Advanced Medical Research

Global Science Publishing

Number 4 (2025), pp. 01–09

Article

Fine-Tuning SAM2 for Generalizable Polyp Segmentation with a Channel Attention-Enhanced Decoder

Yixiao Liu

Sichuan University, Chengdu 610000, China

Received: February 4, 2025; Revised: February 18, 2025; Accepted: February 19, 2025; Published: February 25, 2025

Abstract: Polyp segmentation is a critical task in medical image analysis, particularly in colonoscopy, where it plays a vital role in the early detection and treatment of colorectal cancer. In recent years, advancements in deep learning, especially the application of Convolutional Neural Networks (CNNs) and Transformer models, have significantly improved segmentation performance. Despite these advancements, the generalizability of these models across different datasets is often limited. Recently, Meta released the Segment Anything Model 2 (SAM2), which has demonstrated exceptional performance in both video and image segmentation tasks. This paper aims to develop a universal polyp segmentation model by fine-tuning the pre-trained encoder of SAM2. We introduce a learnable prompt layer within the Transformer blocks and employ a full-scale skip connection structure as a decoder to integrate multi-scale semantic features. Our model outperforms state-of-the-art methods on datasets such as Kvasir-Seg and CVC-ClinicDB. Additionally, our experiments show that the model exhibits excellent transfer learning capabilities on unseen datasets, making it a robust and generalizable model in the field of polyp segmentation.

Keywords: polyp segmentation; vision foundation model; SAM2; fine-tuning; generalizability

1. Introduction

Polyp segmentation plays a vital role in medical image analysis, particularly in colonoscopy, where early detection of colorectal polyps significantly enhances diagnostic accuracy and facilitates timely treatment. Accurate segmentation not only improves physicians' efficiency but also reduces the risk of misdiagnosis. However, the varying shapes of polyps, indistinct boundaries, and the complexity of surrounding intestinal tissues make polyp segmentation a challenging task.

In recent years, deep learning-based methods have achieved remarkable progress in polyp segmentation. Models such as U-Net [1] and SegNet [2] have demonstrated the effectiveness of convolutional neural networks (CNNs) in extracting spatial features for precise localization and segmentation. The emergence of Transformer-based architectures, particularly Vision Transformer (ViT) [3], has further advanced this field by leveraging self-attention mechanisms to capture long-range dependencies and contextual information. Hybrid models that integrate CNNs and Transformers, such as SSFormer-L [4], ColonFormer [5], TransUNet [6], and TransFuse [7], have further improved segmentation performance. However, despite their success on benchmark datasets, many of these models exhibit limited generalization when applied to unseen datasets, underscoring the need for a more adaptable and robust segmentation framework.

Vision foundation models (VFMs) have recently gained attention in computer vision due to their strong generalization capabilities across various visual tasks, including image classification, object detection, and segmentation. The Segment Anything Model (SAM), a pioneering foundational model, has demonstrated impressive segmentation capabilities through its powerful prompting mechanism. Building upon this, recent variants such as FastSAM [8], EfficientSAM [9], and SAM2 [10] have further refined segmentation

performance. SAM2, developed by Meta, introduces Memory Attention and benefits from training on larger datasets, enhancing both video and image segmentation. However, due to its inductive biases from natural image datasets, SAM2's performance in specialized fields like medical imaging remains suboptimal. Additionally, without manual prompts, SAM tends to produce class-agnostic segmentation, limiting its effectiveness in domain-specific tasks such as polyp segmentation.

In this work, we leverage the rich prior knowledge embedded in vision foundation models by fine-tuning the SAM2 encoder for polyp segmentation. Specifically, we introduce a learnable prompt layer within the Transformer encoder blocks and design a full-scale skip connection decoder to effectively integrate multi-scale semantic features. Additionally, we propose a Channel Attention (CA) module that learns critical channel information through multiple pooling operations, enhancing the model's ability to generalize across diverse polyp morphologies. By incorporating deep supervision during training, our method surpasses state-of-the-art approaches on benchmark datasets such as Kvasir-Seg and CVC-ClinicDB. Furthermore, extensive generalization experiments demonstrate that our model maintains strong performance on unseen datasets, highlighting its transferability.

