Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects

Jian Hu[1], Jiayi Lin[1*], Weitong Cai[1], Shaogang Gong[1]*

* Equal contribution; 1 Queen Mary University of London, UK {jian.hu, jiayi.lin, weitong.cai, s.gong}@qmul.ac.uk

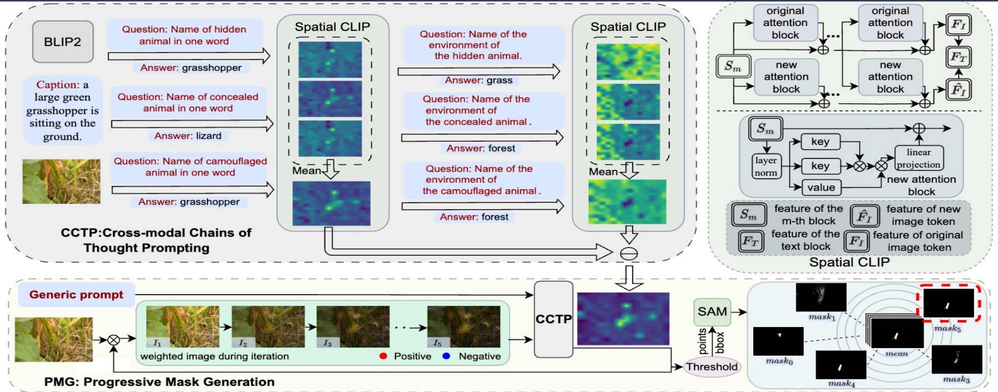
The 38th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Abstract

- Camouflaged object detection (COD) approaches heavily rely on pixel-level annotated datasets. Weakly-supervised COD (WSCOD) approaches use sparse annotations like scribbles or points to reduce annotation efforts, but this can lead to decreased accuracy. The Segment Anything Model (SAM) shows remarkable segmentation ability with sparse prompts like points. But manual prompt is not always feasible, as it may not be accessible in real-world application. And it it only provides localization information instead of semantic one, which can intrinsically cause ambiguity in interpreting targets.
- In this work, we aim to eliminate the need for manual prompt. The key idea is to employ Cross-modal Chains of Thought Prompting (CCTP) to reason visual prompts using the semantic information given by a generic text prompt. To that end, we introduce a test-time instance-wise adaptation mechanism called Generalizable SAM (GenSAM) to automatically generate and optimize visual prompts from the generic task prompt for WSCOD. In particular, CCTP maps a single generic text prompt onto image-specific consensus foreground and background heatmaps, using vision-language models, acquiring reliable visual prompts. Moreover, to test-time adapt the visual prompts, we further propose Progressive Mask Generation (PMG) to iteratively reweight the input image, guiding the model to focus on the targeted region in a coarse-to-fine manner.
- Experiments on three benchmarks demonstrate that GenSAM outperforms point supervision approaches and achieves comparable results to scribble supervision ones, solely relying on general task descriptions.

Methods and Contribution



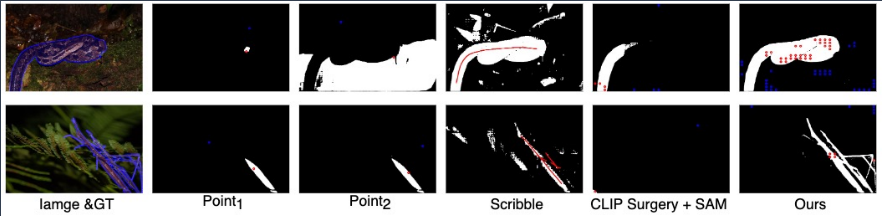
1.Cross-modal Chains of Thought Prompting (CCTP)

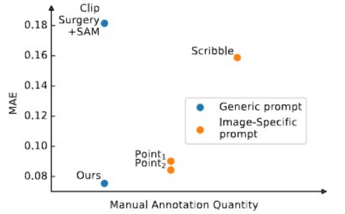
- It takes a generic task prompt as input. BLIP2 generates an image caption for each image with the input generic prompt.
- Based on this prompt and generated caption, three parallel chains of thought are constructed to extract keywords about concealed animals and their corresponding background from unlabelled images.
- These keywords are then fed into our designed spatial CLIP module, which generates heatmaps for locating the camouflaged objects. High-confidence regions selected from these heatmaps serve as prompts to guide the segmentation process.

2. Progressive Mask Generation (PMG)

- The heatmaps generated by CCTP are weighted and utilized as visual prompts in PMG, gradually directing the model's attention towards task-relevant regions.
- In addition, during the adaptation process, the mask generated by a single iteration that is closest to the average mask obtained from multiple iterations is selected as the final output.

Motivation





1. **SAM has limited comprehension on its segmented object.**
- Manual prompts can only provide location information of desired segmentation objects, but lack in semantic information, leading to potential ambiguity.
- Despite prompts in figure above targeting the same object, minor changes in the point prompt's position can make SAM misinterpret the desired object, greatly changing the results.

2. **SAM exhibits a strong bias in interpreting prompts**
- Even with more information from scribble prompts, SAM still misunderstands the target, leading to limited performance.
- SAM still requires instance-specific manual prompts, which may not always be practical in real-world scenarios, and the question of eliminating this need remains unexplored.

