AoP-SAM: Automation of Prompts for Efficient Segmentation

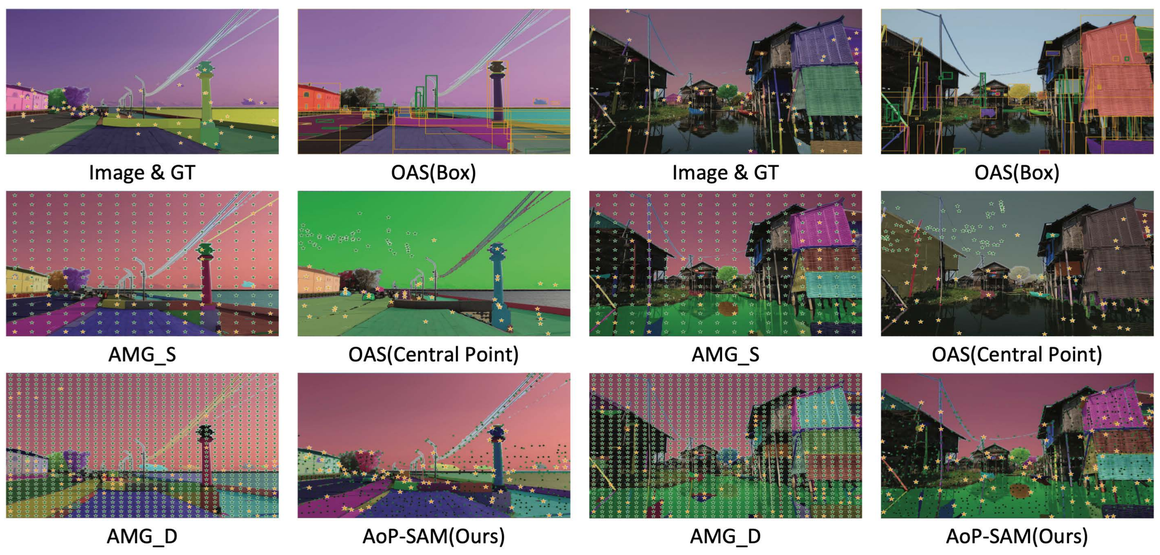
Yi Chen, Muyoung Son, Chuanbo Hua, Joo-Young Kim

Korea Advanced Institute of Science and Technology (KAIST)

Summary

TL; DR. We propose AoP-SAM, a novel approach automatically generate essential prompts for accurate segmentation, eliminating the need for manual prompt provision.

Motivation



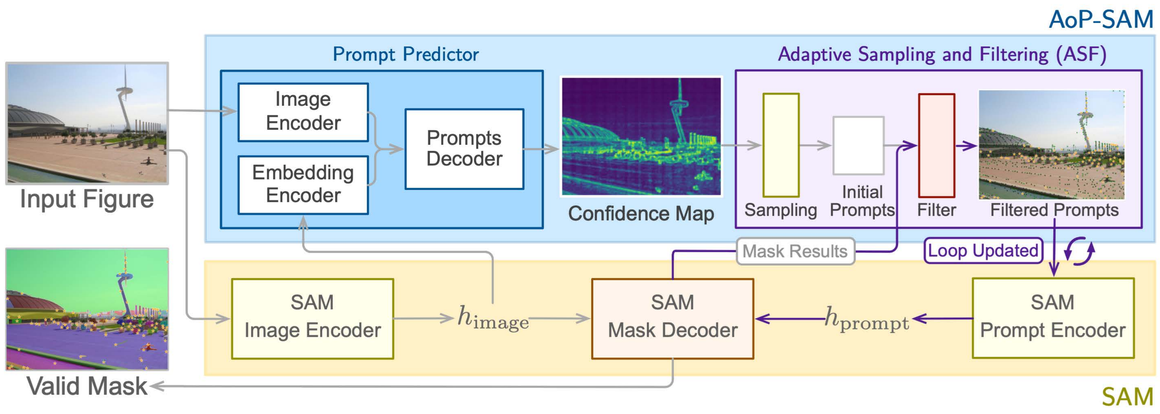
Automating SAM's prompt provision eliminates manual input needs, enhancing mask segmentation efficiency. However, existing approaches face limitations:

- Grid-based prompts (AMG) lead to excessive mask refinements;
- Using extra CV models (OAS) increases computational overhead;
- Both result in increased latency and reduced efficiency.