The main contributions of this paper are summarized as follows:

- We fine-tuned the SAM2 encoder for the polyp segmentation task by freezing its parameters and embedding small, learnable prompt layers. Experiments show that our model outperforms existing state-of-the-art methods and demonstrates strong generalization capabilities, excelling in cross-dataset polyp segmentation tasks.
- We introduce a Channel Attention (CA) module that generates adaptive channel weights through diverse pooling techniques, enabling the model to focus on critical feature channels. Combined with a full-scale skip connection decoder, this approach enhances multi-scale feature integration and improves segmentation robustness across complex polyp structures.

2. Related Work

2.1. SAM2

SAM2 is a foundational model designed to address the challenge of promptable visual segmentation in images and videos. This model utilizes a transformer architecture with streaming memory, incorporating additional components such as a memory encoder, memory bank, and memory attention. These enhancements enable it to effectively process and utilize memory information for real-time video handling. As the successor to SAM, SAM2 features multiple improvements, significantly enhancing its segmentation capabilities:

Video Segmentation Ability: SAM2 introduces support for video segmentation, allowing it to segment objects within videos and perform cross-frame tracking and editing.

Increased Accuracy and Speed: For the same image segmentation tasks, SAM2 achieves a sixfold speed improvement over SAM while maintaining higher segmentation accuracy.

Fine-Grained Segmentation: SAM2 can deliver more precise and finegrained segmentation, extracting deeper semantic information, which enhances its potential for fine-tuning in small object segmentation tasks compared to SAM.

2.2. SAM2 in Medical Image Segmentation

According to studies [11,12], although SAM2 demonstrates exceptional performance in general segmentation tasks, its effectiveness in specialized and complex tasks such as medical image processing remains unsatisfactory. To address this limitation, researchers have undertaken various efforts to explore the potential of SAM2 within the medical field. Reference [11] introduced SAM2-Adapter, which embeds simple adapters into the hierarchical layers of the SAM2 encoder, enabling the model to learn task-specific knowledge and effectively overcoming SAM2's limitations in complex low-level segmentation tasks. In [13], the largest vision encoder (UNI) pretrained on histopathological images was integrated with the original SAM2 encoder, alongside the introduction of a learnable Kolmogorov–Arnold Networks (KAN) classification module to replace the manual prompt process, significantly enhancing SAM2's performance in pathological image segmentation. Ref. [14] combined SAM2 with the YOLOv8 model, utilizing YOLOv8's bounding box predictions to autonomously generate input prompts for SAM2, reducing reliance on manual annotations while achieving high-precision segmentation.

In contrast to the aforementioned works, this paper proposes a refined approach that leverages lightweight Adapters for fine-tuning the SAM2 encoder.

3. Method

The overall structure of the network is shown in Figure 1, with the main components including the SAM2 Encoder, Adapter, Channel Attention Module, Full-Scale Skip Connection Decoder, and Full-Scale Deep Supervision. This section will provide a detailed introduction to each module.

