VRP-SAM: SAM with Visual Reference Prompt

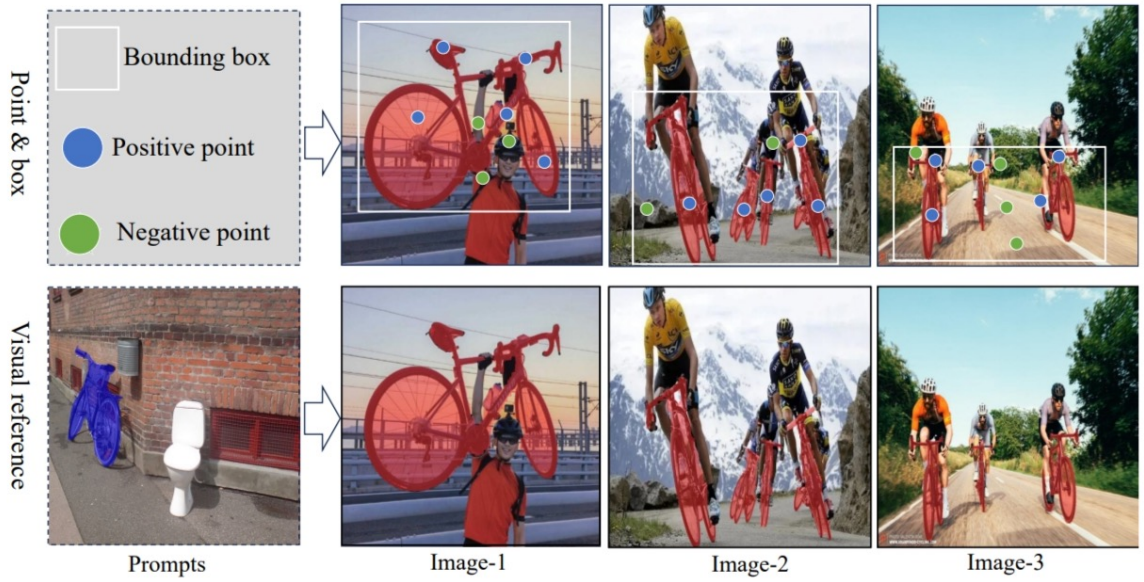
Yanpeng Sun, Jiahui Chen

CVPR 2024

Introduction

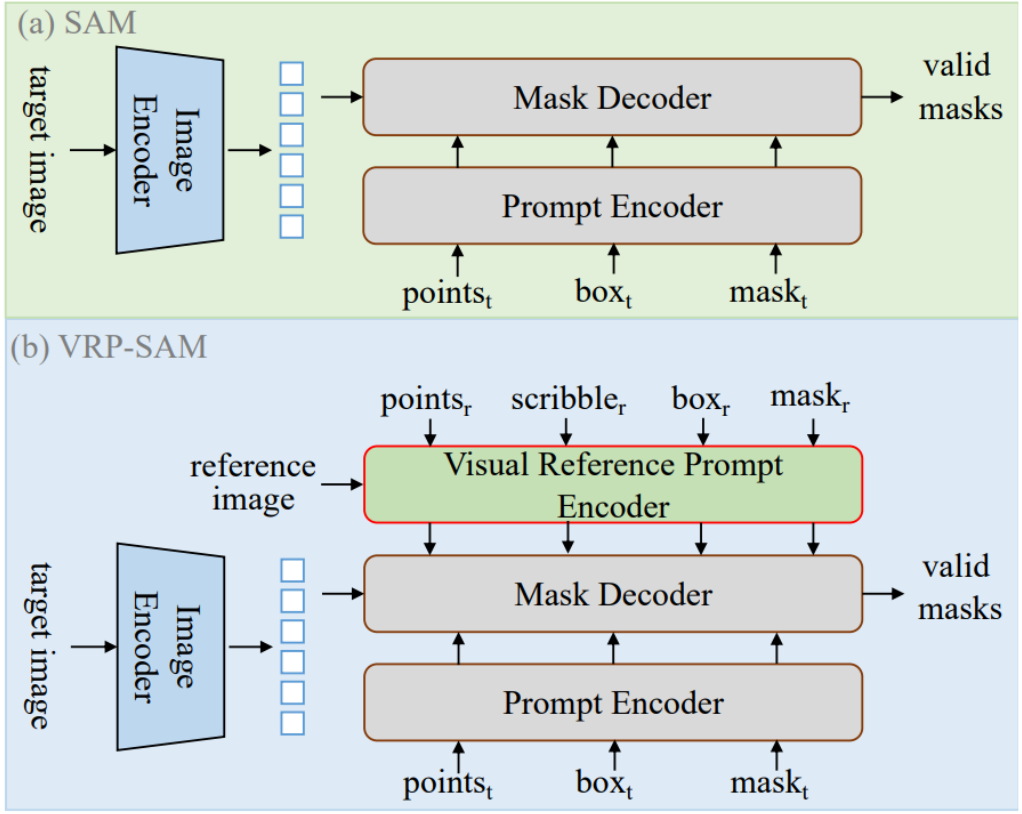
- Motivation
  - The existing prompt formats of SAM present significant challenges in practical applications, especially when dealing with complex scenes and numerous images

