WSPoly-SAM: Weakly-Supervised and Self-Guided Fine-Tuning of SAM for Colonoscopy Polyp Segmentation

Tingting Cai , Hongping Yan * , Kun Ding , Yan Zhang , Yueyue Zhou

Posted Date: 7 May 2024

doi: 10.20944/preprints202405.0389.v1

Keywords: weakly supervised learning; polyp segmentation; segment anything model; pseudo-label generation; deep learning

Article

WSPoly-SAM: Weakly-Supervised and Self-Guided Fine-Tuning of SAM for Colonoscopy Polyp Segmentation

Tingting Cai [1,2], Hongping Yan [1]*, Kun Ding [2], Yan Zhang [1,2] and Yueyue Zhou[1,2]

[1] School of Information Engineering, China University of Geosciences, Beijing 100083, China; caitingting0903@163.com (T.C.)
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; kun.ding@ia.ac.cn (K.D.)
* Correspondence: yanhp@cugb.edu.cn

Abstract: Ensuring precise segmentation of colorectal polyps holds critical importance in the early diagnosis and treatment of colorectal cancer. Nevertheless, existing deep learning-based segmentation methods are fully-supervised, requiring extensive precise manual pixel-level annotation data, which leads to high annotation costs. Additionally, it remains challenging to train large-scale segmentation models when confronted with limited colonoscopy data. To address these issues, we introduce the general segmentation foundation model—Segment Anything Model (SAM) into the field of medical image segmentation. Fine-tuning the foundation model is an effective approach to tackle sample scarcity. However, current SAM fine-tuning techniques still rely on precise annotations. To overcome this limitation, we propose WSPoly-SAM, a novel weakly-supervised approach for colonoscopy polyp segmentation. WSPoly-SAM utilizes weak annotations to guide SAM in generating segmentation masks, which are then treated as pseudo-labels to guide the fine-tuning of SAM, thereby reducing the dependence on precise annotations data. To improve the reliability and accuracy of pseudo-labels, we have designed a series of enhancement strategies to improve the quality of pseudo-labels and mitigate the negative impact of low-quality pseudo-labels. Experimental results on five medical image datasets demonstrate that WSPoly-SAM outperforms current fully-supervised mainstream polyp segmentation networks on the Kvasir-SEG, ColonDB, CVC-300, and ETIS datasets. Furthermore, by using different amounts of training data in weakly-supervised and fully-supervised experiments, it is found that weakly-supervised fine-tuning can save 70% to 73% of annotation time costs compared to fully-supervised fine-tuning. This study provides a new perspective on the combination of weakly-supervised learning and SAM models, significantly reducing annotation time and offering insights for further development in the field of colonoscopy polyp segmentation.

Keywords: weakly supervised learning; polyp segmentation; segment anything model; pseudo-label generation; deep learning;

1. Introduction

Medical image segmentation plays a crucial role in medical image analysis by accurately identifying and delineating regions of interest (ROIs) in medical images, such as organs, lesions, and tissues [1]. This task is pivotal for various clinical applications, including disease diagnosis, treatment planning, and disease progression monitoring [2]. In the field of colonoscopy polyp segmentation, traditional segmentation methods struggle to address the complexity and variability in images due to the diverse locations, sizes, shapes, and textures of polyps, making automated polyp segmentation extremely challenging.

In recent years, substantial advancements have been achieved in both semantic segmentation [3,4] and medical image segmentation [5,6] through the utilization of deep learning methodologies. Among these advancements, methods for colonoscopy polyp segmentation have also been widely explored [7,8]. Nevertheless, existing segmentation models often focus on specific imaging modalities and segmentation targets [9]. Constructing large-scale segmentation models necessitates a considerable volume of medical data. However, acquiring and annotating such data poses significant challenges and expenses,

 doi:10.20944/preprints202405.0389.v1

as they demand skilled physicians to conduct meticulous pixel-level annotations. To mitigate this challenge, fine-tuning foundational models has emerged as an effective approach, particularly in scenarios with limited data availability [10]. In recent times, a noteworthy segmentation foundation model known as the Segment Anything Model (SAM) [11] has emerged as a standout performer, primarily due to its remarkable performance in zero-shot segmentation tasks. Trained on a large-scale visual segmentation dataset, SAM exhibits robust generalization capabilities and adaptability to previously unseen data. Comprising an image encoder, a prompt encoder, and a mask decoder, SAM can generate high-quality object masks from various input prompts such as points, boxes, and masks. Given its exceptional performance across multiple computer vision benchmark tests, numerous researchers have begun fine-tuning SAM to tailor it for specific medical image segmentation domains. For instance, Hu et al. [12] directly fine-tuned SAM (excluding the image encoder) for skin cancer segmentation tasks, while Wu et al. [13] introduced the Medical SAM Adapter (MSA), incorporating domain-specific knowledge by introducing adapters between the SAM encoder and decoder. While these approaches have achieved notable advancements, their reliance primarily on fully-supervised learning entails a considerable investment of time and manpower for precise pixel-level annotations.

To reduce the dependence of SAM fine-tuning on precise pixel-level annotation data, this study proposes a SAM method for colonoscopy polyp segmentation called WSPoly-SAM, based on weak supervision and self-guided fine-tuning. Specifically, the SAM model is guided to generate segmentation masks based on weak annotations (bounding box annotations in this study), which are then treated as pseudo-labels to guide the fine-tuning of SAM. Utilizing self-generated pseudo-labels for fine-tuning not only decreases the dependence on precisely annotated data but also avoids the need to introduce additional segmentation models. High-quality pseudo-labels provide additional information, helping the model in better capturing specific features. To enhance the reliability and accuracy of pseudo-labels, this study designs three enhancement strategies. Firstly, employing a multi-augmentation fusion strategy involves generating multiple augmented views for each image and then fusing their corresponding segmentation masks. This approach highlights reliable prediction results and counteracts biases introduced by image augmentation, thereby enhancing the quality of the final fused mask. Secondly, incorporating a pixel-level weighting strategy involves assigning higher weights to pixels with high certainty predictions through an entropy-based pixel-level weighting mechanism, further improving the accuracy of the segmentation mask. Finally, introducing mask post-processing techniques which can diminish potential noise and inaccuracies in the masks generated by SAM. These three strategies aim to optimize the details of the segmentation masks, improve their consistency with ground-truth labels.

The main contributions of this study include:

- The proposal of a novel weakly-supervised and self-guided fine-tuning method of SAM for colonoscopy polyp segmentation. This method reduces the dependence on precise annotations by fully utilizing SAM's zero-shot capability to use weak annotations for guiding the generation of segmentation masks. These masks are then treated as pseudo-labels, which are then utilized for self-guided fine-tuning, avoiding the need to introduce additional segmentation models.
- The introduction of a series of pseudo-label enhancement strategies to generate high-quality pseudo-labels. These enhancement strategies, including multi-augmentation fusion, pixel-level weighting, and mask post-processing techniques, enable the acquisition of more accurate pseudo-labels.
- Experimental results on five medical image datasets demonstrate that WSPoly-SAM outperforms current fully-supervised mainstream polyp segmentation networks on the Kvasir-SEG, ColonDB, CVC-300, and ETIS datasets. Specifically, on the ColonDB dataset, our method demonstrated an improvement of 9.4% in mDice score and 9.6% in mIoU score compared to state-of-the-art networks, representing a significant breakthrough in the field of colonoscopy polyp segmentation. Furthermore, by using different amounts of training data in weakly-supervised and

 doi:10.20944/preprints202405.0389.v1

fully-supervised experiments, it is found that weakly-supervised fine-tuning can save 70% to 73% of annotation time costs compared to fully-supervised fine-tuning.