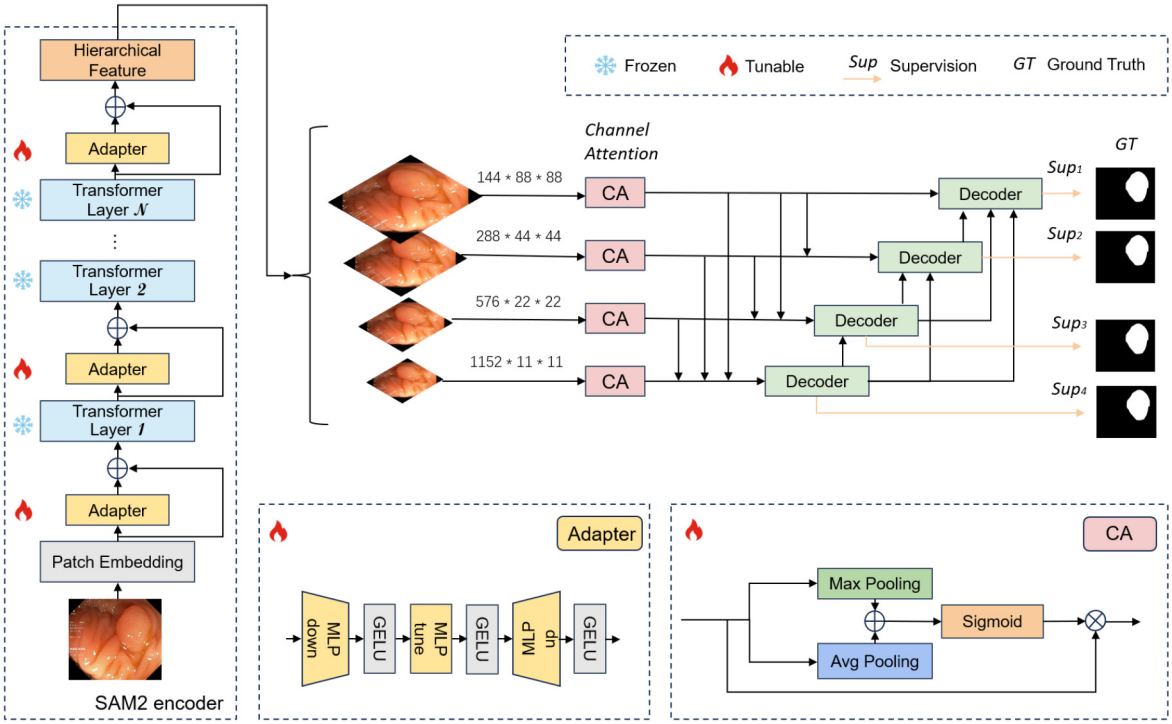

Figure 1. Pipeline of our model. Specifically targeting the core challenges of polyp segmentation—complex target morphology and variable scales—a channel attention module and a full-scale skip connection decoder are introduced to further optimize segmentation outcomes. By harnessing the rich prior knowledge embedded in visual foundation models, our method enhances the generalization capability of the model in polyp segmentation tasks.

SAM2 Encoder and Adapter: Our model uses a fine-tuned SAM2 encoder as the backbone network to extract feature information from polyp images. According to [11,15], introducing task-specific knowledge through the construction of an appropriate prompt layer can enhance the model's generalization in downstream tasks within this domain. We froze the weight parameters in the SAM2 pre-trained encoder and embedded learnable prompt layers between transformer layers, similar to the method used in [16]. This effectively fine-tunes the model for polyp segmentation while reducing training and inference costs and retaining the knowledge learned by SAM2 from large-scale image data. To further reduce computational load, we designed our adapter to consist of only

$$P_i = \mathrm{GELU}(\mathrm{MLP}_{down}(\mathrm{GELU}(\mathrm{MLP}_{tune}\mathrm{GELU}(\mathrm{MLP}_{up}(F_i))))) \quad (1)$$

where P_i and F_i are the input feature map and output prompt, respectively. MLP_{up} and MLP_{down} are the up-projection and down-projection linear layers, respectively, used to adapt the feature dimensions of the transformer layers. MLP_{tune} maintains the feature dimensions unchanged to learn task-specific knowledge for downstream tasks. Notably, all linear layers share parameter information.

