AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder

Tal Shaharabany
shaharabany@mail.tau.ac.il
Aviad Dahan
aviaddahan@mail.tau.ac.il
Raja Giryes
raja@tauex.tau.ac.il
Lior Wolf
wolf@cs.tau.ac.il

Tel-Aviv University, Israel

Abstract

The recently introduced Segment Anything Model (SAM) combines a clever architecture and large quantities of training data to obtain remarkable image segmentation capabilities. However, it fails to reproduce such results for Out-Of-Distribution (OOD) domains such as medical images. Moreover, while SAM is conditioned on either a mask or a set of points, it may be desirable to have a fully automatic solution. In this work, we replace SAM's conditioning with an encoder that operates on the same input image. By adding this encoder and without further fine-tuning SAM, we obtain state-of-the-art results on multiple medical images and video benchmarks. This new encoder is trained via gradients provided by a frozen SAM. For inspecting the knowledge within it, and providing a lightweight segmentation solution, we also learn to decode it into a mask by a shallow deconvolution network. Our code is publicly available at https://github.com/talshaharabany/AutoSAM

1 Introduction

The promptable image segmentation model, SAM [23], is an efficient and practical approach to real-world segmentation tasks that allows for flexibility in prompts, quick mask computation, and ambiguity awareness. However, SAM's performance may not be optimal on medical imaging datasets due to its pre-training on natural images as illustrated in Fig. 1.

In this paper, we propose an end-to-end approach to improve segmentation mask accuracy for medical images without fine-tuning the pretrained SAM network. Our solution involves the training of an auxiliary prompt encoder network, which generates a surrogate prompt for SAM given an input image. While the prompt encoder provided with SAM can accept inputs such as a bounding box, a set of points, or a mask, the one we train has the image itself as its input. We term this overloading, since in object-oriented programming, overloading is a feature that allows a class to have multiple methods with the same name, but with different types of input parameters.

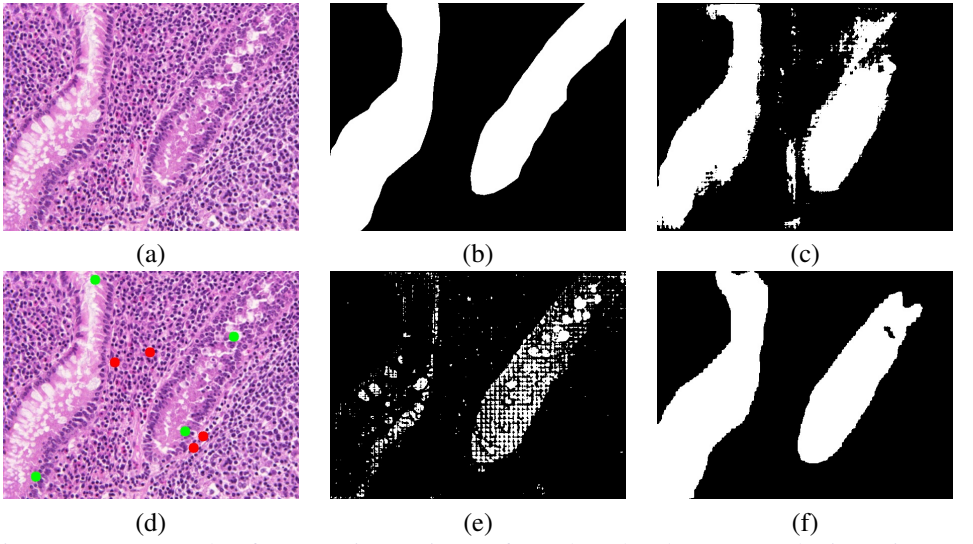


Figure 1: An example of segmenting an image from the Glas dataset. (a) the input image. (b) the ground truth mask. (c) the results of SAM with the GT mask provided to its mask encoder. (d) a point-based prompt. (e) SAM's result based on the point prompt. (f) our result, where the input image itself is given as a prompt to the prompt-encoder we train.

During training, the SAM network propagates gradients to the prompt encoder network from a binary cross-entropy loss and a Dice loss. The encoder network that we train employs the Harmonic Dense Net [5] as its backbone and has significantly fewer learnable parameters than SAM's own decoders. As mentioned, the main SAM network is not modified, which makes our method easy to implement and avoids finding a suitable training schedule for SAM fine-tuning.

We have evaluated our method on multiple publicly available medical images and videos datasets. Our results show a significant improvement in segmentation performance compared to the baseline method and other state-of-the-art approaches.

2 Related Work

Medical image segmentation is an active research area that plays a vital role in diagnosis [7], treatment planning [38], and disease monitoring [31]. U-net [35] has been widely used for various medical image segmentation tasks. Over the years, various modifications and versions have been proposed for the U-net segmentation architecture [32, 37, 49, 53, 58].