The remaining sections of this paper are structured as follows: Section 2 discusses related work, Section 3 outlines the methods employed in our study, Section 4 presents the dataset and analyzes the experimental findings, and finally, Section 5 provides the concluding remarks of our study.

2. Related works

2.1. Polyp Segmentation

Medical image segmentation has been a research hotspot in the field of medical image processing and analysis. The advancement of deep learning has continuously propelled the development of this field, leading to numerous image segmentation methods based on deep neural networks. The Fully Convolutional Neural Network (FCN) [14] is the first end-to-end pixel-level classification method, which opened a new paradigm in the segmentation field. UNet [15], a prominent variant of Fully Convolutional Networks (FCN), is renowned for its symmetric U-shaped encoder-decoder architecture, incorporating skip connections to fuse deep and shallow features across different scales, making it a standard benchmark architecture for polyp segmentation methods. Subsequent variants like UNet++ [16], ResUNet [17] and U2-Net [18] have aimed to enhance feature extraction, receptive field, and multi-scale information integration. Despite the progress in polyp segmentation, there remains a need for comprehensive environmental context around polyps. Various strategies have been explored to enhance polyp segmentation and boundary recognition, including dilated convolutions [19], integration of ASPP modules [20], and utilization of RFB module [21]. Approaches like SFA [22] introduced a shared encoder branch and two decoder branches, along with an innovative edge-sensitive loss to improve polyp region segmentation and boundary recognition. ACSNet [23] leveraged both local and global contextual features to guide encoder modules, with a particular emphasis on the edge region. CCBANet [24] obtained more comprehensive context awareness through cascaded context and balanced attention to improve segmentation performance.

Although these methods have made significant progress, they often require a large amount of training data to produce desirable results. This requirement is challenging for specific tasks, such as colonoscopy polyp segmentation, where the dataset is limited. To address this issue, this work fine-tunes the SAM model pre-trained on large-scale datasets so as to exploit its general knowledge for the specific task.

2.2. Segment Anything Model(SAM) Related

The Segment Anything Model (SAM), proposed by Kirillov et al., serves as a foundational model for universal image segmentation and has been trained on an extensive dataset containing over one billion masks. While SAM demonstrates impressive capabilities in generating accurate object masks using prompts or autonomously, its effectiveness in medical image segmentation is limited due to notable differences between natural and medical images. SAM may encounter challenges, particularly in tasks with faintly defined boundaries commonly found in medical images [25–35]. Recognizing this limitation, Ma et al. [36] introduced MedSAM, which leverages a diverse array of medical image datasets and employs fine-tuning techniques to tailor SAM specifically for medical image segmentation tasks. Cheng et al. [37] introduced SAM-Med2D, a comprehensive exploration of applying SAM to medical 2D images. This approach incorporates learnable adapter layers in the image encoder, fine-tunes the prompt encoder, and updates the mask decoder through interactive training. Moreover, SAM has been employed in weakly-supervised segmentation tasks. For instance, Jiang et al. [38] proposed using SAM to generate pseudo-labels, which were subsequently utilized in training other weakly-supervised semantic segmentation models. Additionally, He et al. [39] utilized sparse annotations as prompts to generate segmentation masks, enhancing the training of other hidden target segmentation models.

 doi:10.20944/preprints202405.0389.v1

Compared to previous studies [36,37], which required full supervision during fine-tuning, and necessitated precise pixel-level annotations, our method circumvents this requirement by using weak annotations to guide SAM in generating pseudo-labels as supervision masks. In contrast to [38,39], which introduced additional segmentation models, we chose to use pseudo-labels to guide the fine-tuning training of SAM.

3. Methods

The Weakly-supervised and Self-guided Fine-tuning of SAM for Colonoscopy Polyp Segmentation (WSPoly-SAM) aims to fine-tune the SAM model from weakly annotated training dataset $\mathcal{S} = \{X_i, Y_i\}_{i=1}^{S}$ and to test it on testing dataset $\mathcal{T} = \{T_i\}_{i=1}^{T}$, where X_i represents the ith training image, T_i represents the ith testing image, and Y_i represents weak annotations, i.e., bounding box annotations used in this study. S and T respectively represent the numbers of training and testing data in the dataset. Specifically, the training image X_i first undergoes a series of augmentation operations, and the augmented images along with their weak annotations Y_i are then inputted into SAM to obtain the corresponding segmentation masks. Multiple masks are subsequently fused, followed by pixel-level weighting and mask post-processing to obtain pseudo-labels. Finally, these pseudo-labels serve as supervision masks, guiding the fine-tuning training of SAM. During this process, only the prompt encoder and mask decoder undergo fine-tuning.

3.1. Background

The Segment Anything Model (SAM) utilizes a transformer-based architecture [40], which has shown effectiveness in natural language processing [41] and image recognition tasks [42]. SAM comprises a visual transformer-based image encoder for feature extraction, a prompt encoder for user interaction integration, and a mask decoder for generating segmentation results and confidence scores based on image embeddings, prompt embeddings, and output tokens. The visual transformer in the image encoder is pre-trained using a masked autoencoder modeling method [43], enabling it to handle high-resolution images (e.g., 1024×1024). The prompt encoder is customizable for different user inputs, supporting four types of prompts: points, boxes, text, and masks. SAM's mask decoder is lightweight, consisting of two transformer layers with dynamic mask prediction heads and Intersection over Union (IoU) score regression heads. The mask prediction head can generate three 4× downsampled masks corresponding to the whole, parts, and subparts of objects.

3.2. Prompt Selection

SAM offers three primary segmentation modes: segment everything mode, bounding box mode, and point mode. However, each of these modes exhibits distinct characteristics, leading to certain biases in their segmentation outcomes. Ma et al. [36] demonstrated these three segmentation modes using typical abdominal computed tomography (CT) images and proved that the bounding box mode has broader practical value than the segment everything mode and point mode.

Considering the significant differences in target morphology, image format, and target region features between abdominal organ images and colonoscopy polyp images, we apply SAM with these three segmentation modes to colonoscopy polyp data to investigate whether the bounding box mode is the optimal choice for our segmentation task. Experimental results are shown in Figure 1. **(1) segment everything mode:** This mode partitions the entire image into multiple regions based on image intensity by defining segmentation grid points (here using 9x9), but the segmentation results often lack semantic labels and cannot clearly focus on the ROI that clinical doctors care about, limiting its utility. In Figure 1c, it can be seen that due to the similarity between polyps and surrounding skin tissue, it is easy to regard the entire polyp area as background in this mode. **(2) bounding box mode:** This mode segments the target objects by providing points for the upper left corner and lower right corner of the bounding box (in green color in Figure 2d). In Figure 1d, it can be seen that this mode can better identify the polyp area and is more similar to the ground truth in Figure 1b. **(3) point mode:** This mode generates

 doi:10.20944/preprints202405.0389.v1

the required mask by continuously adding foreground or background points. In Figure 1e1-4, it can be seen that the segmented area with one point-prompt under this mode will change with the position of the point, leading to segmentation instability, and the contour of the target area does not converge. The target area will only approach the ground truth after repeatedly adding foreground and background points. Overall, the segmentation results with multiple point-prompts in Figure 1f1-4 under the point mode are not as good as that in the box mode in Figure 1d. The point mode usually requires multiple prediction corrections and iterations, which is consistent with the conclusions of the study [36].

