasks, and future research directions is discussed in detail in Section 5. Lastly, in the paper, the conclusions are presented in the Section 6 where the paper summarizes the main findings, highlights clinical importance, and explains the future research opportunities to advance the robust and interpretable segmentation of colon polyps.

2. Related work

2.1. Semantic image segmentation

A fully convolutional (FC) framework was introduced for semantic image segmentation, achieving approximately a 20 % improvement in accuracy on the Pascal VOC 2012 dataset compared to previously traditional methods [21]. To address the limited ability of Fully Convolutional Networks (FCNs) in understanding global context, a model called ParseNet was developed [22], which employs average feature pooling to incorporate broader contextual information at each spatial location. However, although FCNs are typically utilized, they still face challenges, including insufficient understanding of global background, slower inference speeds, and difficulties when used for three-dimensional data. To optimize segmentation precision, [23] integrated convolutional networks (CNs) with fully connected Conditional Random Fields (CRFs) to better capture boundaries. Additionally, an encoder–decoder framework known as SegNet uses pooling indices from the encoder to guide the upsampling process in the decoder [19], resulting in improved segmentation performance.

High-Resolution Networks (HRNet) [24] have gained popularity for preserving high-resolution feature maps throughout the encoding stage by employing parallel multi-resolution convolutions and frequent information exchange between streams. Attention mechanisms have also been widely employed to improve semantic segmentation. To better capture complex contextual relationships, dual attention networks using self-attention mechanisms have been developed [25]. Originally designed for natural language processing tasks [26,27], transformer models have recently been adapted for image analysis [28]. For instance, the Segmentation Transformer (SETR) [29] processes images as sequences of patches and uses positional encoding to retain spatial information. Similarly, fully transformer-based frameworks like Segmenter [30] have been proposed to collect global context and transform feature embeddings into segmentation maps.

Recent research has introduced several advanced segmentation frameworks that significantly extend beyond traditional CNN- or transformer-based designs. LogicSeg [31] incorporates neural logic learning and reasoning to better capture high-level semantic dependencies, enabling more structured scene understanding. Prototype-based segmentation has been revisited in “Rethinking Semantic Segmentation: A Prototype View” [32], where class-specific prototypes guide pixel assignment and improve robustness under intra-class variation. Deep Hierarchical Semantic Segmentation [33] further enhances multi-level contextual modeling by organizing semantic features hierarchically, enabling refined global-to-local reasoning. Additionally, contrastive approaches such as Cross-Image Pixel Contrast [34] leverage cross-image supervision to maximize feature discriminability and improve generalization in complex environments. Together, these recent developments highlight a shift toward reasoning-based, prototype-driven, and contrastive representation frameworks that push the boundaries of modern semantic segmentation.

2.2. Medical visual segmentation

Medical visual segmentation plays a crucial role in accurately identifying and quantifying lesions or abnormalities, which supports diagnosis and treatment planning. For example, boundary-sensitive transformers that incorporate boundary-specific attention have been proposed to enhance lesion segmentation [35]. In dental radiography [36], CNNs have been employed for detecting dental caries, while region-based CNNs are used for precise localization of dental features [37]. Segmentation of organs and tumors [38] has also been studied with hybrid convolution-transformer encoder–decoder models, proving the advantage of integrating both local and global feature extraction segments in the model. Nevertheless, despite these developments, polyp segmentation is not an easy task due to the low contrast between the polyp and surrounding tissues [39–41], as well as the variability of polyp shapes, highlighting the need for further improvements, including domain-specific knowledge.

2.3. Polyp segmentation

Early detection of colonoscopic polyps is essential for preventing colorectal cancer [42,43]. Recent methods for polyp segmentation fall into three main categories based on their visual technique foundations.

1) CNN-based and Ensemble Methods: Fully CNN-based models have been introduced to outperform traditional approaches [44], while variants of Mask R-CNN have also been evaluated for polyp segmentation [45]. The encoder–decoder U-Net architecture [46] and its extensions such as ResUNet++ [45] and DoubleUNet [17,47,48] improve segmentation accuracy through stacked networks. Modified encoder–decoder frameworks leveraging multi-scale contextual information have also shown promising results [49]. Ensemble-based CNN approaches have been proposed to further boost performance. For example, Younas et al. [50] introduced a deep ensemble learning framework with optimized CNN parameters for colorectal polyp classification, and Liew et al. [51] combined modified deep residual CNNs with ensemble learning to improve detection accuracy.

2) Transformer-based Methods: Hybrid CNN-Transformer models capture both local and global features for accurate localization, often using skip connections [52]. Contrastive Transformer networks integrating multi-scale self-attention demonstrate strong feature learning [53]. Ramos and Hortua [54] proposed a Bayesian segmentation framework for colon polyps, enabling well-calibrated uncertainty estimation in predictions. Similarly, Eu et al. [55] employed a modified convolutional encoder–decoder network for polyp segmentation, highlighting the importance of carefully designed architectural modifications for performance. Att-PVT merges CNNs with Pyramid Vision Transformers to integrate context hierarchically and improve boundary delineation. Tharwat et al. [56] reviewed both machine learning and deep learning methods for colorectal cancer diagnosis, providing insights into state-of-the-art segmentation and classification approaches.

3) Lightweight / Real-time Methods: MobileNetV2-based networks with residual blocks [57] and lightweight Transformer designs with efficient encoder replacements and feature fusion modules [58] enable

Table 1
Comparison of architectural and algorithmic novelty of SF-B3 (Ours) with recent transformer-based polyp segmentation methods.

Method	Year	Key architectural / Algorithmic innovation
TransUNet [52]	2021	Hybrid CNN-Transformer encoder-decoder with skip connections; captures local + global features.
TransFuse-L [60]	2021	Combines CNN and transformer features via multi-scale fusion for boundary refinement.
ColonFormer-S [61]	2022	Pyramid transformer encoder with hierarchical feature fusion for polyp segmentation.
Polyp-PVT [62]	2024	Pyramid Vision Transformer for multi-scale global context; emphasizes small polyp detection.
ASPS [63]	2024	Augmented Segment Anything approach; improves generalization on multiple datasets.
SF-B3 (Ours)	2025	SegFormer backbone with MiT encoder capturing multi-scale global context, decoder with neighbor attention for precise boundary delineation; lightweight yet accurate for small/low-contrast polyps.

fast inference suitable for clinical settings. Open-source real-time polyp detection systems have also been evaluated for clinical applicability [59]. Despite these advancements, challenges remain due to subtle variations in polyp appearance, complex pathological characteristics, and limited annotated datasets. To address these issues, this work introduces a multi-level Transformer model that incorporates neighborhood attention mechanisms.

Architectural and Algorithmic Novelty. Compared to existing transformer-based polyp segmentation frameworks, our SF-B3 model introduces several key innovations. It employs a SegFormer backbone with MiT encoder to efficiently capture multi-scale global context, while the neighbor attention mechanism in the decoder enhances boundary precision for small and low-contrast polyps. This design enables the model to remain lightweight yet accurate, generalizes effectively across multiple datasets, and is well-suited for real-world clinical deployment (see Table 1).

3. Methodology

This research introduces a proficient and resource-friendly framework for colonic polyp segmentation employing SF [28]. Unlike conventional approaches that depend on complicated, handcrafted modules or heavy decoders, SF uses a lightweight structure, incorporating a hierarchical Transformer-based encoder with an All-MLP decoder [64]. This design ensures high segmentation accuracy while maintaining low computational complexity, a significant requirement for practical healthcare applications.

3.1. Overall architecture

As demonstrated in Fig. 2, SF contains two main modules: (1) a hierarchical Transformer encoder (MiT) that collects both coarse and fine semantic information across different dimensions, and (2) a compact MLP-based decoder that integrates these characteristics to generate the final segmentation.

The procedure begins with splitting a fed endoscopic image $H \times W \times 3$, into overlapping 4×4 pixels. This fine patch size is particularly beneficial for detecting small and flat polyps. This comparatively small patch size enhances the detection of flat or tiny polyps. The encoder then generates feature maps at $\{1/4, 1/8, 1/16, 1/32\}$ of the original image resolution based on the patch tokens that were created. These are then passed to the decoder to produce a segmentation map at $H/4 \times W/4 \times N_{\text{cls}}$, where N_{cls} is the number of output classes.

3.2. Hierarchical transformer encoder (MiT)

Unlike the standard Vision Transformer (ViT), the encoder utilized in SF is based on the Mix Vision Transformer (MiT), which builds a hierarchical feature representation. This hierarchical design is especially important for polyp segmentation tasks, where both global context and fine-grained visual details such as edges and textures must be accurately captured.

Overlapping Patch Merging: The first major component of MiT is overlapping patch merging, where input images are divided using overlapping convolutions (e.g., kernel size 7×7, stride 4, padding 3). Unlike the non-overlapping approach in ViT, this allows the model to maintain spatial continuity and preserve crucial boundary information, which is essential for accurately segmenting polyps that often have indistinct or fuzzy edges.

Multi-Level Representation: The second key feature of MiT is its multi-level representation, achieved through four sequential encoding stages. At each stage, the network generates feature maps with gradually increasing channel depths and reduced spatial resolutions. This hierarchical design allows reliable segmentation of polyps with varying shapes and sizes by extracting a broad range of semantic information, from fine-grained texture and boundary details to high-level contextual cues.

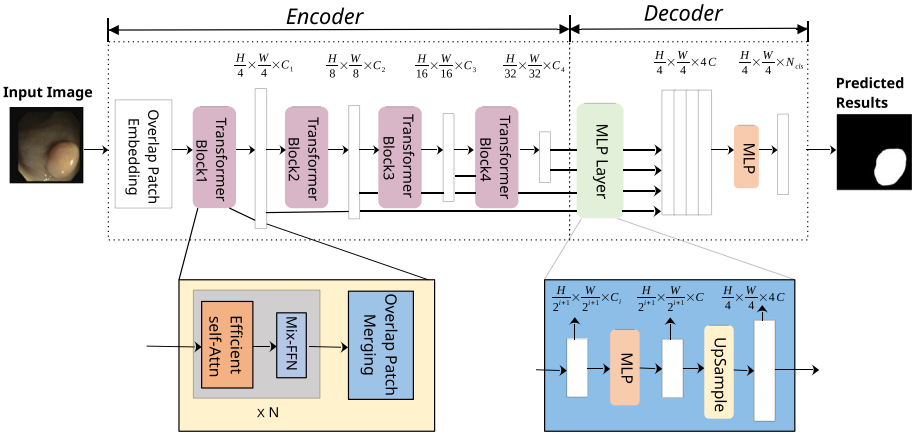


Fig. 2. An outline of the suggested SF design for segmenting colonic polyps. It involves a compact All-MLP decoder for efficient segmentation mask generation and a hierarchical Transformer encoder for multi-level characteristics extraction. Feed-Forward Network is referred to as FFN.