Our solution is based on SAM [23], which is based on a visual-transformer [42], similar to other segmentation architectures [42]. SAM [23] is trained on the largest segmentation dataset reported to date, comprising over 1 billion masks on 11 million licensed and privacy-respecting natural images. The model serves as an effective foundation model and its zero-shot performance is comparable to or better than many fully supervised results in natural image segmentation. Moreover, its modular and promptable design enables transfer learning to new tasks and image distributions. In this work, we harness these properties in order to achieve SOTA results on out-of-distribution (OOD) data by replacing SAM's built-in prompt encoder with our custom encoder.

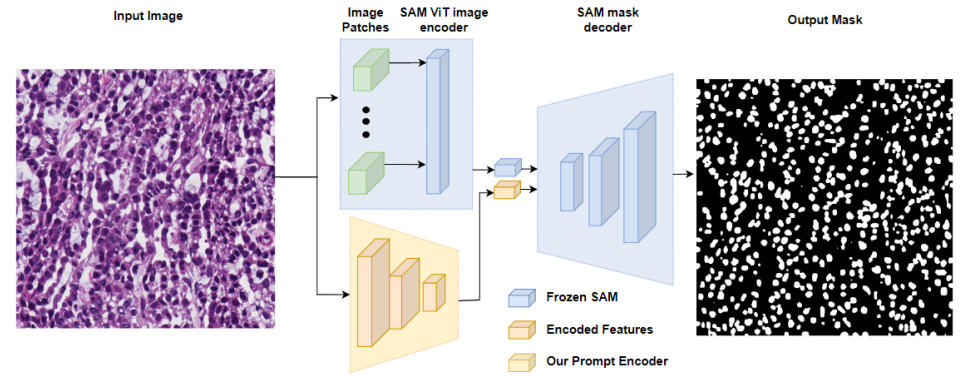


Figure 2: An illustration of AutoSAM. SAM's prompt encoder is replaced with our custom encoder while the image encoder and mask decoder are frozen.

Concurrent with our work, Zu et al. [52] fine-tune SAM's encoder and decoder using adaptation blocks (this technique is used as a baseline in [17]). The prompt encoders of SAM are not tuned and this method, therefore, requires a prompt in the form of positive points. In our method, we replace the prompt encoder. The encoder that we train receives the same input image as the main network, hence the name AutoSAM. Note that we do not fine-tune the encoder and decoder of SAM. Moreover, as we show in Section 4, our encoder can be easily converted to a segmentation network by simply adding a few convolutional layers to it and training them for this task.

In modern Large Language Models [46] and promptable text-image models such as diffusion models [34, 36], a careful prompt can draw the line between a desired outcome and an unusable result. The task of learning the desired prompt for a specific outcome from the model without training its weights can be achieved using various strategies such as prompt-engineering [39] and prompt learning [15, 25, 27, 44, 45, 48]. Our work utilizes a Pseudo-Token optimization method for learning the optimal prompt embeddings OOD samples.

3 Method

SAM, the promptable image segmentation model, is built to be efficient and practical for real-world use. To support flexibility in prompts, quick mask computation, and ambiguity awareness, SAM is designed with three components.

First, a robust image encoder E_s computes an image embedding for an input image I. Second, a prompt encoder E_M embeds prompts for use in the segmentation process. Lastly, a lightweight mask decoder D_s predicts segmentation masks based on the combined information from the image and prompt encoders.

SAM's design allows for the reuse of the same image embedding with different prompts, thereby achieving efficient computation. This separation of components is crucial to enable SAM to support a wide range of prompts and perform computation in real-time.

Since SAM is trained on over 1 billion masks from 11 million natural images, its performance on medical imaging datasets may not be optimal. We present an end-to-end approach to improve segmentation mask accuracy in this domain, without fine-tuning the pretrained SAM network, as presented in Fig. 2.

The SAM network S produces an output segmentation mask M_z by taking the input image I and the prompts' embedding Z:

$$M_z = S(I,Z), \tag{1}$$

The prompts embedding Z can be any representation of different prompts, such as masks, boxes, and points.

Instead of using the original prompts encoder, we introduce a prompts generator network, denoted as g, that generates guidance prompts Z_I for SAM given an input image I. g is the only network trained by our method.

This prompts generator network g takes as input the image I and generates prompts $Z_I = g(I)$ for SAM to improve its segmentation mask output.