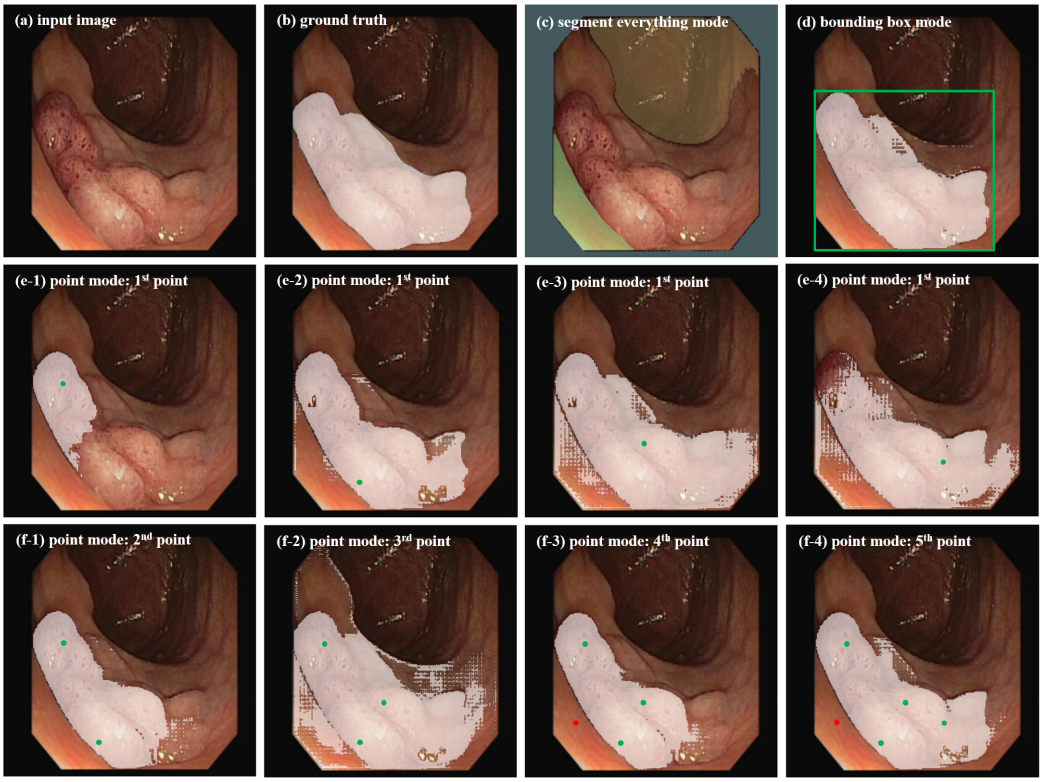

Figure 1. Segmentation results of SAM on a colonoscopy polyp image under different segmentation modes.

In summary, when applying SAM to colonoscopy polyp segmentation, the segment everything mode often leads to irrelevant region divisions, while the point mode can be ambiguous and necessitates multiple prediction correction iterations. In contrast, the bounding box mode effectively specifies the ROI and achieves reasonable segmentation results without requiring multiple trial-and-error attempts. Therefore, in subsequent experiments, bounding boxes are utilized as prompt information for both pseudo-label generation and fine-tuning in the SAM model.

3.3. Self-Guided Pseudo-Label Generation

As a general segmentation foundation model, SAM faces challenges when applied to medical image segmentation, particularly in colonoscopy polyp segmentation. The inherent similarity between polyps and surrounding skin tissues, coupled with their indistinct boundaries, exacerbates the complexity of the segmentation task. Consequently, when weak annotation Y_i guides pretrained SAM in generating segmentation mask as pseudo-labels for the next precise segmentation, the resulting segmentation mask lacks precision due to these complexities. Employing imprecise segmentation masks as pseudo-labels not only fails to provide additional information but also may mislead the model, and therefore could impede SAM's fine-tuning. To address this challenge, we designed three techniques, i.e., multi-augmentation fusion, pixel-level weighting and mask post-processing to enhance

 doi:10.20944/preprints202405.0389.v1

the accuracy of 'pseudo-labels', especially in complex scenarios. For a detailed overview of the model architecture, refer to Figure 2.

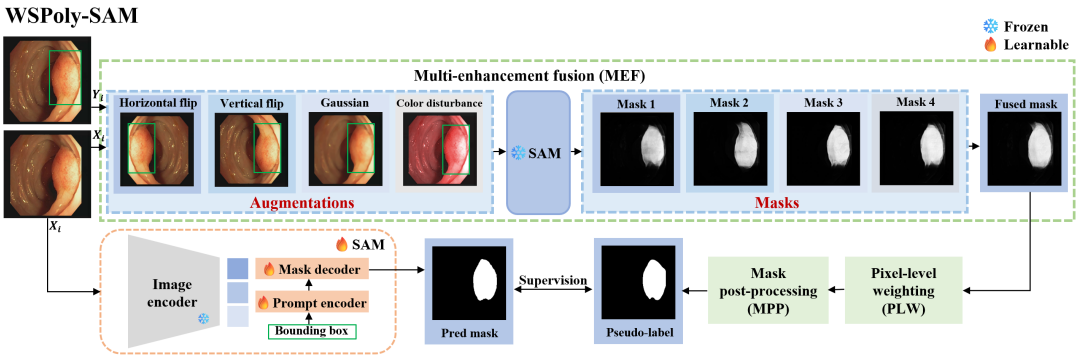

Figure 2. WSPoly-SAM polyp framework under bounding box guidance. Note that the masks corresponding to flipping augmentation should be flipped back before engaging in the fusion process.

Multi-enhancement fusion (MEF). The information content of a single segmentation mask is often insufficient and inaccurate. In order to enhance the information capacity and accuracy of the segmentation mask, we adopt multiple enhancement fusion techniques. For each image in the training set $(X_i, Y_i) \in \mathcal{S}$, random augmentation is applied, including horizontal flipping, vertical flipping, gaussian blurring, and color jittering. Prompt information also changes accordingly after flipping operations, resulting in the generation of K augmented images and corresponding prompts $\{X_i^k, Y_i^k\}_{k=1}^K$. Specifically, color jittering follows the method proposed in [44], where brightness factor is uniformly sampled from the range [0.6, 1.6], contrast is set to 0.2, saturation factor to 0.1, and hue factor to 0.01. Subsequently, we feed $\{X_i^k, Y_i^k\}_{k=1}^K$ into SAM to generate the corresponding segmentation masks $\{\mathcal{M}_i^k\}_{k=1}^K$, where

$$\mathcal{M}_i^k = SAM(X_i^k, Y_i^k). \tag{1}$$

Note that $\mathcal{M}_i^k$ and X_i^k have the same shape but differ from the shape of X_i. To ensure consistency between the mask and the original image, $\mathcal{M}_i^k$ obtained after flipping will be flipped back.

