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

Yu Jiening
Yu Jiening@umac.mo

um 澳大

Cornell University

We gratefully acknowledg

arXiv > cs > arXiv:2306.06370

Search...

Help | Ac

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 10 Jun 2023]

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

Tal Shaharabany, Aviad Dahan, Raja Giryes, Lior Wolf

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.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**
Cite as: arXiv:2306.06370 **[cs.CV]**
(or arXiv:2306.06370v1 **[cs.CV]** for this version)
https://doi.org/10.48550/arXiv.2306.06370

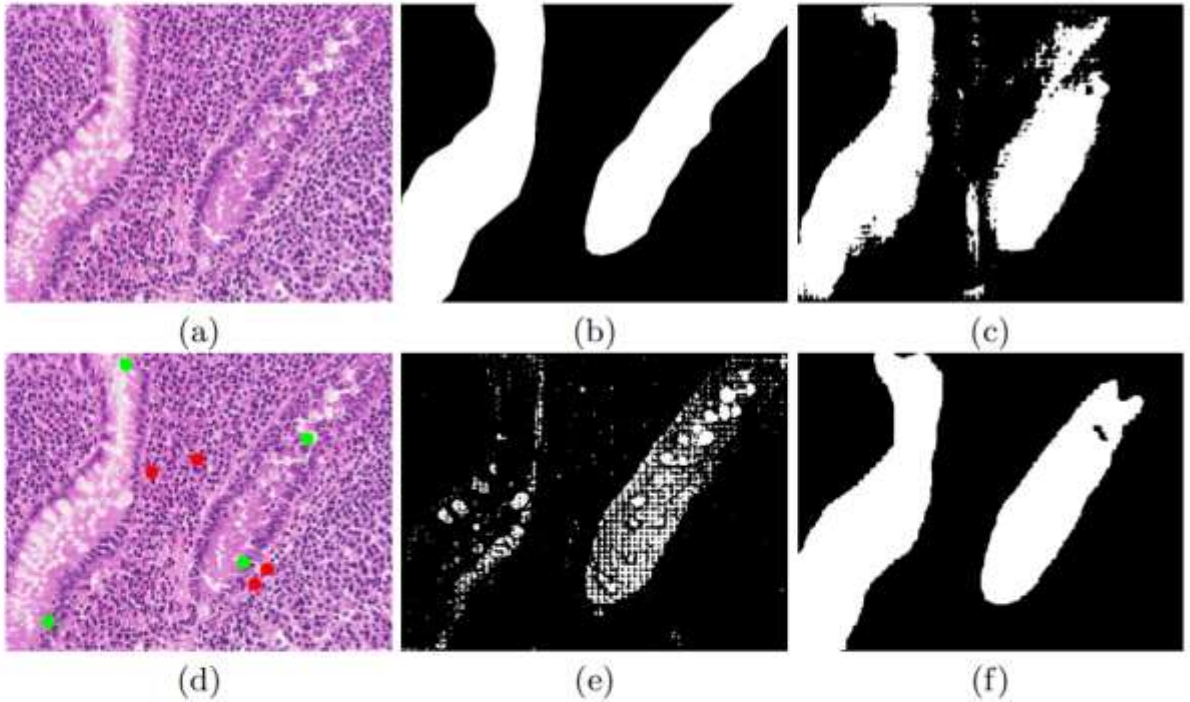
Submission history

From: Tal Shaharabany [view email]
[v1] Sat, 10 Jun 2023 07:27:00 UTC (18,369 KB)

- arXiv preprint arXiv:2306.06370, 2023.

