VP-SAM: Taming Segment Anything Model for Video Polyp Segmentation via Disentanglement and Spatio-temporal Side Network

Zhixue Fang[1], Yuzhi Liu[1], Huisi Wu[1(✉)], and Jin Qin[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract. We propose a novel model (VP-SAM) adapted from segment anything model (SAM) for video polyp segmentation (VPS), which is a challenging task due to (1) the low contrast between polyps and background and (2) the large frame-to-frame variations of polyp size, position, and shape. Our aim is to take advantage of the powerful representation capability of SAM while enabling SAM to effectively harness temporal information of colonoscopic videos and disentangle polyps from background with similar appearances. To achieve this, we propose two new techniques. First, we propose a new semantic disentanglement adapter (SDA) by exploiting amplitude information of the Fourier spectrum to facilitate SAM in more effectively differentiating polyps from background. Second, we propose an innovative spatio-temporal side network (STSN) to provide SAM with spatio-temporal information of videos, thus facilitating SAM in effectively tracking the motion status of polyps. Extensive experiments on SUN-SEG, CVC-612, and CVC-300 demonstrate that our method outperforms state-of-the-art methods. While this work focuses on colonoscopic videos, the proposed method is general enough to be used to analyze other medical videos with similar challenges. Code is available at https://github.com/zhixue-fang/VPSAM.

Keywords: Video polyp segmentation · Segment anything model · Spatio-temporal modeling · Semantic disentanglement

1 Introduction

Colorectal cancer (CRC) is a gastrointestinal malignancy that causes considerable global deaths [5]. Currently, timely screening and removal of polyps (precursors of CRC) using colonoscopy is an effective way for prevention. However, manual screening of colonoscopic videos is labor-intensive, time-consuming, and error-prone [1, 19], usually leading to negligence of some precancerous lesions. In this regard, automated tools for automatic polyp segmentation from colonoscopic videos are highly demanded in clinical practice. It is, however, a very challenging task for the following two reasons. First, polyps usually derive from surrounding diseased tissue and thus naturally have a similar appearance to the

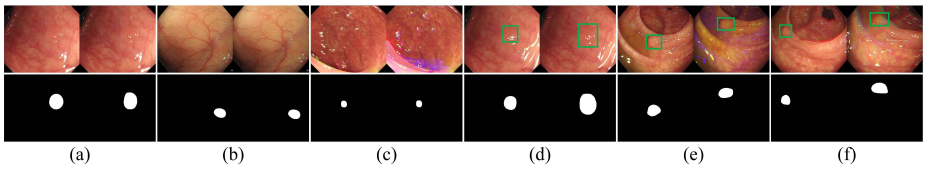

Fig. 1: Challenges of VPS: (a)-(c) low contrast between polyps and background, (d) size variation, (e) position variation, and (f) shape variation.

background in colonoscopic videos, which greatly hinders accurate segmentation, as shown in Figure 1 (a-c). Second, the large frame-to-frame variations of polyp size, position, and shape make it difficult to harness temporal information for precise yet real-time segmentation, as shown in Figure 1 (d-f).

In recent years, many deep-learning methods have been developed for polyp segmentation. Early investigations focus on 2D images/frames instead of videos. Most these methods either utilize the local modeling ability of CNNs [9, 14, 25, 33, 41] or leverage the global modeling ability of transformers [11, 16, 28, 40, 46] to segment polyps from images/frames. However, ignoring temporal information limits segmentation accuracy and efficiency; after all, in clinical practice, doctors operate with colonoscopic videos rather than images. Therefore, some video-based methods are proposed to harness hybrid 2D/3D CNN architectures [35] or normalized attention mechanisms [20, 22] to segment polyps from colonoscopic videos. However, it is still difficult for them to sufficiently meet above challenges due to limited domain-specific training samples and the restricted representation capability caused by inadequate model capacity.

The recently proposed SAM [26], trained on the large-scale dataset SA-1B, has powerful feature extraction capabilities, and its advanced network architecture can accurately focus on semantics of interest based on user prompts (e.g., points and bounding boxes). Hence, some studies propose to employ SAM in medical image segmentation to improve segmentation accuracy [17, 18, 21, 47]. Unfortunately, the dramatic domain variability between medical and natural images makes SAM perform poorly in many medical image segmentation tasks. To the end, effective out-of-domain tuning of SAM is necessary.

To adapt SAM to medical image segmentation tasks, MedSAM [32] and Polyp-SAM [29] focus on tuning parameters of SAM on medical datasets. However, tuning a large model like SAM is very resource intensive and expensive, and hence usually does not applicable in clinical settings. SAMAug [45] uses SAM to augment medical images without re-training SAM. However, the effect of augmentation is limited by the dramatic domain variability between medical and natural images, as SAM is mainly trained on natural images. To this end, most methods [6, 8, 30, 42, 44] propose to introduce additional learnable adapters to eliminate domain variability. While these methods can improve the performance of SAM in medical datasets, most of them focus on images and do not sufficiently harness temporal information of videos, which limits their accuracy and efficacy in video segmentation. Recently, some methods utilize mask trackers [43]

or point trackers [37] to extend SAM into a video segmentation model. However, these methods harness the segmentation of the initial frame as a cue to guide the segmentation of the current frame, and hence still do not sufficiently use temporal information. To address this, MediViSTA-SAM [24] implants cross-frame attention into the SAM to learn spatio-temporal information in videos. However, the fixed step size of cross-frame attention limits spatio-temporal representation capabilities, thus making it challenging to handle the large frame-to-frame variations in colonoscopic videos.