Due to using different augmented images for segmentation, the generated masks $\mathcal{M}_i^k$ may exhibit some shape differences, but they often overlap in certain regions. SAM reliably predicts these overlapping regions, remains unaffected by image transformations, and usually corresponds to the correct foreground areas. Considering this complementarity, we propose a method to fuse the segmentation masks of different augmented images. The fusion process is described as follows:

$$\tilde{\mathcal{M}}_i = \max_k \mathcal{M}_i^k, \tag{2}$$

where $\tilde{\mathcal{M}}_i$ represents the fused mask. We anticipate the fused mask $\tilde{\mathcal{M}}_i$ to be more dependable than individual masks due to its ability to comprehensively reflect the features of the colonoscopy image.

Pixel-level weighting (PLW). The confidence of predictions may fluctuate across pixels. To accentuate the most reliable predictions, we suggest employing entropy for weighting. We compute the entropy of each pixel to generate an entropy map:

$$\tilde{E}_i = -\tilde{\mathcal{M}}_i \log \tilde{\mathcal{M}}_i - (1 - \tilde{\mathcal{M}}_i) \log (1 - \tilde{\mathcal{M}}_i). \tag{3}$$

The entropy map $\tilde{E}_i$ is computed from the fused mask $\tilde{\mathcal{M}}_i$ and quantifies the prediction uncertainty for each pixel across all augmented images. Pixels with high confidence and consistent predictions across all augmented images exhibit low entropy. Thus, we can leverage this entropy map to weight the fused mask, assigning greater weights to more reliable pixels. By applying entropy weights on the fused mask, we obtain a pixel-level weighted mask $\tilde{M}_i$, as follows:

 doi:10.20944/preprints202405.0389.v1

$$\tilde{M}_i = (1 - \tilde{E}_i) \times \tilde{\mathcal{M}}_i. \tag{4}$$

This strategy considers the prediction uncertainty of pixels, allowing more reliable pixels in the fused mask $\tilde{\mathcal{M}}_i$ to receive higher weights. As a result, it enhances the segmentation accuracy of the SAM model in critical regions.

Mask post-processing (MPP). The pixel-level weighted mask $\tilde{M}_i$ may have some issues, such as insufficiently smooth edges or disconnected pathological areas. To address these problems, we perform some post-processing operations. First, the pixel-level weighted mask $\tilde{M}_i$ is binarized as follows:

$$\tilde{M}_i^b = Binary_\alpha(\tilde{M}_i). \tag{5}$$

$\tilde{M}_i^b$ represents the binarized mask, with α representing the threshold which is set to 0.2. Next, we employ gaussian filtering techniques to smooth images and diminish noise, along with morphological closing to connect adjacent regions, thereby ensuring the contiguity of pathological areas and smoother edges. Through these post-processing operations, we reduce noise interference and extraneous information, yielding high-quality pseudo-labels $\tilde{Y}_i$, as follows:

$$\tilde{Y}_i = Smooth(\tilde{M}_i^b). \tag{6}$$

Note that, as shown in Figure 4, multi-enhancement fusion, pixel-level weighting and mask post-processing operations are applied sequentially to make the generated pseudo-labels more accurate.

3.4. Weakly-Supervised Fine-Tuning

In order to adapt SAM to specific colonoscopy images, it is necessary to select appropriate prompt information and network components for fine-tuning. Based on the results analysis in Figure 1, bounding box is the optimal choice as prompt information for generating segmentation masks. It's crucial to highlight that the bounding box is derived from the ground-truth label, with each polyp region aligning with a specific bounding box. SAM's network architecture comprises three key components: the image encoder, prompt encoder, and mask decoder. Due to the significant computational load primarily on the image encoder, which is built on visual transformers, we maintain the image encoder's frozen state and solely fine-tune the prompt encoder and mask decoder components.

Moreover, the choice of loss function is crucial during model training due to the disparities between pseudo-labels and ground-truth labels. We incorporate the weighted intersection over union (IoU) loss from the original SAM loss function, alongside popular segmentation loss functions such as the dice coefficient loss and binary cross-entropy loss, which have demonstrated robustness across various segmentation tasks [45,46] . Figure 2 provides an overview of the WSPoly-SAM framework, where the SAM model's parameters remain frozen during the generation of pseudo-label $\tilde{Y}_i$. Subsequently, we guide SAM's fine-tuning process using $\tilde{Y}_i$ as supervision masks, focusing solely on fine-tuning the prompt encoder and mask decoder. The total loss function is formulated as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{Iou}(Y_i', \tilde{Y}_i) + \mathcal{L}_{BCE}(Y_i', \tilde{Y}_i) + \mathcal{L}_{Dice}(Y_i', \tilde{Y}_i), \tag{7}$$

where Y_i' represents predicted mask, i.e., $Y_i' = SAM(X_i, Y_i)$. Fine-tuning the SAM model helps it better understand and segment the colonoscopy polyp regions, adapting to the image features generated by the pseudo-labels.

4. Experiments

4.1. Datasets

To facilitate a fair comparison, we use the following five frequently-adopted benchmark datasets:

 doi:10.20944/preprints202405.0389.v1

1. Kvasir-SEG [47]: This dataset gathered by the Vestre Viken Health Trust in Norway, comprises 1000 colonoscopy video sequences showcasing polyp images, each accompanied by its respective annotations. The image resolution varies from 332×487 to 1920×1072 pixels. Annotations are meticulously labeled by medical professionals and validated by seasoned gastroenterologists.
2. ClinicDB [48]: This dataset comprises 612 images drawn from 29 colonoscopy video sequences, with a resolution of 288×384 pixels. It is created in partnership with the Hospital Clinic of Barcelona in Spain.
3. ColonDB [49]: This dataset includes 380 polyp images along with their corresponding annotations, with each image having a resolution of 500×570 pixels. These images are extracted from 15 distinct videos, with frames meticulously chosen by experts to showcase diverse perspectives. The annotations are manually crafted with precision.
4. CVC-300 [50]: The dataset contains 60 polyp images, each with a resolution of 500×574 pixels.
5. ETIS [51]: This dataset is released by the MIC-CAI Polyp Detection Subchallenge in 2017, encompassing 196 polyp images extracted from colonoscopy videos, each paired with its respective label. The images boast a resolution of 966×1225 pixels.

These datasets provide polyp images of varying quantities and resolutions, covering diverse scenarios and perspectives. All images have been manually annotated and verified by medical professionals, serving as benchmarks for evaluating algorithm performance and conducting fair comparisons.

4.2. Implementation Details

Data Split. To ensure an equitable comparison, we employ an identical data partitioning methodology to that utilized in the experiments conducted by PraNet [54]. Specifically, the Kvasir-SEG dataset consists of 1000 polyp images, while ClinicDB comprises 612 images. From the Kvasir-SEG and ClinicDB datasets, we allocate 900 and 550 images, respectively, for the training set, while the remaining 100 and 62 images are designated for the test set. Furthermore, our model undergoes evaluation on three additional datasets not previously encountered: ColonDB, CVC-300, and ETIS.