- Problems



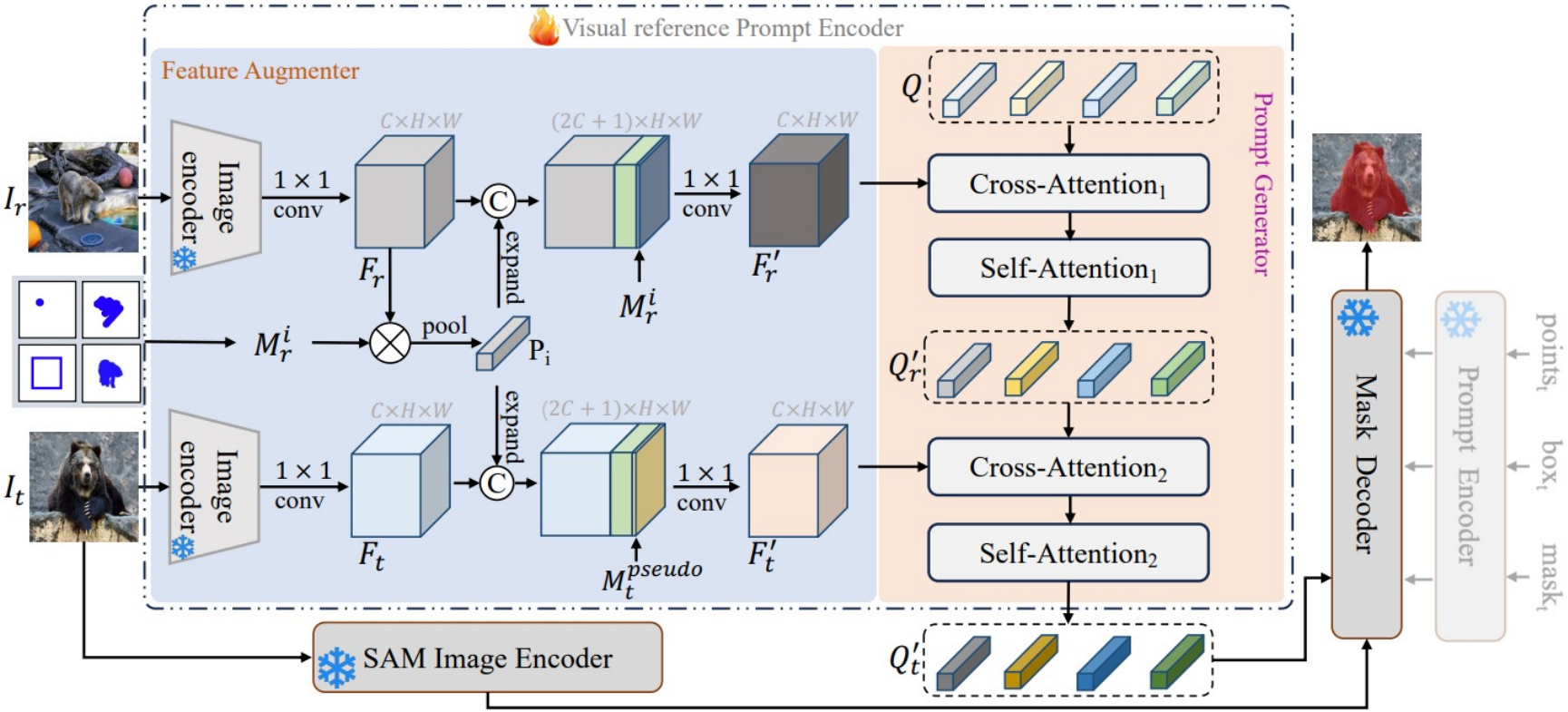
Method

- SAM vs. VRP-SAM
  - Visual reference image
  - Visual reference prompt encoder



Method

- VRP-SAM
  - Feature Augmenter
  - Prompt Generator



Experiments

- Experiment settings
  - Few-shot setting on COCO / Pascal datasets
- Comparison with other foundation models