In this paper, we propose a novel model adapted from SAM for VPS to comprehensively meet its challenges. Our model has two innovative components: a semantic disentanglement adapter (SDA) and a spatio-temporal side network (STSN); we call it *VP-SAM*. In the SDA, we exploit amplitude information of the Fourier spectrum to facilitate the SAM to recognize and erase background interference, which is one of the main reasons that SAM performs poorly on medical data like colonoscopic videos, where the contrast between targeting objects and background is very low. In the STSN, we efficiently supplement SAM with spatio-temporal information to facilitate the SAM to track the motion status of polyps, and hence further improve the segmentation accuracy. We conduct extensive experiments on three benchmarking datasets, SUN-SEG [22], CVC-612 [3], and CVC-300 [4], with comprehensive comparison with task-specific SOTA methods, as well as recently proposed SAM-based models. Experimental results show that our VP-SAM significantly outperforms existing methods. Our major contributions are summarized as follows:

- We propose a novel model adapted from SAM for VPS to comprehensively meet its challenges; our model takes advantage of SAM in terms of representation capability while adapting SAM to medical videos with low object-background contrast and large frame-to-frame variations.
- We propose a new semantic disentanglement adapter (SDA) and an innovative spatio-temporal side network (STSN), where the former facilities SAM to disentangle targeting objects from similar surroundings, and the latter supplements SAM with spatio-temporal information in videos; working together, they achieve SOTA performance in VPS.
- Our method significantly outperforms SOTA methods on three famous colonoscopic video datasets: SUN-SEG, CVC-612, and CVC-300, demonstrating the effectiveness of our proposed method.

2 Related Work

2.1 Polyp Segmentation

With the development of deep learning, some CNN-based methods [9, 14, 25, 33, 41] rely on excellent local information modeling capabilities for image polyp segmentation. However, the lack of global information limits these methods. Therefore, some researchers [11, 16, 28, 40, 46] combine Transformer [39] and CNN to simultaneously model global and local information, thus improving performance.

However, these image-based methods fail to exploit the temporal information in colonoscopic videos. To this end, some methods use hybrid 2/3D architectur [35] or normalized self-attention mechanisms [20, 22] to learn spatio-temporal information in videos, achieving performance improvement. However, limited domain-specific training samples and inadequate network capacity hinder fine segmentation. Polyp-sam [29] retrain SAM on polyp datasets, while SAMAug [45] uses SAM to augment polyp images without training SAM. However, both methods target images and thus also ignore valuable temporal information.

2.2 Foundation Models

Different from task-specific models, foundation models are essentially large pre-trained models on large-scale datasets, which usually yield strong feature extraction abilities. Recently, SAM [26] pre-trained on SA-1B, as a foundation model for computer vision, has achieved impressive performance. SAM can handle multiple prompts, such as points, bounding boxes, and coarse masks, to produce high-quality segmentation. Such a segmentation paradigm shows the potential to segment any object. However, some recent attempts have shown that SAM exhibits significant performance degradation on some special downstream tasks [17,18,21,23,47], such as medical segmentation. This discovery has stimulated some studies [7, 8, 15, 29, 30, 32, 42, 44] to explore how to improve the performance of SAM on these tasks.

2.3 SAM-based Medical Segmentation

Due to the performance degradation of the foundation model in some downstream tasks, various tuning strategies are used to improve performance. Most prior tuning strategies either retrain part or whole parameters of SAM [29, 32], or introduce and train additional learnable adapters [6,8,15,30,42,44]. However, achieving fine VPS is extremely challenging for these tuning strategies due to the inability to extract temporal information in videos. To this end, MediViSTA-SAM [24] implants cross-frame attention into each transformer block of SAM. However, fixing the step size of cross-frame attention makes the spatio-temporal representation sensitive to large frame-to-frame variations in colonoscopic videos. Different from these tuning strategies, we propose a new semantic disentanglement adapter (SDA) to overcome low-contrast interference, and propose an innovative spatio-temporal side network (STSN) to provide more precise spatio-temporal information for SAM.

3 Method

3.1 Overall Architecture

The original SAM framework consists of three key components: an image encoder for extracting image embedding, a prompt encoder for extracting prompt

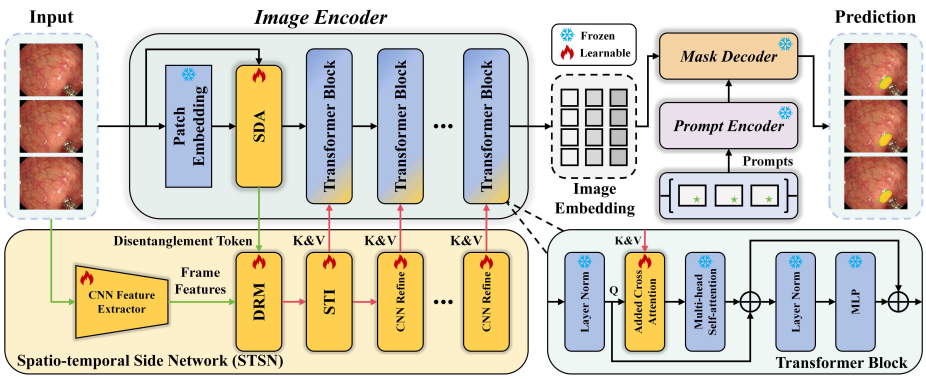

Fig. 2: Overview of our method, which mainly introduces a semantic disentanglement adapter (SDA) and a spatio-temporal side network (STSN). DRM is the disentanglement reference mechanism, and STI represents the spatio-temporal injector.

embedding, and a mask decoder for combining these embedding to generate segmentation masks. To effectively adapt SAM for VPS, we must address two critical challenges. First, it is crucial to identify semantic entanglement states for better distinguishing polyps from background tissue. Second, it is necessary to incorporate spatio-temporal information into the network to make SAM compatible with video data, as it was initially trained on pure 2D images. Based on these insights, we introduce two essential modules into the SAM framework, which are the semantic disentanglement adapter (SDA) and the parallel spatio-temporal side network (STSN), as shown in Figure 2. The semantic disentanglement adapter aims to eliminate entanglement signals between polyps and background, and the spatio-temporal side network is responsible for guiding SAM to focus on spatio-temporal information in videos. Given a colonoscopic video clip, both the SDA and STSN modules interact with the image encoder to obtain high-quality image embedding, which is then forwarded to the mask decoder to generate segmentation results. During tuning process, only our introduced modules are updated, while all the original modules in SAM remain frozen.