Channel Attention (CA): For the output feature map of the i-th layer of the SAM2 encoder, $F_i \in R^{C\times H\times W}$ (where i = 1, 2, 3, 4) global average pooling and global max pooling are first performed. Here, C, H and W represent the number of channels, height, and width, respectively. The features extracted through these two pooling methods represent the average and maximum values of each channel. After combining them, a Sigmoid

activation function is used to generate the weights for each channel $f_i^* \in R^{C\times1\times1}$. This module achieves down-sampling and feature compression through the pooling process, thereby reducing the dimensionality of the feature map and eliminating redundant data. Global max pooling helps extract the most representative information from each channel, while global average pooling reflects all data evenly to minimize excessive information loss. Finally, the CA module multiplies the generated channel weights element-wise with the input feature map to obtain the filtered feature map, effectively extracting and compressing information while retaining key information. The core purpose of the CA module is to enhance the most important channels in the input feature map through weight adjustment. This aids in highlighting the key features shared between the source and target domains in transfer learning. The process can be expressed as:

$$f_i^* = \text{Sigmoid}(\text{MaxPool}(f_i) + \text{AvgPool}(f_i)) \tag{2}$$

$$f_i = f_i^* * f_i \tag{3}$$

where MaxPool and AvgPool are the max pooling and average pooling operations, respectively.

Full-Scale Skip Connection Decoder: Unlike SAM, SAM2 is based on a hierarchical vision transformer, allowing it to extract multi-layer information at different scales. This naturally provides advantages for recognizing polyps of various sizes across different datasets. To effectively combine high-level and low-level semantic information from different scales, we selected a full-scale skip connection network structure as the decoder. For the *i*-th layer node in the decoder, information is sourced from three places: shallower layers, the same layer, and deeper layers of the decoder feature maps. MaxPooling and bilinear interpolation are used to unify the dimensions of shallow and deep features, respectively, and finally, all layer information is concatenated. The convolution in the decoding layer is performed in two steps: first, individual convolutions on to integrate and extract information. The design of the full-scale skip connection decoder aims to effectively integrate features from different levels to enhance the model's performance in multi-scale tasks. For specific implementation details of this network, refer to [17], which will not be elaborated here. Only the modifications made to this model are explained: (1) We replaced the encoder outputs in the original model with the SAM2 encoder outputs processed by the CA module. (2) To accommodate the multi-scale feature layers output by the SAM2 encoder, we modified the decoder to have four layers.

Full-scale Deep Supervision: To better integrate multi-scale information, we train the model using deep supervision. In the full-scale skip connection structure, each decoder corresponds to a side output SO_i (i = 1, 2, 3, 4), which is supervised using ground truth. For deep supervision, the last layer of each decoder is passed through a standard 3 × 3 convolution layer, followed by bilinear upsampling and a sigmoid function. For each side output, binary cross-entropy loss L_{bce} and Dice loss L_{Dice} are computed with respect to the ground truthGT. The losses from the four side outputs are summed and averaged to obtain the total loss L_T for backpropagation, expressed as:

$$L_T = \frac{1}{n}\sum_{i=1}^{n}[L_{bce}(SO_i, \text{GT}) + L_{Dice}(SO_i, \text{GT})] \tag{4}$$

4. Experiments

4.1. Datasets

To evaluate the performance of our model, we selected the following two popular polyp segmentation datasets for experiments:

Kvasir-SEG is an open-source dataset for gastrointestinal polyp images and corresponding segmentation masks, manually annotated and verified by experienced gastroenterologists. It contains 1000 polyp images from the Kvasir dataset v2, along with their respective ground truth annotations. The images in Kvasir-SEG vary in resolution from 332 × 487 to 1920 × 1072 pixels, with polyps covering 100% of the images. The number of pixels in the images ranges from 849 to 1,094,201. This variety in scale makes Kvasir-SEG a challenging dataset for model segmentation, as it includes polyps of different sizes. The dataset is intended for research and development of new and improved methods for polyp segmentation, detection, localization, and classification, offering cutting-edge solutions for polyp-related tasks.

CVC-ClinicDB is a medical imaging dataset designed for colorectal cancer detection and research, particularly for the early detection and diagnosis of colorectal cancer. It consists of 612 high-resolution

colonoscopy images from rectal cancer patients, with each image annotated by experts to mark the cancerous lesions. This dataset is widely used in the training and validation of colorectal cancer models, image segmentation, medical image processing, and clinical research.