Table 2
Encoder stages and feature resolutions.

Encoder stage	Feature resolution	Reduction ratio (R)
Stage 1	$\frac{1}{4}$ of input size	64
Stage 2	$\frac{1}{8}$ of input size	16
Stage 3	$\frac{1}{16}$ of input size	4
Stage 4	$\frac{1}{32}$ of input size	1 (no reduction)

Efficient Self-Attention Mechanism: One key computational challenge in Transformer encoders comes from the self-attention step. Input feature maps with spatial dimensions H by W lead to a sequence length N that equals $H \times W$ Standard multi-head self-attention uses query Q, keys K, and values V matrices, all sized N by C, where C stands for the feature channels count. The output from self-attention follows this formula.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V, \tag{1}$$

Here, d_{head} means the size of each attention head. That setup brings a complexity of $\mathcal{O}(N^2)$, and it becomes really costly for high-resolution images that segmentation tasks often require. SF tackles this with Sequence Reduction Attention, or SRA, which reduces the sequence length for keys and values by a ratio R. People call R the reduction ratio. This significantly cuts down the computational load a lot, yet it retains the key contextual information for achieving good segmentation results. They accomplish this by reshaping and projecting K, just as the description below illustrates.

$$\hat{K} = \text{Reshape}\left(K, \frac{N}{R}, C \cdot R\right),$$
$$K_{\text{reduced}} = \text{Linear}\left(C \cdot R, C\right)(\hat{K}), \tag{2}$$

where $\text{Reshape}(\cdot)$ rearranges K into a smaller sequence, and $\text{Linear}(\cdot)$ is a learnable linear projection. This reduces the key sequence dimension from $N \times C$ to $\frac{N}{R} \times C$. Consequently, the computational intensity of the attention module decreases from $\mathcal{O}(N^2)$ to $\mathcal{O}\left(\frac{N^2}{R}\right)$. This reduction preserves the quality of feature representations while significantly lowering the processing requirements. In our implementation with MiT-B3, the reduction ratios are set $R = [64, 16, 4, 1]$ for the four encoding steps, respectively, as shown in Table 2. This structure effectively balances computational resources with the preservation of fine spatial details, enabling the model to successfully capture both the global context and local features required for accurate colonic polyp segmentation.

The streamlined self-attention part helps SF handle high-resolution endoscopic images. It keeps things suitable for real time, while also drawing precise polyp boundaries. These setups strike a good balance between efficient computation and preserving spatial details sharp. This makes it highly effective for accurate, real-time polyp segmentation in the colon.

Mix-FFN: MiT employs a Mix Feed-Forward Network (Mix-FFN) in place of positional encoding, which enables it to be more effective for the model to adapt across varying input sizes. Within the FFN block, this module consists of a 3×3 depthwise convolution. This facilitates the learnable, data-driven injection of spatial locality, which helps the network to better define fine boundaries in complex scenarios such as polyp areas.

3.3. Lightweight all-MLP decoder

The decoder in SF is designed to be compact and efficient, relying mainly on the rich hierarchical features extracted by the encoder. Due to its simplicity and reduced computational load, it is well-suited for real-time applications and deployment on resource-constrained hardware.

To begin with, multi-dimensional features F_i obtained from different encoder stages are projected into a unified channel dimension C through linear projection operators. This ensures that all feature maps, regardless of their original dimensions, are brought into a consistent representation space:

$$\hat{F}_i = \mathcal{P}_i(F_i), \qquad i = 1, \ldots, 4, \tag{3}$$

where $\mathcal{P}_i$ denotes the linear projection applied independently at each scale. Each projected feature map $\hat{F}_i$ is then upsampled to a fixed spatial size, typically $\frac{H}{4} \times \frac{W}{4}$, to maintain spatial alignment among features from different resolutions:

$$\tilde{F}_i = \mathcal{U}\left(\hat{F}_i\right), \qquad i = 1, \ldots, 4, \tag{4}$$

where $\mathcal{U}$ represents an interpolation-based upsampling operator. The spatially aligned multi-scale feature maps are concatenated along the channel dimension and then fused through another linear transformation to effectively aggregate information across different encoder levels:

$$F = \mathcal{F}\left(\tilde{F}_1 \parallel \tilde{F}_2 \parallel \tilde{F}_3 \parallel \tilde{F}_4\right), \tag{5}$$

with $\parallel$ denoting channel-wise concatenation. Finally, the fused feature representation F is mapped to the segmentation mask M with N_{cls} output channels, corresponding to the number of semantic classes:

$$M = \mathcal{S}(F). \tag{6}$$

Here, $\mathcal{S}$ denotes the segmentation prediction operator that converts the unified feature representation into class-wise logits. This streamlined decoder efficiently integrates multi-scale encoder features, enhances boundary localization, and improves segmentation performance, particularly in challenging low-contrast regions.

3.4. Suitability for colonic polyp segmentation

SF is highly suitable for colonic polyp segmentation for several important reasons, supported by both quantitative performance and qualitative observations. First, its Transformer-based encoder effectively captures global as well as local contextual information through long-range dependency modelling. This property is crucial for distinguishing polyps from visually similar mucosal regions. In our experiments, SF demonstrated improved boundary preservation and reduced false positives, particularly in cases where polyps exhibit low contrast against the surrounding tissue.

Second, the use of small patch sizes together with the hierarchical attention structure enables the detection of small, flat, or partially occluded polyps. Empirically, SF delivered higher recall on small-polyp subsets compared to CNN-based baselines, indicating its ability to retain fine-grained details that conventional architectures often miss.

Third, the decoder of SF is entirely MLP-based and lightweight, resulting in low computational overhead. In practice, we observed a significant reduction in inference time while maintaining competitive accuracy. This makes SF suitable for real-time clinical workflows such as colonoscopy, where timely feedback is essential.

Moreover, SF demonstrated strong generalization across different datasets and image resolutions without requiring fixed positional encodings. This flexibility is particularly relevant for medical imaging, where acquisition devices and imaging conditions vary widely. In cross-dataset evaluations, SF retained robust performance with minimal degradation, highlighting its adaptability to diverse clinical environments.

Taken together, these empirical findings confirm that SF is not only theoretically well-designed but also practically effective and reliable for colonic polyp segmentation across varying polyp sizes, shapes, and imaging conditions.

4. Experiment and analysis

4.1. Datasets

The efficiency and generalization capability of the introduced SF-based technique are measured comprehensively using four well-known and publicly accessible gastrointestinal polyp segmentation datasets: Kvasir-SEG [65], CVC-ClinicDB [66], ETIS-Larib [67], and Kvasir-Capsule-SEG [68]. These datasets are ideal for serving as a standard for assessing segmentation methods since they cover a wide range of imaging circumstances, resolutions, and polyp features. Each dataset is described in detail below:

Kvasir-SEG dataset: This dataset consists of 1000 polyp images collected during the colonoscopy process. The images vary in resolution from 720×576 to 1920×1072 pixels and are labelled with high-quality pixel-wise binary masks. The dataset featuring polyps with different sizes, shapes, and lighting conditions, offers a challenging environment for testing segmentation algorithms. It is one of the most commonly employed standard datasets in polyp segmentation research.

CVC-ClinicDB dataset: This dataset, which involves 612 images, was extracted from 29 different colonoscopy video sequences. Every image is offered at a set resolution of $\times$ 288 pixels. The dataset attributes an accurate binary label, and every image has at least one visible polyp. Because it is video-derived, it follows practical healthcare circumstances by including temporal duplication and slight motion blur.

ETIS-Larib dataset: This dataset comprises 196 high-resolution polyp images (1225×966 pixels) acquired in a healthcare setting. It contains tiny and flat polyps with precise texture differences from the surrounding mucosa, leading to increased segmentation difficulty. ETIS is often used to evaluate a model's ability to generalize to difficult and rare cases.

Kvasir-Capsule-SEG dataset: This dataset differs from usual colonoscopy ones, as it comes from WCE procedures. The set includes thousands of frames overall. More than 4000 of them carry detailed pixel-level masks for polyps. Every single frame comes at 336×336 pixels in size. Still, this dataset presents distinct hurdles. Motion blur appears often. Lighting varies significantly. Intestinal folds become complex too. All those elements align naturally with capsule imaging setups.

4.2. Dataset setup and preprocessing

To keep things consistent in how we evaluated everything, all the experiments followed an 80/10/10 split for training, validation, and testing sets. This applied across those four main benchmark datasets, Kvasir-SEG, CVC-ClinicDB, ETIS-Larib, and Kvasir-Capsule-SEG. Every input image was resized to 384×384 pixels. That way, it all worked smoothly with the SegFormer setup.

We put together a standard data augmentation pipeline for the process. It included multi-scale resizing, random cropping, photometric distortion, horizontal and vertical flipping, and small-angle random rotation. Those steps really help the model get better at generalizing. They do this by showing it all sorts of visual changes in polyp shapes, scales, and lighting conditions. When we did combined training, we merged all four datasets first. Then we applied the 80/10/10 split to the whole thing. This ensured no single dataset ended up in both the training and the unseen evaluation parts at the same time. We also fixed the random seed values for every run. That ensured the results remained reproducible across the board.

As detailed in Table 3, the data from each dataset was split with 80 % allocated for training, 10 % reserved for validation, and the final 10 % set aside for testing. Additionally, we conducted a combined training scenario where all four datasets were merged and trained in combination to assess the suggested method's resilience in greater detail.

As shown in Fig. 3, representative samples from each dataset are displayed, including the original endoscopic image, ground truth segmentation mask, and overlay results. These examples highlight variations in polyp appearance, from well-lit and clearly defined polyps in Kvasir-SEG to low-contrast and challenging cases in ETIS-Larib and Kvasir-Capsule.