While training our method, the SAM network S propagates gradients to the prompts generator network g from two segmentation losses that we employ: the binary cross-entropy loss (BCE) and the Dice loss. The BCE loss is given by the negative log-likelihood of the ground truth mask M and the SAM output $S(I,Z_I)$, while the Dice loss measures the overlap between the predicted and ground truth masks. Formally, the losses are expressed as:

$$L_{seg}(I) = L_{BCE}(I,Z_I,M) + L_{dice}(I,Z_I,M), \tag{2}$$

where the BCE loss is defined as:

$$L_{BCE}(I,Z,M) = -M * log(S(I,Z)) - (1-M) * log(1-S(I,Z)). \tag{3}$$

The Dice loss is defined as:

$$\mathcal{L}_{dice}(I,Z,M) = 1 - \frac{2TP(S(I,Z),M)+1}{2TP(S(I,Z),M)+FN(S(I,Z),M)+FP(S(I,Z),M)+1}, \tag{4}$$

where TP, FN, and FP denote the true positive, false negative, and false positive, respectively, between the ground truth mask M and the output mask $S(I,Z)$. To simplify the implementation, we do not use weighting for the loss terms.

Architecture The proposed architecture for g employs the Harmonic Dense Net [5] as its backbone. This network comprises six "HarD" blocks, each with output channels of 192, 256, 320, 480, 720, and 1280, respectively. We initialize the network with pretrained ImageNet weights.

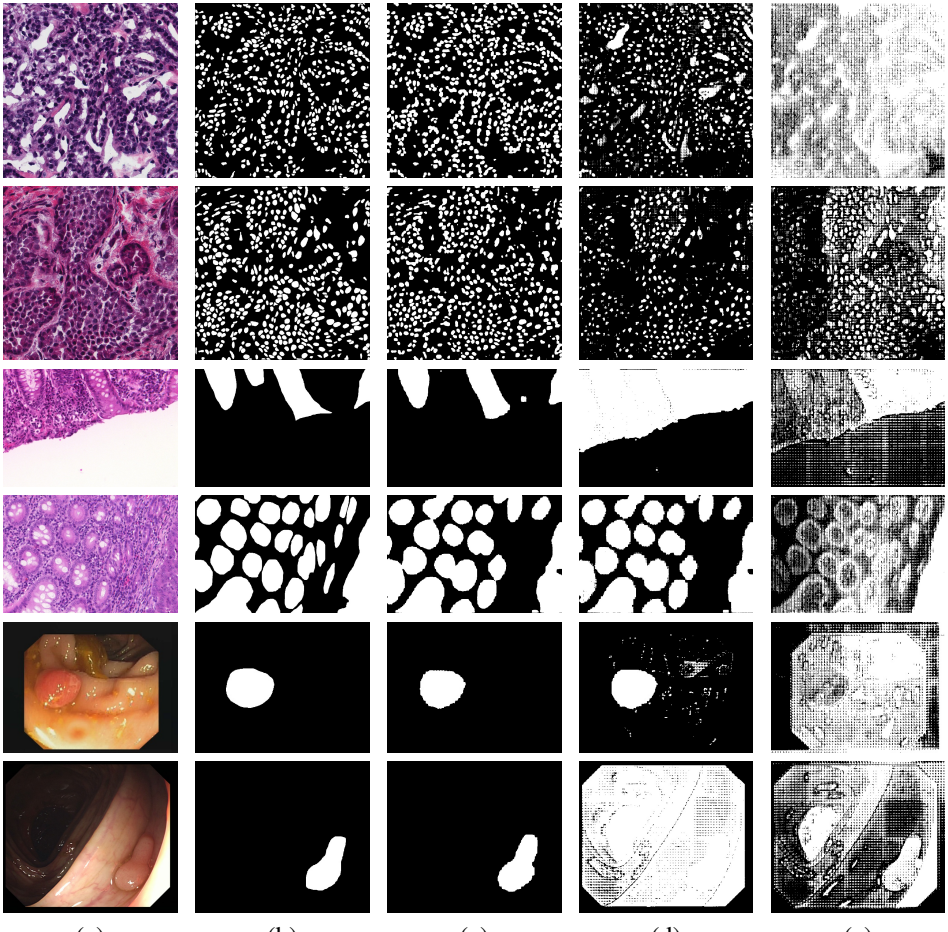
The decoder of g includes two upsampling blocks that produce a resolution of 64×64 with 256 output channels. Each block consists of two convolutional layers with a kernel size of 3 and zero padding of one. Additionally, we apply batch normalization after the last convolution layer and before the activation function. The activation function of the first layer is ReLU, while the second layer uses *tanh*. Each layer receives a skip connection from the encoder block with the same spatial resolution. Notably, our decoder requires significantly fewer learnable parameters than a regular decoder, and fewer skip connections are used in the encoder since only two blocks are employed there.

In terms of FLOPs, our model uses 25.11 GMACs for an image size of 256^2, whereas SAM uses 2733.31 GMACs for an image size of 1024^2 (fixed size of the ViT) only for the image encoder. The peak memory consumption is 371MB for our model and 6006MB for SAM image encoder. Therefore, the overhead of our encoder is almost negligible. The number of parameters of g is 41.56M while SAM ViT has 637M.

A surrogate decoder for $g(I)$ To gain insight into the information provided by the encoder we train, we decode $g(I)$ as a mask. For this purpose, we learn a mapping h from the space of encoded images $g(I)$ to the corresponding ground truth mask M.

