VRP-SAM: SAM with Visual Reference Prompt – Supplementary Material

Yanpeng Sun[1,2], Jiahui Chen[3], Shan Zhang[4], Xinyu Zhang[2], Qiang Chen[2]
Gang Zhang[2], Errui Ding[2], Jingdong Wang[2], Zechao Li[1*]

[1]Nanjing University of Science and Technology,
[2]Baidu VIS, [3]Beihang University, [4]Australian National University
{yanpeng_sun, zechao.li}@njust.edu.cn

Table 1. Details of the data split for PASCAL-5^i. Each row consists of 5 classes designated for testing, with the remaining 15 classes utilized for training.

Fold	Test classes
0	Aeroplane, Bicycle, Bird, Boat, Bottle
1	Bus, Car, Cat, Chair, Cow
2	Dining table, Dog, Horse, Motorbike, Person
3	Potted plant, Sheep, Sofa, Train, TV/monitor

Table 2. Details of the data split for COCO-20^i. Each row consists of 20 classes designated for testing, with the remaining 60 classes utilized for training.

Fold	Test classes
0	Person, Airplane, Boat, Parking meter, Dog, Elephant, Backpack, Suitcase, Sports ball, Skateboard, Wine glass, Spoon, Sandwich, Hot dog, Chair, Dining table, Mouse, Microwave, Sink, Scissors
1	Bicycle, Bus, Traffic light, Bench, Horse, Bear, Umbrella, Frisbee, Kite, Surfboard, Cup, Bowl, Spoon, Orange, Pizza, Couch, Toilet, Remote, Oven, Book, Teddy bear
2	Car, Train, Fire hydrant, Bird, Sheep, Zebra, Handbag, Skis, Baseball bat, Tennis racket, Fork, Banana, Broccoli, Donut, Potted plant, TV, Keyboard, Toaster, Clock, Hair drier
3	Motorcycle, Truck, Stop sign, Cat, Cow, Giraffe, Tie, Snowboard, Baseball glove, Bottle, Knife, Apple, Carrot, Cake, Bed, Laptop, Cell phone, Sink, Vase, Toothbrush

Appendix

A. Additional Implementation Details

Training: To capture semantic correlations between reference and target images, we introduced a semantic relevance model in the Visual Prompt Encoder. Specifically, we validated the effectiveness of VRP-SAM on VGG-16 [1], ResNet-50 [2], and DINOv2 [6]. Following prior work [10, 11], we utilized mid-level features from the image encoder to retain finer details, with high-level features used to generate a pseudo mask for the target image. For instance, in the case of ResNet-50, mid-level features corresponded to the 3^th and 4^th blocks, while high-level features were extracted from the 5^th block.

Within the feature augmenter of the Visual Reference Encoder, the pseudo mask for the target was computed by evaluating the pixel-wise similarity map through the comparison of high-level features of reference and target images. We retained the maximum similarity at each pixel and normalized the similarity map to the [0, 1] range using min-max normalization. This similarity map serves as the pseudo mask for the target image. We followed the SEEM [17] approach, where point reference prompts involve randomly sampling $(1, 20)$ points from the ground truth (GT), scribble reference prompts are generated randomly using a free-form training mask generation algorithm proposed in [14], resulting in $(1, 20)$ scribbles, and box reference prompts are obtained by extracting object bounding boxes from the GT.

Evaluation: For quantitative evaluation on the employed benchmark, we employed the same approach as our training strategy to obtain visual reference prompts about

Table 3. Details of the data split for PASCAL-5^i in domain shift scenario. Each line represents non-overlapping classes in the training set corresponding to the respective fold of COCO.

Fold	Test classes
0	Aeroplane, Boat, Chair, Dining table, Dog, Person
1	Horse, Sofa, Bicycle, Bus
2	Bird, Car, Potted plant, Sheep, Train, TV/monitor
3	Bottle, Cow, Cat, Motorbike

*Corresponding author.

Table 4. Compare with the State-of-the-arts. Results of one-shot semantic segmentation on COCO-20^i.Gray indicates the model is trained by in-domain datasets. The red colors respectively represent the optimal results.

