Research Square

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

A Pixel-based Confidence Weighted Transfusion model for Colonoscopy and Colon Capsule Endoscopy image-based Polyp Segmentation

Smith K. Khare

smkh@mmmi.sdu.dk

Southern Denmark University

Jan-Matthias Braun

Southern Denmark University

Benedicte Schelde-Olesen

Odense University Hospital

Anastasios Koulaouzidis

Odense University Hospital

Esmaeil S. Nadimi

Southern Denmark University

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11012682/v1

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

Additional Declarations: No competing interests reported.

A Pixel-based Confidence Weighted Transfusion model for Colonoscopy and Colon Capsule Endoscopy image-based Polyp Segmentation

Smith K. Khare[1,*], Jan-Matthias Braun[1], Benedicte Schelde-Olesen[2,3], Anastasios Koulaouzidis[3,4], Esmaeil S. Nadimi[1], on behalf of the AICE Consortium

[1*]Applied AI and Data Science Unit, Mærsk Mc-Kinney Møller Institute, Faculty of Engineering, University of Southern Denmark, Denmark.
[2]The Department of Surgery, Odense University Hospital, Odense, Denmark.
[3]The Department of Clinical Research, University of Southern Denmark, Odense, Denmark.
[4]The Department of Medicine, Odense University Hospital, Svendborg, Denmark.

*Corresponding author(s). E-mail(s): smkh@mmmi.sdu.dk;
Contributing authors: j-mb@mmmi.sdu.dk;
Benedicte.Schelde-Olesen@rsyd.dk; Anastasios.Koulaouzidis@rsyd.dk;
esi@mmmi.sdu.dk;

Abstract

Polyps one of the most frequently occurring lesions in the human gastrointestinal system and play a critical role in colorectal cancer development, often serving as early indicators or precursors of malignant transformation. Timely and early detection of polyps is critical for lowering the risk of colorectal cancer and increasing survival. Polyp segmentation is crucial in improving computer-aided detection and diagnosis systems, which may support early intervention. CNN-based segmentation models often struggle to capture global contextual information, leading to suboptimal performance. These limitations are further amplified in endoscopic imaging scenarios, where variable resolution in colon capsule endoscopies (CCE) and conventional colonoscopy images—introduces inconsistencies in feature representation and contextual awareness. Therefore, we develop a comprehensive TransFuseNet model that captures local and global receptive fields at the encoder and decoder levels. The pixel-based confidence-weighted fusion combines features from two models, which facilitates knowledge and domain shift features that allows segmentation of

low- and high-resolution images. We validated the model's performance on seven diverse optical colonoscopies, CCE, and synthetic datasets. Analysis reveals that our developed model has yielded an intersection over the union of 84.67%, 89.05%, 88.91%, 89.46%, 89.58%, 82.07%, and 86.41% on the Kvasir, CVC-300, CVC-ColonDB, CVC-ClinicDB, ETIS, CCE CareForColon2015 dataset, and Synthetic datasets, respectively. The external validation also confirms the robustness and effectiveness of our developed model, providing a new pathway for polyp segmentation using optical colonoscopy and CCE images with explainable characteristics and polyp pixel measurement. Thus, our proposed TransFuseNet effectively handles both low- and high-resolution endoscopic images by fusing local and global contextual features. This supports more reliable polyp segmentation across diverse imaging conditions, enhancing diagnostic accuracy in clinical practice.

Keywords: Colon capsule endoscopy, colonoscopy, deep learning, vision transformer, pixel-based confidence weighted fusion, semantic segmentation.

1 Introduction

Colorectal cancer (CRC) is one of the most frequently diagnosed types of cancer globally, just after lung and breast cancers, with 1.93 million new cases and about 935,000 deaths in 2020 [1, 2]. The age-standardized incidence rate (ASR) of CRC is around 19.5 per 100,000 person-years. In Europe, CRC constitutes a major public health concern as it accounted for approximately 362,000 new cases and 161,000 deaths in 2022, which was 13% of all new cancers and 12% of deaths by cancers in Europe [3]. Europe possesses one of the highest CRC incidence rates globally, with an ASR of 30.4 per 100,000 [4]. The EU region had a standardized death rate for CRC of 27 per 100,000 inhabitants in 2021 [5]. Denmark has some of the most advanced rates of CRC in Europe, with an ASR of 48.1 per 100,000 inhabitants in 2022, with Danish women having the highest CRC incidence in the EU at 95.7 per 100,000 inhabitants [6, 7]. Statistics from Denmark indicate the increasing trend for serrated polyps, especially sessile serrated lesions (SSLs), has risen over the past two decades, with more than a 100-fold increase in SSL incidence from 2006 to 2021, indicating more attention toward early diagnosis and prevention [8]. Studies show that polyps contribute to around 95% of the CRC cases, which progresses from mucosa to adenomatous polyps that later transform to precancerous polyps followed by adenocarcinoma that finally results in a cancer [9]. Adenomatous and serrated colorectal polyps are types of polyps that can develop into CRC. Studies show that adenomatous and hyperplastic polyps have global prevalence of 18.4% and 16.4% in autopsy series, respectively [10]. Colorectal polyps are usually found during the routine screening procedures, and one European study reported that among the patients who had undergone colonoscopy, 76.99% had neoplastic polyps, out of which 44.31% had advanced adenomas [11]. Figure 1 shows the progression of mucosa to CRC in four stages. Figure 1 also indicates that the survival rate is greater than 90% if detected and treated at the early stage but drops significantly to less than 5% in stage 4 [12, 13]. Therefore, regular screening for early detection of polyps plays a key role in faster diagnosis and timely treatment. Colonoscopy, considered as the gold standard for identifying colorectal polyps, uses an invasive technique, where a camera-equipped flexible tube is inserted into the rectum and advanced through to the colon, allowing direct visualization of the colon and rectum's interior surfaces [14, 15].

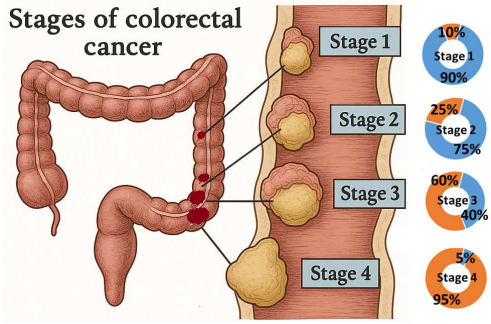


Fig. 1: Stages of CRC development from normal mucosa along with approximate survival rate in different stages.

Colon capsule endoscopy (CCE), in contrast, is a non-invasive procedure, consisting of a pill-sized capsule containing two cameras, light sources, and a transmitter [16, 17]. Optical colonoscopy and CCE have massive differences in camera type, lighting, cleansing, and perspective. In both cases, experienced clinicians screen the images and diagnose polyps. However, manually inspecting and segmenting polyps is time-consuming, labor-intensive, and subject to human error due to hidden polyps or deficiencies in endoscopy equipment [18, 19]. Accurate estimation of pixel length and the relative size of polyps in endoscopic images is essential for malignancy risk assessment, given that polyp size is a well-established indicator of colorectal cancer potential [20, 21]. About 25% to 30% of polyps can be missed during colonoscopies, thereby increasing the risk of cancer [22–24]. As a result, there is an urgent need for a computer-aided, automated, and accurate diagnostic and segmentation of polyps using artificial intelligence (AI) techniques.