This surrogate decoder h minimizes a segmentation loss very similar to Eq. 2, except that it compares $h(g(I))$ with M, for a fixed g. The architecture of h comprises two deconvolution layers that produce a map with a resolution of 256×256, making it a lightweight alternative to SAM.

As it turns out, despite its size, $h(g(I))$ is often a reasonable segmentation mask, see Sec. 4. However, it is not as powerful as AutoSAM, which applies SAM to $g(I)$.



(a) (b) (c) (d) (e)

Figure 3: Sample results of the proposed method on the Nucleus challenges (MoNuSeg) - rows 1,2. The gland segmentation dataset (Glas) rows 3,4. The Kvasir polyp segmentation dataset rows 5,6 where (a) Input image. (b) Ground truth segmentation. (c) The final segmentation map M_z. (d) output of SAM with our mask as input to the mask prompt encoder. (e) output of SAM with the ground truth mask as input to the same prompt encoder.

4 Experiments

In this study, we evaluate our proposed method on multiple medical datasets. We compare our results with state-of-the-art methods and present a number of exloratory results.

Datasets The MoNuSeg dataset [24] comprises 30 microscopic images from seven organs in the training set, with annotations of 21,623 individual nuclei, and 14 similar images in the test set. To be consistent with previous work, we resize the images to 512×512[47] and employ an encoder-decoder architecture based on the HarDNet-85[5] backbone.

The Gland segmentation (GlaS) challenge [41] comprises 85 images for training and 80 for testing, with all images resized to 224×224 following [29].

We also evaluated our algorithm on four Polyp datasets: Kvasir-SEG [19], ClinicDB [3], ColonDB [43], and ETIS [40], following [12]. We split the data into a training set of 1448 images, comprising 900 images from ClinicDB and 548 from Kvasir, and a test set comprising 100 images from ClinicDB, 64 from Kvasir, 196 from ETIS, and 380 from ColonDB.

Lastly, our method was tested on the SUN-SEG Video-Polyp-Segmentation database, based on [22, 30]. The colonoscopy videos are from Showa University and Nagoya University database (also named SUN-database) [30]. The initial classification information and bounding box annotations are provided by three research assistants and examined by two expert endoscopists with professional domain knowledge. The SUN dataset is then extended by Ji et al. [22] to have various annotations such as object masks, boundaries, scribbles, and polygons. The original SUN database has 113 colonoscopy videos, including 100 positive cases with 49,136 polyp frames and 13 negative cases with 109,554 non-polyp frames. in their work Ji et al. [22] manually trim them into 378 positive and 728 negative clips while maintaining their consecutive intrinsic relationship. Such data preprocessing ensures that each clip has around 3-11s duration at a real-time frame rate (i.e., 30 fps), promoting the fault-tolerant margin for various algorithms and devices. Overall, the SUN-SEG database contains 1,106 short video clips with 158,690 video frames total. Although being a video-segmentation task, we have chosen to use our architecture without any modification, using a single frame at a time as the input without relying on temporal data whatsoever. This image-based architecture achieved SOTA performance in almost every metric, competing with video-based methods as shown in table 3.

Training details During the training of our network, we employ the ADAM optimizer with an initial learning rate of 0.0003, and a weight decay regularization parameter set to $1 \cdot 10^{-5}$. A batch size of 10 is utilized, and we conduct training on NVIDIA A6000 with 48GB GPU RAM. The maximum number of epochs for network training was set to 200. The SAM pre-trained weights that we utilized were based on the ViT 'huge' architecture. SAM received an input image size of 1024×1024 as per the original algorithm.

To ensure fairness in comparison with the state-of-the-art method 3P-SEG [37], we employed identical data augmentations during training. For the GlaS dataset, we applied a set of augmentations that included: (i) color jitter with the parameters of brightness sampled uniformly between $[0, 0.2]$, contrast in the range $[0, 0.2]$, saturation in the range $[0, 0.2]$, and hue in the range $[0, 0.1]$; (ii) a random horizontal flip; and (iii) a random affine transformation with a translation of 5 and scale of $(0, 0.2)$. For the MoNu dataset, we utilized (i) a random rotation augmentation of ± 20 degrees and a scale range of $[0.75, 1.25]$; (ii) a random horizontal flip with a probability of 0.5; and (iii) random color jitter with a maximal value of

0.4 for brightness, 0.4 for contrast, 0.4 for saturation, and 0.1 for hue.

During the training of the lightweight decoder h, we utilized the ADAM optimizer with an initial learning rate of 0.0003 and set the weight decay regularization parameter to $1 \cdot 10^{-5}$. We trained with a batch size of 24 on an NVIDIA A5000 with 24GB GPU RAM and set the maximum number of iterations for network training to 60.