We addresses these challenges by efficiently generating essential prompts for accurate mask generation without human intervention:

- Orange labels (stars/boxes): Prompts generating valid masks;
- Green labels: Prompts generating invalid masks;
- Black stars: Filtered prompts processed by AoP-SAM.

AoP-SAM



Prompt Predictor Utilizes a dual-encoder architecture (CNN + ViT) to process both original image and SAM's embeddings. Processes inputs through CNN layers with ReLU activation and generates a Prompt Confidence Map (PCM) using Sigmoid activation, highlighting the optimal regions for the following prompt placements.

ASF Coarse Processing Applies Gaussian filtering to the PCM to reduce noise and identify local maxima. These maxima, identified as initial prompt candidates, are then mapped back to the original image coordinates to ensure precise placement of potential prompts.

ASF Fine Filtering Creates a Prompt Elimination Map (PEM) using cosine similarity between image features and generated reference masks. Applies adaptive threshold to remove redundant prompts in PCM, ensuring only essential ones are retained for final mask generation.

Training Leverages subsets of the SA-1B dataset, containing about 200K masks and prompts. Uses point prompts and their corresponding masks as ground truth, with MSELoss and Adam optimization employed over 1000 epochs. This approach maintains SAM's robust generalization capabilities while adding efficient prompt generation.

Experiments

Experiment Setup We evaluate AoP-SAM on *SA-1B*, *COCO*, and *LVIS* datasets, comparing against baseline methods including *AMG-S* (Sparse grid), *AMG-D* (Dense grid), and *OAS* (using YOLOv8). Performance is measured through mean IoU (mIoU) using greedy IoU algorithm, along with Prompt Inference Latency (Inf_{Lat}) and Peak Memory (Peak_{Mem}) for efficiency metrics.

Automating Prompts Methods	SA-1B mIoU ↑	SA-1B Inf_{Lat} ↓	SA-1B Peak_{Mem} ↓	SA-1B #P	COCO mIoU ↑	COCO Inf_{Lat} ↓	COCO Peak_{Mem} ↓	COCO #P	LVIS mIoU ↑	LVIS Inf_{Lat} ↓	LVIS Peak_{Mem} ↓	LVIS #P
MobileSAM Image Encoder												
AMG-S [13]	29.8	-	4.5	38.6	56.0	-	**1.9**	33.5	56.2	-	**1.9**	33.4
AMG-D [13]	46.9	-	9.1	71.0	60.9	-	**1.9**	55.9	61.1	-	**1.9**	55.5
OAS(Box) [23]	50.7	0.191	7.3	100	55.5	0.187	4.2	44	55.7	0.188	4.0	38
OAS(Central Point) [23]	48.7	0.188	7.7	141.0	53.9	0.167	4.3	69.0	54.5	0.164	4.3	68.1
AoP-SAM	**51.4**	**0.101**	**4.1**	71.7	**61.5**	**0.096**	2.1	58.1	**62.3**	**0.094**	2.1	57.5
ViT-L Image Encoder												
AMG-S [13]	40.0	-	5.7	55.5	61.4	-	4.4	48.8	63.2	-	**4.3**	49.5
AMG-D [13]	65.6	-	10.3	108.9	67.7	-	**4.3**	86.0	69.2	-	**4.3**	86.5
OAS(Box) [23]	65.8	0.150	9.1	100	63.3	0.152	5.4	44	62.9	0.151	5.3	38
OAS(Central Point) [23]	67.6	0.149	9.7	199.3	64.2	0.133	5.5	98.4	63.5	0.132	5.5	98.9
AoP-SAM	**71.1**	**0.120**	**5.4**	118.3	**68.4**	**0.116**	4.4	97.0	**69.8**	**0.117**	4.4	97.2
ViT-H Image Encoder												
AMG-S [13]	40.8	-	7.1	56.3	63.3	-	5.7	49.8	64.9	-	5.6	50.5
AMG-D [13]	66.8	-	11.8	109.6	69.5	-	5.7	87.4	71.0	-	5.6	88.0
OAS(Box) [23]	66.9	0.160	10.4	100	64.1	0.152	6.8	44	63.3	0.153	6.6	38
OAS(Central Point) [23]	68.3	0.154	11.1	207.6	65.1	0.134	6.9	102.1	63.0	0.134	6.8	102.4
AoP-SAM	**70.6**	**0.122**	**6.6**	107.8	**70.1**	**0.120**	**5.5**	90.0	**71.9**	**0.122**	**5.5**	89.7