2 Literature survey

Over the years, researchers have developed various deep learning models based on convolutional layers (CL) and transformer layers to segment polyps. Ronneberger et al. [25] designed the U-Net model that uses an encoder and decoder composed of CL for segmentation. Patel et al. [26] enhanced the U-Net architecture performance using a semantic feature enhancement module and an adaptive global context module in the decoder for polyp extraction. Zhou et al. [27] used enhanced skip connections techniques for feature fusion to obtain faster and more accurate segmentation. Fang et al. [28] proposed a selective feature aggregation model by embedding selective kernel modules into CL for polyp segmentation. Chen et al. [29] implemented an adaptive segment anything model (SAM), which uses task-specific visual prompts for effective segmentation. Zhang et al. [30] proposed an extension of SAM called SAM-path, which uses cues for trainable class prompts and an additional encoder for domain-specific information. Fang et al. [31] developed another variant of SAM called VP-SAM that uses amplitude information of the Fourier spectrum and spatiotemporal side network for segmentation purposes. Li et al. [32] proposed a fine-tuned transfer SAM called Polyp-SAM that optimizes the prompt embedding strategies for polyp segmentation. Risab Biswas [33] proposed a modification of Polyp-SAM called Polyp-SAM++, which is a text prompt-aided

SAM. Dong et al. [34] developed the SAM2 model that uses multi-plane feature aggregation and cross-slice attention mechanisms along prompts for polyp segmentation. Bo et al. [35] introduced a transformer-based polyp-PVT combining cascaded fusion, camouflage identification, and similarity aggregation modules for polyp segmentation. Fan et al. [36] investigated feature aggregation using a parallel partial decoder with guidance area and boundary cues using reverse attention to develop the PraNet model. Jha et al. [37] proposed a ResUNet++ model leveraging atrous spatial pyramidal pooling along with residual, squeeze excitation, and attention blocks for poly segmentation. The work proposed by Mahmud et al. [38] used sequential depth dilated inception, deep fusion skip module, and deep reconstruction module at the encoder and decoder to develop the PolypSegNet model. Kim et al. [39] modified the PraNet model by aggregating the feature map with three area maps, namely parallel axial attention encoder, parallel axial attention decoder, and uncertainty augmented context attention to obtain a UACANet model for polyp segmentation. Lin et al. [40] introduced a bit slicing context attention mechanism to enhance boundary detail extraction in polyp segmentation. It combines encoder–decoder modules to capture multi-scale contextual features and suppress noise for polyp segmentation. Weng et al. [41] developed the INet model that used fixed-sized feature maps, customized residual shortcuts, and a fusion of multilevel semantics for medical image segmentation. Wang et al. [42] framed a DeepNeXt model that comprised depthwise separable CL, multiscale attention feature encoding using deep strip convolution and coordinate attention and jumping multistage feature fusion for polyp segmentation. Lee et al. [43] proposed shallow and reverse attention-based dual attention modules for channel-wise vectors for polyp segmentation. Wei et al. [44] coined a shallow attention network combining color exchange and a probability correction strategy to balance biased predictions to improve polyp segmentation. Xia et al. [45] utilized a pyramid vision transformer with a parallel hierarchical aggregation decoder and feature fusion using a refinement pyramid module. These features are combined with a reverse coordinate attention module for polyp segmentation. Zhao et al. [46] developed a multi-scale subtraction network that utilizes subtraction units with varying receptive fields to obtain multi-scale difference information. Zhang et al. [47] used cross-semantic attention, hybrid semantic complementarity, and multi-scale prediction modules to capture long-range dependencies and local details for polyp segmentation. Tashk et al. [48] developed an AID-U-Net semantic segmentation model composed of direct contracting, expansive paths, and distinguishing features having sub-contracting and sub-expansive paths. Zhou et al. [49] proposed a feature aggregation and propagation model that combines boundary guidance, cross-level fusion, and propagation modules to extract boundary-enhanced features and feature fusion for polyp segmentation. Yin et al. [50] presented a model to capture image and cross-image contextual relations to improve contextual region embedding within and across images. Zhou et al. [51] presented a cross-level feature fusion and boundary aggregated module to capture feature fusion at different scales to deal with scale variations and the boundary context information for polyp segmentation. Pan et al. [52] et al. introduced a multi-scale boundary prediction network for accurate polyp segmentation, which combines boundary prediction modules, multi-scale attention, and a shallow filtering module to refine polyp edges. Zhang et al. [53] developed a context-based segmentation that uses an encoder and decoder block composed of local context and global context modules to extract local and global context features and then aggregates with adaptive selection modules. In another work, Zhang et al. proposed [54] implemented a self-attention

module to capture global attributes and a lesion-aware cross-attention module to improve feature contrast between polyp and background for automated polyp segmentation. Hu et al. [55] developed an autoSAM model consisting of a transposed CL and a lightweight cross-attention module for medical image segmentation. The literature review shows that most of the models suffer from the following limitations:

- CNN-based architectures face challenges quantifying the disparate contributions of features collected at various levels, frequently assigning equal significance to low-level and high-level data despite their distinct semantic and spatial characteristics.
- Existing models generally lack an effective method for fusing multi-scale information, leading to a sub-optimal fusion of intricate details and overarching context, which limits the segmentation accuracy.

To overcome these challenges, we aim to develop an effective and robust model that

- uses a combination of CNN and transformer layer at encoder, which captures low-level, high-resolution features from the input image to retain fine-grained spatial information.
- uses a transformer encoder to capture the long-range dependencies and global features, which CNNs lack.
- uses a pixel-based weighted fusion strategy to adaptively fuse the predictions from different models that facilitate knowledge transfer and domain shift, which applies to optical colonoscopy and CCE images.

To summarize, the proposed TransFuseNet model is a modular fusion framework that integrates the outputs of two independently trained segmentation models-one optimized for optical colonoscopy and the other for CCE images. The pixel level confidence weighted fusion allows adaptive prediction by assigning higher weight to the model that demonstrates greater confidence, which effectively captures subtle changes in variable resolution. Thus, the TransFuseNet model dynamically adapt to modality-specific strengths and weaknesses, improving polyp delineation performance across heterogeneous imaging conditions. The paper is structured as: Section 3 presents Methodology, Results presented in Section 4, Section 6 covers Discussion, and Section 7 concludes the paper.

3 Methodology

The schematic overview of the proposed TransFuseNet model is shown in Fig. 2. It combines a fine-tuned Trans-UNet model with a pixel-based weighted fusion module to process high- and low-resolution images for polyp segmentation.

3.1 Dataset

To validate the performance of our proposed model we used seven datasets from CCE, colonoscopy, and synthetically generated images. Following subsections presents the details of the datasets.

3.1.1 CCE dataset

The CCE CareForColon2015 (AICE) dataset is derived from with CCE recordings, containing at least one colorectal polyp of any size or morphology. This dataset originates from

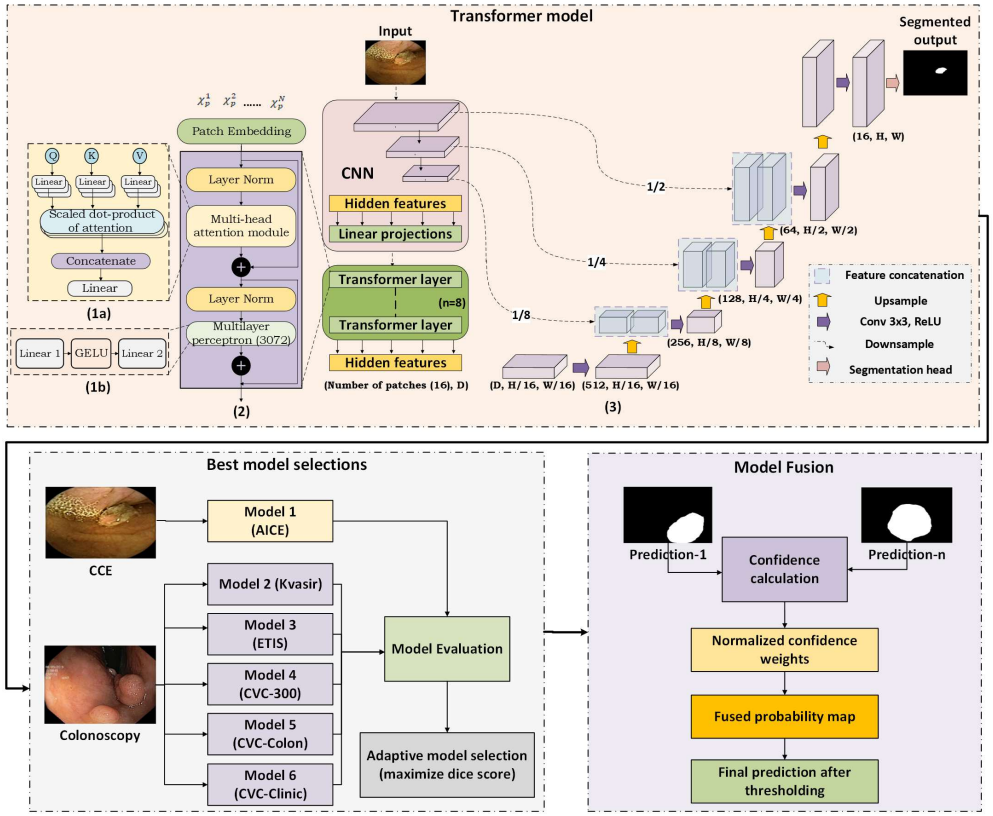


Fig. 2: Schematic representation of the proposed TransFuseNet model for CCE and Colonoscopy images.