3.2 Semantic Disentanglement Adapter

The quality of features crucial for the task of VPS is notably degraded due to the low-contrast between polyps and background caused by semantic entanglement in colonoscopic videos. Therefore, it is necessary to explicitly model and eliminate these entanglement states in order to tailor SAM for effective VPS. To achieve this, as shown in Figure 3 (a), our SDA module consists of an entanglement states encoder and a semantic disentanglement identification process. The entanglement states encoder leverages amplitude information from the Fourier spectrum to explicitly model entanglement states between polyps and background. Subsequently, the semantic disentanglement identification process involves constructing a difference map between the original image embedding

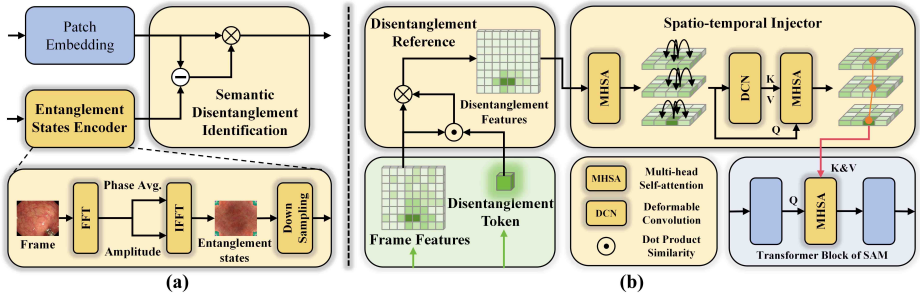

Fig. 3: (a) SDA mainly consists of a entanglement states encoder (ESE) and a semantic disentanglement identification (SDI). (b) STSN mainly includes the disentanglement reference mechanism (DRM) and the spatio-temporal injector (STI).

and entanglement states to obtain disentanglement embedding, thus improving the quality of features.

Entanglement States Encoder. The low contrast between polyps and background is caused by entanglement states in colonoscopic videos, which often consist of color and texture [11, 13, 27, 31]. Fortunately, it is known that the amplitude in the Fourier spectrum can represent these low-level color and texture information. Therefore, we can effectively capture entanglement states using the amplitude information present in the Fourier spectrum. Specifically, given a colonoscopic sequence clip $x \in \mathbb{R}^{L\times H\times W\times 3}$, we first perform the following fast Fourier transform (FFT) to obtain the Fourier spectrum:

$$\mathcal{F}(x)_{u,v} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j} e^{-J2\pi(\frac{ui}{H}+\frac{vj}{W})}, \tag{1}$$

where L is the clip length, $H \times W$ is the resolution of the clip, and J represents the imaginary unit. Then, the amplitude component can be extracted as:

$$A(x)_{u,v} = \sqrt{R^2(x)_{u,v} + I^2(x)_{u,v}}, \tag{2}$$

where $R(x)$ represents the real part of $\mathcal{F}(x)$, while $I(x)$ is the imaginary part. Then, in order to focus only on low-level entanglement states, we need to reconstruct the clip using only amplitude information, since low-level information in images is usually represented by amplitude in frequency domain space, while phase information usually represents contour information. In this way, we can fix all phase values to be constant (e.g. average value) while keeping the amplitude information unchanged when reconstructing the clip, thus allowing the reconstructed clip to focus only on low-level entanglement states as follows:

$$\widetilde{x} = \widetilde{\mathcal{F}}[A(x)_{u,v} e^{-Jc}], \tag{3}$$

where $\widetilde{\mathcal{F}}(\cdot)$ is the inverse fast Fourier transform (IFFT), and c is the average of phase components. In this way, the reconstructed $\widetilde{x}$ can represent the entanglement states between polyps and background in RGB domain. Following this, $\widetilde{x}$

passes through a CNN down-sampling module, which consists of max pooling and convolutional layers, to embed $\widetilde{x}$ into entanglement state F_e for subsequent semantic disentanglement identification.

Semantic Disentanglement Identification. Given the entanglement states $F_e \in \mathbb{R}^{L\times h\times w\times C}$ obtained from the entanglement states encoder and the clip patch embedding $F \in \mathbb{R}^{L\times h\times w\times C}$ obtained from patch embedding of SAM, we use the difference map between F_e and F as a weight map to emphasize the features related to polyps while weakening the interference of entanglement states, where $h \times w$ is the resolution of embedding. Specifically, we compute the difference map between F_e and F to obtain the disentanglement embedding. Then, the difference map is multiplied as a per-pixel weight with the clip patch embedding to locate more precise semantic regions. The whole process can be formulated as follows:

$$F_d = F \cdot (F - F_e)^2, \tag{4}$$

where F_d represents the final disentanglement patch embedding. Through these operations, semantic disentanglement identification adapts SAM to the distribution of entanglement states between polyps and background.

3.3 Spatio-temporal Side Network

The spatio-temporal information plays an important role for VPS. First, maintaining spatial consistency across frames can improve object perception and boost the robustness of segmentation. Second, temporal consistency allows define the motion state of polyps, which facilitates the localization of fine-grained semantic regions. Hence, enabling SAM to perceive spatio-temporal information in colonoscopic videos is essential for effective VPS. To achieve this, as shown in Figure 3 (b), we first use the disentanglement reference mechanism to attenuate the interference of entanglement states on spatio-temporal information in colonoscopic videos. Then, we introduce a spatio-temporal injector to supplement spatio-temporal information into the transformer block of SAM.