Data Augmentation. The colonscopy polyp dataset have the following three characteristics: (1) significant variations in polyp size, including both large and small polyps; (2) close similarity in color between most polyps and the background, resulting in low contrast between foreground and background and making some samples challenging to segment; (3) blurry imaging resulting from the rotating movement of the camera within the intestine, significantly increasing the difficulty of polyp detection. To tackle these issues, we introduced a series of random augmentation operations during model training: (1) random scaling by a factor of 0.7 to 1.3; (2) color augmentation following the method in [32]; (3) contrast enhancement sampled randomly within the range of [0.8, 1.2]; (4) gaussian filtering using a 5x5 convolutional kernel. The purpose of gaussian filtering is to simulate the blurriness caused by lens rotation, aiming to improve the model's ability to adapt to blurred samples. It is worth noting that random scaling involves operations on the shape of the image, so the corresponding pseudo-labels and bounding boxes also follow this operation.

Training Details. Our algorithm is trained using the PyTorch framework. For hardware, we leverage an NVIDIA Tesla A100 GPU equipped with 40GB of GPU memory to train the segmentation model. Throughout the network training phase, mini-batch training iterations are conducted with a batch size set to 4. The model parameters are optimized utilizing the Adam [36] optimization algorithm, initialized with a learning rate of 1e-4 and a weight decay of 1e-5. The entirety of the training process spans a total of 25 epochs.

Evaluation Metrics. In line with the evaluation metrics used in Polyp-PVT [52] and HSNet [53], we employ the mean intersection over union (mIoU) and mean dice score (mDice) to assess the segmentation results in our experiments.

 doi:10.20944/preprints202405.0389.v1

4.3. Results

4.3.1. Quantitative Results

In this section, we assessed our model's learning capacity using the Kvasir-SEG and ClinicDB datasets. To ascertain the model's ability to generalize, we conducted evaluations on three additional datasets: ColonDB, CVC-300, and ETIS. We compared our model with the current mainstream fully-supervised polyp segmentation networks.

Table 1 reports the comparison results of our approach with fully-supervised networks (including UNet [15], UNet++ [16], SFA [22], PraNet [54], C2FNet [55], Polyp-PVT [52], and HSNet[53]). From the comparison in Table 1, WSPoly-SAM exhibits superior performance over existing mainstream baseline networks across the Kvasir-SEG, ColonDB, CVC-300, and ETIS datasets. Specifically, our approach demonstrates a 0.7% enhancement in mDice score and a 0.2% boost in mIoU score compared to HSNet on the Kvasir-SEG dataset. In ColonDB, it achieves a notable 9.4% improvement in mDice score and a substantial 9.6% increase in mIoU score relative to HSNet. Furthermore, on the CVC-300 dataset, it delivers a 3% increase in mDice score and a 3.9% elevation in mIoU score compared to HSNet. On ETIS, it achieves an 8% improvement in mDice score and a 7.9% improvement in mIoU score compared to HSNet. Compared to the original SAM, our method demonstrates significant improvements after fine-tuning on all five datasets, demonstrating that SAM is more adaptable to specific domains after fine-tuning.

Table 1. Quantitative results of WSPoly-SAM and SAM on Kvasir-SEG, ClinicDB, ColonDB, CVC-300 and ETIS datasets. Here, 'B' represents the use of ViT-B, 'Δ' represents the increase relative to SAM.

Models	Kvasir-SEG		ClinicDB		ColonDB		CVC-300		ETIS	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	0.818	0.746	0.823	0.755	0.504	0.436	0.710	0.627	0.398	0.335
U-Net++	0.821	0.744	0.794	0.729	0.482	0.408	0.707	0.624	0.401	0.344
SFA	0.723	0.611	0.700	0.607	0.456	0.337	0.467	0.329	0.297	0.271
PraNet	0.898	0.840	0.899	0.849	0.712	0.640	0.871	0.797	0.628	0.567
C2FNet	0.886	0.831	0.919	0.872	0.724	0.650	0.874	0.801	0.699	0.624
Polyp-PVT	0.917	0.864	**0.948**	**0.905**	0.808	0.727	0.900	0.833	0.787	0.706
HSNet	0.926	0.877	0.937	0.887	0.810	0.735	0.903	0.839	0.808	0.734
SAM-B	0.892	0.832	0.884	0.812	0.856	0.777	0.922	0.862	0.873	0.794
WSPoly-SAM-B	**0.933**	**0.879**	0.920	0.858	**0.904**	**0.831**	**0.933**	**0.878**	**0.888**	**0.813**
Δ	0.041	0.047	0.036	0.046	0.048	0.054	0.011	0.016	0.015	0.019

4.3.2. Qualitative Results

To visually illustrate the superiority of our proposed method, Figure 3 shows the predicted results of our model compared to the competing models. From Figure 3, our method more accurately locates and identifies the contours of polyps, generating prediction masks that closely resemble ground-truth labels compared to fully-supervised models. This improvement can be attributed to the enhanced learning of complex polyp features in SAM after fine-tuning, with the incorporation of prompts helping to reduce misjudgments in erroneous areas. Additionally, we provide visualizations of the original prediction masks by SAM, revealing issues such as misjudgments, jagged contours, and unclear edges before fine-tuning. Overall, our method effectively captures global contextual information and restores detailed appearance features, thereby better delineating the polyp regions.

 doi:10.20944/preprints202405.0389.v1

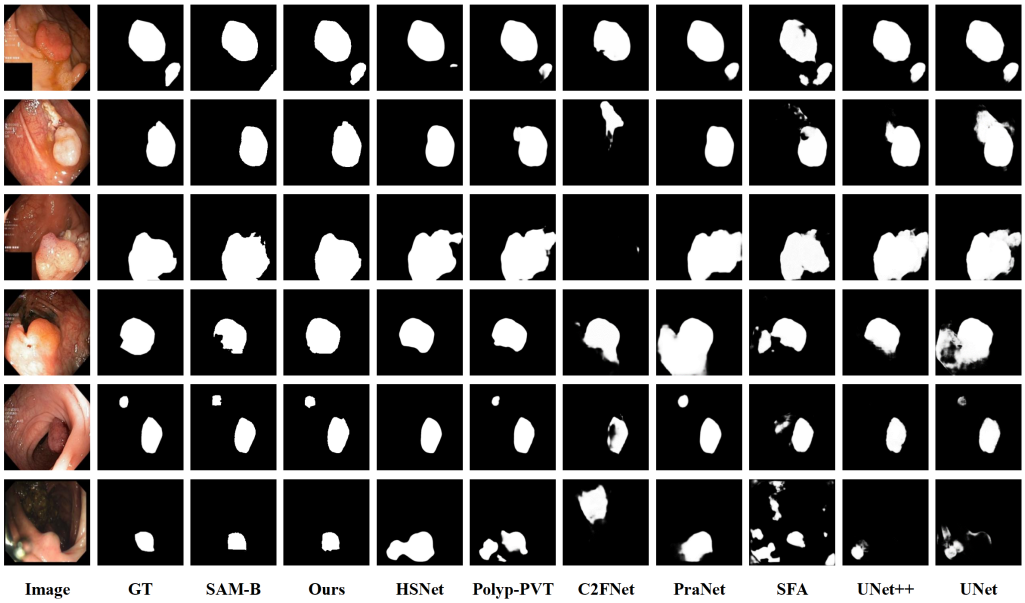

Figure 3. Visualization results with different models.