Evaluation Metrics For evaluating the performance of our network on image-based segmentation tasks, we employed the widely-used evaluation metrics of Mean Intersection-over-Union (IoU) and Dice-Score. Specifically, we computed the IoU by dividing the area of overlap between the ground truth (GT) masks and the network's output mask by the area of union between the two masks. Moreover, we computed the Dice-Score as a measure of the overlap between the two masks, by taking twice the area of overlap and dividing it by the sum of the areas of the two masks. Both metrics were computed after thresholding the network's output mask to obtain a binary mask that separates the foreground and background regions.

As for the video segmentation task, following [22], we use six different metrics for model evaluation between prediction P_s and ground-truth G_s at timestamp s. These metrics are as follows: (a) Dice coefficient (Dice $= \frac{2\times|P_s \cap G_s|}{|P_s \cup G_s|}$). The operators $\cap$, $\cup$, and $|\cdot|$ denote the intersection, union, and the number of pixels in an area, respectively. (b) Pixel-wise sensitivity (Sen $= \frac{|P_s \cap G_s|}{|G_s|}$). (c) F-measure [1]. The harmonic mean of precision and recall, weighted by β, ($F_\beta = \frac{(1+\beta^2)\times \mathrm{Prc} \times \mathrm{Rcl}}{\beta^2 \times (\mathrm{Prc}+\mathrm{Rcl})}$). This metric is widely used in measuring binary masks by combining both precision (Prc $= \frac{|P_s \cap G_s|}{|P_s|}$) and recall (Rcl $= \frac{|P_s \cap G_s|}{|G_s|}$) for more comprehensive evaluation. (d) Weighted F-measure [29] ($F^w_\beta = \frac{(1+\beta^2)\times \mathrm{Prc}^w \times \mathrm{Rcl}^w}{\beta^2 \times (\mathrm{Prc}^w+\mathrm{Rcl}^w)}$). This metric, suggested by [6, 13] amends the "Equal importance flaw" in Dice and F_β, providing more reliable evaluation results. As for β^2, we set this factor of F_β and F^w_β to be 0.3 and 1, respectively,

Method	Monu		GlaS	
	Dice	IoU	Dice	IoU
FCN [3]	28.84	28.71	-	-
U-Net [35]	79.43	65.99	86.05	75.12
U-Net++ [58]	79.49	66.04	87.36	79.03
Res-UNet [53]	79.49	66.07	-	-
Axial Attention [50]	76.83	62.49	-	-
MedT [47]	79.55	66.17	88.85	78.93
FCN-Hardnet85 [5]	79.52	66.06	89.37	82.09
UCTransNet [49]	79.87	66.68	89.84	82.24
3P-SEG [37]	80.30	67.19	91.19	84.34
MedAdaptor-SAM [52] (conditioned on GT points)	80.34	67.33	92.02	85.88
AutoSAM (ours)	**82.43**	**70.17**	**92.82**	**87.08**
Lightweight decoder $h(g(I))$	76.75	62.32	91.51	84.80
SAM w/ GT point prompt	29.65	17.52	61.67	46.40
SAM w/ GT mask as prompt	30.24	18.21	58.46	42.81
SAM w/ AutoSAM output as the mask prompt	58.10	41.26	87.71	79.92

Table 1: MoNu and GlaS results. Our method achieves SOTA results on both datasets. MedAdaptor-SAM requires point input as a prompt. GT=ground truth.

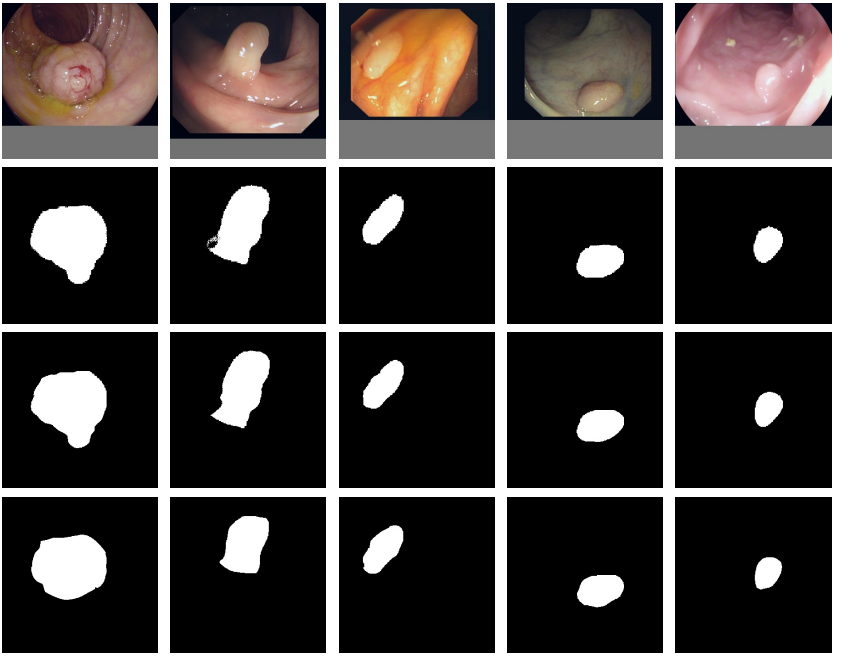


Figure 4: The results of the lightweight decoder h on sample test images. The first row shows the input image I, the second row shows $h(g(I))$, which is the segmentation mask obtained with the surrogate decoder h, the third depicts the results of AutoSAM using the same $g(I)$, and the last row shows the ground-truth segmentation mask M.