4.3. Experimental setting and performance metrics

The experiments ran on a powerful PC with an Intel Core i9-14900 K CPU at 3.20 GHz. It had an NVIDIA GeForce RTX 3080 GPU with 10 GB VRAM too. The setup included 24 GB RAM and a 2 TB Samsung 990 PRO SSD for quick data pulls. Windows 10 64-bit served as the OS. Programming used Python 3.10, plus key libraries such as Matplotlib, NumPy, TensorFlow, OpenCV, and scikit-learn. These handled data preparation, model runs, and visuals. For pushing segmentation models forward, we used MMSegmentation version 1.2.2. It built on MMCV 2.1.0 and PyTorch 2.5.1 with CUDA 12.1. This setup provided smooth GPU boosted training and inference. It ensured top performance and compatibility with current deep learning tools.

4.4. Training setup

The model was trained using AdamW, with a weight decay of 0.01, base learning rates (LR) 6×10^{-5}, and betas $(0.9, 0.999)$. A layer-based learning rate modification was used, in which positional encoding blocks and normalization layers were not subject to weight decay, and the classification head also used a learning rate multiplier of 10 to speed up convergence. It employed a layer-wise learning rate style, and positional encoding blocks and normalization layers were not weight-decayed. The classification head used a learning rate that was 10 times lower, to expedite convergence. The LR schedule was designed as a linear warm up from 0.001 times the base LR to the full base rate during the first 500 iterations followed by a decay from 0 to full base rate in the last 39,500 iterations. The training was conducted over a total of 40,000 iterations in an iteration-based loop. The validation was performed after every 3000 iterations to evaluate the performance of the model and prevent overfitting.

Data were loaded in 4 batches at a time with 4 worker threads during training along with persistent workers and pinned memory to optimize the throughput of the GPU. To validate and test, a stable and efficient batch size of 1 and 2 worker threads was applied. The entire training hyperparameters applicable in the SF-based architecture are outlined in Table 4.

Table 3
Preprocessing and data split configuration across datasets.

Setting	Kvasir-SEG	CVC-ClinicDB	ETIS-Larib	Kvasir-Capsule-SEG	Combined training
Image Size	384×384	384×384	384×384	384×384	384×384
Augmentation	Random Resize Random Crop Photometric Distortion H/V Flip Random Rotation	Same	Same	Random Crop Same	Random Resize Photometric Distortion H/V Flip Random Rotation
Split (Train/Val/Test)	80/10/10	80/10/10	80/10/10	80/10/10	80/10/10 combined across all four datasets
Reproducibility	Fixed Seed	Fixed Seed	Fixed Seed	Fixed Seed	Fixed Seed

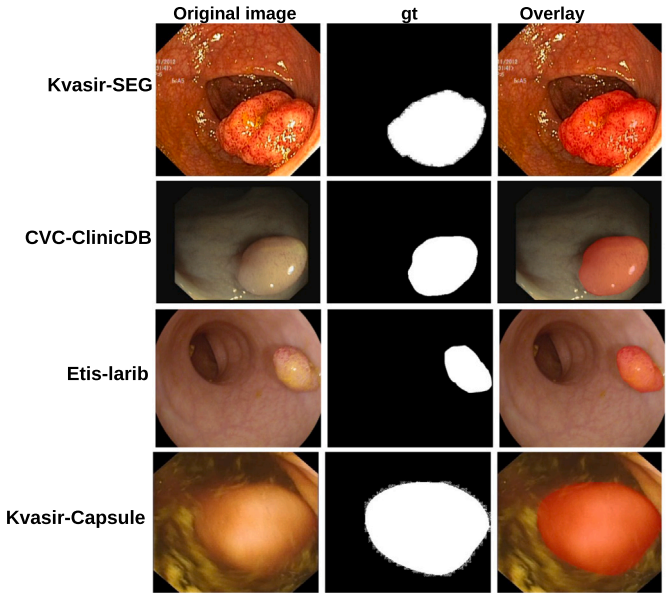


Fig. 3. Example images, ground truth (gt) masks, and overlay visualizations from the four datasets used: Kvasir-SEG, CVC-ClinicDB, ETIS-Larib, and Kvasir-Capsule. The first column shows original endoscopic images, the second column presents binary ground truth masks, and the third column displays the segmentation mask overlaid on the original image. This visual comparison highlights the diversity in polyp shape, size, and appearance across datasets.

Table 4
Hyperparameters used in training the proposed SegFormer-based model for polyp segmentation.

Hyper-parameters	Values
Backbone (all datasets)	SegFormer-B3
Image size	384×384
Patch size	16×16
Projection dimension (k)	64
Number of heads	8
Transformer units	[128, 64]
Transformer layers	13
Batch size	4 (train) / 1 (validation/test)
Precision	Automatic Mixed Precision (FP16)
Optimizer	AdamW
Base learning rate	6×10^{-5}
Learning rate schedule	Linear warm-up + Polynomial decay
Weight decay rate	0.01
Preprocessing	RGB conversion, normalization, flip, rotation, distortion
Post-processing	Mask resizing, overlay visualization
Inference speed (FPS)	Measured without dataloader time
Hardware	NVIDIA RTX 3080 (10 GB VRAM)
Number of iterations	40,000

Mixed-precision training (AMP, FP16) was enabled to improve computational efficiency and reduce GPU memory consumption on the NVIDIA GeForce RTX 3080. Each input image was resized to 384×384 pixels and normalized using ImageNet mean and standard deviation. Data preprocessing included RGB conversion, random horizontal flipping (probability = 0.5), random rotation (degree = 10), and photometric distortion. Post-processing involved resizing predictions to their original resolution and overlaying the segmentation mask for visualization. Inference speed (FPS) was computed by averaging over 100 single-image forward passes at 384×384 resolution, excluding dataloader time to isolate model inference performance.

Fig. 4 illustrates the complete training and evaluation pipeline. The SegFormer-B3 backbone was trained using the MMSegmentation framework. After training, the model weights were saved and evaluated using the same test split to assess performance. During inference, test-time augmentation (TTA) was applied to improve robustness, and post-processing ensured output mask alignment with the input image dimensions. Quantitative metrics, including the Dice coefficient, Intersection over Union (IoU), precision, recall, and FPS, were used for comprehensive evaluation.

4.5. Evaluation metrics

For the experimental evaluation, we use several important performance measures: accuracy, precision, recall, intersection over union (IoU), and the dice coefficient.

Dice Coefficient is a widely used performance measure for assessing the similarity between two sets, such as a predicted segmentation mask and the corresponding ground truth mask. It combines precision and recall by calculating their harmonic mean, providing a balanced measure of how accurately the prediction matches the baseline data. The F1 score, often known as the Dice coefficient, has the following mathematical definition:

$$\text{Dice} = \frac{2 \times (P \cap G)}{\|P\| + \|G\|} \tag{7}$$

Here, P represents the set of pixels predicted as positive, and G represents the set of actual positive pixels from the ground truth. The Dice coefficient can also be expressed in terms of true positives TP, false positives FP, and false negatives FN as:

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

The Dice coefficient ranges from 0 to 1, where 1 means perfect overlap between the predicted and true masks, and 0 means no overlap at all. This metric is fundamental in segmentation tasks where accurately outlining object boundaries is critical.

Intersection over Union (IoU) (also called the Jaccard Index) quantifies the overlap between the predicted polyp region and the ground truth by calculating the ratio of their intersection to their combined area. It is defined as:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{9}$$

Precision indicates the accuracy of positive predictions. It is the ratio of correctly predicted polyp pixels out of all pixels predicted as polyps. High precision means fewer false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall measures how well the model detects all actual polyp pixels. It is the ratio of correctly predicted polyp pixels out of all actual polyp pixels in the ground truth. High recall means fewer false negatives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

Accuracy evaluates the segmentation's overall correctness. It calculates the proportion of pixels correctly classified, including both polyp (positive) and non-polyp (negative) pixels. High accuracy means most pixels are labelled correctly.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{12}$$

In segmentation evaluation, True Positives TP represent pixels that are correctly detected as belonging to the polyp area. True Negatives TN represent the pixels that were correctly identified as belonging to non-polyp areas. On the other hand, False Positives FP are pixels inaccurately classified as polyp pixels, even though they belong to the background or non-polyp region. Similarly, False Negatives FN are polyp pixels that the model fails to detect, mistakenly labelling them as non-polyp. These four categories form the basis for calculating important performance metrics in segmentation tasks.

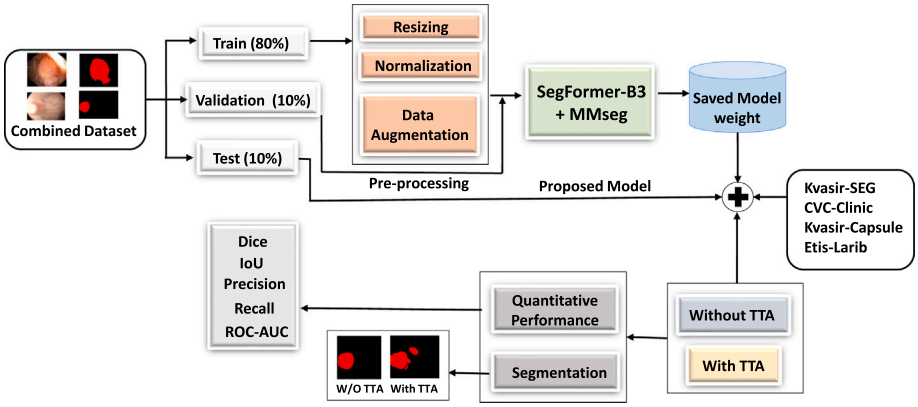


Fig. 4. The overall training and evaluation pipeline uses four benchmark polyp segmentation datasets: Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS-Larib. Each dataset undergoes preprocessing (resizing, normalization, and data augmentation), followed by training the SF-B3 model using the MMSegmentation framework. Quantitative performance is evaluated both on individual datasets and on a combined dataset created by merging all four. Post-processing with Test-Time Augmentation (TTA) is applied for enhanced segmentation performance.