4.3.3. Ablation Study

Our method includes a key component, which is pseudo-label generation based on SAM. It consists of three parts: Multi-enhancement fusion (MEF), Pixel-level weighting (PLW), and Mask post-processing (MPP). In this section, we conducted comprehensive ablation experiments on five colonoscopy polyp datasets, with results shown in Table 2. The 1^{st} row presents the results of directly using SAM to generate segmentation masks on the five datasets. As for the baseline, we used SAM to generate a segmentation mask for training image and fine-tuned SAM (2^{nd} row) as the baseline without data augmentation. We then gradually added MEF (3^{rd} row), PLW (4^{th} row), and MPP (5^{th} row) on top of the baseline, ultimately achieving the best evaluation scores.

Table 2. Ablation study of WSPoly-SAM's Pseudo-labels.

Baseline	MEF	PLW	MPP	Kvasir-SEG		ClinicDB		ColonDB		CVC-300		ETIS	
				mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
-	-	-	-	0.892	0.832	0.884	0.812	0.856	0.777	0.922	0.862	0.873	0.794
√				0.928	0.873	0.914	0.850	0.895	0.820	0.931	0.875	0.876	0.796
√	√			0.928	0.875	0.914	0.851	0.897	0.822	0.931	0.875	0.880	0.797
√	√	√		0.932	0.878	0.916	0.851	0.903	0.830	0.933	0.878	0.885	0.805
√	√	√	√	**0.933**	**0.879**	**0.920**	**0.858**	**0.904**	**0.831**	**0.933**	**0.878**	**0.888**	**0.813**

Additionally, we conducted visual analysis to assess the impact of each module on the quality of model predictions (refer to Figure 4). It can be observed that in the baseline scenario, due to unclear boundaries between polyps and surrounding mucosa and low contrast between polyp foreground and background information, 'Baseline' exhibits issues such as unclear predicted edges and missed detections. After introducing 'MEF', the fusion of multiple enhancement results provided additional predictive information, significantly reducing instances of missed detections. Subsequently, with the introduction of 'PLW', assigning higher weights to reliable pixels further reduced the probability of error detection. Finally, with the inclusion of 'MPP', better noise reduction during prediction and improvement in edge clarity were achieved, further enhancing the segmentation results. In summary, our method enhances the detection coverage of polyp regions, reduces segmentation errors, and minimizes missed detections.

 doi:10.20944/preprints202405.0389.v1

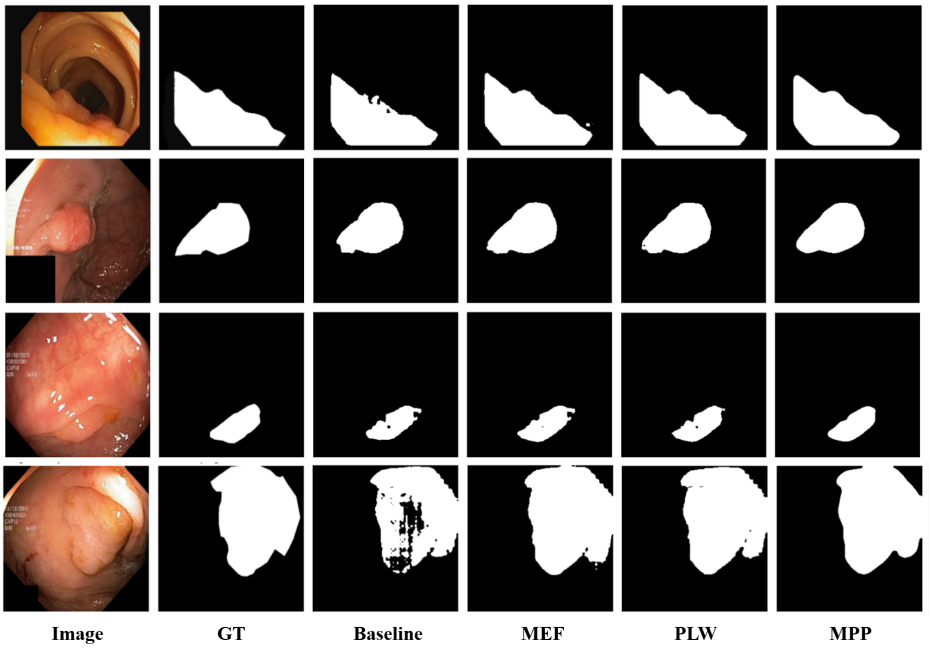

Figure 4. Visualization of the ablation study results.

4.3.4. Further Analysis

Different Visual Backbones. We also investigated the segmentation capabilities of the SAM model with different Vision Transformer (ViT) for encoding image features, as shown in Table 3. Table 3 (1^{st} row, 2^{nd} row, and 3^{rd} row) shows the predictions of the original SAM model without fine-tuning. It can be observed that on the ETIS dataset, although SAM-H and SAM-L are more computationally complex, their performance is not as good as SAM-B. To fully explore the transfer learning capabilities of SAM on different ViTs, we conducted fine-tuning. Considering memory limitations, in this work, we only fine-tuned SAM-B and SAM-L. From Table 3 (4^{th} row and 5^{th} row), it can be observed that under weakly-supervised strategies, SAM-B outperforms SAM-L on the Kvasir-SEG, ClinicDB, and ETIS datasets, while performing less effectively on ColonDB and CVC-300 compared to SAM-L. This indicates that the lightweight SAM-B demonstrates better learning capabilities compared to the higher-complexity SAM-L.

Table 3. Results of SAM with different visual backbones. Model-ViT represents SAM with different ViTs. Mask indicates the supervision mask used during fine-tuning, where 'Pseudo' represents pseudo-labels and 'GT' represents ground-truth labels.

Model-ViT	Mask	Kvasir-SEG		ClinicDB		ColonDB		CVC-300		ETIS	
		mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
SAM-B	-	0.892	0.832	0.884	0.812	0.856	0.777	0.922	0.862	0.873	0.794
SAM-L	-	0.921	0.865	0.884	0.817	0.897	0.823	0.926	0.868	0.855	0.782
SAM-H	-	0.923	0.868	0.890	0.826	0.895	0.822	0.914	0.852	0.866	0.732
SAM-B	Pseudo	0.933	0.879	0.920	0.858	0.904	0.831	0.933	0.878	0.888	0.813
SAM-L	Pseudo	0.932	0.884	0.925	0.867	0.906	0.837	0.934	0.879	0.881	0.812
SAM-B	GT	0.936	0.882	0.935	0.882	0.907	0.836	0.934	0.879	0.895	0.822
SAM-L	GT	0.934	0.886	0.926	0.876	0.908	0.842	0.934	0.880	0.890	0.815

To verify this point, we performed full supervision fine-tuning on SAM-B and SAM-L using ground-truth labels. As shown in Table 3 (6^{th} row and 7^{th} row), under full supervision, the trends across datasets align with those observed under weak supervision. In general, although SAM-L entails significantly greater computational complexity, it does not surpass the lightweight SAM-B model on three out of five public datasets. This finding suggests that the lightweight SAM-B model is well-suited

 doi:10.20944/preprints202405.0389.v1

for adoption in medical image applications. Additionally, we observed that under the same ViT, weak supervision performs only 0.3% lower in average mDice than full supervision on the Kvasir-SEG, ColonDB, and CVC-300 datasets. This result indicates that while weakly-supervised methods reduce dependence on precise annotations, they can still maintain comparable segmentation precision to fully-supervised methods.