following [4, 22]. (e) Structure measure [9] ($\mathcal{S}_\alpha = \alpha \times \mathcal{S}_o(P_s, G_s) + (1-\alpha) \times \mathcal{S}_r(P_s, G_s)$). This metric is used to measure the structural similarity at object-aware $\mathcal{S}_o$ and region-aware $\mathcal{S}_r$, respectively. we set $\alpha = 0.5$. (f) Enhanced-alignment measure, proposed by [10] is a human visual perception-based metric,: $E_\phi = \frac{1}{W \times H} \sum_x^W \sum_y^H \phi(P_s(x,y), G_s(x,y))$, where ϕ is the enhanced-alignment matrix. W and H are the width and height of ground-truth G_s.

Results The MoNu dataset results are reported in Tab. 1. We outperform all baselines, including the latest Axial attention Unet [50], Medical transformer [47], 3P-Seg [37] and MedAdaptorSAM [52] , for both the Dice score and Mean-IoU. Our algorithm also performs better than the fully convolutional segmentation network with the same backbone Hardnet-85 (3P-Seg also uses the same backbone). Sample results for this and other datasets are presented in Fig. 3.

The results for GlaS are also shown in Tab. 1. Our algorithm outperforms the Medical transformer by almost 10% IoU [37], 3P-SEG by almost 3%, and MedAdaptor-SAM by more than a percent, despite the latter utilizing additional information in the form of ground truth points that are placed on the desired objects. Fig. 5 shows a visual comparison between our solution for SAM, with another MedAdaptor-SAM [52].

We also compare our algorithm using different types of prompts with the original prompt encoders of SAM. Tab. 1 shows that our solution for medical prompts improves dramatically the performance of SAM, without any fine-tuning for SAM. Fig. 3 illustrates the gap in the accuracy between our solution and the one that uses the original prompt encoders of SAM.

It is intriguing that SAM encounters difficulties segmenting accurately medical images

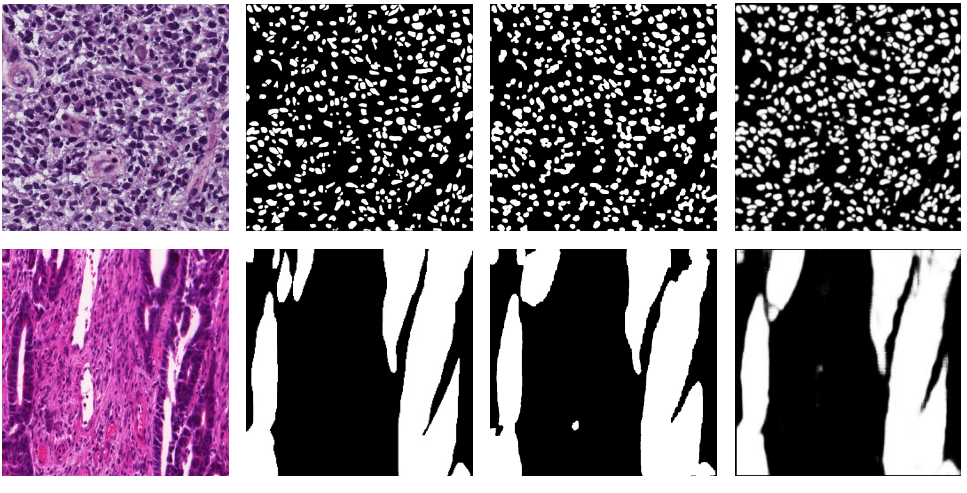
Method	Kvasir33 [19]		Clinic [3]		Colon [43]		ETIS [40]	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
U-Net [33]	81.8	74.6	82.3	75.5	51.2	44.4	39.8	33.5
U-Net++ [58]	82.1	74.3	79.4	72.9	48.3	41.0	40.1	34.4
SFA [12]	72.3	61.1	70.0	60.7	46.9	34.7	29.7	21.7
MSEG [18]	89.7	83.9	90.9	86.4	73.5	66.6	70.0	63.0
DCRNet [54]	88.6	82.5	89.6	84.4	70.4	63.1	55.6	49.6
ACSNet [56]	89.8	83.8	88.2	82.6	71.6	64.9	57.8	50.9
PraNet [17]	89.8	84.0	89.9	84.9	71.2	64.0	62.8	56.7
EU-Net [30]	90.8	85.4	90.2	84.6	75.6	68.1	68.7	60.9
SANet [51]	90.4	84.7	91.6	85.9	75.3	67.0	75.0	65.4
Polyp-PVT [8]	91.7	86.4	93.7	88.9	80.8	72.7	78.7	70.6
FCN-Hardnet85 [6]	90.0	84.9	92.0	86.9	77.3	70.2	76.9	69.5
3P-SEG [37]	**91.8**	86.5	**93.8**	89.0	80.9	73.4	79.1	71.4
Lightweight decoder $h(g(I))$	86.5	79.6	88.5	82.0	80.7	72.4	71.5	63.0
AutoSAM (ours)	91.0	**87.0**	92.8	**89.3**	**83.0**	**76.7**	**79.7**	**74.0**