a new study designed in alignment with the protocols and methodologies of the "Danish CareForColon2015 trial (cfc2015)," which began in 2021 as part of the national Danish Colorectal Cancer Screening Program. As the largest randomized controlled trial on CCE [56, 57]; this dataset is independently generated and follows the rigorous procedures and ethical standards. Its primary objective is to assess differences in detected colorectal cancers and intermediate- to high-risk adenomas between intervention and control cohorts. Secondary objectives include analysis of patient acceptability, adverse events, completion and interval CRC rates, patient-reported outcomes (PROs), long-term cancer incidence, socioeconomic disparities, clinical applicability of CCE, and cost-effectiveness. The detailed dataset description can be found in [58, 59]. The detection of polyps was carried out as a part of a clinical study in the screening population [48, 60]. The following manual segmentation of the polyps was carried out by student workers and that was specifically used for AI training and validation purposes. Each image in the dataset had 576×576 pixels. Since the originally annotated image was composed to timestamp and device name, we have pre-processed and cropped the original image to remove these timestamps and device name. Therefore, after pre-processing, each image was resized to 512×512.

Table 1: Summary of different datasets used in our work.

Dataset	AICE [48, 60]	Kvasir [61]	CVC-300 [62]	CVC-CoDB [9, 63]	CVC-CDB [64, 65]	ETIS [66, 67]	Synth-Colon [68, 69]
Images	11,956	1000	60	379	612	196	20014
Image size	512×512	720×576 to 1920×1072	574×500	574×500	288×384	966×1225	500×500
Training	9436	800	48	303	488	156	16010
Validation	2520	200	12	76	124	40	4004
Type	CCE	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy	Synthetic
Modality	RGB	RGB	RGB	RGB	RGB	RGB	RGB

3.1.2 Colonoscopy datasets

We used Kvasir-SEG, CVC-300, CVC-ColonDB (CVC-CoDB), CVC-ClinicDB (CVC-CDB), and ETIS-Larib datasets, which are referred to as optical colonoscopy. The Kvasir-SEG dataset consists of gastrointestinal polyp images of healthy and diseased tissues, which have been verified and annotated by experienced gastroenterologist. The dataset is composed of 1000 images having varying image dimensions and corresponding groundtruth (GT) [61]. The CVC-300 dataset has 60 high-resolution colonoscopy images for polyp segmentation, with a pixel size of 500×574 [62]. Another publicly available dataset is CVC-ColonDB, which is composed of 379 colonoscopy images with 500×574 pixels extracted from 15 different videos [9, 63]. CVC-ClinicDB dataset has a total of 612 images are extracted from 29 different colonoscopy videos, each having a pixel resolution of 288×384 [64, 65]. Finally, the ETIS-Larib dataset is composed of 196 high-resolution colonoscopy images, each with 966×1225 pixels.

3.1.3 Synth-Colon dataset

Synth-Colon dataset is composed of more than 20,000 images, each having a size of 500×500 pixels [68, 69]. Table 1 shows a detailed summary for each dataset.

3.2 Trans-UNet model

The Trans-UNet model encoder is derived from patch embedding and transformer layers, while the decoder uses skip connections with convolutional blocks.

3.2.1 Transformer Encoder

The segmentation begins by partitioning an input image I into non-overlapping patches of size $p \times p$. Thus, the total number of patches N for an image with height h and width w is given by $\frac{h}{p} \times \frac{w}{p}$. Finally, these patches are flattened to obtain $\mathscr{Z}$ using the projection matrix $\mathscr{E} \in \mathbb{R}^{p^2 c \times d}$:

$$\mathscr{Z} = \mathscr{F}(I)\mathscr{E} \tag{1}$$

where $\mathscr{F}$ is the flattening function, c are the number of channels, and the vectorized patches x_p into a latent D-dimensional embedding space using a trainable linear projection. After this, positional encoding is appended to maintain the spatial position using $p_{pos} \in \mathbb{R}^{1 \times n \times d}$ given by [70]:

$$\mathscr{Z} = [x_p^1\mathscr{E}; x_p^2\mathscr{E}; \ldots; x_p^N\mathscr{E}] + p_{pos} \tag{2}$$

where The transformer layer is composed of multiple (L) multihead self-attention (MSA) and multilayer perceptron (MLP) layers having skip connections, which is given by [70]:

$$\mathscr{Z}_l' = MSA(L_{\text{norm}}(\mathscr{Z}_{l-1}) + \mathscr{Z}_{l-1} \tag{3}$$

$$\mathscr{Z}_l = MLP(L_{\text{norm}}(\mathscr{Z}_l') + \mathscr{Z}_l' \tag{4}$$

where L_{norm} is a layer normalization, $\mathscr{Z}_l'$ is the output of lth layer of MSA, and $\mathscr{Z}_l$ is the output of lth transformer layer. The MSA is obtained by the following steps:

$$Q = \mathscr{Z}W_Q, \quad K = \mathscr{Z}W_K, \quad V = \mathscr{Z}W_V \tag{5}$$

$$\text{Attention}(Q,K,V) = \phi\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{6}$$

$$MSA = \Theta(\mathscr{H}_1,\ldots,\mathscr{H}_h)W_O \tag{7}$$

where W_Q, W_K, and W_V are the learnable weight matrices for computing the queries (Q), keys (K), and values (V), ϕ is the softmax function, Θ is a concatenation block, $\mathscr{H}$ are the attention heads, and W_O is the output projection matrix obtained after the concatenation of all the attention heads.

3.2.2 Trans-UNet decoder

The decoder of Trans-UNet follows a U-Net-like architecture, which upsamples the feature maps and refines the upsampled features using skip connections.

Skip Connections The upsampled feature maps are concatenated with encoder feature maps at each level of the decoder as shown in Fig 2. The mathematical formulation of upsampled feature concatenation ($\mathscr{D}_i'$) is given by:

$$\mathscr{D}_i' = \Theta(\mathscr{D}_i, E_{F_i}) \tag{8}$$

where $\mathscr{D}_i$ is the feature maps at ith decoder block and E_F is the feature maps from the ith encoder block.

Upsampling and Convolution The upsampling of decoder features is implemented using transposed convolutions and CL given by:

$$\mathscr{D}_O = \phi(\text{US}(\mathscr{D}_i')) \tag{9}$$

where ϕ is a convolutional function, US is upsampling, and $\mathscr{D}_O$ is the output of the decoder.
Final Output The Trans-UNet model computes the final segmentation $\mathscr{Y}_O$ by using sigmoid activation function given by:

$$\mathscr{Y}_O = \sigma(\mathscr{D}_O) \tag{10}$$

3.3 Pixel-wise Confidence-weighted Fusion

Ensemble-based model fusion is widely adopted for classification tasks to boost performance. It uses the aggregation of the predictions yielded by multiple independent trained models,

either by using majority voting or uniform averaging. In object detection, non-maximum suppression draws boxes around certain classes if the intersection-over-union (IoU) and overlap exceed a predefined threshold. However, box-based fusion may not be effective for segmentation tasks as it can discard valuable probabilistic confidence information due to a hard threshold. Therefore, we used a pixel-based confidence weighted fusion (PBCWF) that calculates the confidence scores from each model to construct a final averaged confidence score, which is based on weighted-boxes fusion [71]. For N independently trained segmentation models $\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_N$, we obtain a probability map $\mathscr{P}_m(\mathbf{x}) \in [0,1]^{H \times W}$ for an input image $\mathbf{x}$ for each model $\mathscr{M}_m$. For these predicted pixel probabilities, we define a confidence map for each model, defined by:

$$C_m(i,j) = 2 \cdot |\mathscr{P}_m(i,j) - 0.5| \tag{11}$$

The above formulation assigns higher confidence to 0 or 1, and lower confidence to predictions near 0.5, indicating the uncertainty. In the next step, we compute the normalized confidence matrix using:

$$\hat{w}_m(i,j) = \frac{C_m(i,j)}{\sum_{k=1}^{M} C_k(i,j) + \varepsilon} \tag{12}$$

where ε is a small constant to avoid division by zero. The fused probability map is obtained by weighted average:

$$\mathscr{P}_{\text{fused}}(i,j) = \sum_{m=1}^{M} \hat{w}_m(i,j) \cdot \mathscr{P}_m(i,j) \tag{13}$$

Finally, binary predictions are obtained by hard-threshold:

$$\hat{Z}(i,j) = \begin{cases} 1 & \text{if } \mathscr{P}_{\text{fused}}(i,j) > \tau \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where τ is the threshold set to 0.5 in our case. To identify the most effective fusion, we systematically explored all possible combinations of models, evaluating Dice score for respective dataset. Through this exhaustive evaluation, we adaptively selected the best-performing fusion combination with a maximum Dice score. The final predictions are obtained using pixel-wise weighted averaging of the probability maps from the selected models, with confidence scores computed using proposed confidence map. This adaptive strategy ensures that the fusion is data driven based on dataset characteristics, improving robustness and overall performance.