Catalogue

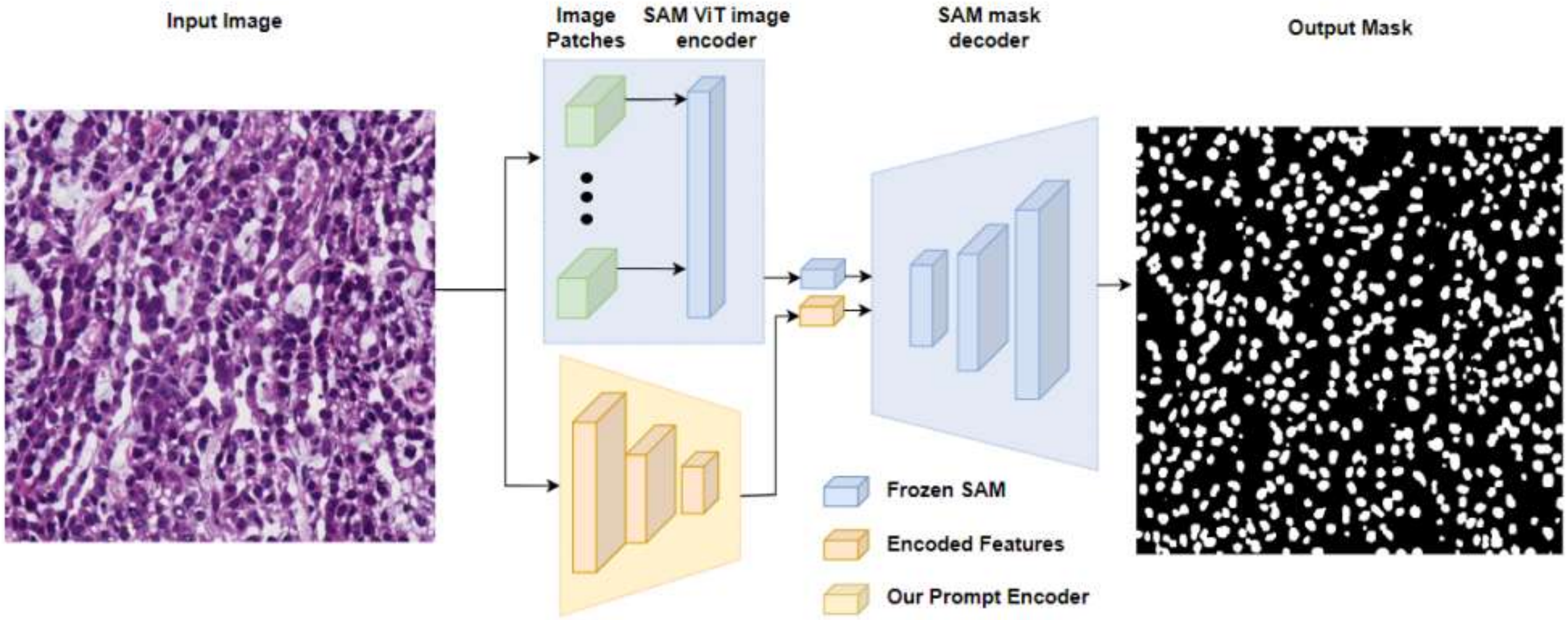
Introduction



- SAM’s performance may not be optimal on medical imaging datasets due to its pretraining on natural images

- 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.

Method


Input Image
Image Patches
SAM ViT image encoder
SAM mask decoder
Output Mask
Frozen SAM
Encoded Features
Our Prompt Encoder

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

$$M_z = S(I, Z),$$

- 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.

- 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.
- The architecture of surrogate decoder h comprises two deconvolution layers that produce a map with a resolution of 256 × 256, making it a lightweight alternative to SAM.