4.6. Training performance

Fig. 5 presents the training progress of the SF-B3 model on the combined dataset. The plots illustrate both the evolution of the loss functions and the corresponding accuracy metrics throughout the training process. In the left panel of the figure, the total loss, is shown along with its components, Dice loss and Focal loss. The curves reveal a consistent downward trend, suggesting effective optimization and convergence of the hybrid loss function. The smooth decline in both Dice and Focal losses indicates that the model is simultaneously improving region-wise overlap and focusing on hard-to-classify pixels, as intended. The right panel shows the decode accuracy and auxiliary accuracy across training iterations. The decode accuracy refers to the model's primary output, while the auxiliary accuracy typically reflects the performance of intermediate layers (used for deep supervision). Both curves exhibit a steady upward trend, further validating stable learning and improved segmentation capability as training progresses. Overall, these plots confirm that the model trains reliably on the combined dataset, with no signs of divergence or instability.

4.7. Quantitative results

To learn the dataset-specific performance, we trained and validated four distinct SF-B3 models on single datasets (CVC-ClinicDB, Kvasir-SEG, Kvasir-Capsule-SEG, and ETIS-Larib) all at once to comprehend the dataset-specific performance of a particular model Fig. 7. All the models were trained with the same hyperparameters, augmentation techniques, and train validation test-split. A fifth model was trained on a mixed dataset (all four), and this along with other categories enables us to test the generalizability and strength across clinical states. This was implemented to provide a complete measure of generalization in clinical conditions across a variety of conditions as shown in Fig. 6. Dice loss, Focal loss as well as a combination of Dice and Focal loss were experimented with to determine the most effective loss function. The hybrid structure was found to offer the optimal trade-off between pixel-level and region-level accuracy. The joint model achieved Dice score of 0.969 and mIoU of 0.942, which demonstrates good segmentation results. Although the Dice loss alone marginally increased recall on Kvasir-SEG, the hybrid loss still yielded generally better results across several metrics, including F1-score and mIoU. Hence, Dice + Focal loss was chosen as the default setup in all the experiments. Of the single datasets, the model trained on ETIS-Larib had the lowest Dice and Focal losses (0.280 and 0.023, respectively), which may be explained by the smaller and less dispersed image distribution. Nevertheless, the overall performance was the highest when trained on the combined dataset: Dice Loss of 0.286, Focal Loss of 0.024, mIoU of 0.942 %, Precision of 0.950 %, and Recall of 0.941 %. These findings indicate the advantage of multi-source training in enhancing segmentation strength across various clinical environments. Further the inference speed was maintained

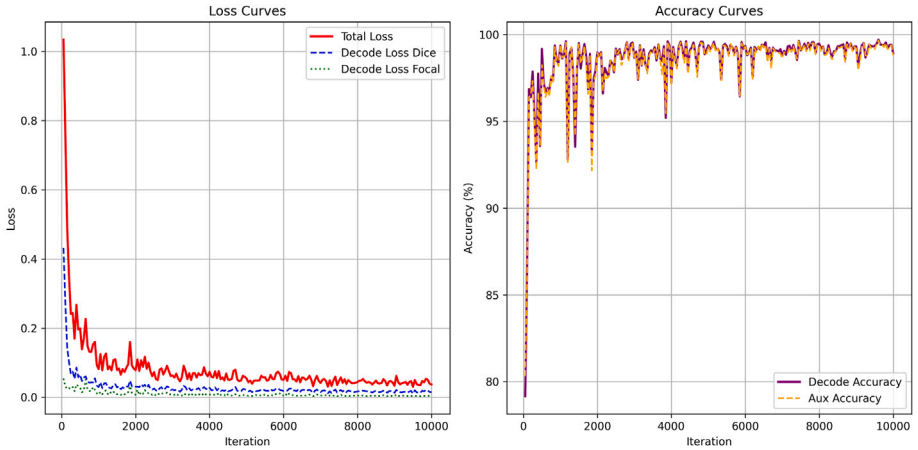


Fig. 5. Training loss and accuracy curves for the combined dataset. Left: total, Dice, and Focal loss curves. Right: Decode and auxiliary accuracy curves over iterations.

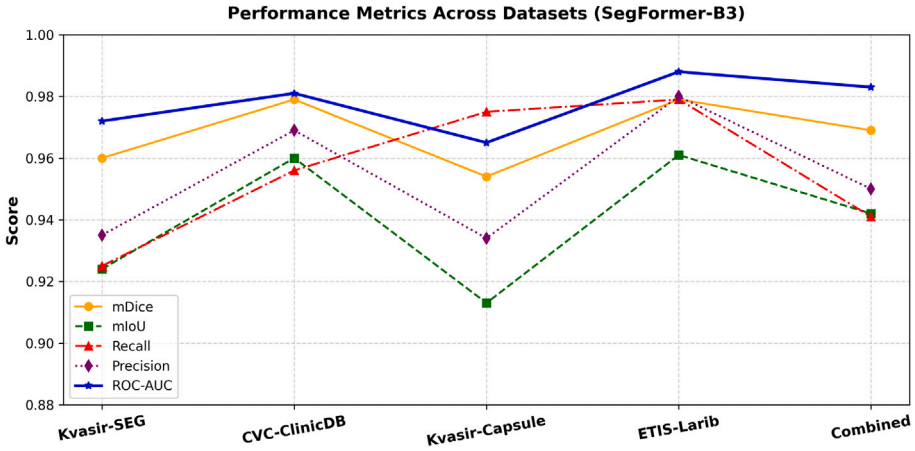


Fig. 6. Validation metric comparison across individual and combined datasets, including mDice, mIoU, precision, and recall.

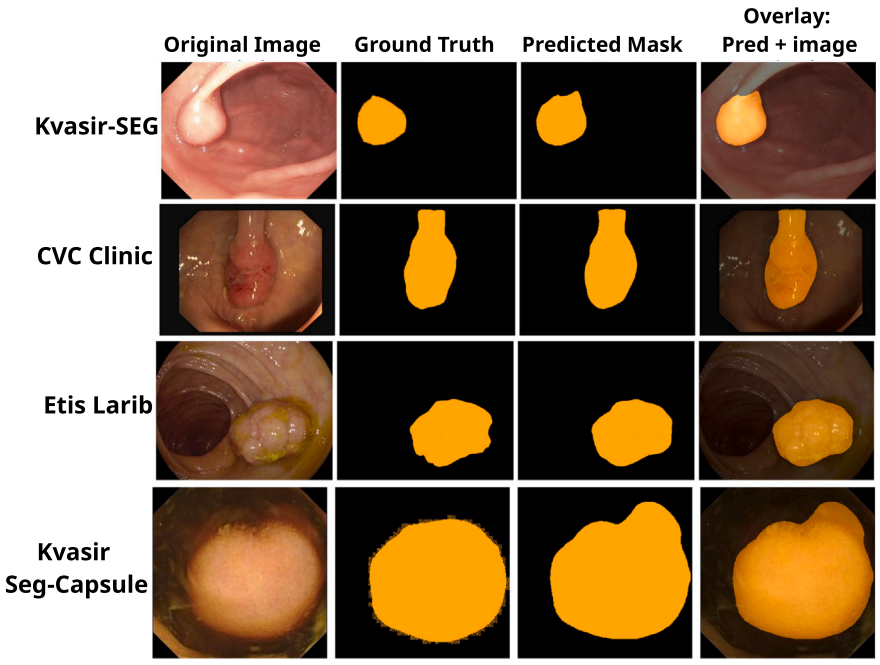


Fig. 7. Qualitative results of the SF-B3 model on polyp segmentation. From left to right: original image, ground truth mask, predicted mask, and overlay of the prediction on the original image.

Table 5
Performance comparison across individual datasets and the combined dataset using the SegFormer-B3 model.

Dataset	Dice Loss	Focal Loss	Precision (%)	Recall (%)	mIoU (%)	ROC-AUC (%)	FPS
Kvasir-SEG	0.295	0.025	0.935	0.925	0.924	0.972	69
CVC-ClinicDB	0.310	0.029	0.969	0.956	0.960	0.981	79
Kvasir-Capsule	0.301	0.027	0.934	0.975	0.913	0.965	90
ETIS-Larib	0.280	0.023	0.980	0.979	0.961	0.988	92
Combined	**0.286**	**0.024**	**0.950**	**0.941**	**0.942**	**0.983**	**94**

at the optimum and the combined model could reach 94 FPS, which is acceptable in terms of real-time clinical application.

The full analysis of the SF-B3 model performance on the separate and combined datasets is provided in Table 5. The model, which was trained on the combined dataset, performed better or was equally as good as the models that were trained on the individual datasets, which highlights its superior and stronger generalization and resistance to training through multi-source learning.

4.8. Computational efficiency and model complexity

To demonstrate the practicality of the proposed SF-B3 model, we also present more research on the model complexity, memory consumption, and inference performance (Table 6). SF-B3 model can inquire in real-time at 94 FPS, with moderate memory usage and computational energy, which implies that the model is resource-efficient and can be used in clinical practice. These metrics point to the fact that the model

Table 6
Model complexity and computational metrics for the proposed SF-B3 model.

Metric	Value
Model	SegFormer-B3 (MiT-B3 backbone)
Image Input Size	384×384
Parameters (M)	46.3 M
FLOPs (G)	~42.8 G (scaled from 52 G @384×384)
Training Time	~11.5 h (40 k iterations, RTX 3080)
Peak GPU Memory Usage	3415 MB
Average Training Accuracy	99.14 %
Average Validation Accuracy (mIoU)	94.27 %
Inference Time per Image	0.0106 s (≈94 FPS)

can operate on high-resolution endoscopic images in real-time, which is fundamental when implementing it in real-time in clinical colonoscopy systems.

The complexity regarding parameters and FLOPs is on the one hand moderate when compared to larger SegFormer variants (B4 and B5), but it also has high accuracy in segmentation. Such complexity/performance balance is aimed at the ability of the SF-B3 model to run on resource-limited hardware without compromising clinical utility. The highest memory usage of the GPU during the training process does not exceed 3.5 GB, which is why it is possible to use commonly available cards to perform the training with a NVIDIA RTX 3080. The fact that the training of the model took about 11.5 h, to run 40 k iterations, further reinforces the productivity of the model and makes the experimentation and fine-tuning of the model possible within sensible timeframes.

Also, the model is highly generalized across various benchmark datasets and still makes fast inferences which is essential in clinical applications where quick feedback is needed. The moderate computational requirements, high inference rate and good accuracy make the SF-B3 a convenient and scalable model to use for automated polyp detection in a variety of automated endoscopic imaging settings. The proposed SF-B3 is almost twice as fast as other state-of-the-art models like U-Net (67.9 GFLOPs, 48 FPS) and TransFuse (88.1 GFLOPs, 40 FPS) yet retains the similar or higher accuracy rates, which proves the fact that it is lightweight and designed for real-time use.