4 Experimental settings

Before feeding the images to the Trans-UNet model, we pre-processed all the images from each dataset using the same settings. We used a transformation strategy during the training, where images were resized, flipped, rotated, scaled, and distorted. Images were resized to 256×256 with a horizontal and vertical flip. We used random rotation by 90^0 and shift scale rotation with a shift limit of 0.0625, scale limit of 0.2, rotate limit of 20, and the probability of applying a transform is set to 0.4. Additionally, we also applied optical distortion, hue saturation, and random brightness contrast. It is noteworthy that all the parameters were set

empirically and kept the same for all the datasets. We used the ResNet-50 model as a backbone to create a feature map of local features and a vision transformer to extract evidence for global reasoning. We used a ViT base model with 86M parameters as an encoder with a patch size of 16×16. The number of skips is 3 and the number of heads is 4. The MLP has 3072 input units and one hidden layer with 768 units. For experimentation, we used a batch size of 2. The used Adam optimizer using an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-6}, was ran over maximum 40 epochs with early stopping over 7 epochs, cosine annealing warm restarts scheduler, and Dice binary cross-entropy loss.

5 Results

First, we started by selecting the best hyper-tuning parameters for the CCE AICE dataset. Table 2 shows the performance obtained on the CCE AICE dataset with the Trans-UNet model. As evident from Table 2, variation in the number of transformer layers and dropout changes IOU and Dice score by 3% and 2%. As a result, we selected the parameters that yielded the highest performance on our CCE AICE dataset. Therefore, further model simulations are carried out on other datasets using the same experimental setting as mentioned earlier, but with the values of dropout and number of transformer layers as 0.2 and 8, respectively. In the next phase, we evaluated the model's performance on different datasets and

Table 2: Hypertuning of Trans-UNet model on the AICE dataset.

AICE	Dropout	Layers	IOU (%)	Dice (%)
Dropout	0	8	80.67	87.26
Dropout	0.3	12	80.97	88.48
Dropout	0.4	8	79.24	86.54
Dropout	0.1	12	81.22	88.61
Dropout	0.2	12	81.44	89.09
Dropout	**0.2**	**8**	**82.07**	**89.29**

computed the accuracy (ACC), Dice score, IoU, precision (PRC), and recall parameters. Table 3 summarizes the performance report of the Trans-UNet model on Synth-colon, Kvasir, ETIS, CVC-300, CVC-CoDB, CVC-CDB, and CCE AICE datasets. As shown in Table 3, the Trans-UNet model maintains high performance in all the performance matrices. The highest performance is obtained for the Synth-Colon dataset. The Kvasir dataset has yielded the highest performance for optical colonoscopy image-based datasets, while for the CCE AICE dataset, the values of accuracy, IoU, Dice, precision, and recall are 98.22%, 82.07%, 89.29%, 89.91%, and 88.68%, respectively. The performance report demonstrates the model's ability to extract the polyp region accurately. Also, the consistency in the model's performance across various datasets highlights its ability to adapt to various domains. To further validate the effective of Trans-UNet model, we conducted five independent runs using the same experimental protocol to report mean (μ) and standard deviation (STD). The results indicate that the model achieves consistent performance across all datasets, with relatively low standard deviation

values. On the Kvasir dataset, the model obtained an IoU of 87.11±1.08 and Dice score of 92.88±0.67, demonstrating stable segmentation performance. Similarly, on the CVC-CoDB and CVC-CDB datasets, the IoU values of 86.41±1.21 and 85.96±1.56, respectively, indicate limited variation across runs. For the ETIS and AICE datasets, the model also maintains stable performance, achieving IoU scores of 82.45±1.59 and 81.76±0.74, respectively, with corresponding Dice scores of 89.63±1.10 and 88.55±0.54. These results further confirm the robustness of the model across challenging and diverse data distributions. Although slightly higher variance is observed in certain cases, such as the CVC-300 dataset with an IoU of 82.18±3.46, the overall performance remains consistent without significant variations. Additionally, Synth-Colon dataset exhibit extremely low variance, which confirms the reliability of the model. Overall, the low STD values across all datasets demonstrate that the observed performance improvements are not due to random fluctuations, but instead reflect the stability and generalization capability of the model. Fig. 3 shows the exemplary comparison of true masks (second column) and the predicted masks (third column) from the Trans-UNet model. The proposed model can effectively detect the smaller and larger polyps from the images with high accuracy. However, it is also seen from Fig. 3 that the Trans-UNet model sometimes predicts false polyps as highlighted with the yellow region. The error comparison in Fig. 3 shows that our proposed model extracts the polyps in a fine manner with little error around the boundaries, as shown in the fourth column. The error also increases when the model fails to predict the larger polyps, as shown in the figure from the Kvasir dataset example. Overall, our model provides accurate segmentation of smaller, larger, and multiple polyps with significantly low errors. In addition to the error, we evaluated the relative size of the predicted regions. The fifth column in Fig. 3 presents the estimate of pixel length and relative polyp size derived by drawing a contour with yellow boundaries. We cannot calculate the exact size of polyps because of the unavailability of information on pixel width in the datasets. To calculate the relative size, we first drew the contour around the predicted mask. In the next step, we found out the maximum distance across the mask and divided it by the diagonal length of the mask.

Table 3: Performance (%) obtained on various datasets using the fine-tuned Trans-UNet model.

Datasets	ACC	IoU	Dice	PRC	Recall
Synth-Colon	99.85	98.6	99.29	98.72	99.88
Kvasir	96.36	87.78	93.3	92.66	93.97
ETIS	97.45	83.66	90.5	89.26	91.84
CVC-300	99.05	82.38	89.41	84.29	96.48
CVC-CoDB	98.52	85.5	91.65	91.85	91.45
CVC-CDB	97.47	86.2	92.21	93.68	90.86
AICE	98.22	82.07	89.29	89.91	88.68

Table 4: Performance (%) obtained on various datasets by from multiple runs using Trans-UNet model.

	ACC	IoU	DICE	PRC	Recall
Datasets	$\mu \pm STD$	$\mu \pm STD$	$\mu \pm STD$	$\mu \pm STD$	$\mu \pm STD$
Kvasir	96.21 ± 0.31	87.11 ± 1.08	92.88 ± 0.67	93.54 ± 0.58	92.3 ± 1.24
CVC-300	99.07 ± 0.27	82.18 ± 3.46	89.17 ± 2.52	85.72 ± 3.77	93.67 ± 1.53
CVC-CoDB	98.62 ± 0.11	86.41 ± 1.21	92.24 ± 0.79	92.12 ± 0.46	92.42 ± 1.79
CVC-CDB	97.28 ± 0.35	85.96 ± 1.56	92.05 ± 0.99	91.28 ± 1.3	92.93 ± 1.53
ETIS	97.43 ± 0.28	82.45 ± 1.59	89.63 ± 1.1	91.74 ± 1.97	87.89 ± 1.9
AICE	98.19 ± 0.06	81.76 ± 0.74	88.55 ± 0.54	91.5 ± 0.78	86.07 ± 1.27
Synth-Colon	99.34 ± 0.02	98.01 ± 0.13	98.75 ± 0.07	98.17 ± 0.08	99.34 ± 0.06

Table 5: IoU obtained from external validation of the proposed Trans-UNet model.