4.2 Evaluation Metrics

The performance of the our model is evaluated using several key metrics: Dice Score provides a balanced view of the model's effectiveness by measuring the overlap between the predicted changes and the ground truth. Dice Score, Intersection over Union (IoU), Precision and Recall. Precision measures the accuracy of the model's positive predictions, while Recall assesses its ability to identify all actual changes. IoU evaluates the spatial overlap between the predicted changes and the ground truth. These metrics are defined as follows:

$$\text{Dice Score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

where *TP*, *FP*, *FN*, and *TN* are the true positive, false positive, false negative, and true negative counts, respectively.

4.3. Implementation Detail

We trained our model to predict binary segmentation maps for RGB images. Following the conventions in [5,6,18–20], we resized the input images to a resolution of 352 × 352. To ensure the objectivity and fairness of the experiment, all models evaluated in this study used the same dataset preprocessing and splitting rules. Specifically, we partitioned the dataset into training, validation, and test sets with an 8:1:1 ratio, following the practices outlined in [5,8,18,19]. During training, several data augmentation techniques were applied: First, we applied Gaussian blurring to the images. Next, we randomly adjusted the brightness, contrast, and hue of the images. The images were also subjected to horizontal and vertical flipping. Additionally, affine transformations were applied, including rotations, translations, scaling, and shearing. Gaussian blurring and color jitter were applied only to the images, while flipping and affine transformations were applied to both the images and their corresponding segmentation maps.

We selected several existing state-of-the-art (SOTA) methods, including DUCK-Net [20], FCN-Transformer [18], Pra-Net [19], MSRF-Net [21], ResUnet++ [22] and U-Net [2], to conduct segmentation performance comparison experiments based on the Kvasir-SEG and CVC-ClinicDB datasets. Furthermore, to assess the generalizability of our model on unseen datasets, we designed a cross-dataset generalization experiment. In this experiment, we trained the model on one dataset and tested the results on another. Specifically, we used the pre-trained weights obtained from training on the Kvasir-SEG dataset and applied them to the test set of CVC-ClinicDB to evaluate performance, and vice versa. This process ensured that the model has the capability to transfer across different datasets and effectively adapt to various types of medical imaging data. We conducted the same experiment with the existing SOTA methods, and through comparison of the segmentation results from different models, we demonstrated the superior generalization and robustness of our model on unseen data.

The model was built using the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU with 24 GB of VRAM. The SAM2 encoder is initialized with the sam2 hiera large weights of size 224.4 M. During training, we used the AdamW optimizer with an initial learning rate of 1×10^{-4}, a batch size of 4, and a total of 150 epochs. The loss function combined Dice loss and binary cross-entropy loss. A learning rate scheduler reduced the learning rate by half when the Dice score did not improve, with a minimum learning rate of 1×10^{-6}. Model performance was evaluated using metrics such as Dice score, IoU, precision, recall, and accuracy, and the model was saved whenever the Dice score improved.

4.4. Evaluation

Qualitative Results: Figure 2 compares the segmentation results of different SOTA models on three test images from the Kvasir-SEG dataset. It is evident that our model leverages the multi-scale features extracted by SAM2 and the decoder's full-scale skip connection structure to effectively fuse and learn high-level and low-level semantics. In the second image, where large and small polyps overlap, and in the third image, where a small polyp is hidden in a shadow, most models capture only the more prominent polyps while ignoring the lower-layer or obscured ones. In contrast, our model excels at capturing global features, enabling it to accurately predict polyps in all these challenging scenarios.