Method	Venue	Image encoder	F-0	F-1	F-2	F-3	Mean
few-shot method.							
BAM+SVF [10]	NeurIPS'22	ResNet-50	46.9	53.8	48.4	44.8	48.5
VAT [3]	ECCV'23		39.0	43.8	42.6	39.7	41.3
HDMNet [7]	CVPR'23		43.8	55.3	51.6	49.4	50.0
FPTrans [15]	NeurIPS'22	Deit-B/16	44.4	48.9	50.6	44.0	47.0
DCAMA [9]	ECCV'23	Swin-B	49.5	52.7	52.8	48.7	50.9
based foundation methods.							
Painter [12]	CVPR'23	Painter	31.2	35.3	33.5	32.4	33.1
PerSAM [16]	arXiv'23	-	23.1	23.6	22.0	23.4	23.0
Matcher [4]	arXiv'23	DINOv2-L	52.7	53.5	52.6	52.1	52.7
SegGPT [13]	ICCV'23	Painter	56.3	57.4	58.9	51.7	56.1
VRP-SAM	This work	VGG-16	43.6	51.7	50.0	46.5	48.0
		ResNet-50	48.1	55.8	60.0	51.6	53.9
		DINOv2-B	56.8	61.0	64.2	59.7	60.4

Table 5. Ablation study based different label types on COCO-20^i. The red and blue colors respectively represent the optimal and suboptimal results.

Method	Image encoder	Label Type	F-0	F-1	F-2	F-3	Mean
VRP-SAM	VGG-16	*point.*	24.6	34.4	35.1	36.7	32.7
		scribble.	32.7	49.6	46.8	39.5	42.2
		box.	36.5	49.7	49.7	43.2	44.8
		mask.	43.6	51.7	50.0	46.5	48.0
	ResNet-50	*point.*	32.0	39.2	43.0	39.3	38.4
		scribble.	40.2	52.0	52.4	44.4	47.3
		box.	44.5	49.3	55.7	49.1	49.7
		mask.	48.1	55.8	60.0	51.6	53.9

point, scribble, and box. For points and scribbles, we randomly selected $(1, 20)$ as prompts. In visual experiments, we showcased the performance of VRP-SAM using only one point or scribble. When inferring with N visual reference prompts, we utilize the visual reference prompt encoder to generate N sets of queries. Subsequently, these N sets of queries are concatenated and fed into the mask decoder.

B. Datasets Setting

To quantify the generalization of VRP-SAM to unseen objects, we adopt the data setup of few-shot segmentation, organizing all classes from the COCO and PASCAL datasets into four folds. For each fold, PASCAL-5^i [8] comprises 15 base classes for training and 5 novel classes for testing, while COCO-20^i [5] includes 60 training base classes and 20 testing novel classes. Table 1 and Table 2 provide a detailed breakdown of the testing classes for PASCAL-5^i and COCO-20^i in each fold, where the training classes are composed of combinations of testing classes from other folds. Additionally, to assess the performance of VRP-SAM on domain shift, we trained on COCO-20^i and tested on PASCAL-5^i. To ensure that there is no overlap between training and testing classes, we performed a new partition of the PASCAL dataset, and Table 3 provides a detailed description of the newly segmented folds for PASCAL.

C. More experiments

C.1. Comparison with the State-of-the-art

To demonstrate the superior performance of VRP-SAM, we conducted experiments on the COCO-20^i dataset, comparing it with state-of-the-art methods. As shown in Table 4, we observed outstanding results when employing DINOv2-B [6] as the image encoder in VRP-SAM, particularly surpassing SegGPT [13] for the first time — a method utilizing an image encoder trained on in-domain datasets. Furthermore, utilizing ResNet-50 as the image encoder still

Table 6. Ablation study on COCO-20^i base set. The red and blue colors respectively represent the optimal and suboptimal results.

Methods	Image encoder	Label type	Base set					Novel set
			F-0	F-1	F-2	F-3	Means	Means
VRP-SAM	ResNet-50	*point.*	30.1	36.8	44.5	42.4	38.5	38.2
		scribble.	44.7	50.7	49.8	53.1	49.6	47.2
		box.	46.4	50.4	49.8	57.3	51.0	49.7
		mask.	52.4	56.4	58.8	61.3	57.2	53.9
	DINOv2-B	*mask.*	66.1	69.0	70.0	65.2	67.6	60.4

achieved a noteworthy Mean IoU of 53.9, outperforming the majority of current approaches. These findings substantiate the versatility and robustness of VRP-SAM across different encoders, solidifying its superiority in visual reference segmentation tasks.