	CVC-300	**CVC-CDB**	**CVC-CoDB**	**Kvasir**	**AICE**	**ETIS**	**Synth-Colon**
CVC-300	–	84.96	**90.71**	85.19	**83.85**	84.22	47.01
CVC-CDB	64.14	–	72.79	**76.44**	**74.85**	74.26	48.7
CVC-CoDB	76.78	**87.55**	–	73.34	**75.07**	79.88	44.95
Kvasir	65.14	72.61	73.19	–	**71.05**	**74.45**	50.49
AICE	57.25	**62.49**	60.05	56.14	–	56.17	51.07
ETIS	71.11	**75.23**	73.89	74.86	**76.07**	–	52.34
Synth-Colon	81.8	**83.67**	79.36	80.74	**80.15**	80.91	–

As a next evaluation strategy, we externally validated the model trained on one dataset on other datasets. Table 5 presents the summary of external validation performance. The first column indicates the name of a dataset on which we validated other trained models, while columns 2-8 indicate the names of the models trained on respective datasets. It is noteworthy that we used the entire datasets for external validation, as the trained model on one dataset has never seen images from other datasets. It is evident from Table 5 that the model trained on optical colonoscopy images performed well when validated on other optical colonoscopy-based image datasets. The comparison report shows that the model trained on the synthetic Synth-Colon fails significantly to cope with its performance when validated on real-world optical colonoscopy-based image datasets and the AICE CCE dataset. The Synth-Colon data has advantages in terms of geometric ground truth information, but seemingly falls short in producing realistic imagery, missing relevant information which allows transfer to real-world data sets. This results in a performance drop due to its ability to adapt to the changes in real-world colonoscopy images. However, when models trained on other datasets were validated on the Synth-Colon dataset, they obtained decent performance in terms of all the evaluation matrices. Similarly, when a trained model on one of the optical colonoscopy-based image datasets is validated on other optical colonoscopy-based image datasets, the performance report drops little but without significant degradation. It is noteworthy that the

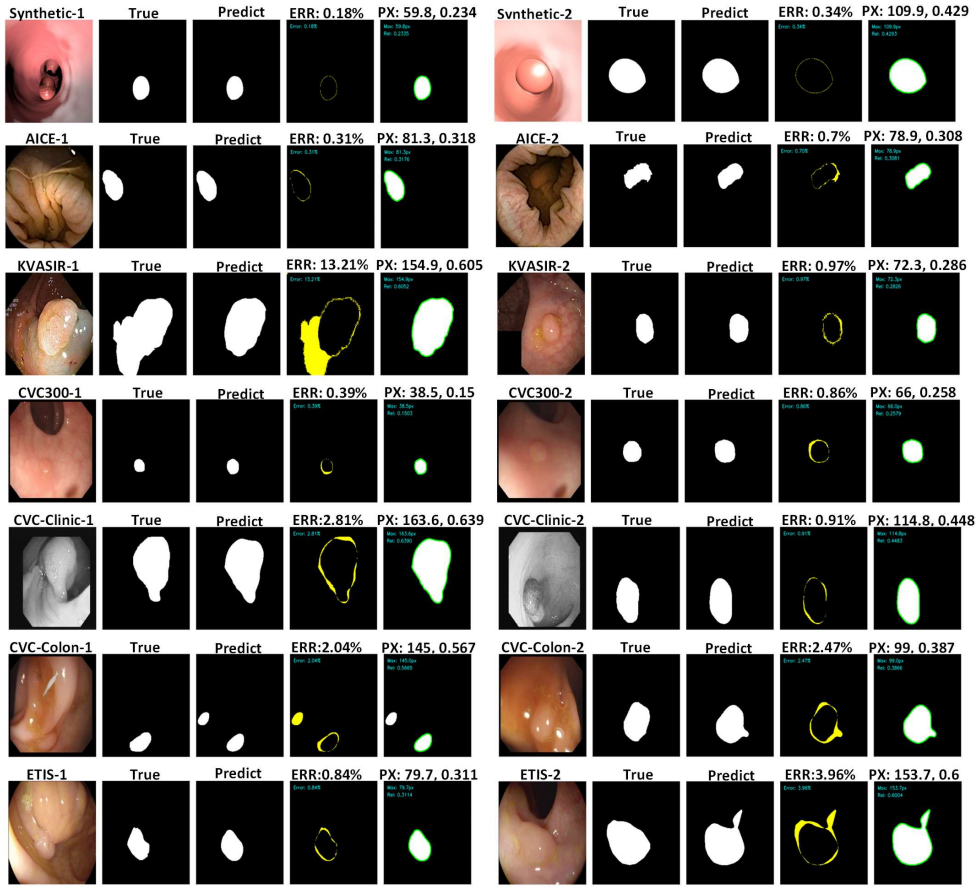


Fig. 3: Exemplary comparison of true mask, predicted mask, and error (ERR) obtained by taking a difference between true and predicted mask from the proposed Trans-UNet model.

Table 6: Performance (%) obtained on various datasets by our TransFuseNet model.

Datasets	ACC	IoU	Dice	PRC	Recall	TT (s)	ATI(s)
Kvasir	95.46	84.67	90.73	95.62	87.31	2.56	0.022
CVC-300	99.52	89.05	93.89	93.16	94.64	0.4	0.023
CVC-CoDB	98.94	88.91	93.83	95.77	92.06	1.72	0.023
CVC-CDB	98.07	89.46	94.22	94.51	93.94	3.48	0.028
ETIS	98.57	89.58	94.27	96.34	92.4	1.53	0.022
AICE	98.22	82.07	89.29	89.91	88.68	38.82	0.015
Synth-Colon	98.41	86.41	92.27	93.91	90.77	93.5	0.023

TT: Total time; ATI: Avg Time/Image

model trained on the CCE AICE dataset shows little degradation in performance reports but maintains uniformity in its performance when validated on optical colonoscopy-based image datasets. However, the models trained on optical colonoscopy-based image datasets fail to maintain their performance when validated on the CCE AICE dataset. The performance drop observed when applying optical colonoscopy-trained models to CCE data highlights a persistent and clinically significant challenge in medical image analysis namely, domain shift between imaging modalities. Although both optical colonoscopy and CCE aim to observe the gastrointestinal system, they differ significantly in terms of image acquisition settings such as resolution, illumination, viewing angles, field of view, color consistency, and the degree of bowel preparation. Optical colonoscopy often benefits from light control, increased spatial resolution, and active scope mobility, which results in sharper anatomical boundaries and greater textural depth. In contrast, CCE images are collected passively, typically under inadequate illumination and with minimal control over motion, orientation, or bowel cleaning, resulting in reduced contrast, fluctuating lighting aberrations, and motion blur. These intrinsic discrepancies have a major impact on the feature distributions acquired by deep learning models, explaining the poor cross-domain generalizability of CE-trained networks to CCE data. This lack of knowledge transfer and domain adaptation from optical colonoscopy to CCE limits the model's generalizability. Therefore, we aim to combine the trained model to facilitate effective learning, domain shift, and knowledge transfer when applied to either optical colonoscopy, CCE, or synthetic images, respectively. For this, we tested the developed TransFuseNet model that fuses the weights from different models to provide effective and accurate polyp segmentation. Table 6 indicates the performance report on various datasets using the TransFuseNet model. It is noteworthy that while validating the performance of the TransFuseNet model, we only used the validation set of different datasets. The rationale for selecting the validation folders for performance evaluation is due to the exposure of images from the train folder during training, which can cause data leakage to provide biased decisions. We fused all the models to obtain the overall effective model. However, we aim to develop a model that yields overall improvement across other datasets. Therefore, we started our simulation to validate the performance of all the datasets. The fusion strategy follows the adaptive model selection framework described earlier, where the best-performing combination is selected based on Dice score and integrated using pixel-wise confidence-weighted averaging. The performance report in Table 6 shows that post-fusion improves model performance and yields higher evaluation metrics across multiple datasets. The summary indicates that the TransFuseNet model achieved strong performance on the CVC-300, CVC-CoDB, CVC-CDB, and ETIS datasets, with IoU values of 89.05%, 88.91%, 89.46%, and 89.58%, respectively, and corresponding Dice scores exceeding 93%. In comparison to individual models, the analysis reflects consistent improvements in IoU of approximately 7%, 4%, 3%, and 6% on the CVC-300, CVC-CoDB, CVC-CDB, and ETIS datasets, respectively. However, a slight degradation in performance is observed for the Kvasir and Synth-Colon datasets, where IoU values are 84.67% and 86.41%, respectively. This variation can be attributed to dataset-specific characteristics and domain differences, which influence the adaptive fusion process. In terms of computational performance, the model shows efficient inference with an average time per image ranging from 0.015s to 0.028s, corresponding to an average processing speed of approximately 46 frame per second (FPS) across all datasets. This satisfies

real-time requirements for clinical applications, as typical endoscopic video streams operate $\geq$ 30 FPS [72]. Therefore, the proposed model is capable of processing frames at or above clinical video rates, making it suitable for real-time polyp segmentation in practical deployment scenarios.