Pseudo-label Generation vs. WSPoly-SAM. We conducted a series of experiments (refer to Table 4) comparing our method WSPoly-SAM with directly using pseudo-label generation, aiming to evaluate whether our fine-tuning approach is superior to the direct adoption of pseudo-label generation. The experimental results from five test datasets indicate that the mDice and mIoU scores obtained through the pseudo-label generation strategy are between the original predictions of SAM and those after fine-tuning, and significantly lower than the results after fine-tuning. This indicates that while the pseudo-label generation method can optimize the predictions of the original SAM to some extent, there is still a certain gap compared to the predictions after fine-tuning. Additionally, the pseudo-label generation process requires multiple invocations of SAM, incurring high time costs, making it challenging for direct use in predictions. In comparison, using WSPoly-SAM is more feasible.

Table 4. Comparison between Pseudo-label generation and WSPoly-SAM.

Methods	Kvasir-SEG		ClinicDB		ColonDB		CVC-300		ETIS	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
SAM	0.892	0.832	0.884	0.812	0.856	0.777	0.922	0.862	0.873	0.794
Pseudo-label Generation	0.899	0.844	0.912	0.844	0.890	0.814	0.932	0.876	0.884	0.810
WSPoly-SAM	0.933	0.879	0.920	0.858	0.904	0.831	0.933	0.878	0.888	0.813

Effectiveness of weakly-supervised fine-tuning in reducing annotation costs. We further investigated the impact of different amounts of training data for supervision masks on the results of fine-tuned SAM, fully demonstrating SAM's efficiency in utilizing fine-tuning data. Figure 5 illustrates the comparison between weakly-supervised and fully-supervised approaches when the training data amounts from 10% to 100%, using SAM-B model for fine-tuning. As the amount of data increases, the overall accuracy gradually improves. On the CVC-300 dataset, weakly-supervised strategy with 100% training data achieves results comparable to fully-supervised strategy with 90% to 100% training data. On the Kvasir-SEG dataset, weakly-supervised strategy with 100% training data achieves results comparable to fully-supervised strategy with 80% training data. On the ColonDB dataset, weakly-supervised strategy with 100% training data achieves results comparable to fully-supervised strategy with 50% training data. However, performance is relatively poorer on the other two datasets. On the ETIS dataset, weakly-supervised strategy with 100% training data achieves results comparable to fully-supervised strategy with 40% to 50% training data. On the ClinicDB dataset, weakly-supervised strategy with 100% training data only achieves results comparable to fully-supervised strategy with 20% to 30% training data. Overall, using 100% training data under the weakly-supervised strategy achieves 56% to 62% of the training results of fully-supervised learning. We simulated the scenario of annotating polyp data and found that the time ratio between bounding box annotation and rough pixel-level annotation is approximately 1:6 (which would be even higher if precise pixel-level annotation is done by professional annotators). Based on this calculation, for example, if bounding box annotation takes 1 second and pixel-level annotation takes 6 seconds, to achieve the training data results of 56% to 62% under fully-supervised learning with 100 images, the annotation time ratio would be approximately 27% to 30%, saving 70% to 73% of the time cost. This demonstrates the effectiveness of our method in reducing annotation costs.

 doi:10.20944/preprints202405.0389.v1

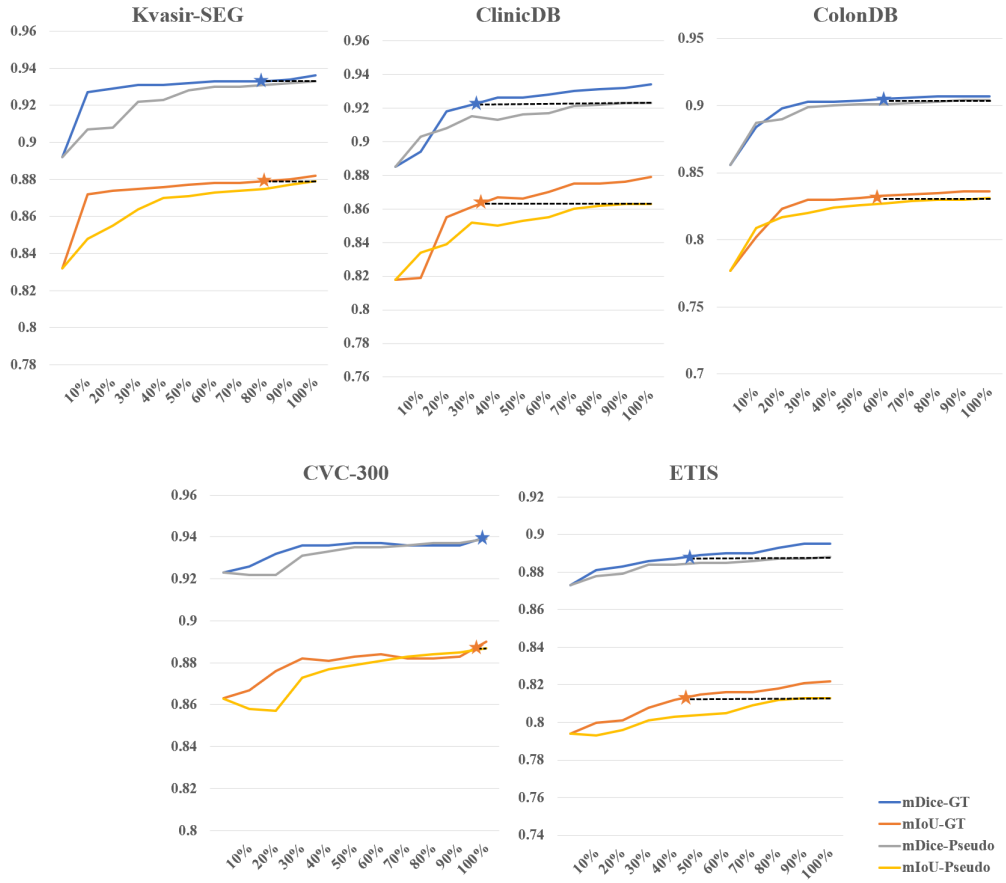

Figure 5. Comparison between weakly-supervised and fully-supervised fine-tuning under different amounts of supervision masks.

5. Conclusions

This study proposes a novel method for colon polyp segmentation called WSPoly-SAM, which utilizes weak annotations to guide SAM in generating pseudo-labels for self-guided fine-tuning, reducing the dependence on precise annotation data. The experimental results demonstrate a competitive performance of WSPoly-SAM in colon polyp segmentation compared to current fully-supervised learning methods, bringing a new technological breakthrough to this field. Additionally, through experiments on the segmentation performance of the SAM model under different versions of the ViT and comparing different amounts of training data under different supervision strategies (weakly-supervised vs. fully-supervised), several key findings can be obtained.

Firstly, despite the higher computational complexity of SAM-L, its performance on most public datasets is not superior to the lightweight SAM-B. This finding demonstrates that SAM-B is more suitable for efficient deployment in medical imaging applications.

Secondly, under the weakly-supervised strategy, using 100% of the training data can achieve training results of 56% to 62% compared to fully-supervised learning. This means that 70% to 73% of the annotation time costs can be saved. These findings not only affirm the technical efficacy of our proposed approach but also emphasize its potential value in reducing annotation costs and improving work efficiency.