4.9. Comprehensive ablation study

To thoroughly evaluate the proposed SF-B3 model, we conducted extensive ablation experiments on the combined polyp segmentation dataset, covering SegFormer variants, loss functions, and augmentation strategies. We analyzed SegFormer variants B2–B5 to examine the trade-offs among segmentation accuracy, computational cost (FLOPs), and inference speed (FPS). Table 7 summarizes the parameters, FLOPs, and performance metrics for each variant. While larger models (B4, B5) achieved slightly higher mDice and mIoU, the improvements were marginal compared to B3, whereas FLOPs and memory usage increased substantially, and FPS decreased. This justifies the selection of B3 as the optimal backbone, balancing high segmentation accuracy with real-time efficiency suitable for clinical deployment.

We evaluated Dice, Focal, and Hybrid Dice + Focal loss functions. As shown in Table 8, the hybrid loss consistently achieved the best overall performance across SegFormer variants, capturing both region-level and boundary-level accuracy while maintaining a balance between precision and recall. For example, in B3, Dice-only achieved mDice 0.969 and mIoU 0.942, whereas hybrid achieved mDice 0.968 and mIoU 0.943, with slightly higher precision and balanced recall. This demonstrates that the hybrid loss is preferred when optimizing both segmentation accuracy and boundary delineation.

Table 7
SegFormer Variants: parameters, FLOPs, and performance trade-offs for input size 384 × 384.

Model	Parameters (M)	FLOPs (G)	FPS	mDice	mIoU	Input size
SegFormer-B2	27.0	33.0	75	0.957	0.931	384 × 384
SegFormer-B3	46.3	52.0	94	**0.960**	**0.935**	384 × 384
SegFormer-B4	64.0	81.0	68	0.962	0.936	384 × 384
SegFormer-B5	86.7	131.0	54	0.963	0.937	384 × 384

Table 8
Comparison of loss functions across SegFormer variants (combined dataset, decimal format). The highlighted color show the highest performance of our proposed model.

Model	Loss	mDice	mIoU	Precision	Recall
B2	Dice	0.955	0.929	0.930	0.941
B2	Focal	0.953	0.927	0.928	0.939
B2	Hybrid	0.957	0.931	0.931	0.942
B3	Dice	0.969	0.942	0.941	0.950
B3	Focal	0.966	0.940	0.939	0.948
B3	**Hybrid**	**0.970**	**0.943**	**0.972**	**0.968**
B4	Dice	0.960	0.934	0.935	0.947
B4	Focal	0.961	0.935	0.936	0.947
B4	Hybrid	0.962	0.936	0.937	0.948
B5	Dice	0.961	0.935	0.936	0.948
B5	Focal	0.962	0.936	0.937	0.948
B5	Hybrid	0.963	0.937	0.938	0.949

4.9.1. Impact of test-time augmentation (TTA) on model performance

In order to further assess the robustness of the model, we examined the impact of TTA (test-time augmentation), i.e., multi-scale resizing and horizontal flipping, on mDice and mIoU scores on a dataset and the aggregate dataset. Table 9 demonstrates mDice and mIoU scores with and without TTA on a dataset and the combined dataset.

The findings reveal that TTA only results in slight improvement (0.001–0.005 in mDice and mIoU), which means that the model is already a powerful one. This is due to the various augmentations used during training, such as multi-scale resizing, random cropping, photometric distortion, flipping, and small-angle rotation, which support the model in being generalized to differences in scale, orientation, and lighting. This means that the extra value of TTA in inference is marginal, so it is optional and does not substantially affect the speed at which predictions can be made, which is a major benefit when clinical or real-time prediction is required.

Lastly, we evaluated the contribution made by individual data augmentation ingredients by choosing to remove random cropping, photometric distortion, horizontal/vertical flipping, and small-angle rotation. The elimination of any of these augmentations resulted in a significant reduction in segmentation performance especially for small or low-contrast polyps, which emphasizes the significance of the complete augmentation pipeline. Altogether, these ablation experiments justify the choice of SegFormer-B3 as the backbone, the hybrid Dice + Focal loss, and the augmentation plan suggested, showing the best balance

Table 9
Evaluation with and without Test-Time Augmentation (TTA) across test datasets. TTA gains are minor due to diverse training augmentations, highlighting model robustness and making TTA optional for faster inference.

Dataset	TTA	mDice (%)	mIoU (%)
Kvasir-SEG	No	0.960	0.924
	Yes	**0.97**	**0.925**
CVC-ClinicDB	No	0.979	0.960
	Yes	**0.980**	**0.965**
ETIS-Larib	No	0.979	0.961
	Yes	**0.981**	**0.964**
Kvasir-Capsule	No	0.954	0.913
	Yes	**0.956**	**0.920**

Table 10
Effect of different loss functions on combined dataset performance (decimal format).

Loss Function	mDice	mIoU	Precision	Recall
Dice	0.969	0.942	0.941	0.950
Focal	0.966	0.940	0.939	0.948
Dice + Focal (Hybrid)	0.968	0.943	0.940	0.949

between the success of the segmentation, efficiency, and robustness factors to be implemented in clinical practice.

4.9.2. Loss function comparison

Table 10 indicates that the hybrid loss delivered highest mDice and mIoU, and thus represented the best trade-off between region-level and boundary-level segmentation accuracy. Loss based on Dice only produced slightly improved recall but at the cost of the other metrics, especially on harder datasets. In order to build the hybrid loss, we weighted Dice and Focal losses by a weighted sum as follows:

$$\mathcal{L}_{\text{Hybrid}} = \alpha \cdot \mathcal{L}_{\text{Dice}} + \beta \cdot \mathcal{L}_{\text{Focal}} \tag{13}$$

where the weights are set to $\alpha = 0.5$ and $\beta = 0.5$, giving equal importance to both losses in the hybrid formulation, as previously described in hybrid loss studies [81]. This is a composite method combining the sensitivity of Dice loss to the region-overlap with the hard-sample-focused Focal loss. The Dice component is a better class imbalance control method than the Focal component, as it maximizes the spatial overlap, whereas the Focal component focuses on learning challenging pixels especially on object boundaries.

Nonetheless, this approach is not free of limitations. Equal weighting might not be the best with all datasets and it may be necessary to tune the best performance of a dataset by adjusting the α and β. In addition, Focal loss also raises the computational load and can decrease the speed of training convergence. Nevertheless, the hybrid loss exhibits healthy and balanced performance, especially when dealing with small and irregularly shaped objects in the process of complex segmentation.

Furthermore, this method possesses certain drawbacks. Equal weights may not necessarily be the best choice for all datasets, and the parameters alpha and beta can be adjusted to achieve the best results. Moreover, Focal loss would not only delay the convergence of training but also increase the cost of resources. Despite these, the hybrid loss method is quite fair and has high performance especially in challenging segmentation problems with small or irregular objects.

4.10. Comparison with state-of-the-art methods

This section provides a comprehensive comparison of the suggested SF-based polyp segmentation model and a wide variety of the cutting-edge proposed methods that have been developed over the last ten years (2015-2025). The evaluation includes traditional CNN and hybrid CNN models that combine convolutional and transformer modules such as TransUNet or TransFuse as well as some transformer-based or transformer-based-enhanced models like NASF, PolypSegTrack, and MANet.

The evaluation is carried out on the five popular publicly available benchmark datasets Kvasir-SEG, CVC-ClinicDB, Kvasir-Capsule, ETIS-Larib, and on a Combined Dataset comprising of the four previously cited datasets. The challenges presented by every dataset are different in terms of polyp appearance, size, shape, and imaging conditions, and it will serve as a comprehensive platform to evaluate model robustness and generalization performance.

A graphical comparison of the results of segmentation is given in Fig. 8, demonstrating the performance of various models on the sample images of four popular polyp datasets. It provides the original endoscopic images, as well as the related ground truth masks and segmentation predictions of various approaches, such as conventional CNNs, hybrid networks, and transformer-based networks. The SF-B3 model is one of them, and it has been shown to have accurate and clean polyp segmentation and is capable of picking up fine details in changing imaging conditions. The results of these models are summarized in Table 11 with four common segmentation metrics, which are mDice, mIoU, recall, and precision. These measures are all evaluations of how well the estimated polyp masks model the ground truth, with higher values representing performance. Some of the observations are as follows:

Superior Performance of SegFormer-based Models: Our SegFormer variant B3, which uses a hierarchical transformer backbone

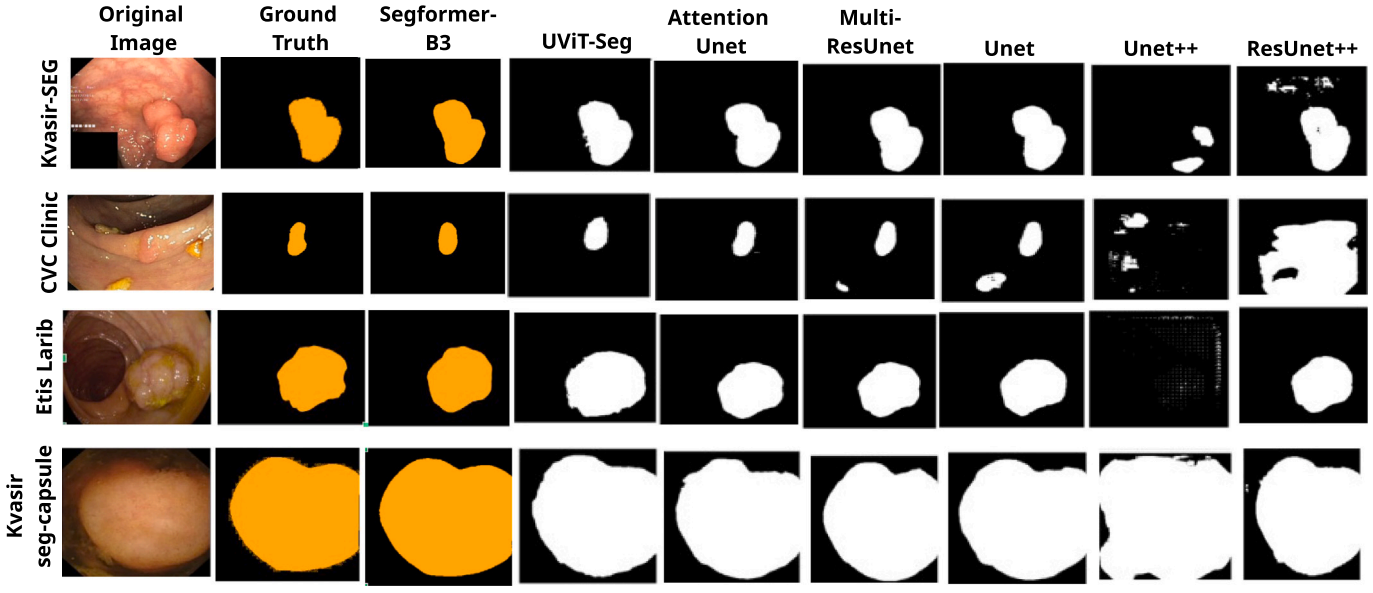


Fig. 8. Polyp segmentation outcomes from four benchmark datasets (Kvasir-SEG, CVCClinicDB, ETIS-Larib, and Kvasir-Capsule) are qualitatively compared. Only those models and datasets that were commonly available across all four datasets are included for a fair and consistent comparison. The visualizations include the original images, ground truth masks, and predicted masks from several cutting-edge models, encompassing classical CNNs (Unet [69], UNet++ [70], ResUNet [71]), hybrid architectures (Attention Unet [73], MultiResUnet [72]), and transformer-based approaches (UviT-Seg [76], SegFormer-B3).