Disentanglement Reference Mechanism. The low contrast caused by entanglement states in colonoscopic videos may diminish object perception and make the motion state of polyps difficult to capture. Therefore, it is necessary to mitigate the interference caused by entanglement states on spatio-temporal information to ensure accurate and reliable processing. Specifically, we first introduce a CNN-branch [30] to extract features $F_{s,t} \in \mathbb{R}^{L\times h\times w\times C}$ with spatio-temporal information. Then, in order to mitigate the interference caused by entanglement states on spatio-temporal information, we can establish a consensus $F_{d,s,t}$ between F_d and $F_{s,t}$, which identifies global-pixel disentanglement level and also maintains spatio-temporal information in videos. For simplicity, we use the i-th frame $F^i_{s,t}$ and F^i_d in the clip as an illustration, where $i \in 1, 2, \cdots, L$. In particular, we first use F^i_d and $F^i_{s,t}$ to build the affinity correlation as:

$$s_{i,j,k} = \frac{d(F^i_d, F^i_{s,t})}{\sqrt{C}} \in \mathbb{R}^{h\times w}, \tag{5}$$

where $d(\cdot,\cdot)$ represents the dot product similarity [39], C is the scaling factor, $j \in 1, 2, \ldots, h$, and $k \in 1, 2, \ldots, w$. Then, we use the normalized $s_{i,j,k}$ as a weight map to update $F^i_{s,t}$ using the following equation:

$$F^i_{d,s,t} = \frac{F^i_{s,t} \cdot \exp(s_{i,j,k})}{\sum_{j=1}^{h} \sum_{k=1}^{w} \exp(s_{i,j,k})}. \tag{6}$$

Finally, we concatenate $F^i_{d,s,t}$ into $F_{d,s,t}$. In this way, $F_{d,s,t}$ can represent the consensus between spatio-temporal information and disentanglement token, which eliminates the interference of entanglement states without losing spatio-temporal information, thus providing high-quality features for subsequent injection of spatio-temporal consistency.

Spatio-temporal Injector. Given the disentanglement spatio-temporal feature $F_{d,s,t}$, we employ vanilla self-attention [39] to capture spatio-temporal consistency. To model spatio-temporal consistency more efficiently, we follow [38] to decouple the modeling of the spatial and temporal consistency. In our framework, we first introduce self-attention modules to independently model the global spatial dependence within each frame, ensuring precise spatial alignment. Given the potential for varying semantics in pixels at identical positions across frames caused by the large frame-to-frame variations in videos, we incorporate deformable convolution [10] to learn the motion offset associated with each pixel, aiming at overcoming large frame-to-frame variations. Then, we model the temporal dependencies within the spatial dimension by using the aligned spatial features as queries and the extracted offset as keys and values. The entire modeling of spatio-temporal consistency can be described as follows:

$$F^{'}_{d,s,t} = \mathcal{E}_2(\Delta, \mathcal{D}(\Delta)), \Delta = \mathcal{E}_1(F_{d,s,t}), \tag{7}$$

where $\mathcal{E}_1(\cdot)$ and $\mathcal{E}_2(\cdot)$ represent self-attention modules, and $\mathcal{D}(\cdot)$ is deformable convolution layers. In this way, $F^{'}_{d,s,t}$ can model accurate spatio-temporal consistency under disentanglement condition.

Furthermore, in order to enable SAM to capture frame-to-frame spatio-temporal information, we introduce a simple yet effective cross-attention module to integrate the spatio-temporal information from $F^{'}_{d,s,t}$ into SAM. Specifically, we use the embedding F_{sam} derived from SAM as queries, and $F^{'}_{d,s,t}$ as keys and values to supplement the missing spatio-temporal information in the transformer blocks of SAM. The supplementary process of spatio-temporal information for the first transformer block is described as:

$$S(F_{sam}, F^{'}_{d,s,t}) = Softmax(\frac{F_{sam}E_q(F^{'}_{d,s,t}E_k)^{\mathsf{T}}}{\sqrt{C}} + B)(F^{'}_{d,s,t}E_v), \tag{8}$$

where $E_q, E_k, E_v \in \mathbb{R}^{C\times C}$ are the learnable weight matrices, and $B \in \mathbb{R}^{hw\times hw}$ is position embedding. In order to minimize the computational overhead, we reuse $F^{'}_{d,s,t}$ to supplement the spatio-temporal information of the remaining transformer blocks. Considering that different transformer blocks have different

mean attention distance (pixels) [36], the 3×3 convolutional layer that expands the receptive field is used to bind the gradually larger mean attention distance in the transformer blocks, thus efficiently adapting spatio-temporal information to different transformer blocks.

To sum up, with the aid of our STSN and SDA, the image embedding extracted by the image encoder have spatio-temporal information and strong disentanglement perception, which is the high quality features for VPS and can be used to obtain the final mask prediction z as follows:

$$z = \mathcal{M}(\mathcal{I}(x), \mathcal{P}(y)), \tag{9}$$

where $\mathcal{M}(\cdot)$ is the mask decoder, $\mathcal{I}(\cdot)$ is the image encoder, $\mathcal{P}(\cdot)$ is the prompt encoder, and y is the prompt information.