C.2. Different label type

We conducted ablation study on label types of reference image, exploring their impact on VRP-SAM results using both VGG-16 and ResNet-50. Table 5 presents the findings, demonstrating a gradual performance improvement with increasing annotation precision, particularly from points to masks. Specifically, transitioning from point to mask annotations significantly boosted VRP-SAM performance by approximately 15 miou. This underscores the crucial influence of annotation granularity on VRP-SAM performance.

C.3. Comparison on base set

To assess the performance of VRP-SAM on trained categories, we randomly selected 1000 reference-target pairs from the training set of COCO-20^i. This subset was used to test VRP-SAM's performance on categories it had been trained on. We refer to the test set composed of categories used for training in each fold as the base set. The experimental results, as shown in Table 6, demonstrate that VRP-SAM performs better on the base set than on the novel set. This improvement is attributed to VRP-SAM acquiring specific knowledge about the base classes during training, contributing to enhanced performance on those classes. Moreover, employing DINOv2 as the image encoder significantly enhances VRP-SAM's performance on the base set. This improvement is attributed to DINOv2's feature representation being more adept at capturing intricate semantic relationships, thereby boosting the model's performance on known classes.

C.4. Compare with text-guided SAM

The text-guided SAM can also segment target objects based on category names, thus achieving a functionality similar to VRP-SAM. To verify this, we conducted a comparison with existing text-guided SAM in the PASCAL-5^i setting. The experimental results, as shown in Table 7. indicate that

Table 7. Compare with text-guided SAM on PASCAL-5^i novel set. The red colors respectively represent the optimal results.

model	F-0	F-1	F-2	F-3	means
text-guided SAM:					
FastSAM$_{(\text{ViT-B + YOLOv8x})}$	18.9	29.1	24.4	32.0	26.1
CLIP-SAM$_{(\text{ViT-B + SAM-H})}$	25.9	51.8	33.5	51.3	40.6
visual-guided SAM:					
VRP-SAM$_{(\text{RN50 + SAM-H})}$	73.9	78.3	70.6	65.1	71.9

the performance of text-guided SAM is poor. This not only demonstrates that SAM is a semantically agnostic segmentation model, but also underscores the superiority of VRP-SAM.

C.5. Results on more application scenarios

In the aforementioned experiments focusing on typical object segmentation tasks, we conducted a comprehensive evaluation of VRP-SAM. Next, we validate the effectiveness of VRP-SAM in atypical object segmentation tasks. Specifically, we conducted experiments on both part segmentation and video object segmentation tasks. The results are shown in Table 8 and 9. The results affirm VRP-SAM's promising performance in the domain of Part Segmentation and Video Object Segmentation. We believe these findings further support the versatility of VRP-SAM across diverse applications.

Table 8. Result of part segmentation on PASCAL-PART. The red colors respectively represent the optimal results.

model	animals	indoor	person	vehicles	means
Painter [12]	20.2	49.5	17.6	34.4	30.4
SegGPT [13]	22.8	50.9	31.3	38.0	35.8
PerSAM [16]	19.9	51.8	18.6	32.0	30.1
VRP-SAM$_{(\text{RN50})}$	23.4	56.6	25.8	35.6	35.4
VRP-SAM$_{(\text{DINOv2-B})}$	30.3	52.1	25.8	36.7	36.2

D. Limitation and future works

Currently, we only demonstrate the effectiveness of VRP-SAM in few-shot semantic segmentation. Extending VRP-SAM to more vision tasks, such as video object segmentation and object tracking, needs more investigation. We leave

Table 9. Result of video object segmentation on DAVIS 2017 dataset. The red colors respectively represent the optimal results.

model	*J&F*	*J*	*F*
Painter [12]	34.6	28.5	40.8
PerSAM [16]	60.3	56.6	63.9
VRP-SAM	64.8	62.1	67.4

it for future work.

References

[1] Yoshua Bengio and Yann LeCun. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[3] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126, 2022. 2

[4] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023. 2

[5] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *IEEE International Conference on Computer Vision*, pages 622–631, 2019. 2

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2

[7] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 2

[8] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 2

[9] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168, 2022. 2

[10] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. In *Advances in Neural Information Processing Systems*, pages 37484–37496, 2022. 1, 2

[11] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020. 1

[12] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2, 3, 4

[13] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *IEEE International Conference on Computer Vision*, pages 1130–1140, 2023. 2, 3

[14] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 1

[15] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. In *Advances in Neural Information Processing Systems*, pages 6575–6588, 2022. 2

[16] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 2, 3, 4

[17] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 1