Methods	Label type	F-0	F-1	F-2	F-3	Means
Painter [34]	*mask.*	31.2	35.3	33.5	32.4	33.1
SegGPT [35]		56.3	57.4	58.9	51.7	56.1
PerSAM† [43]		23.1	23.6	22.0	23.4	23.0
PerSAM-F† [43]		22.3	24.0	23.4	24.1	23.5
Matcher† [18]		52.7	53.5	52.6	52.1	52.7
VRP-SAM†	*point.*	30.1	39.2	43.0	40.4	38.2
	scribble.	40.2	52.0	52.4	44.4	47.2
	box.	44.5	49.3	55.7	49.1	49.7
	mask.	48.1	55.8	60.0	51.6	53.9

Experiments

- Comparison with few-shot methods

Method	Image encoder	Learnable params	COCO-20^i					Pascal-5^i				
			F-0	F-1	F-2	F-3	Mean	F-0	F-1	F-2	F-3	Mean
PFENet [31]	VGG-16	10.4M	35.4	38.1	36.8	34.7	36.3	56.9	68.2	54.5	52.4	58.0
BAM [14]		4.9M	36.4	47.1	43.3	41.7	42.1	63.2	70.8	66.1	57.5	64.4
HDMNet [25]		4.2M	40.7	50.6	48.2	44.0	45.9	64.8	71.4	67.7	56.4	65.1
VRP-SAM		1.0M	43.6	51.7	50.0	46.5	48.0	70.0	74.7	68.3	61.9	68.7
PFENet [31]	ResNet-50	10.4M	36.5	38.6	34.5	33.8	35.8	61.7	69.5	55.4	56.3	60.8
HSNet [21]		2.6M	36.3	43.1	38.7	38.7	39.2	64.3	70.7	60.3	60.5	64.0
CyCTR [41]		15.4M	38.9	43.0	39.6	39.8	40.3	65.7	71.0	59.5	59.7	64.0
SSP [7]		8.7M	35.5	39.6	37.9	36.7	37.4	60.5	67.8	66.4	51.0	61.4
NTRENet [17]		19.9M	36.8	42.6	39.9	37.9	39.3	65.4	72.3	59.4	59.8	64.2
DPCN [16]		-	42.0	47.0	43.3	39.7	43.0	65.7	71.6	69.1	60.6	66.7
VAT [10]		3.2M	39.0	43.8	42.6	39.7	41.3	67.6	72.0	62.3	60.1	65.5
BAM [14]		4.9M	39.4	49.9	46.2	45.2	45.2	69.0	73.6	67.6	61.1	67.8
HDMNet [25]		4.2M	43.8	55.3	51.6	49.4	50.0	71.0	75.4	68.9	62.1	69.4
VRP-SAM		1.0M	48.1	55.8	60.0	51.6	53.9	73.9	78.3	70.6	65.0	71.9
FPTrans [42]	DeiT-B/16	-	44.4	48.9	50.6	44.0	47.0	72.3	70.6	68.3	64.1	68.8
DCAMA	Swin-B	47.7M	49.5	52.7	52.8	48.7	50.9	72.2	73.8	64.3	67.1	69.3

Experiments

- Comparison with Geometric Prompts
  - Geometric prompts are randomly sampled from the pseudo-mask generated by the image encoder.

Method	Image Encoder	Prompts	Mean IoU
GP-SAM	ResNet-50	*box.*	19.7
		point.	21.7
		box. + point.	23.2
	DINOv2	*box.*	31.3
		point.	36.6
		box. + point.	37.8
		box. + point. † [18]	52.7
VRP-SAM	ResNet-50	*point.*	38.4
		scribble.	47.3
		box.	49.7
		mask.	53.9

Experiments

- Visualization