3.4 Loss Function

We apply a binary cross-entropy loss $\mathcal{L}_{bce}$ like [20,22] to supervise our adaptation process. In order to further weaken the impact of inconsistent distributions of targets in different clips, a dice loss $\mathcal{L}_{dice}$ is chosen as another part of our loss. In this way, our tuning objective can be defined as follows:

$$\min_{\mathbb{P}} \mathcal{L}_{bce}(z, G) + \mathcal{L}_{dice}(z, G), \tag{10}$$

where G is the ground truth, and $\mathbb{P}$ is the learnable parameters.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate our method on three video-based datasets: SUN-SEG [22], CVC-612 [3] and CVC-300 [4]. (1) SUN-SEG includes 49,136 frames from 285 sequences, which consists of three subsets: training set contains 19,544 frames from 112 sequences; test set SUN-SEG-Easy contains 17,070 frames from 119 sequences; test set SUN-SEG-Hard includes 12,522 frames from 54 sequences. (2) CVC-612 consists of 612 frames from 31 colonoscopic sequences at 576×768 resolution. (3) CVC-300 contains 15 cases, each with 20 frames at 500×574 resolution.

Evaluation Metrics. We apply mean Dice (mDice), mean Interaction over Union (mIoU), and mean Hausdorff Distance (mHD) to quantify our method.

Implementation Details. The tuning process is performed on a single RTX 3090 GPU with batch size of 2, using AdamW and poly learning rate schedule with an initial learning rate of 0.0001 and a weight decay of 0.1 for 10 epochs. For SUN-SEG, we separate 20% from the training set as the validation set. For CVC-612, we split the training set, validation set, and test set with a ratio of $7:1:2$. CVC-300 is used to test generalization ability. All frames are resized to a resolution of 256×256. Clip length L is set to 3, which can be changed based on GPU memory. ViT-B [12] is used as the backbone for all experiments. For fair comparison, all SAM-based methods use the same prompts (1 pt/frame) generated by a random algorithm [30], except for manual prompts.

Table 1: Comparison with *task-specific* and *SAM-based* methods on three datasets.

Method	Year	Backbone	SUN-SEG-Easy			SUN-SEG-Hard			CVC-612			CVC-300		
			mDice	mIoU	mHD	mDice	mIoU	mHD	mDice	mIoU	mHD	mDice	mIoU	mHD
DCNet	2023	ViT-B	78.36	68.63	29.24	75.04	65.55	28.48	88.21	80.85	25.77	86.24	78.25	18.26
TarVIS	2023	ViT-B	79.16	68.95	28.13	76.44	66.96	28.31	89.73	82.84	22.24	86.97	79.77	17.52
PNS+	2022	ViT-B	79.26	70.24	26.38	76.51	68.11	28.67	90.06	83.43	21.79	86.59	78.24	15.98
META-UNet	2023	ViT-B	81.17	72.43	25.71	80.16	70.55	26.27	90.64	84.39	20.49	86.64	78.55	16.02
MSAF	2023	ViT-B	81.33	73.18	25.19	80.52	71.33	25.82	90.47	84.58	21.29	86.83	78.72	16.53
Ours (w/o prompts)	2024	ViT-B	**85.62**	**78.16**	**21.22**	**85.28**	**77.16**	**21.38**	**92.33**	**86.79**	**19.34**	**88.26**	**80.17**	**14.06**
SAM	2023	ViT-B	54.67	46.42	312.97	55.53	47.09	296.93	59.68	49.54	286.37	65.01	56.26	205.38
MedSAM	2023	ViT-B	69.04	60.29	220.35	68.23	58.71	207.92	76.08	66.82	182.58	79.02	70.69	113.32
Polyp-SAM	2023	ViT-B	70.80	61.37	151.36	70.42	60.31	151.24	77.76	68.58	135.33	80.96	71.33	91.52
SAMed	2023	ViT-B	78.23	67.99	28.58	76.94	66.78	28.99	89.92	83.33	23.69	86.41	78.98	16.38
SAM-Med2D	2023	ViT-B	81.99	73.37	25.55	80.41	71.37	26.44	90.23	83.68	23.98	86.81	79.23	16.12
MediViSTA-SAM	2023	ViT-B	84.34	77.21	22.36	83.25	73.34	24.86	90.47	85.06	21.15	87.09	79.52	14.98
SAMUS	2023	ViT-B	84.83	77.48	21.72	84.11	75.04	22.52	91.12	85.15	20.24	87.17	79.41	14.33
Ours (1 pt/frame)	2024	ViT-B	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**19.64**	**93.54**	**88.83**	**17.86**	**89.93**	**82.38**	**12.38**

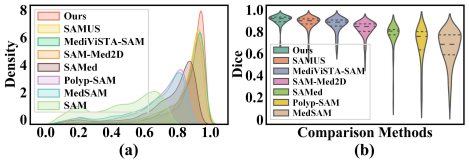

Fig. 4: (a) Density estimation on SUN-SEG-Easy. (b) Violin plot on CVC-612.

Table 2: Comparison results with SOTA spatio-temporal strategies or adapters on SUN-SEG dataset.

Method	SUN-SEG-Easy			SUN-SEG-Hard		
	mDice	mIoU	mHD	mDice	mIoU	mHD
SAM + Point-Tracker [37]	58.85	47.96	308.93	57.36	48.59	290.87
SAM + Mask-Tracker [43]	60.31	48.13	310.48	58.74	47.67	294.56
SAM + ST-Adapter [34]	84.17	76.38	22.65	83.44	75.32	23.15
Ours	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**19.64**

4.2 Comparison with State-of-the-art Methods

To demonstrate the superiority of our proposed method, we compare our method with some SOTA methods, including task-specific methods PNS+ [22], DCNet [31], TarVIS [2], MSAF [38], and META-UNet [40] and SAM-based methods SAM [26], MedSAM [32], Polyp-SAM [29], SAMed [44], SAM-Med2D [8], MediViSTA-SAM [24], and SAMUS [30].