Performance Hightlights

- Achieves highest mIoU scores across all datasets and encoders, outperforming methods using grid-based and object detection models.
- Demonstrates excellent computational efficiency with fast inference (0.122s latency) and low memory usage (6.6MB peak).
- Surpasses both baseline methods (AMG-S, AMG-D) and advanced approaches (OAS), achieving better balance between segmentation accuracy and prompt generation efficiency.
- Successfully maintains high performance while keeping resource usage within practical limits, suitable for real-world applications.

Method's variants	SA-1B mIoU ↑	SA-1B Inf_{Lat} ↓	SA-1B Peak_{Mem} ↓	SA-1B #P	COCO mIoU ↑	COCO Inf_{Lat} ↓	COCO Peak_{Mem} ↓	COCO #P	LVIS mIoU ↑	LVIS Inf_{Lat} ↓	LVIS Peak_{Mem} ↓	LVIS #P
AoP-SAM w/o AS, AF	57.2	0.059	7.2	106.4	67.9	0.078	5.7	70.4	60.9	0.075	5.7	60.6
AoP-SAM w/o AF	72.8	0.130	10.1	120.1	70.5	0.122	5.7	97.9	71.7	0.121	5.7	97.5
AoP-SAM	71.3	0.122	6.6	107.8	70.1	0.112	5.7	91.1	71.9	0.122	5.7	89.7

Ablation Study Component analysis shows each module's crucial role: Prompt Predictor provides the foundation and Adaptive Sampling (AS) significantly boosts mIoU; the combination with Adaptive Filtering (AF) achieves optimal performance by efficiently removing redundant prompts.

(a) Sampling Smoothing Factor

Factor	mIoU ↑	Inf_{Lat} ↓	Peak_{Mem} ↓	#P
1	**72.4**	0.124	51.5	114.4
2	70.4	0.122	42.2	107.8
3	67.3	**0.118**	32.7	100.3
4	63.4	0.122	**24.6**	91.2

(b) Confidence Intensity Threshold

Thr.	mIoU ↑	Inf_{Lat} ↓	Peak_{Mem} ↓	#P
0.1	**70.9**	0.121	10.1	109.2
0.2	70.4	0.122	9.75	107.8
0.3	68.7	**0.116**	**9.52**	103.9
0.4	66.4	0.117	9.60	99.5

(c) Prompt Spacing Factor

Factor	mIoU ↑	Inf_{Lat} ↓	Peak_{Mem} ↓	#P
4	**72.7**	0.123	10.0	114.8
5	71.6	0.123	9.88	111.4
6	70.4	0.122	**9.75**	107.8
7	68.9	**0.117**	9.82	104.2

(d) Prompt Elimination Threshold

Thr.	mIoU ↑	Mask_{Lat} ↓	$\text{Ratio}_{\text{elim.}}$ ↑	#P
1.25	68.4	**0.671**	**51.5**	100.9
1.3	70.4	0.799	42.3	107.8
1.35	71.6	0.930	32.7	113.2
1.4	**72.2**	1.041	24.6	116.4

Parameter Analysis

- *Sampling Smoothing Factor* impacts the coverage area of Gaussian filtering - larger factors provide stronger smoothing and reduce memory usage during PCM generation.
- *Confidence Intensity Threshold* and *Prompt Spacing Factor* optimize point prompt generation from PCM, ensuring essential and accurate point selection for critical areas.
- *Prompt Elimination Threshold* controls the balance between efficiency and accuracy - lower thresholds increase prompt removal ratio for faster mask generation with minimal accuracy trade-off.

Conclusion

We propose AoP-SAM, a novel approach designed to efficiently generate essential prompts for accurate mask generation in SAM. Our method features a lightweight Prompt Predictor, trained to predict optimal prompt locations, and a test-time ASF mechanism for automatic prompt generation. Evaluated on three segmentation datasets with three SAM-based models, AoP-SAM improves both accuracy and efficiency, making it ideal for automated segmentation tasks with SAM.

Paper