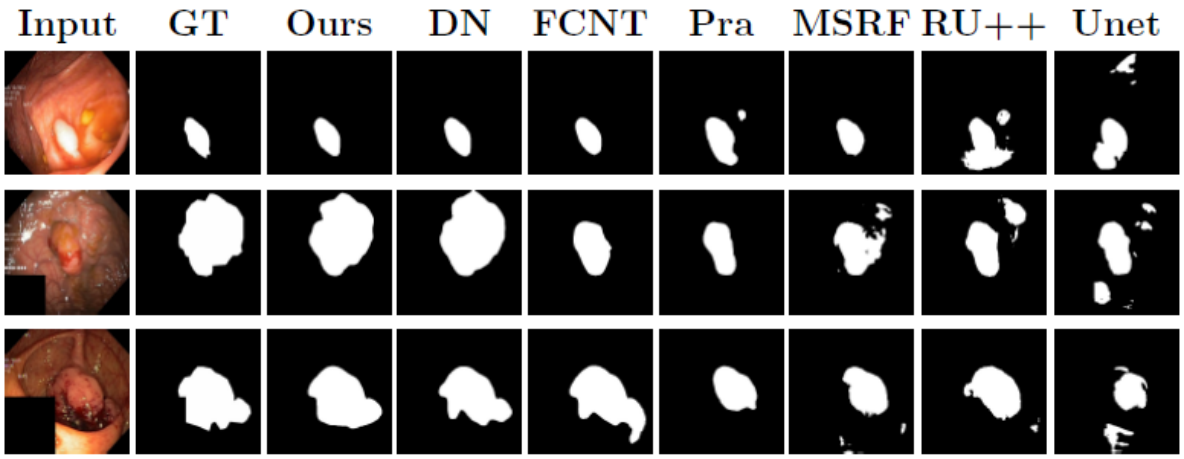

Figure 2. Vision Comparison of Models on Different Datasets. Note: **Unet** refers to U-Net, **RU**++ refers to ResUnet++, **MSRF** refers to MSRF-Net, Pra refers to Pra-Net, **FCNT** refers to FCN-Transformer, and **DN** refers to Duck-Net.

Learning Capability: Tables 1 and 2 compare the performance of different methods on the Kvasir-SEG and CVC-ClinicDB datasets, respectively. The evaluation metrics include mDice, mIoU, mPrecision, and mRecall, where m represents the average performance of the model on the test set. For methods with training setups and dataset splits similar to ours, we cite the experimental results reported in their original papers. We trained and tested methods that did not use the same datasets (e.g., U-Net) using our experimental settings.

Table 1. Model Evaluation on Kvasir-SEG.

	Kvasir-SEG			
Model	**mDice**	**mIoU**	**mPrec**	**mRec**
U-Net	0.7796	0.7152	0.7041	0.7688
ResUnet++	0.8133	0.7927	0.8774	0.7064
MSRF-Net	0.9217	0.8914	0.9666	0.9198
Pra-Net	0.898	0.840	-	-
FCN-Transformer	0.9385	0.8903	0.9459	0.9401
DUCK-Net	0.9502	-	0.9628	0.9379
Our Model	0.9489	0.9064	0.9565	0.9470

Table 2. Model Evaluation on CVC-ClinicDB Dataset.

	CVC-ClinicDB			
Model	**mDice**	**mIoU**	**mPrec**	**mRec**
U-Net	0.7902	0.7427	0.7183	0.8065
ResUnet++	0.7955	0.7962	0.8785	0.7022
MSRF-Net	0.9420	0.9043	0.9427	0.9567
Pra-Net	0.899	0.849	-	-
FCN-Transformer	0.9469	0.9020	0.9525	0.9441
DUCK-Net	0.9478	-	0.9468	0.9489
Our Model	0.9482	0.9049	0.9540	0.9443

The results demonstrate that on the Kvasir-SEG dataset, our model achieved a mDice of 0.9489 and a mIoU of 0.9064, showing superior performance compared to existing SOTA models. Although DUCK-Net, currently a leading polyp segmentation network, slightly outperformed our model in mDice (0.9502), it is worth

noting that DUCK-Net is specifically designed for polyp segmentation. In contrast, our model, derived from fine-tuning a larger foundational model, may not excel in all individual metrics. However, our model delivers more balanced performance across multiple key metrics, indicating a more comprehensive segmentation capability. On the CVC-ClinicDB dataset, our model achieved nearly the best performance across all metrics, particularly excelling in core indicators such as mDice (0.9482) and mIoU (0.9049). It was only slightly outperformed by DUCK-Net in mRec. These results highlight the excellent performance of our model on individual datasets, demonstrating its robust learning capability and adaptability.