Experiments and Example Results

1. Main Results

Results on COD with point supervision and scribble supervision

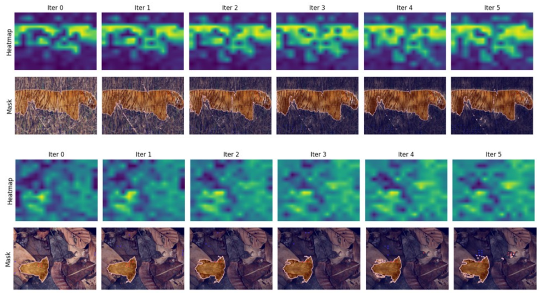
Methods	Venue	CHAMELEON $M\downarrow$	CHAMELEON $F_\beta\uparrow$	CHAMELEON $E_\phi\uparrow$	CHAMELEON $S_\alpha\uparrow$	CAMO $M\downarrow$	CAMO $F_\beta\uparrow$	CAMO $E_\phi\uparrow$	CAMO $S_\alpha\uparrow$	COD10K $M\downarrow$	COD10K $F_\beta\uparrow$	COD10K $E_\phi\uparrow$	COD10K $S_\alpha\uparrow$
Scribble Supervision Setting													
WSSA(Zhang et al. 2020)	CVPR20	0.067	0.692	0.860	0.782	0.118	0.615	0.786	0.696	0.071	0.536	0.770	0.684
SCWS(Yu et al. 2021)	AAAI21	0.053	0.758	0.881	0.792	0.102	0.658	0.795	0.713	0.055	0.602	0.805	0.710
TEL(Zhang et al. 2020)	CVPR22	0.073	0.708	0.827	0.785	0.104	0.681	0.797	0.717	0.057	0.633	0.826	0.724
SCOD(He et al. 2023b)	AAAI23	**0.046**	**0.791**	**0.897**	**0.818**	**0.092**	**0.709**	**0.815**	**0.735**	0.049	0.637	**0.832**	0.733
SAM(Kirillov et al. 2023)	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730
SAM-S(Kirillov et al. 2023)	ICCV23	0.076	0.729	0.820	0.650	0.105	0.682	0.774	0.731	**0.046**	**0.695**	0.828	**0.772**
Point Supervision Setting													
WSSA(Zhang et al. 2020)	CVPR20	0.105	0.660	0.712	0.711	0.148	0.607	0.652	0.649	0.087	0.509	0.733	0.642
SCWS(Yu et al. 2021)	AAAI21	0.097	0.684	0.739	0.714	0.142	0.624	0.672	**0.687**	0.082	0.593	0.777	0.738
TEL(Zhang et al. 2020)	CVPR22	0.094	**0.712**	**0.751**	**0.746**	0.133	**0.662**	0.674	0.645	0.063	0.623	**0.803**	0.727
SCOD(He et al. 2023b)	AAAI23	**0.092**	0.688	0.746	0.725	0.137	0.629	0.688	0.663	**0.060**	0.607	0.802	0.711
SAM(Kirillov et al. 2023)	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730
SAM-P(Kirillov et al. 2023)	ICCV23	0.101	0.696	0.745	0.697	**0.123**	0.649	**0.693**	0.677	0.069	**0.694**	0.796	**0.765**
Task-Generic Prompt Setting													
CLIP Surgery+SAM(Li et al. 2023)	Arxiv2023	0.180	0.557	0.710	0.637	0.206	0.466	0.666	0.573	0.187	0.448	0.672	0.601
GenSAM	Ours	**0.090**	**0.680**	**0.807**	**0.764**	**0.113**	**0.659**	**0.775**	**0.719**	**0.067**	**0.681**	**0.838**	**0.775**

Results on Polyp Image Segmentation and Shadow Detection with generic task prompt.

Datasets	Methods	M↓	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
ETIS (Silva et al. 2014)	CLIP Surgery+SAM	0.537	0.047	0.296	0.272
(Polyp Image Segmentation)	GenSAM	**0.205**	**0.090**	**0.554**	**0.430**
SBU(Vicente et al. 2016)	CLIP Surgery+SAM	0.331	0.336	0.517	0.442
(Shadow Detection)	GenSAM	**0.215**	**0.421**	**0.621**	**0.529**

Highly competitive performance and less supervision:
- Achieve compete performance with only one prompt
- Up to **0.174** boost on COD10K dataset;

2. Visualization



3. Ablation Study

BLIP2 keyword	chain foreground	PMG	kkv self-attention	chain background	CHAMELEON $M\downarrow$	CHAMELEON $F_\beta\uparrow$	CHAMELEON $E_\phi\uparrow$	CHAMELEON $S_\alpha\uparrow$	CAMO $M\downarrow$	CAMO $F_\beta\uparrow$	CAMO $E_\phi\uparrow$	CAMO $S_\alpha\uparrow$	COD10K $M\downarrow$	COD10K $F_\beta\uparrow$	COD10K $E_\phi\uparrow$	COD10K $S_\alpha\uparrow$
					0.180	0.557	0.710	0.637	0.206	0.466	0.666	0.573	0.187	0.448	0.672	0.601
✓					0.106	0.689	0.803	0.749	0.200	0.503	0.676	0.602	0.146	0.556	0.735	0.673
✓	✓				0.094	0.687	0.800	0.754	0.198	0.521	0.687	0.613	0.143	0.569	0.740	0.681
✓	✓	✓			0.098	0.659	0.779	0.741	0.161	0.554	0.719	0.642	0.086	0.616	0.797	0.731
✓	✓	✓	✓		**0.078**	**0.711**	**0.817**	**0.776**	0.147	0.583	0.746	0.666	0.069	0.660	0.820	0.760
✓	✓	✓	✓	✓	0.090	0.680	0.807	0.764	**0.113**	**0.659**	**0.775**	**0.719**	**0.067**	**0.681**	**0.838**	**0.775**

(Method's variant; settings on camouflaged object detection)

Proposed two components bring a significant performance boost :
- Up to **0.09** MAE improvements.

Proposed k-k-v self attention substantially helps the text mapping and improves the model performance with a very limited computational costs.