However, our research also has some limitations. For instance, WSPoly-SAM currently relies on ground-truth bounding boxes as supplementary prompt information. Moving forward, we aim to explore the development of an end-to-end model that eliminates the requirement for prompts altogether. Furthermore, despite the prevalence of 3D medical images, SAM operates exclusively in 2D mode, which is a notable constraint within the medical imaging domain. We hope to conduct a

 doi:10.20944/preprints202405.0389.v1

more comprehensive evaluation of segmentation tasks from different modalities and objectives in the future, further enhancing the flexibility of our approach.

Author Contributions: Conceptualization, T.C., H.Y. and K.D.; methodology, T.C., H.Y. and K.D.; software, T.C.; validation, T.C.; resources, T.C.; data curation, T.C.; supervision, H.Y. and K.D.; writing—original draft preparation, T.C.; writing—review and editing, T.C., H.Y., K.D., Y.Z. and Y.Z.; funding acquisition, K.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under Grant 62306310.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are available in this link at https://github.com/DengPingFan/PraNet.

Acknowledgments: The authors wish to thank the reviewers for their valuable comments and suggestions concerning this manuscript.

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

References

1. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **2021**, *18*, 203–211.
2. Ouyang, D.; He, B.; Ghorbani, A.; Yuan, N.; Ebinger, J.; Langlotz, C.P.; Heidenreich, P.A.; Harrington, R.A.; Liang, D.H.; Ashley, E.A.; others. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **2020**, *580*, 252–256.
3. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
4. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *40*, 834–848.
5. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Medical Image Analysis* **2017**, *42*, 60–88.
6. Asgari Taghanaki, S.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review* **2021**, *54*, 137–178.
7. Zhang, R.; Lai, P.; Wan, X.; Fan, D.J.; Gao, F.; Wu, X.J.; Li, G. Lesion-aware dynamic kernel for polyp segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2022, pp. 99–109.
8. Zhou, T.; Zhou, Y.; Gong, C.; Yang, J.; Zhang, Y. Feature aggregation and propagation network for camouflaged object detection. *IEEE Transactions on Image Processing* **2022**, *31*, 7036–7047.
9. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nature Communications* **2024**, *15*, 654.
10. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.M.; Chen, W.; others. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* **2023**, *5*, 220–235.
11. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; others. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
12. Hu, M.; Li, Y.; Yang, X. Skinsam: Empowering skin cancer segmentation with segment anything model. *arXiv preprint arXiv:2304.13973* **2023**.
13. Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; Arbel, T. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620* **2023**.
14. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

15. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
16. 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* **2019**, *39*, 1856–1867.
17. Yang, X.; Li, X.; Ye, Y.; Lau, R.Y.; Zhang, X.; Huang, X. Road detection and centerline extraction via deep recurrent convolutional neural network U-Net. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*, 7209–7220.
18. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition* **2020**, *106*, 107404.
19. Sun, X.; Zhang, P.; Wang, D.; Cao, Y.; Liu, B. Colorectal polyp segmentation by U-Net with dilation convolution. 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2019, pp. 851–858.
20. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
21. Liu, S.; Huang, D.; others. Receptive field block net for accurate and fast object detection. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.
22. Fang, Y.; Chen, C.; Yuan, Y.; Tong, K.y. Selective feature aggregation network with area-boundary constraints for polyp segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer, 2019, pp. 302–310.
23. Zhang, R.; Li, G.; Li, Z.; Cui, S.; Qian, D.; Yu, Y. Adaptive context selection for polyp segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. Springer, 2020, pp. 253–262.
24. Nguyen, T.C.; Nguyen, T.P.; Diep, G.H.; Tran-Dinh, A.H.; Nguyen, T.V.; Tran, M.T. CCBANet: Cascading context and balancing attention for polyp segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, 2021, pp. 633–643.
25. Deng, R.; Cui, C.; Liu, Q.; Yao, T.; Remedios, L.W.; Bao, S.; Landman, B.A.; Wheless, L.E.; Coburn, L.A.; Wilson, K.T.; others. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155* **2023**.
26. Hu, C.; Li, X. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506* **2023**.
27. He, S.; Bao, R.; Li, J.; Grant, P.E.; Ou, Y. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324* **2023**.
28. Roy, S.; Wald, T.; Koehler, G.; Rokuss, M.R.; Disch, N.; Holzschuh, J.; Zimmerer, D.; Maier-Hein, K.H. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396* **2023**.
29. Zhou, T.; Zhang, Y.; Zhou, Y.; Wu, Y.; Gong, C. Can sam segment polyps? *arXiv preprint arXiv:2304.07583* **2023**.
30. Mohapatra, S.; Gosai, A.; Schlaug, G. Brain extraction comparing segment anything model (sam) and fsl brain extraction tool. *arXiv preprint arXiv:2304.04738* **2023**.
31. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis* **2023**, *89*, 102918.
32. Chen, J.; Bai, X. Learning to "segment anything" in thermal infrared images through knowledge distillation with a large scale dataset satir. *arXiv preprint arXiv:2304.07969* **2023**.
33. Tang, L.; Xiao, H.; Li, B. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709* **2023**.
34. Ji, G.P.; Fan, D.P.; Xu, P.; Cheng, M.M.; Zhou, B.; Van Gool, L. SAM Struggles in Concealed Scenes–Empirical Study on" Segment Anything". *arXiv preprint arXiv:2304.06022* **2023**.
35. Ji, W.; Li, J.; Bi, Q.; Li, W.; Cheng, L. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750* **2023**.

36. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nature Communications* **2024**, *15*, 654.
37. Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; others. Sam-med2d. *arXiv preprint arXiv:2308.16184* **2023**.
38. Jiang, P.T.; Yang, Y. Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.01275* **2023**.
39. He, C.; Li, K.; Zhang, Y.; Xu, G.; Tang, L.; Zhang, Y.; Guo, Z.; Li, X. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems* **2024**, *36*.
40. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Transformer neural network for weed and crop classification of high resolution UAV images. *Remote Sensing* **2022**, *14*, 592.
41. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in Neural Information Processing Systems* **2020**, *33*, 1877–1901.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
43. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
44. 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* **2022**, *150*, 106173.
45. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **2021**, *18*, 203–211.
46. Ma, J.; Chen, J.; Ng, M.; Huang, R.; Li, Y.; Li, C.; Yang, X.; Martel, A.L. Loss odyssey in medical image segmentation. *Medical Image Analysis* **2021**, *71*, 102035.
47. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; De Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. Springer, 2020, pp. 451–462.
48. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* **2015**, *43*, 99–111.
49. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging* **2015**, *35*, 630–644.
50. Vázquez, D.; Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; López, A.M.; Romero, A.; Drozdzal, M.; Courville, A. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering* **2017**, *2017*.
51. 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* **2014**, *9*, 283–293.
52. 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**.
53. 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* **2022**, *150*, 106173.
54. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 263–273.
55. Sun, Y.; Chen, G.; Zhou, T.; Zhang, Y.; Liu, N. Context-aware cross-level fusion network for camouflaged object detection. *arXiv preprint arXiv:2105.12555* **2021**.

 doi:10.20944/preprints202405.0389.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.