Table 2: Polyp Segmentation benchmarks results

despite the availability of various prompts, including those based on ground truth. Nevertheless, as our method demonstrates, SAM is capable of delivering state-of-the-art segmentation outcomes without altering the core encoder and decoder modules for the learned prompts.

This phenomenon may have originated from two possible causes. One potential factor is the precision at which SAM incorporates the information encoded by g. Alternatively, a latent signal, analogous to adversarial noise, could be present, which alters the classification of the image without causing significant changes in its appearance.

The results for the Polyp datasets are listed in Tab. 2. In terms of IOU metric, our method outperforms the state-of-the-art on this benchmark 3P-SEG [37] and Polyp-PVT [8]. For all the four dataset Kvasir-SEG and ClinicDB, ColonDB, and ETIS our algorithm achieved



(a) (b) (c) (d)

Figure 5: A visual comparison of our solution to MedAdapterSAM [52] for Glas and Monu datasets, where (a) input image (b) ground-truth mask (c) our solution (d) MedAdapterSAM [52] output.

	Method	SUN-SEG-Easy						SUN-SEG-Hard					
		$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	F_β^{mn}	Dice	Sen	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	F_β^{mn}	Dice	Sen
Image-based	UNet [65]	0.669	0.677	0.459	0.528	0.530	0.420	0.670	0.679	0.457	0.527	0.542	0.429
	UNet++ [69]	0.684	0.687	0.491	0.553	0.559	0.457	0.685	0.697	0.480	0.544	0.554	0.467
	ACSNet [56]	0.782	0.779	0.642	0.688	0.713	0.601	0.783	0.787	0.636	0.684	0.708	0.618
	PraNet [11]	0.733	0.753	0.572	0.632	0.621	0.524	0.717	0.735	0.544	0.607	0.598	0.512
	SANet [61]	0.720	0.745	0.566	0.634	0.649	0.521	0.706	0.743	0.526	0.580	0.598	0.505
	AutoSAM(ours)	**0.815**	**0.855**	**0.716**	**0.774**	0.753	**0.672**	**0.822**	**0.866**	**0.714**	**0.764**	**0.759**	**0.726**
Video-based	COSNet [28]	0.654	0.600	0.431	0.496	0.596	0.359	0.670	0.627	0.443	0.506	0.606	0.380
	MAT [67]	0.770	0.737	0.575	0.641	0.710	0.542	0.785	0.755	0.578	0.645	0.712	0.579
	PCSA [16]	0.680	0.660	0.451	0.519	0.592	0.398	0.682	0.660	0.442	0.510	0.584	0.415
	2/3D [53]	0.786	0.777	0.652	0.708	0.722	0.603	0.786	0.775	0.634	0.688	0.706	0.607
	AMD [26]	0.474	0.533	0.133	0.146	0.266	0.222	0.472	0.527	0.128	0.141	0.252	0.213
	DCF [55]	0.523	0.514	0.270	0.312	0.325	0.340	0.514	0.522	0.263	0.303	0.317	0.364
	FSNet [21]	0.725	0.695	0.551	0.630	0.702	0.493	0.724	0.694	0.541	0.611	0.699	0.491
	PNSNet [20]	0.767	0.744	0.616	0.664	0.676	0.574	0.767	0.755	0.609	0.656	0.675	0.579
	VPS+ [22]	0.806	0.798	0.676	0.730	**0.756**	0.630	0.797	0.793	0.653	0.709	0.737	0.623

Table 3: Quantitative results of two test sub-datasets from the SUN-SEG [22] dataset. Although being image-based, our method competes with video-based approaches, achieving SOTA performance in almost every benchmark. The best values are highlighted in **bold**.

state-of-the-art results with a gap of 0.5, 0.3, 3.3 and 2.6 respectively. With respect to the DICE metric, our method outperforms other methods in two out of four datasets.