6 Discussion

Table 7: Performance comparison (%) of the proposed TransFuseNet model with state-of-the-art techniques.

Dataset	**CVC-CDB**		**Kvasir**		**CVC-CoDB**		**ETIS**		**CVC-300**		**AICE**	
Model	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
UNet [25]	82.3	75.5	81.8	74.6	50.4	43.6	39.8	33.5	71.0	62.7	81.19	68.34
UNet++ [27]	79.4	72.9	82.1	74.4	48.2	40.8	40.1	34.4	70.7	62.4	87.64	78.0
SFA [28]	70.0	60.7	72.3	61.1	45.6	33.7	29.7	21.7	46.7	32.9	–	–
PraNet [36]	89.9	84.9	89.8	84.0	70.9	64.0	62.8	56.7	87.1	79.7	–	–
ACSNet [53]	88.2	82.6	89.8	83.8	71.6	64.9	57.8	50.9	86.3	78.7	84.95	77.87
DCRNet [50]	86.9	80.0	90.4	84.4	66.1	57.6	50.9	43.2	75.3	67.0	–	–
SANet [34]	91.6	85.9	90.4	84.7	75.2	66.9	75.0	65.4	88.8	81.5	–	–
MSNet [46]	91.8	86.9	90.5	84.9	74.7	66.8	72.0	65.0	86.2	79.6	–	–
UACANet [39]	91.6	87.0	90.5	85.2	78.3	70.4	69.4	61.5	90.2	83.7	–	–
LDNet [54]	88.1	82.5	88.7	82.1	79.4	71.5	77.8	70.7	89.3	82.6	–	–
SAM-Adapter [29]	77.4	67.3	84.7	76.3	67.1	56.8	59.0	47.6	81.5	72.5	–	–
AutoSAM [55]	75.1	64.2	78.4	67.5	53.5	41.8	40.2	30.8	82.9	73.9	–	–
SAMPath [30]	75.0	64.4	82.8	73.0	63.2	51.6	55.5	44.2	84.4	75.6	–	–
PolypSAM [32]	91.7	87.4	90.1	**86.4**	88.9	83.8	92.9	88.9	90.3	85.2	–	–
Polyp-SAM++ [33]	91.0	86.0	90.0	86.0	–	–	–	–	73.0	69.0	–	–
AID-U-Net [48]	74.08	58.83	–	–	–	–	–	–	–	–	88.12	78.76
TransFuseNet	**94.22**	**89.46**	**90.73**	84.67	**93.83**	**88.91**	**94.27**	**89.58**	**93.89**	**89.05**	**89.29**	**82.07**

We compared the performance of the developed TransFuseNet model with state-of-the-art techniques on different datasets. Table 7 summarizes performance with the state-of-the-art on the Kvasir, CVC-ColonDB, CVC-300, CVC-ClinicDB, ETIS, and CCE AICE datasets. The comparison report shows that our developed framework has surpassed the performance of the benchmark models on the CVC-ClinicDB, CVC-ColonDB, CVC-300, and CCE AICE datasets. However, for the Kvasir and ETIS datasets, the value obtained for the Dice score using our proposed model is the highest compared to other models, while values of IoU are the second best after PolypSAM [32]. Therefore, our developed TransFuseNet model captures subtle changes within the image, resulting in a higher IoU score of about 2% to 6% compared to the state-of-the-art on various datasets, thereby our TransFuseNet model outperforms the state-of-the-art. Finally, we plotted the prediction made by the TransFuseNet model

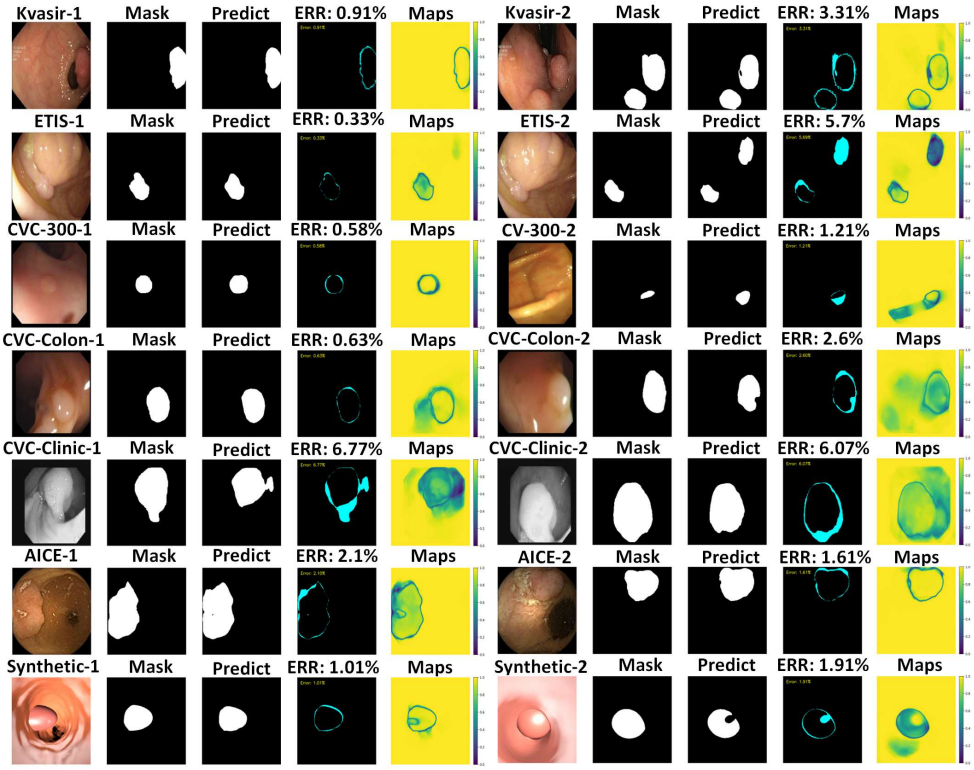


Fig. 4: Exemplary comparison of true and predicted mask along with confidence matrix obtained from the fusion of models.

as shown in Fig. 4. As shown in Fig. 4, the predictions yielded by the model are close to the GT, making our model effective and accurate. The error matrix shows that the proposed model effectively predicts smaller, larger, medium, and multiple polyps. However, the model captures extra details as false polyps, resulting in higher errors. We also plotted the confidence maps indicating the normalized confidence score obtained from the predicted masks. The yellow region in the images of the fifth column denotes the predictions with high confidence, whereas the purple color indicates the low-confidence regions. Typically, the model presented low confidence in detecting the boundaries of polyps. The green portion in the image indicates a region where the model has uncertainty. This shows a region where the model provides predictions close to the threshold. Our developed model mainly outperforms CNN-based segmentation models as it captures local features using CNN and transformer layers to capture global context and long-range dependencies. Additionally, the PBCWF method adaptively fuses predictions based on pixel-level certainty, hence improving boundary accuracy and overall resilience. Therefore, our developed model provides a multi-dimensional framework that provides accurate and effective segmentation with high-performance measures, with error measurement, external validation, pixel detection with relative polyp size detection,

and confidence in its prediction, making our model robust, effective, and explainable. Importantly, we evaluated the model using real-world endoscopic images, proving its suitability for the data commonly used in clinical practice. While this work focuses on methodological advancement, the methodology might aid in real-time clinical decision-making by offering more consistent and precise polyp delineation. Our model, with a total size of approximately 293 MB, ≈50 GFLOPS, ≈77 million parameters, and an average processing speed of approximately 46 FPS, the proposed solution remains deployable on standard clinical-grade GPUs. Further compression or pruning strategies could support lightweight deployment in capsule endoscopy review systems where hardware resources are constrained. Finally, validation in prospective or intra-procedural condition is required to determine its significance in typical diagnostic processes.