Experiments

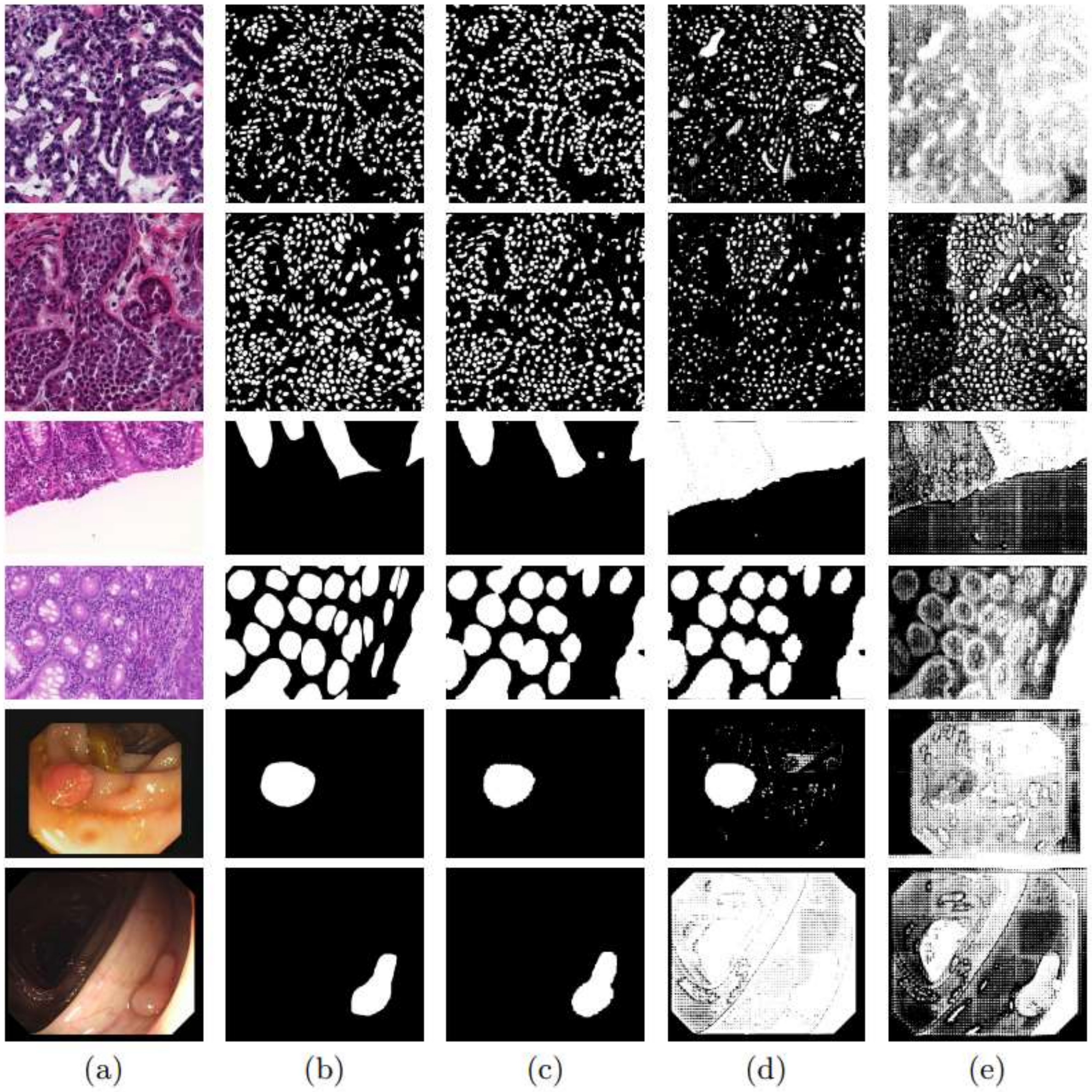
- The MoNuSeg dataset(显微图像) comprises 30 microscopic images from seven organs in the training set, with annotations of 21,623 individual nuclei, and 14 similar images in the test set.
- The Gland segmentation (GlaS) challenge comprises 85 images for training and 80 for testing.
- Four Polyp(息肉) datasets: Kvasir-SEG, ClinicDB, ColonDB, and ETIS
- Tested on the SUN-SEG Video-Polyp-Segmentation database(视频息肉分割)