The results for the SUN-SEG video dataset are listed in Tab. 3. For the SUN-SEG-Hard (unseen) dataset our method outperforms the state-of-the-art [22] on every metric tested, i.e. $\mathcal{S}_\alpha$, E_ϕ^{mean} , F_β^w , F_β^{mean} , Dice & Sen with a margin of 2.5, 7.3, 6.1, 5.5, 2.2, 10.3 respectively. For the SUN-SEG-Easy (unseen) dataset our method outperforms State-Of-The-Art methods in every metric except for the Dice-Score, which achieves 0.3 below VPS+ [22]. Note that we are using an image-based method and outperforming the video-based methods that significantly outperform any other image-based method.

Finally, we measure the performances of the lightweight decoder h for all the medical image datasets. As can be seen in Tab. 1 and Tab. 2, $h(g(I))$ achieves a reasonable mask, although not as good as the output of SAM with g prompts. A visual comparison of $h(g(I))$ and AutoSAM on the same $g(I)$ is shown in Fig. 4.

5 Conclusions

SAM is a powerful segmentation model for natural images. It has the potential to become a prominent foundation model, i.e., be effective for downstream tasks such as medical image analysis. We show that this may only require "the right guidance" in the form of a dedicated conditioning signal that is provided by an auxiliary network g that replaces the prompt embedding. As no prompt is required, our method turns SAM into a fully automatic method. In future work, we plan to learn one g network for multiple medical imaging domains. It would be interesting to learn how well this "universal-AutoSAM" generalizes to new tasks without further training.

Acknowledgment This research of RG was supported by ERC-StG SPADE grant no. 757497

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, Miami, FL, USA, 2009. IEEE.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12), 2017.

[3] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43, 2015.

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.

[5] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *ICCV*, 2019.

[6] Ming-Ming Cheng and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129(9):2622–2638, 2021.

[7] Sindhu Devunooru, Abeer Alsadoon, PWC Chandana, and Azam Beg. Deep learning neural networks for medical image segmentation of brain tumours for diagnosis: a recent review and taxonomy. *Journal of Ambient Intelligence and Humanized Computing*, 12:455–483, 2021.

[8] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv*, 2021.

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, Venice, Italy, 2017. IEEE.

[10] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, Stockholm, Sweden, 2018. IJCAI.

[11] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. pages 263–273. Springer, 2020.

[12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*. Springer, 2020.

[13] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 6:6, 2021.

[14] Yuqi Fang, Cheng Chen, et al. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI*, 2019.

[15] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

[16] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, volume 34, pages 10869–10876. AAAI Press, 2020.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[18] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. Hardnet-mseg: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *arXiv*, 2021.

[19] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MMM*. Springer, 2020.

[20] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. pages 142–152. Springer, 2021.

[21] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *ICCV*, pages 4922–4933. IEEE, 2021.

[22] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[24] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, et al. A multi-organ nucleus segmentation challenge. *TMI*, 2019.

[25] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*, 2021.

[26] Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. In *NeurIPS*, [Online], 2021. Curran Associates, Inc.

[27] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021.

[28] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, pages 3623–3632. IEEE, 2019.

[29] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255. IEEE, 2014.

[30] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy*, 93(4): 960–967, 2021.

[31] Berk Norman, Valentina Pedoia, and Sharmila Majumdar. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. *Radiology*, 288(1):177–185, 2018.

[32] Krushi Patel, Andrés M Bur, and Guanghui Wang. Enhanced u-net: A feature enhancement network for polyp segmentation. In *CRV*. IEEE, 2021.

[33] Juana González-Bueno Puyal, Kanwal K Bhatia, Patrick Brandao, Omer F Ahmad, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney, and Danail Stoyanov. Endoscopic polyp segmentation using a hybrid 2d/3d cnn. pages 295–305. Springer, 2020.

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[37] Tal Shaharabany and Lior Wolf. End-to-end segmentation of medical images via patch-wise polygons prediction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 308–318. Springer, 2022.

[38] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India*, 35(1): 3, 2010.

[39] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *EMNLP*, 2020.

[40] Juan Silva, Aymeric Histace, Olivier Romain, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *IJCARS*, 2014.

[41] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35, 2017.

[42] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[43] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *TMI*, 35(2), 2015.

[44] Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. Zero-shot video captioning with evolving pseudo-tokens, 2022.

[45] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[47] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv*, 2021.

[48] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. *EMNLP*, 2019.

[49] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *arXiv*, 2021.

[50] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*. Springer, 2020.

[51] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *MICCAI*. Springer, 2021.

[52] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[53] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *ITME*. IEEE, 2018.

[54] Zijin Yin, Kongming Liang, Zhanyu Ma, and Jun Guo. Duplex contextual relation network for polyp segmentation. *arXiv*, 2021.

[55] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, pages 1553–1563. IEEE, 2021.

[56] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. Adaptive context selection for polyp segmentation. In *MICCAI*. Springer, 2020.

[57] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE TIP*, 29:8326–8338, 2020.

[58] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *DLMIA*. Springer, 2018.

[59] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. 39(6):1856–1867, 2019.