Comparison with SOTA Methods. As shown in Table 1, compared with SOTA task-specific methods, our method (*w/o prompts*) shows significant performance improvements in all metrics, which is attributed to our specific SDA, STSN, and the inherent design advantages of the SAM. Furthermore, SAM trained on SA-1B, without fine-tuning, shows significant performance degradation on three datasets, which indicates that low polyp-background contrast in colonoscopic videos and the inability to extract temporal information severely hinder SAM from performing fine segmentation. Among all tuning methods, our method (1 pt/frame) achieves better performance, which shows that our method possesses strong semantic disentanglement abilities while leveraging valuable temporal information. As shown in Figure 4, the high probability density of Dice intervals (e.g., [0.8, 1.0]) in the kernel density estimation and violin plot demonstrates that our method is more robust.

Comparison with Spatio-temporal Adapters/Strategies. As shown in Table 2, to more comprehensively evaluate the spatio-temporal modeling

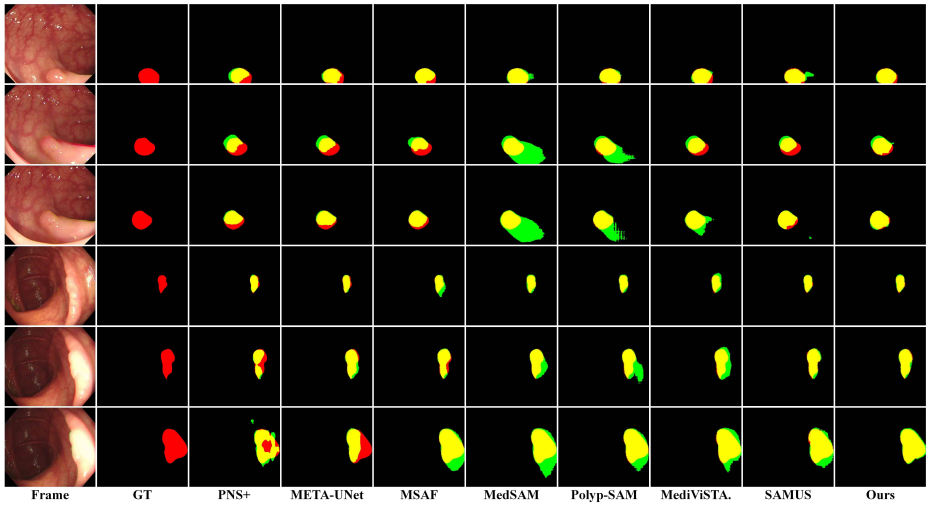

Fig. 5: Visual comparison with SOTA methods on SUN-SEG. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively. (*w/o prompts*)

capabilities of our method, we also compare our method with other spatio-temporal adapters, including spatio-temporal adapters (ST-Adapter) [34], point-tracker [37], and mask-tracker [43]. Our method still achieves better performance. For ST-Adapter, the kernel size of 3D convolution limits its spatio-temporal representation, making it difficult to cope with large frame-to-frame variations in colonoscopic videos. For point/mask tracker, the segmentation quality of the initial frame will seriously affect the segmentation of subsequent frames.

Visual Comparison with SOTA. As shown in Figure 5, our proposed method is able to identify more accurate target semantics in colonoscopic videos through the introduced SDA and STSN, thereby achieving fine segmentation. Compared to previous methods, our method achieves better target boundary separability, which shows that SDA enables SAM to have strong capabilities of semantic disentanglement, while STSN enables SAM to utilize the valuable spatio-temporal information in colonoscopic videos.

4.3 Ablation Studies

Effectiveness of SDA. Our SDA mainly consists of a entanglement states encoder (ESE) and a semantic disentanglement identification (SDI). As shown in Table 3, the performance degradation caused by removing both components suggests that our SDA is necessary for SAM to adapt to polyp segmentation from colonoscopic videos. The improvement brought by introducing the ESE and SDI verifies that Fourier spectral amplitude enables SAM to explicitly perceive entanglement states between polyps and background. We also conduct visual comparison. As shown in the Figure 6, the complete SDA makes the localization of prediction closer to GT, indicating the effectiveness of the components. The

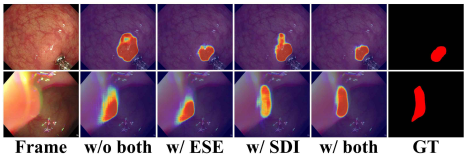

Fig. 6: Visual comparison of image embedding on SDA components.

Table 3: Component ablation experiments of SDA on SUN-SEG.

ESE	SDI	SUN-SEG-Easy			SUN-SEG-Hard		
		mDice	mIoU	mHD	mDice	mIoU	mHD
		85.76	77.92	21.29	85.68	77.53	21.46
✓		86.73	78.56	20.44	86.41	78.26	20.11
	✓	86.57	78.45	20.23	86.38	77.93	20.23
✓	✓	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**19.64**

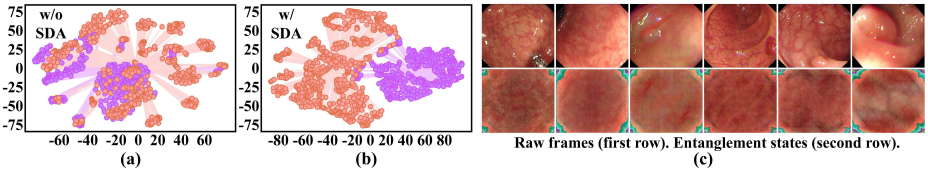

Fig. 7: (a-b) t-SNE visualization of image embedding, where purple represents polyps and red represents background. (c) Visualization of modeled entanglement states.

modeled entanglement states in Figure 7 (c) has many similar low-level characteristics to the background in the original frame, which can be used as entanglement states. Furthermore, as shown in Figure 7 (a-b), SDA eliminates entanglement states, allowing for a clearer demarcation between polyps and background, which once again proves the effectiveness of SDA.