Table 11
Comparison of polyp segmentation performance (mDice, mIoU, recall, and precision) across multiple models and datasets, including recent 2015–2025 methods and combined datasets.

Dataset	Method	Year	mDice	mIoU	Recall	Precision
Kvasir-SEG	Unet [69]	2015	0.818	0.781	0.827	0.786
	UNet++ [70]	2019	0.803	0.716	0.833	0.716
	ResUNet [71]	2018	0.803	0.759	0.820	0.772
	MultiResUnet [72]	2020	0.821	0.797	0.826	0.804
	AttentionUnet [73]	2018	0.815	0.786	0.804	0.799
	HarDNet-MSEG [74]	2021	0.912	0.857	–	–
	TransUNet [52]	2021	0.913	0.857	–	–
	TransFuse-L [60]	2021	0.920	0.870	0.921	0.927
	ColonFormer-S [61]	2022	0.924	0.875	0.900	0.910
	Polyp-PVT [62]	2024	0.917	0.864	–	–
	IECFNet [75]	2024	0.907	0.856	0.905	0.917
	ASPS [63]	2024	0.920	0.858	0.821	0.814
	UviT-Seg [76]	2024	0.835	0.801	–	–
	HiFiSeg [77]	2024	0.933	0.886	0.925	0.960
	NASegformer [78]	2024	0.943	0.920	0.905	0.902
	MANet [79]	2025	0.911	0.854	0.905	0.902
	PolypSegTrack [80]	2025	0.947	0.910	**0.970**	**0.980**
	SF-B3 (Ours)	**2025**	**0.960**	**0.924**	0.958	0.961
CVC-ClinicDB	Unet [69]	2015	0.799	0.796	0.812	0.808
	UNet++ [70]	2019	0.786	0.745	0.801	0.788
	ResUNet [71]	2018	0.857	0.840	0.864	0.840
	MultiResUnet [72]	2020	0.879	0.909	0.871	0.904
	AttentionUnet [73]	2018	0.784	0.812	0.834	0.821
	HarDNet-MSEG [74]	2021	0.932	0.882	–	–
	TransUNet [52]	2021	0.935	0.887	–	–
	TransFuse-L [60]	2021	0.942	0.897	–	–
	ColonFormer-S [61]	2022	0.948	0.904	0.950	0.941
	Polyp-PVT [62]	2024	0.937	0.889	0.905	0.915
	IECFNet [75]	2024	0.924	0.873	–	–
	ASPS [63]	2024	0.951	0.914	0.900	0.910
	UviT-Seg [76]	2024	0.907	0.902	–	–
	HiFiSeg [77]	2024	0.942	0.897	–	–
	MANet [79]	2025	0.922	0.863	0.915	0.918
	PolypSegTrack [80]	2025	0.833	0.769	0.917	0.888
	SF-B3 (Ours)	**2025**	**0.979**	**0.960**	**0.956**	**0.969**
ETIS-Larib	Unet [69]	2015	0.817	0.781	0.827	0.786
	UNet++ [70]	2019	0.803	0.716	0.833	0.716
	ResUNet [71]	2018	0.808	0.759	0.820	0.772
	MultiResUnet [72]	2020	0.821	0.797	0.826	0.804
	AttentionUnet [73]	2018	0.815	0.786	0.804	0.799
	TransUNet [52]	2021	0.731	0.660	–	–
	TransFuse-L [60]	2021	0.737	0.663	–	–
	HarDNet-MSEG [74]	2021	0.677	0.613	–	–
	Polyp-PVT [62]	2024	0.787	0.706	–	–
	IECFNet [75]	2024	0.707	0.444	0.701	0.750
	ASPS [63]	2024	0.861	0.769	0.914	0.919
	UviT-Seg [76]	2024	0.913	0.913	0.771	0.768
	MANet [68]	2025	0.776	0.690	0.938	0.942
	PolypSegTrack [80]	2025	0.914	0.853	0.938	0.942
	SF-B3 (Ours)	**2025**	**0.979**	**0.961**	**0.979**	**0.980**
Kvasir-Capsule	Unet [69]	2015	0.528	0.474	0.543	0.480
	UNet++ [70]	2018	0.428	0.367	0.426	0.465
	ResUNet [71]	2019	0.514	0.446	0.531	0.460
	AttentionUnet [73]	2018	0.526	0.486	0.535	0.501
	MultiResUnet [72]	2020	0.509	0.443	0.524	0.463
	UviT-Seg [76]	2024	0.537	0.491	0.541	0.500
	NASegformer [78]	2024	0.827	0.692	0.978	0.716
	SF-B3 (Ours)	**2025**	**0.972**	**0.946**	**0.982**	**0.963**

on top of an efficient self-attention implementation, consistently outperforms all baseline and recent SOTA models across all datasets and metrics. One such example is on the Kvasir dataset, where SF-B3 attains an astounding mDice of 0.950, which is significantly higher in comparison to PolypSegTrack 0.947 and MANet 0.911. The trend is also consistent in CVC-ClinicDB and other datasets.

Effectiveness Across Diverse Data: The generalization capability of our model is highlighted by the combined dataset results. In the real world, the deployment requires resiliency to diverse image sources, types of polypys and the high scores of our model (mDice of 0.969, mIoU of 0.942) suggest that our model can use heterogeneous data to learn better.

Transformer-based Architectures Leading the Field: It has been well developed over the years, with more conventional convolution-based networks like UNet (2015) and UNet++ (2018) showing moderate performance, and hybrid and fully transformer-based ones beginning to improve in quality of segmentation results, as published in 2021. This underlines the increased effectiveness of the

attention mechanisms and global context modelling in medical image segmentation [82].

High Recall and Precision Balance: The models perform well in identifying real positives and minimizing false positives as evidenced by their high recall and precision balance in addition to segmentation accuracy. This becomes critical in clinical practices where false alarms or missed polyps can have a significant impact on the diagnosis and outcome of patients. The performance of our model is superior to any existing method, particularly on the pooled dataset, which can be attributed to its capability to learn from a variety of training samples and ensure a high level of generalization across a variety of real-world datasets.

5. Discussion

Limitations and Failure Cases: Although the proposed SF-B3 model shows good performance and results across a wide range of datasets for polyp segmentation, it has significant limitations and failure modes to consider. Specifically, the model is sometimes non-performing when the polyps are subjected to extreme artifacts including, motion blur, mirrored lights, or poor lighting. These artifacts may confuse the boundaries of polypes causing partial segmentation or slight misclassification. Likewise, polyps in anatomically challenging areas, e.g., close to folds or sharp angles in the colon, are difficult to delineate accurately. Reduced segmentation accuracy may also be caused by small polyps, flat lesions, and low-contrast lesions when compared to surrounding tissue even though the model is generally robust.

Clinical Relevance: From a clinical perspective, these limitations highlight the need for cautious interpretation of automated segmentation findings, especially in high-stakes diagnostic scenarios. Missing or partially segmented polyps can impact patient outcomes if used as a standalone diagnostic tool. However, the model maintains high recall and precision overall, and the incorporation of diverse training augmentations, multi-scale features, and hybrid loss functions helps mitigate many common challenges in polyp detection.

Architectural Advantages of SegFormer-B3: In addition to the performance gains as indicated above, SegFormer-B3 architecture owes its success to a number of key design elements. The hierarchical MiT encoder is effective in capturing multi-scale contextual features, which are useful in accurately segmenting polyps of different sizes, shapes, and textures. Its lightweight and effective self-attention system summarizes global context with minimal computational cost, which is essential in the detection of small polyps or low-contrast polyps. The MLP-based decoder is an effective multi-level combination of encoder multi-level features without losing fine-grained boundary data. Moreover, the design of SegFormer-B3 is more efficient in nature compared to other transformer-based models like TransUNet and ColonFormer-S which either use a heavier attention mechanisms or more involved encoder–decoder interactions. Together, these architectural advancements are why SegFormer-B3 not only achieves higher metric scores but is also robust and generalizes across a variety of datasets and challenging clinical cases, ensuring it can be practically applicable in real-world colonoscopy.

Generalizability to Other Medical Segmentation Tasks: This paper concentrates more on the process of colonic polyp segmentation but the design principles of the SF-B3 model can be easily applied to the other medical image segmentation processes. It is very versatile in anatomical imaging of the liver, kidney or pancreas, being able to segment a wide range of anatomical structures using a hierarchical transformer encoder, a hybrid loss function, and efficient feature aggregation. The ability of the model to balance contextual knowledge on the global scale and boundary precision on a fine scale is especially useful in complex areas that have ambiguous or subtle organ boundaries. In addition, hybrid loss (Dice + Focal) is introduced to increase the resistance to class imbalance and small lesion segmentation, which are typical challenges in medical imaging outside of endoscopy. The framework may be used in future work, where it might be modified to be domain-general and utilize domain-generalization strategies to apply knowledge learned on one dataset to another dataset of different modalities (e.g., CT, MRI, and ultrasound). This kind of extension would demonstrate the greater extent to which the proposed model may be applied, achieving the long-term aim of creating a single and effective segmentation strategy for heterogeneous medical imaging tasks.