7 Conclusion

The research demonstrates a promising segmentation framework for polyps' detection using both optical colonoscopy and CCE images. Our developed TransFuseNet model captures subtle variations within the images by extracting local receptive and global receptive fields, capturing fine-grained spatial and long-range dependencies. Additionally, fusion refines the final predictions by adaptively emphasizing more reliable outputs at each pixel, leading to sharper boundaries and greater robustness in challenging or ambiguous regions. The meticulous combination of encoder, decoder, and fusion resulted in a robust knowledge transfer and domain shift to detect and capture subtle changes in high-resolution optical colonoscopy and low-resolution CCE images. The proposed model achieves efficient inference with real-time processing capability (approximately 46 FPS), making it suitable for practical clinical deployment. Our multi-characteristics and multi-dimensional framework provide accurate segmentation with high values of performance matrices, pixel size detection, and confidence evaluation. This method is especially useful in clinical contexts where picture quality, resolution, and visual features vary dramatically between modalities, resulting in more robust and generalizable segmentation for real-world endoscopic applications.

8 Acknowledgments

8.1 Funding

This research is part of AICE project (number 101057400) funded by the European Union, and it is part-funded by the United Kingdom government. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

8.2 Ethical Statement

The study was approved by the Regional Health Research Ethics committee (journal number S-20190100), was registered with the Regional Data Protection Agency (journal number 19/29858) and with ClinicalTrials (identifier NCT04049357). All participants received verbal and written study information prior to participation and signed informed consent was

obtained from each individual. The study was conducted in accordance with the declaration of Helsinki.

8.3 Declaration of interest

The authors declare no competing interests.





References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71**(3), 209–249 (2021) https://doi.org/10.3322/caac.21660

[2] Luo, C., Wang, Y., Deng, Z., Lou, Q., Zhao, Z., Ge, Y., Hu, S.: Colonic polyp segmentation based on transformer-convolutional neural networks fusion. Pattern Recognition **170**, 112116 (2026) https://doi.org/10.1016/j.patcog.2025.112116

[3] Uncovering inequalities: Colorectal cancer screening in europe. European Cancer Inequalities Registry (ECIR) (2024) https://doi.org/https://cancer-inequalities.jrc.ec.europa.eu/sites/default/files/ECIR-inequalities-factsheet-colorectal-cancer-screening-Mar2024.pdf

[4] Roshandel, G., Ghasemi-Kebria, F., Malekzadeh, R.: Colorectal cancer: Epidemiology, risk factors, and prevention. Cancers **16**(8) (2024) https://doi.org/10.3390/cancers16081530

[5] Eurostat: Cancer statistics – specific cancers. Statistics Explained (2024)

[6] Colorectal cancer statistics | World Cancer Research Fund — wcrf.org. https://www.wcrf.org/preventing-cancer/cancer-statistics/colorectal-cancer-statistics/. [Accessed 07-04-2025] (2022)

[7] Santucci, C., Mignozzi, S., Malvezzi, M., Boffetta, P., Collatuzzo, G., Levi, F., La Vecchia, C., Negri, E.: European cancer mortality predictions for the year 2024 with focus on colorectal cancer. Annals of Oncology **35**(3), 308–316 (2024)

[8] Andrea, M., Jepsen, R.K., Klein, M.F., Gögenur, I., Kuhlmann, T.P.: Colorectal serrated lesions and polyps in the danish population: A large nationwide register-based cohort study. Endoscopy International Open **11**(12), 1116–1122 (2023)

[9] Bernal, J., Sánchez, J., Vilariño, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012) https://doi.org/10.1016/j.patcog.2012.03.002 . Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011)

[10] Wang, S., Yang, Z., Sha, F., Qi, X., He, Z., Szeto, C.-H., Yang, Z., Tang, J.: Prevalence of incidental colorectal cancer and polyps in autopsies of different populations: a systematic review with meta-regression analysis. European Journal of Epidemiology **38**(9), 939–955 (2023)

[11] Li, X., Hu, M., Wang, Z., Liu, M., Chen, Y.: Prevalence of diverse colorectal polyps and risk factors for colorectal carcinoma in situ and neoplastic polyps. Journal of Translational Medicine **22**(1), 361 (2024)

[12] Levin, B., Lieberman, D.A., McFarland, B., Andrews, K.S., Brooks, D., Bond, J., Dash, C., Giardiello, F.M., Glick, S., Johnson, D., *et al.*: Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology. Gastroenterology **134**(5), 1570–1595 (2008)

[13] Ji, G.-P., Xiao, G., Chou, Y.-C., Fan, D.-P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)

[14] Moriyama, T., Uraoka, T., Esaki, M., Matsumoto, T.: Advanced technology for the improvement of adenoma and polyp detection during colonoscopy. Digestive Endoscopy **27**, 40–44 (2015)

[15] Demirbaş, A.A., Üzen, H., Fırat, H.: Spatial-attention convmixer architecture for classification and detection of gastrointestinal diseases using the kvasir dataset. Health Information Science and Systems **12**(1), 32 (2024)

[16] Sahafi, A., Wang, Y., Rasmussen, C., Bollen, P., Baatrup, G., Blanes-Vidal, V., Herp, J., Nadimi, E.: Edge artificial intelligence wireless video capsule endoscopy. Scientific reports **12**(1), 13723 (2022)

[17] Nadimi, E.S., Braun, J.-M., Schelde-Olesen, B., Khare, S., Gogineni, V.C., Blanes-Vidal, V., Baatrup, G.: Towards full integration of explainable artificial intelligence in colon capsule endoscopy's pathway. Scientific Reports **15**(1), 5960 (2025)

[18] Li, L., Yang, H., Zhang, J., Zheng, C., Chen, X., He, X.: Hierarchical contextual information aggregation for polyp segmentation. Scientific Reports (2026)

[19] Şener, A., Ergen, B.: Automatic detection of gastrointestinal system abnormalities using deep learning-based segmentation and classification methods. Health Information Science and Systems **13**(1), 37 (2025)

[20] Lieberman, D., Moravec, M., Holub, J., Michaels, L., Eisen, G.: Polyp size and advanced histology in patients undergoing colonoscopy screening: Implications for CT colonography. Gastroenterology **135**(4), 1100–1105 (2008) https://doi.org/10.1053/j.gastro.2008.06.083

[21] Pickhardt, P.J., Pooler, B.D., Kim, D.H., Hassan, C., Matkowskyj, K.A., Halberg, R.B.: The natural history of colorectal polyps: Overview of predictive static and dynamic features. Gastroenterology Clinics of North America **47**(3), 515–536 (2018) https://doi.org/10.1016/j.gtc.2018.04.004 . Radiologic Gastrointestinal Imaging

[22] Wang, S., Lin, S., Sun, F., Li, X.: Multi-feature fusion for accurate polyp segmentation using pyramid visual transformers. Expert Systems with Applications, 127558 (2025) https://doi.org/10.1016/j.eswa.2025.127558

[23] Ahamed, M.F., Islam, M.R., Nahiduzzaman, M., Chowdhury, M.E., Alqahtani, A., Murugappan, M.: Automated colorectal polyps detection from endoscopic images using multiresUNet framework with attention guided segmentation. Human-Centric Intelligent Systems **4**(2), 299–315 (2024)

[24] Xue, H., Yonggang, L., Min, L., Lin, L.: A lighter hybrid feature fusion framework for polyp segmentation. Scientific Reports **14**(1), 23179 (2024)

[25] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[26] Patel, K., Bur, A.M., Wang, G.: Enhanced U-Net: A feature enhancement network for polyp segmentation. In: 2021 18th Conference on Robots and Vision (CRV), pp. 181–188 (2021). https://doi.org/10.1109/CRV52889.2021.00032

[27] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2020) https://doi.org/10.1109/TMI.2019.2959609

[28] Fang, Y., Chen, C., Yuan, Y., Tong, K.-y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, pp. 302–310 (2019). Springer

[29] Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: SAM-adapter: Adapting segment anything in underperformed scenes. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3359–3367 (2023). https://doi.org/10.1109/ICCVW60793.2023.00361