Effectiveness of STSN. Our STSN mainly consists of a disentanglement reference mechanism (DRM) and a spatio-temporal injector (STI). As shown in Table 4, the performance degradation caused by removing STSN is very intuitive, because it ignores valuable spatio-temporal information in videos. Moreover, we observe that the improvement brought by each component working individually is not as significant as the improvement achieved by working together, which verifies that directly modeling spatio-temporal consistency without reference to disentanglement token suffers from the interference of pixel similarity brought by low polyp-background contrast. It also shows that consensus between disentanglement and spatio-temporal information can promote purer spatio-temporal consistency. Furthermore, we also conduct visual comparison on components of STSN. As shown in Figure 9, each component working individually makes the separation between target and background more difficult, while working together allows for a clearer division.

Effect of Different Manual Prompts. Based on SAM and our method, we qualitatively compare the results obtained with different manual prompts (e.g. different number of point prompts, different prompt modes), as shown in Figure 8. In most cases, bounding boxes can facilitate the model to obtain more accurate semantic regions, but the visual results obtained by our bbox-based method are clearer and closer to GT. In the cases of point prompts, SAM struggles to locate polyp regions, resulting in some false positives and false negatives. Compared

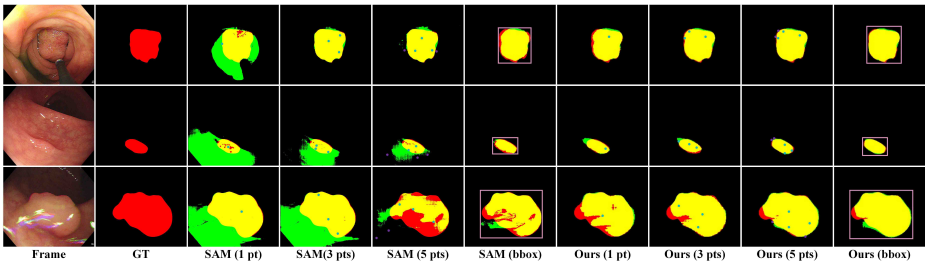

Fig. 8: Comparison of results on different *manual prompts*. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively. Blue points represent the positive, and purple points represent the negative. (*best viewed on zoom*)

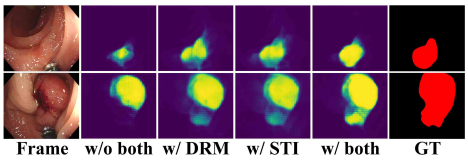

Fig. 9: Visual comparison of image embedding on STSN components.

Table 4: Component ablation experiments of STSN on SUN-SEG.

DRM	STI	SUN-SEG-Easy			SUN-SEG-Hard		
		mDice	mIoU	mHD	mDice	mIoU	mHD
		82.61	74.43	26.04	82.22	73.95	28.69
✓		83.43	75.31	24.16	82.96	74.88	25.66
	✓	85.87	78.02	20.56	85.73	77.61	20.46
✓	✓	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**19.64**

with SAM, the results obtained by our point-based method are more stable and superior, among which the results based on 1 pt prompt are closer to GT.

Effect of Number of Point Prompts. As shown in Figure 10, we also evaluate the performance with different point prompts created by random algorithm [30]. In general, our approach is robust to the number of point prompts. However, more points does not necessarily lead to performance gains. We analyze the possible reason is that more point prompts increase the probability of random points being close to the blurred boundary, which may increase the confidence of false positives.

Effect of Clip Length L. As shown in Figure 11 (a-b), we also explore the impact of different clip length L. The performance improves greatly when L increases from 1 to 3 because more spatio-temporal information is obtained. However, larger L values (e.g., 5 and 7) cause performance degradation. Longer clips can theoretically bring more spatio-temporal information, which is effective for clips composed of frames with high boundary discrimination. However, for colonoscopic videos with low boundary discrimination, we analyze the possible reason is that establishing spatio-temporal information between frames with a long temporal distance may bring redundant information that interferes with effective spatio-temporal information.

4.4 Discussions and Limitations

While we only conduct experiments on colonoscopic video datasets, we believe that our VP-SAM is general enough to be used to analyze other medical videos

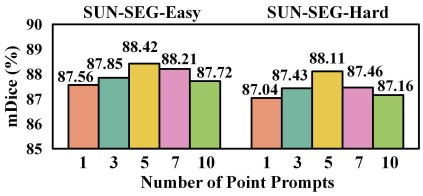

Fig. 10: Ablation experiments on the number of point prompts on SUN-SEG.

Table 5: Performance vs. efficiency on SUN-SEG-Hard based on a single RTX 3090 GPU and batch size equal to 1.

Method	Resolution	SUN-SEG-Hard			
		mDice	Params	GFLOPs	FPS
SAM [26]	1024 × 1024	55.53	90.49M	1115.96	12.18
SAMed [44]	256 × 256	76.94	90.36M	284.17	53.63
SAM-Med2D [8]	256 × 256	80.41	270.99M	196.11	35.68
MediViSTA-SAM [24]	256 × 256	83.25	135.02M	102.29	48.56
SAMUS [30]	256 × 256	84.11	130.10M	429.77	42.45
Ours	256 × 256	87.04	142.27M	470.59	38.23