Future Work and Improvement Opportunities: The constraints also bring about areas of improvement. Future research may involve artifact-conscious training methods to deal with motion blur, specular reflection and low light, attention systems to address with difficult body parts or multi-task training where segmentation and polyp classification are performed simultaneously. Furthermore, patient-specific datasets and real-time feeds on endoscopies can also be adjusted to increase clinical reliability. Lastly, the SF-B3 model can be implemented effectively, but it needs to be combined with endoscopic hardware and software pipelines, such as latency optimization and user-interface-related considerations to implement it in actual clinical workflows.

All in all, even though the failure cases are identified, the SF-B3 framework provides a solid and realistic approach to automated polyp detection, and it has been shown to generalize across different datasets. The identification of its limitations and how they can be mitigated is the key to effective integration of the model into clinical practice which will assist physicians in the screening and diagnosis of colorectal cancer.

5.1. Implications

The superior performance of our SF-based model suggests that the integration of efficient transformers into medical image segmentation frameworks can significantly enhance polyp detection accuracy. The results also indicate that our architectural design effectively captures multi-scale features and contextual information, enabling precise delineation of polyps that vary widely in size, shape, and texture. Moreover, the consistent gains across datasets with varying characteristics demonstrate the robustness of our approach, making it a promising candidate for real-world clinical deployment. This robustness is especially important when models trained on one dataset need to generalize well to others without extensive retraining or domain adaptation.

5.2. Future directions

Building on these promising results, future work could explore further model compression and acceleration techniques to enable real-time deployment on resource-constrained devices such as endoscopy systems. Additionally, extending the model to handle polyp classification and risk stratification jointly with segmentation could provide a more comprehensive diagnostic tool. Other potential directions include incorporating artifact-aware training strategies to handle motion blur, specular reflections, and poor illumination, as well as attention mechanisms tailored to anatomically complex regions. Multi-task or multimodal learning approaches that integrate segmentation with other imaging modalities, such as chromoendoscopy or optical coherence tomography, could further enhance diagnostic accuracy. Finally, implementing continual learning for patient-specific adaptation and developing user-friendly clinical interfaces will be essential to facilitate safe and efficient adoption in real-world clinical workflows.

6. Conclusion

This study presented an efficient and lightweight SegFormer-B3 (SF-B3)–based framework for medical image segmentation using the MMSegmentation platform. The proposed model demonstrated strong capability in accurately segmenting challenging polyp cases, including small, flat, and low-contrast regions that are frequently missed by conventional CNN-based methods. Experimental evaluations confirmed that the SF-B3 framework achieves high segmentation accuracy while maintaining low computational complexity and real-time inference efficiency. The hierarchical transformer structure and hybrid Dice–Focal

loss design contributed to enhanced feature representation and improved generalization across multiple datasets. The proposed framework shows potential for extension to other medical image segmentation tasks, such as organ and lesion delineation in CT and MRI scans. Overall, SF-B3 offers a promising balance between accuracy, efficiency, and adaptability, making it a viable candidate for integration into clinical workflows. Future work will focus on optimizing hardware deployment and expanding the framework toward multimodal and multi-task learning for comprehensive medical image analysis.

CRediT authorship contribution statement

Gul E Arzu: Writing – original draft, Methodology, Investigation, Conceptualization. **Muhammad Fayaz:** Visualization, Data analysis. **Usman Ali:** Writing – review & editing, Investigation. **L. Minh Dang:** Writing – review & editing, Formal analysis. **Hyeonjoon Moon:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

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.

Acknowledgements

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Technology Commercialization Support Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2025-02218444), and by the “Regional Innovation System & Education (RISE)” through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2025-RISE-01–019-04) and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Training Global Talent for Copyright Protection and Management of On-Device AI Models, Project Number: RS-2025-02221620, Contribution Rate: 100 %).

Data availability

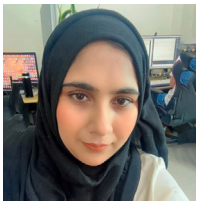
Data will be made available on request.

References

[1] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer Statistics, 2023, CA: Cancer J. Clin. 73 (1) (2023).
[2] R.L. Siegel, N.S. Wagle, A. Cercek, R.A. Smith, A. Jemal, Colorectal Cancer statistics, 2023, CA. Cancer J. Clin. 73 (3) (2023) 233–254.
[3] Y. Hazewinkel, E. Dekker, Colonoscopy: basic principles and novel techniques, Nat. Rev. Gastroenterol. Hepatol. 8 (10) (2011) 554–564.
[4] R.G. Holzheimer, J.A. Mannick, Surgical Treatment: Evidence-Based and Problem-Oriented, Zuckschwerdt, 2001.
[5] C.V. Tranquillini, W.M. Bernardo, V.O. Brunaldi, E.T.D. MOURA, S.B. Marques, E.G.H.D. MOURA, Best polypectomy technique for small and diminutive colorectal polyps: a systematic review and meta-analysis, Arq. Gastroenterol. 55 (2018) 358–368.
[6] G. Costamagna, S.K. Shah, M.E. Riccioni, F. Foschia, M. Mutignani, V. Perri, A. Vecchioli, M.G. Brizi, A. Picciocchi, P. Marano, A prospective trial comparing small bowel radiographs and video capsule endoscopy for suspected small bowel disease, Gastroenterology 123 (4) (2002) 999–1005.
[7] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, Wireless capsule endoscopy, Nature 405 (6785) (2000) 417.
[8] T. Omori, T. Hara, S. Sakasai, H. Kambayashi, S. Murasugi, A. Ito, S. Nakamura, K. Tokushige, Does the Pillcam SB3 capsule endoscopy system improve image reading efficiency irrespective of experience? A pilot study, Endosc. Int. Open. 6 (6) (2018) E669–E675.
[9] S. Gross, T. Stehle, A. Behrens, R. Auer, T. Aach, R. Winograd, C. Trautwein, J. Tischendorf, A comparison of blood vessel features and local binary patterns for colorectal polyp classification, in: Medical Imaging 2009: Computer-Aided Diagnosis, vol. 7260, SPIE, 2009, pp. 758–765.
[10] Y. Iwahori, A. Hattori, Y. Adachi, M.K. Bhuyan, R.J. Woodham, K. Kasugai, Automatic detection of polyp using hessian filter and hog features, Proc. Comput. Sci. 60 (2015) 730–739.
[11] A. Amber, Y. Iwahori, M.K. Bhuyan, R.J. Woodham, K. Kasugai, Feature point based polyp tracking in endoscopic videos, in: 2015 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence, IEEE, 2015, pp. 299–304.
[12] P. Sasmal, M.K. Bhuyan, Y. Iwahori, K. Kasugai, Colonoscopic polyp classification using local shape and texture features, IEEE Access 9 (2021) 92629–92639.
[13] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, P. Halvorsen, Deep learning and hand-crafted feature based approaches for polyp detection in medical videos, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2018, pp. 381–386.
[14] M. Hmoud Al-Adhaileh, E. Mohammed Senan, F.W. Alsaade, T.H.H. Aldhyani, N. Alsharif, A. Abdullah Alqarni, M.I. Uddin, M.Y. Alzahrani, E.D. Alzain, M.E. Jadhav, Deep learning algorithms for detection and classification of gastrointestinal diseases, Complexity 2021 (1) (2021) 6170416.
[15] D. Jha, S. Ali, N.K. Tomar, H.D. Johansen, D. Johansen, J. Rittscher, M.A. Riegler, P. Halvorsen, Real-time polyp detection, localization and segmentation in colonoscopy using deep learning, IEEE Access 9 (2021) 40496–40510.
[16] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, P. Baldi, Deep learning localizes and identifies polyps in real time with 96 % accuracy in screening colonoscopy, Gastroenterology 155 (4) (2018) 1069–1078.
[17] S. Jain, A. Seal, A. Ojha, A convolutional neural network with meta-feature learning for wireless capsule endoscopy image classification, J. Med. Biol. Eng. 43 (4) (2023) 475–494.
[18] N. Goel, S. Kaur, D. Gunjan, S.J. Mahapatra, Dilated CNN for abnormality detection in wireless capsule endoscopy images, Soft Comput. 26 (3) (2022) 1231–1247.
[19] 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.
[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
[21] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[22] W. Liu, A. Rabinovich, A.C. Berg, Parsenet: Looking wider to see better, arXiv preprint arXiv:1506.04579, 2015.
[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062, 2014.
[24] Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 173–190.
[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, Segformer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[29] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H.S. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[30] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272.
[31] L. Li, W. Wang, Y. Yang, Logicseg: parsing visual semantics with neural logic learning and reasoning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4122–4133.
[32] T. Zhou, W. Wang, E. Konukoglu, L. Van Gool, Rethinking semantic segmentation: a prototype view, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2582–2593.
[33] L. Li, T. Zhou, W. Wang, J. Li, Y. Yang, Deep hierarchical semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1246–1257.
[34] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, L. Van Gool, Exploring cross-image pixel contrast for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7303–7313.
[35] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-aware transformers for skin lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 206–216.
[36] J.-H. Lee, D.-H. Kim, S.-N. Jeong, S.-H. Choi, Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm, J. Dent. 77 (2018) 106–111.
[37] J.-H. Lee, S.-S. Han, Y.H. Kim, C. Lee, I. Kim, Application of a fully deep convolutional neural network to the Automation of tooth segmentation on panoramic radiographs, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 129 (6) (2020) 635–642.