Generalization ability: Generalization remains a critical challenge in medical image segmentation, where models trained on a specific dataset often struggle when applied to unseen datasets. To evaluate our model's cross-dataset generalization, we conduct experiments by training on one dataset and testing on another.

Training on Kvasir-SEG, *Testing on CVC-ClinicDB*. Table 3 presents the results when models are trained on the Kvasir-SEG dataset and tested on the CVC-ClinicDB dataset. Our model achieves an mDice of 0.9182 and an mIoU of 0.8637, significantly outperforming all other methods. Compared to FCN-Transformer (mDice of 0.8760) and DUCK-Net (mDice of 0.8464), our model maintains higher segmentation accuracy across all metrics, underscoring its strong transfer learning capability.

Table 3. Trained on Kvasir-SEG and tested on CVC-ClinicDB.

Trained on Kvasir-SEG, Tested on CVC-ClinicDB				
Model	**mDice**	**mIoU**	**mPrec**	**mRec**
U-Net	0.7132	0.6192	0.7240	0.7691
ResUnet++	0.5732	0.4763	0.6954	0.5811
MSRF-Net	0.6914	0.6280	0.6973	0.7811
Pra-Net	0.7815	0.7223	0.8266	0.8074
FCN-Transformer	0.8760	0.7828	0.8863	0.8947
DUCK-Net	0.8464	0.7724	0.8953	0.8260
Our Model	0.9182	0.8637	0.9240	0.9185

Training on CVC-ClinicDB, *Testing on Kvasir-SEG*. Table 4 evaluates the reverse scenario, where models are trained on CVC-ClinicDB and tested on Kvasir-SEG. Our model achieves the highest performance, with an mDice of 0.8944 and an mIoU of 0.8375, surpassing FCN-Transformer (mDice of 0.8764) and DUCK-Net (mDice of 0.8373). These results highlight our model's ability to adapt to different polyp segmentation datasets while maintaining stable and high-performance segmentation. Our model's strong generalization can be attributed to the rich feature extraction capabilities of SAM2's pre-trained encoder, which enables effective adaptation to varying polyp distributions. This characteristic is highly valuable in medical image segmentation, as it reduces the need for dataset-specific retraining while maintaining high segmentation quality. Thus, it is a cost-effective and practical solution for real-world clinical applications.

Table 4. Trained on CVC-ClinicDB and tested on Kvasir-SEG.

Trained on CVC-ClinicDB, Tested on Kvasir-SEG				
Model	**mDice**	**mIoU**	**mPrec**	**mRec**
U-Net	0.6014	0.4413	0.5398	0.6616
ResUnet++	0.5236	0.4315	0.6673	0.4659
MSRF-Net	0.7417	0.6320	0.8159	0.7556
Pra-Net	0.7851	0.6914	0.7790	0.8335
FCN-Transformer	0.8764	0.8132	0.9211	0.8660
DUCK-Net	0.8373	0.7754	0.7631	0.8550
Our Model	0.8944	0.8375	0.9483	0.8853

Ablation study: To verify the effectiveness of the Channel Attention module (CA), we conducted an ablation study on the Kvasir-SEG dataset, with results shown in Table 5. After removing CA, the model's mDice dropped from 0.9489 to 0.9402, and the mIoU decreased from 0.9064 to 0.8893. This indicates that CA significantly enhances the model's ability to capture detailed features, thereby improving segmentation accuracy and consistency. These findings demonstrate that CA plays a critical role in enhancing feature representation and optimizing segmentation performance.

Table 5. Ablation study of the proposed—on the Kvasir-SEG dataset.