Training details

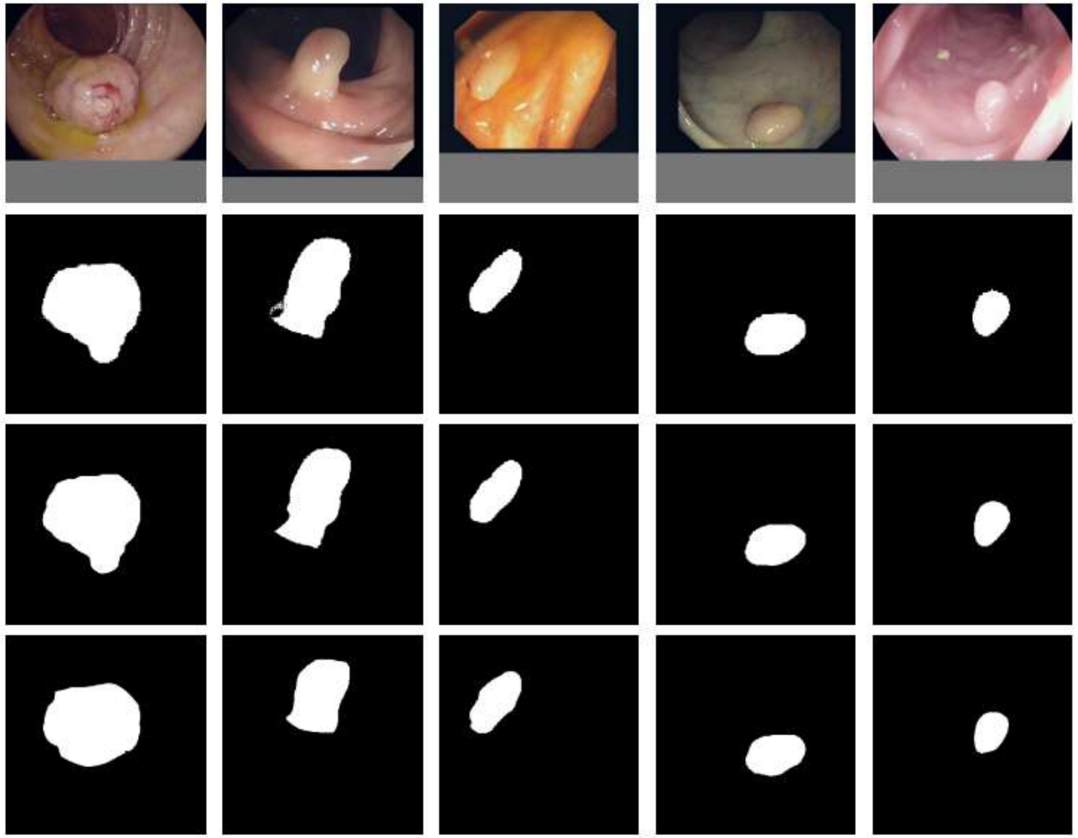
- ADAM optimizer with an initial learning rate of 0.0003 and set the weight decay regularization parameter to $1 \cdot 10^{-5}$
- Batch size:10
- NVIDIA A6000 with 48GB GPU RAM
- Epoch:200
- input image size: 1024×1024