[38] Z. Shen, H. Yang, Z. Zhang, S. Zheng, Automated kidney tumor segmentation with convolution and transformer network, in: International Challenge on Kidney and Kidney Tumor Segmentation, Springer, 2021, pp. 1–12.
[39] X. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, J. Si, Real-time gastric polyp detection using convolutional neural networks, PloS One 14 (3) (2019) e0214133.
[40] M. Misawa, S.-E. Kudo, Y. Mori, T. Cho, S. Kataoka, A. Yamauchi, Y. Ogawa, Y. Maeda, K. Takeda, K. Ichimasa, et al., Artificial intelligence-assisted polyp detection for colonoscopy: initial experience, Gastroenterology 154 (8) (2018) 2027–2029.
[41] H.A. Qadir, Y. Shin, J. Bergsland, I. Balasingham, Accurate real-time polyp detection in videos from concatenation of latent features extracted from consecutive frames, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, pp. 2461–2466.
[42] A. Krenzer, A. Hekalo, F. Puppe, Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks, in: EndoCV@ ISBI, 2020, pp. 58–63.
[43] X. Guo, Z. Chen, J. Liu, Y. Yuan, Non-equivalent images and pixels: confidence-aware resampling with meta-learning mixup for polyp segmentation, Med. Image Anal. 78 (2022) 102394.
[44] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S.M.R. Soroushmehr, N. Karimi, S. Samavi, K. Najarian, Polyp segmentation in colonoscopy images using fully convolutional network, in: 2018 40th Annual International Conference of the Ieee Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 69–72.
[45] H.A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, I. Balasingham, Polyp detection and segmentation using mask r-CNN: does a deeper feature extractor CNN always perform better? in: 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), IEEE, 2019, pp. 1–6.
[46] 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, Springer, 2015, pp. 234–241.
[47] 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: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 263–273.
[48] A. Galdran, G. Carneiro, M.A.G. Ballester, Double encoder-decoder networks for gastrointestinal polyp segmentation, in: International Conference on Pattern Recognition, Springer, 2021, pp. 293–307.
[49] 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.
[50] F. Younas, M. Usman, W.Q. Yan, A deep ensemble learning method for colorectal polyp classification with optimized network parameters, Appl. Intell. 53 (2) (2023) 2410–2433.
[51] W.S. Liew, T.B. Tang, C.-H. Lin, C.-K. Lu, Automatic colonic polyp detection using integration of modified deep residual convolutional neural network and ensemble learning approaches, Comput. Methods Programs Biomed. 206 (2021) 106114.
[52] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, 2021.
[53] B. Xiao, J. Hu, W. Li, C.-M. Pun, X. Bi, Ctnet: contrastive transformer network for polyp segmentation, IEEE Trans. Cybern. 54 (9) (2024) 5040–5053.
[54] D.L. Ramos, H.J. Hortua, Deep Bayesian segmentation for colon polyps: well-calibrated predictions in medical imaging, Biomed. Signal Process. Control 104 (2025) 107383.
[55] C.Y. Eu, T.B. Tang, C.-H. Lin, L.H. Lee, C.-K. Lu, Automatic polyp segmentation in colonoscopy images using a modified deep convolutional encoder-decoder architecture, Sensors 21 (16) (2021) 5630.
[56] M. Tharwat, N.A. Sakr, S. El-Sappagh, H. Soliman, K.-S. Kwak, M. Elmogy, Colon cancer diagnosis based on machine learning and deep learning: modalities and analysis techniques, Sensors 22 (23) (2022) 9250.
[57] D. Jha, N.K. Tomar, S. Ali, M.A. Riegler, H.D. Johansen, D. Johansen, T. de Lange, P. Halvorsen, Nanonet: real-time polyp segmentation in video capsule endoscopy and colonoscopy, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 37–43.
[58] L. Lin, G. Lv, B. Wang, C. Xu, J. Liu, Polyp-lvt: polyp segmentation with lightweight vision transformers, Knowl.-Based Syst. 300 (2024) 112181.
[59] A. Krenzer, M. Banck, K. Makowski, A. Hekalo, D. Fitting, J. Troya, B. Sudarevic, W.G. Zoller, A. Hann, F. Puppe, A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks, J. Imaging 9 (2) (2023) 26.
[60] J. Zhang, et al., Transfuse: fusing transformers and CNNS for medical image segmentation, arXiv preprint arXiv:2102.08005, 2021.
[61] N.T. Duc, N.T. Oanh, N.T. Thuy, T.M. Triet, V.S. Dinh, Colonformer: an efficient transformer based method for colon polyp segmentation, IEEE Access 10 (2022) 80575–80586.
[62] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, arXiv preprint arXiv:2108.06932, 2021.
[63] H. Li, D. Zhang, J. Yao, L. Han, Z. Li, J. Han, Asps: augmented segment anything model for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 118–128.
[64] Y. Li, L.M. Dang, H. Wang, M. Fayaz, S. Danish, J. Shang, H.-K. Song, H. Moon, Transformer-based detection of abnormal rice growth using drone-based multispectral imaging, Comput. Electron. Agric. 239 (2025) 111055.
[65] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, P. Halvorsen, T. Lange, T. de Lange, Kvasir-Seg: a segmented polyp dataset, Int. Conf. Multimed. Model. 11962 (2019) 451–462.
[66] J. Bernal, N. Tajkbaksh, F.J. Sánchez, B. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, M. Wallace, C. Sanchez-Montes, W. Li, Q. Wang, K. Pogorelov, M.A. Riegler, S. Choi, G. Debard, V. Iglovikov, D. Jha, P. Brandao, D. Stoyanov, M. Angermann, S. Realdon, Wm-Dova maps for accurate polyp highlighting in colonoscopy: validation VS. Saliency maps from physicians, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, Springer, 2015, pp. 429–437.
[67] J. Silva, A. Histace, et al., Toward automated polyp detection in colonoscopy videos, Expert Syst. Appl. 41 (2014) 7436–7446.
[68] P.H. Smedsrud, V. Thambawita, S.A. Hicks, H. Gjestang, O.O. Nedrejord, E. Næss, H. Borgli, D. Jha, T.J.D. Berstad, S.L. Eskeland, et al., Kvasir-Capsule, a video capsule endoscopy dataset, Sci. Data 8 (1) (2021) 142.
[69] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, MICCAI (2015) https://arxiv.org/abs/1505.04597.
[70] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNET + +: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.
[71] 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, 2018, pp. 327–331.
[72] N. Ibtehaz, M.S. Rahman, Multiresunet: rethinking the U-net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[73] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, 2018.
[74] 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, arXiv preprint arXiv:2101.07172, 2021.
[75] J. Liu, W. Zhang, Y. Liu, Q. Zhang, Polyp segmentation based on implicit edge-guided cross-layer fusion networks, Sci. Rep. 14 (1) (2024) 11678, https://www.nature.com/articles/s41598-024-62331-5.
[76] Y. Oukdach, A. Garbaz, Z. Kerkaou, M. El Ansari, L. Koutti, A.F. El Ouafdi, M. Salihoun, Uvit-Seg: an efficient VIT and U-net-based framework for accurate colorectal polyp segmentation in colonoscopy and wce images, J. Imaging Inform. Med. 37 (5) (2024) 2354–2374.
[77] J. Ren, X. Zhang, L. Zhang, HiFiSeg: high-frequency information enhanced polyp segmentation with global-local vision transformer, IEEE Access 13 (2025) 38704–38713.
[78] D. Liu, C. Lu, H. Sun, S. Gao, Na-Segformer: a multi-level transformer model based on neighborhood attention for colonoscopic polyp segmentation, Sci. Rep. 14 (1) (2024) 22527.
[79] M. Jian, N. Yang, C. Zhu, MANet: multi-attention network for polyp segmentation, Med. Eng. Phys. 143 (2025) 104396, https://www.sciencedirect.com/science/article/pii/S1350453325001158.
[80] N.K. Tomar, D. Jha, K. Biswas, U. Bagci, Transformer-enhanced iterative feedback mechanism for polyp segmentation, in: ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2025, pp. 1–5.
[81] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention U-net for lesion segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 683–687, https://doi.org/10.1109/ISBI.2019.8759329
[82] B. Urooj, M. Fayaz, S. Ali, L.M. Dang, K.W. Kim, Large language models in medical image analysis: a systematic survey and future directions, Bioengineering 12 (8) (2025) 818.

Author biography



Gul E Arzu received her B.S. degree in Software Engineering from the National Textile University, Pakistan, in 2024. She is currently pursuing her M.S. degree in the Department of Computer Science and Engineering at Sejong University, Seoul, South Korea, and works as a Research Assistant at the Computer Vision and Pattern Recognition (CVPR) Lab. Her research interests include medical image processing, artificial intelligence, computer vision, deep learning, and machine learning.

Muhammad Fayaz completed his Bachelor's degree from Islamia College University, Peshawar, and a Master's in Computer Engineering from Cyprus International University, specializing in computer vision, deep learning, and machine learning. He is currently a Research Assistant at the Computer Vision and Pattern Recognition (CVPR) Lab, Sejong University. His research focuses on medical image analysis and land cover mapping, developing deep learning algorithms for improved diagnostic precision and sustainable environmental monitoring.

Usman Ali received his B.Sc. degree in Electrical Engineering from the University of Engineering and Technology (UET), Pakistan, in 2009, and his M.Sc. degree in Electrical Engineering from the University of Management and Technology (UMT), Pakistan, in 2013. He earned his Ph.D. degree in Computer Engineering from the Korea University of Technology and Education (Koreatech), South Korea, in 2020. From 2020 to 2022, he worked as a Postdoctoral Researcher at the School of Computer Science and Engineering, Koreatech. He then served as a Research Professor in the Department of Electrical and Computer Engineering at Sungkyunkwan University (SKKU), South Korea, from September 2022 to February 2024. He is currently an Assistant Professor in the Department of Computer Science and Engineering at Sejong University, South Korea. His research interests include computer vision, pattern recognition, optimization, machine learning, and deep learning.

L. Minh Dang received a B.S. degree in information systems from the University of Information Technology, VNU HCMC, Vietnam, in 2016. He is currently pursuing a PhD degree in computer science at Sejong University, Seoul, South Korea. In 2017, he joined the Computer Vision Pattern Recognition Laboratory. His current research interests include computer vision, natural language processing, and artificial intelligence.

Hyeonjoon Moon received the B.S. degree in electronics and computer engineering from Korea University, in 1990, and the M.S. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo, in 1992 and 1999, respectively. From January 1996 to October 1999, he was a Senior Researcher at the Electro-Optics/Infrared Image Processing Branch at the U.S. Army Research Laboratory (ARL), Adelphi, MD, USA. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) Program. From November 1999 to February 2003, he was a Principal Research Scientist at Viisage Technology, Littleton, MA, USA. His main interest in research and development is real-time facial recognition systems for access control, surveillance, and big database applications. He has an extensive background in still image and real-time video-based computer vision and pattern recognition. Since March 2004, he has been with the Department of Computer Science and Engineering, Sejong University, where he is currently a Professor and the Chairperson. His current research interests include image processing, biometrics, artificial intelligence, and machine learning.