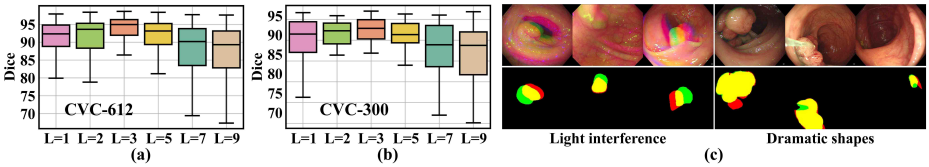

Fig. 11: (a-b) Ablation studies of clip length L. (c) Failure cases, where red, green and yellow represent the GT, prediction and their overlapping regions, respectively.

with similar challenges. Furthermore, as shown in Table 5, we also conduct a performance-efficiency comparison with SOTA methods. Compared with these methods, while we are somewhat less efficient, our method achieves significant performance improvements with a real-time inference speed (38.23 FPS), which can be a trade-off. In fact, the main overhead is caused by considering the spatio-temporal information (modeling and injection) in colonoscopic videos, which is inevitable since SAM is not a video segmentation model. It is known that modeling spatio-temporal information in videos is a computationally intensive process. In order for SAM to perceive the valuable spatio-temporal information in videos, we must additionally do this computationally intensive process since SAM is trained on pure 2D images, thus causing a loss in efficiency. Furthermore, our method still has some limitations. As shown in Figure 11 (c), the light interference and dramatic shapes may limit our method.

5 Conclusion

In this paper, we propose a novel model VP-SAM for video polyp segmentation, which is adapted from SAM via a semantic disentanglement adapter (SDA) and a spatio-temporal side network (STSN). The SDA allows SAM to explicitly sense and eliminate entanglement states between polyps and background in colonoscopic videos. Besides, the STSN supplements SAM with valuable spatio-temporal information in colonoscopic videos. With the aid of our SDA and STSN, our method enhances the semantic disentanglement capabilities of SAM and solves the defect of SAM being unable to extract valuable spatio-temporal information in colonoscopic videos. Extensive experimental results on SUN-SEG, CVC-612 and CVC-300 demonstrate the effectiveness of our proposed method.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of China (No. 62273241), the Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), the Shenzhen Institute of Artificial Intelligence and Robotics for Society, and the General Research Fund of Hong Kong Research Grants Council (No. 15218521).

References

1. Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut and liver **6**(1), 64 (2012)
2. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.: Tarvis: A unified approach for target-based video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18738–18748 (2023)
3. 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 **43**, 99–111 (2015)
4. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
5. Center, M.M., Jemal, A., Smith, R.A., Ward, E.: Worldwide variations in colorectal cancer. CA: a cancer journal for clinicians **59**(6), 366–378 (2009)
6. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
7. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
8. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
9. Cheng, M., Kong, Z., Song, G., Tian, Y., Liang, Y., Chen, J.: Learnable oriented-derivative 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. 720–730. Springer (2021)
10. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
11. 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)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
13. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2777–2787 (2020)

14. 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. Springer (2019)
15. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv preprint arXiv:2306.13465 (2023)
16. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Transactions on Medical Imaging (2023)
17. 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)
18. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? arXiv preprint arXiv:2304.14660 (2023)
19. Hurlstone, D., Cross, S., Slater, R., Sanders, D., Brown, S.: Detecting diminutive colorectal lesions at colonoscopy: a randomised controlled trial of pan-colonic versus targeted chromoscopy. Gut **53**(3), 376–380 (2004)
20. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)
21. 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)
22. 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)
23. 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)
24. Kim, S., Kim, K., Hu, J., Chen, C., Lyu, Z., Hui, R., Kim, S., Liu, Z., Zhong, A., Li, X., et al.: Medivista-sam: Zero-shot medical video analysis with spatio-temporal sam adaptation. arXiv preprint arXiv:2309.13539 (2023)
25. Kim, T., Lee, H., Kim, D.: Uacanet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2167–2175 (2021)
26. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
27. Le, T.N., Cao, Y., Nguyen, T.C., Le, M.Q., Nguyen, K.D., Do, T.T., Tran, M.T., Nguyen, T.V.: Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. IEEE Transactions on Image Processing **31**, 287–300 (2021)
28. Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R.: Medical image segmentation using squeeze-and-expansion transformers. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 807–815. International Joint Conferences on Artificial Intelligence Organization (8 2021). https://doi.org/10.24963/ijcai.2021/112, https://doi.org/10.24963/ijcai.2021/112, main Track

29. Li, Y., Hu, M., Yang, X.: Polyp-sam: Transfer sam for polyp segmentation. arXiv preprint arXiv:2305.00293 (2023)
30. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 (2023)
31. Luo, N., Pan, Y., Sun, R., Zhang, T., Xiong, Z., Wu, F.: Camouflaged instance segmentation via explicit de-camouflaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17927 (2023)
32. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
33. 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. 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. 633–643. Springer (2021)
34. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems **35**, 26462–26477 (2022)
35. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 295–305. Springer (2020)
36. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems **34**, 12116–12128 (2021)
37. Rajič, F., Ke, L., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Segment anything meets point tracking. arXiv preprint arXiv:2307.01197 (2023)
38. Su, J., Yin, R., Zhang, S., Luo, J.: Motion-state alignment for video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3570–3579 (2023)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
40. Wu, H., Zhao, Z., Wang, Z.: Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. IEEE Transactions on Automation Science and Engineering (2023)
41. Wu, H., Zhao, Z., Zhong, J., Wang, W., Wen, Z., Qin, J.: Polypseg+: A lightweight context-aware network for real-time polyp segmentation. IEEE Transactions on Cybernetics **53**(4), 2610–2621 (2022)
42. 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)
43. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968 (2023)
44. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
45. Zhang, Y., Zhou, T., Wang, S., Liang, P., Zhang, Y., Chen, D.Z.: Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 129–139. Springer (2023)

46. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image 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. 14–24. Springer (2021)
47. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can sam segment polyps? arXiv preprint arXiv:2304.07583 (2023)