Training of the lightweight decoder h

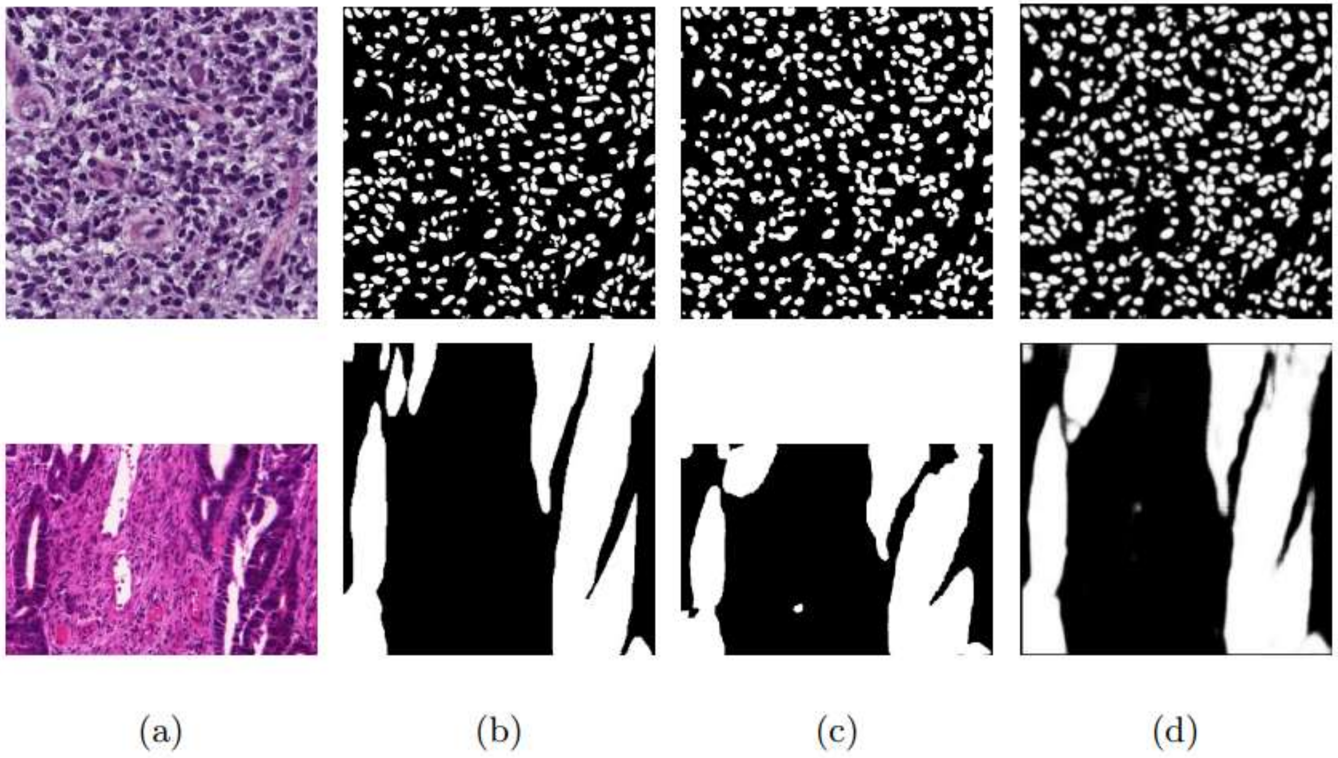
- ADAM optimizer with an initial learning rate of 0.0003 and set the weight decay regularization parameter to $1 \cdot 10^{-5}$
- Batch size:24
- NVIDIA A5000 with 24GB GPU RAM
- set the maximum number of iterations for network training to 60



- 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 Mz.
- (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.



- 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.



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

Method	Monu		GlaS	
	Dice	IoU	Dice	IoU
FCN [2]	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

- MoNu and GlaS results. Our method achieves SOTA results on both datasets. MedAdaptor-SAM requires point input as a prompt.
- Our algorithm outperforms the Medical transformer by almost 10% IoU , 3P-SEG by almost 3%

Method	Kvasir33 [19]		Clinic [3]		Colon [43]		ETIS [40]	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
U-Net [35]	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 [14]	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 [12]	89.8	84.0	89.9	84.9	71.2	64.0	62.8	56.7
EU-Net [32]	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 [5]	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**

- Polyp Segmentation benchmarks results
- For all the four dataset， our algorithm achieved SOTA 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.

	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 [35]	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++ [59]	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 [13]	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 [51]	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 [57]	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 [33]	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

- Quantitative results of two test sub-datasets from the SUN-SEG dataset.

Conclusion

- SAM is a powerful segmentation model for natural images.
- 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.

Thank You!