Model	mDice	mIoU	mPrec	mRec
without CA	0.9402	0.8893	0.9328	0.9436
(Ours)	0.9489	0.9064	0.9565	0.9470

5. Conclusion

In this paper, we proposed a novel polyp segmentation model by fine-tuning the Segment Anything Model 2 (SAM2) encoder, combined with a learnable prompt layer and a full-scale skip connection decoder. Our approach effectively leverages the pre-trained knowledge embedded in SAM2, making it both computationally efficient and capable of producing high-quality segmentation results. Additionally, we evaluated the model's generalization performance by testing it on multiple unseen datasets. Our results show that the model maintains robust segmentation capabilities across these diverse datasets, further highlighting its strong generalization ability and potential for real-world clinical applications.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

1. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
2. Badrinarayanan V, Kendall A, Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2017; **39**(**12**): 2481–2495.
3. Dosovitskiy A. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* 2020, arXiv:2010.11929.
4. Wang J, Huang Q, Tang F, *et al*. Stepwise Feature Fusion: Local guides Global. In Proceedings of the 2022 International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; Springer: Cham, Switzerland, 2022; pp. 110–120.
5. Duc NT, Oanh NT, Thuy NT, *et al*. Colonformer: An Efficient Transformer Based Method for Colon Polyp Segmentation. *IEEE Access* 2022; **10**: 80575–80586.
6. Chen J, Lu Y, Yu Q, *et al*. Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* 2021, arXiv:2102.04306.
7. Zhang Y, Liu H, Hu Q. Transfuse: Fusing Transformers and CNNS for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; Springer: Cham, Switzerland, 2021; pp. 14–24.
8. Zhao X, Ding W, An, Y, *et al*. Fast Segment Anything. *arXiv* 2023, arXiv:2306.12156.
9. Xiong Y, Varadarajan B, Wu L, *et al*. Efficient sam: Leveraged masked image pretraining for efficient segment anything. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16111–16121.
10. Ravi N, Gabeur V, Hu YT, *et al*. Sam2: Segment Anything in Images and Videos. *arXiv* 2024, arXiv:2408.00714.
11. Chen T, Lu A, Zhu L, *et al*. Sam2-Adapter: Evaluating & Adapting Segment Anything 2 in Downstream Tasks: Camouflage, Shadow, Medical image Segmentation, and More. *arXiv* 2024, arXiv:2408.04579.
12. Ma J, Kim S, Li F, *et al*. Segment Anything in Medical Images and Videos: Benchmark and Deployment. *arXiv* 2024, arXiv:2408.03322.

13. Zhang M, Wang L, Chen Z, *et al*. Path-sam2: Transfer sam2 for Digital Pathology Semantic Segmentation. *arXiv* 2024, arXiv:2408.03651.
14. Mansoori M, Shahabodini S, Abouei J, *et al*. Self-Prompting Polyp Segmentation in Colonoscopy Using Hybrid yolo-sam2 Model. *arXiv* 2024, arXiv:2409.09484.
15. Chen T, Zhu L, Deng C, *et al*. Sam-Adapter: Adapting Segment Anything in Underperformed Scenes. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3367–3375.
16. Qiu Z, Hu Y, Li H, *et al*. Learnable Ophthalmology sam. *arXiv* 2023, arXiv:2304.13425.
17. Huang H, Lin L, Tong R, *et al*. Unet 3+: A Full-Scale Connected UNET for Medical Image Segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
18. Sanderson E, Matuszewski BJ. FCN-Transformer Feature Fusion for Polyp Segmentation. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis, Cambridge, UK, 27–29 July 2022; pp. 892–907.
19. Fan D-P, Ji G-P, Zhou T, *et al*. Pranet: Parallel Reverse Attention Network for polyp Segmentation. In Proceedings of the 2020 International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; pp. 263–273.
20. Dumitru R-G, Peteleaza D, Craciun C. Using Duck-Net for Polyp Image Segmentation. *Scientific Reports* 2023; **13**(**1**): 9803.
21. Srivastava A, Jha D, Chanda S, *et al*. MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation. *IEEE Journal of Biomedical and Health Informatics* 2021; **26**(**5**): 2252–2263.
22. Jha D, Smedsrud PH, Riegler MA, *et al*. Resunet++: An Advanced Architecture for Medical Image Segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 225–2255.