[30] Zhang, J., Ma, K., Kapse, S., Saltz, J., Vakalopoulou, M., Prasanna, P., Samaras, D.:

SAM-path: A segment anything model for semantic segmentation in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 161–170 (2023). Springer

[31] Fang, Z., Liu, Y., Wu, H., Qin, J.: VP-SAM: Taming segment anything model for video polyp segmentation via disentanglement and spatio-temporal side network. In: European Conference on Computer Vision, pp. 367–383 (2024). Springer

[32] Li, Y., Hu, M., Yang, X.: Polyp-SAM: Transfer SAM for polyp segmentation. In: Medical Imaging 2024: Computer-Aided Diagnosis, vol. 12927, pp. 759–765 (2024). SPIE

[33] Biswas, R.: Polyp-SAM++: Can a text guided SAM perform better for polyp segmentation? arXiv preprint arXiv:2308.06623 (2023)

[34] Dong, H., Gu, H., Chen, Y., Yang, J., Chen, Y., Mazurowski, M.A.: Segment anything model 2: an application to 2D and 3D medical images. arXiv preprint arXiv:2408.00756 (2024)

[35] Dong, B., Wang, W., Fan, D.-P., Li, J., Fu, H., Shao, L.: Polyp-PVT: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)

[36] Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: PraNet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 263–273 (2020). Springer

[37] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., Lange, T.D., Halvorsen, P., D. Johansen, H.: ResUNet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), pp. 225–2255 (2019). https://doi.org/10.1109/ISM46123.2019.00049

[38] Mahmud, T., Paul, B., Fattah, S.A.: PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Computers in Biology and Medicine **128**, 104119 (2021) https://doi.org/10.1016/j.compbiomed.2020.104119

[39] Kim, T., Lee, H., Kim, D.: UACANet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21, pp. 2167–2175. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475375 . https://doi.org/10.1145/3474085.3475375

[40] Lin, Y., Wu, J., Xiao, G., Guo, J., Chen, G., Ma, J.: BSCA-Net: Bit slicing context attention network for polyp segmentation. Pattern Recognition **132**, 108917 (2022) https://doi.org/10.1016/j.patcog.2022.108917

[41] Weng, W., Zhu, X.: INet: Convolutional networks for biomedical image segmentation.

IEEE Access **9**, 16591–16603 (2021) https://doi.org/10.1109/ACCESS.2021.3053408

[42] Wang, C., Wang, S., Shao, S., Zhai, J.: DeepNeXt: a lightweight polyp segmentation algorithm based on multi-scale attention. Quantitative Imaging in Medicine and Surgery **14**(12), 8551 (2024)

[43] Lee, G.-E., Cho, J., Choi, S.-I.: Shallow and reverse attention network for colon polyp segmentation. Scientific Reports **13**(1), 15243 (2023)

[44] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, pp. 699–708 (2021). Springer

[45] Xia, M., Zhang, L., Yin, G., Dong, Y.: Polyp segmentation achieved through multi-scale feature fusion based on attention mechanism. In: 2024 11th International Conference on Behavioural and Social Computing (BESC), pp. 1–8 (2024). https://doi.org/10.1109/BESC64747.2024.10780683

[46] Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I, pp. 120–130. Springer, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-87193-2_12 . https://doi.org/10.1007/978-3-030-87193-2_12

[47] Zhang, W., Fu, C., Zheng, Y., Zhang, F., Zhao, Y., Sham, C.-W.: HSNet: A hybrid semantic network for polyp segmentation. Computers in Biology and Medicine **150**, 106173 (2022) https://doi.org/10.1016/j.compbiomed.2022.106173

[48] Tashk, A., Herp, J., Bjørsum-Meyer, T., Koulaouzidis, A., Nadimi, E.S.: AID-U-Net: An innovative deep convolutional architecture for semantic segmentation of biomedical images. Diagnostics **12**(12) (2022) https://doi.org/10.3390/diagnostics12122952

[49] Zhou, T., Zhou, Y., Gong, C., Yang, J., Zhang, Y.: Feature aggregation and propagation network for camouflaged object detection. IEEE Transactions on Image Processing **31**, 7036–7047 (2022) https://doi.org/10.1109/TIP.2022.3217695

[50] Yin, Z., Liang, K., Ma, Z., Guo, J.: Duplex contextual relation network for polyp segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761402

[51] Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognition **140**, 109555 (2023) https://doi.org/10.1016/j.patcog.2023.109555

[52] Pan, X.-L., Ding, J.-R., Li, X., Liu, S., Wang, J., Hua, B., Tang, G.-Z., Zhong, C.-H.:

MSBP-Net: A multi-scale boundary prediction network for automated polyp segmentation. Pattern Recognition **170**, 112101 (2026) https://doi.org/10.1016/j.patcog.2025.112101

[53] Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, pp. 253–262 (2020). Springer

[54] Zhang, R., Lai, P., Wan, X., Fan, D.-J., Gao, F., Wu, X.-J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 99–109 (2022). Springer

[55] Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (SAM) to medical images. arXiv preprint arXiv:2306.13731 (2023)

[56] Care for Colon 2015 (CFC2015). Morten Kobaek-Larsen, Odense University Hospital. [Accessed 06-08-2025]. https://clinicaltrials.gov/study/NCT04049357#more-information

[57] Deding, U., Bjørsum-Meyer, T., Kaalby, L., Kobaek-Larsen, M., Thygesen, M.K., Madsen, J.B., Kroijer, R., Baatrup, G.: Colon capsule endoscopy in colorectal cancer screening: Interim analyses of randomized controlled trial careforcolon2015. Endoscopy International Open **9**(11), 1712–1719 (2021)

[58] Deding, U., Herp, J., Havshoei, A.-L., Kobaek-Larsen, M., Buijs, M., Nadimi, E., Baatrup, G.: Colon capsule endoscopy versus CT colonography after incomplete colonoscopy. application of artificial intelligence algorithms to identify complete colonic investigations. United European Gastroenterology Journal **8**(7), 782–789 (2020) https://doi.org/10.1177/2050640620937593

[59] Kjølhede, T., Ølholm, A.M., Kaalby, L., Kidholm, K., Qvist, N., Baatrup, G.: Diagnostic accuracy of capsule endoscopy compared with colonoscopy for polyp detection: systematic review and meta-analyses. Endoscopy **53**(07), 713–721 (2021)

[60] Nadimi, E.S., Buijs, M.M., Herp, J., Kroijer, R., Kobaek-Larsen, M., Nielsen, E., Pedersen, C.D., Blanes-Vidal, V., Baatrup, G.: Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy. Computers & Electrical Engineering **81**, 106531 (2020) https://doi.org/10.1016/j.compeleceng.2019.106531

[61] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling, pp. 451–462 (2020). Springer

[62] Wang, F., hong: Polyp DataSet (2022) https://doi.org/10.6084/m9.figshare.21221579.v2

[63] Bernal, J., Sánchez, J., Vilariño, F.: CVC-colonDB: A database for assessment of polyp detection. Database (2012)

[64] Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: CVC-ClinicDB. https://polyp.grand-challenge.org/CVCClinicDB/

[65] Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S.R., Fernández-Esparrach, G., Dray, X., Liang, J., Histace, A.: Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge. IEEE Transactions on Medical Imaging **36**(6), 1231–1249 (2017) https://doi.org/10.1109/TMI.2017.2664042

[66] Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**, 283–293 (2014)

[67] Haj-Manouchehri, A., Mohammadi, H.M.: Polyp detection using cnns in colonoscopy video. IET Computer Vision **14**(5), 241–247 (2020)

[68] Moreu, E., McGuinness, K., O'Connor, N.E.: Synthetic data for unsupervised polyp segmentation. arXiv preprint arXiv:2202.08680 (2022)

[69] Moreu, E., McGuinness, K., O'Connor, N.E.: Synth-Colon dataset — enric1994.github.io. https://enric1994.github.io/synth-colon/. [Accessed 22-04-2025] (2022)

[70] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[71] Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021) https://doi.org/10.1016/j.imavis.2021.104117

[72] Zhang, C., Cui, R., Wang, G., Gao, T., Yan, J., Xian, W., Wei, X., Qin, Y.: Hardware-aware neural architecture search for real-time video processing in FPGA-accelerated endoscopic imaging. Applied Sciences **15**(20) (2025) https://doi.org/10.3390/app152011200