

# ProSAM: Enhancing the Robustness of SAM-based Visual Reference Segmentation with Probabilistic Prompts

Xiaoqi Wang<sup>1,2</sup>, Clint Sebastian<sup>2</sup>, Wenbin He<sup>1,2</sup>, Liu Ren<sup>1,2</sup>

<sup>1</sup>Bosch Research North America, <sup>2</sup>Bosch Center for Artificial Intelligence (BCAI)

xiaoqi.wang@us.bosch.com clint.sebastian@de.bosch.com wenbin.he2@us.bosch.com liu.ren@us.bosch.com

## Abstract

The recent advancements in large foundation models have driven the success of open-set image segmentation, a task focused on segmenting objects beyond predefined categories. Among various prompt types (such as points, boxes, texts, and visual references), visual reference segmentation stands out for its unique flexibility and strong zero-shot capabilities. Recently, several SAM-based methods have made notable progress in this task by automatically generating prompts to guide SAM. However, these methods often generate prompts at boundaries of target regions due to suboptimal prompt encoder, which results in instability and reduced robustness. In this work, we introduce ProSAM, a simple but effective method to address the stability challenges we identified in existing SAM-based visual reference segmentation approaches. By learning a variational prompt encoder to predict multivariate prompt distributions, ProSAM avoids generating prompts that lie in unstable regions, overcoming the instability caused by less robust prompts. Our approach consistently surpasses state-of-the-art methods on the Pascal-5<sup>i</sup> and COCO-20<sup>i</sup> datasets, providing a more robust solution for visual reference segmentation.

## 1. Introduction

Open-set image segmentation methods have gained considerable attention for their ability to segment objects beyond a fixed set of categories. These methods incorporate diverse prompts, including points, boxes, texts, and visual references, effectively addressing the limitations of closed-set approaches. The introduction of the Segment Anything Model (SAM) series [14, 28] has notably advanced open-set segmentation performance using point and box prompts. However, when segmenting the same type of object across multiple images, using SAM can be tedious and time-consuming because it requires custom prompts for each image individually. Image segmentation methods us-

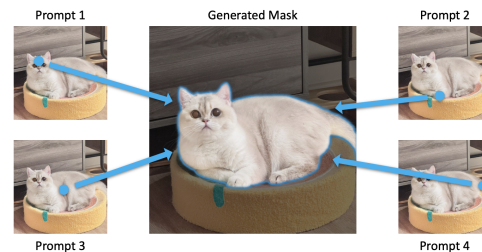


Figure 1. The same mask can be generated by SAM using various prompts in a region.

ing text prompts [8, 37, 39] offer a solution to this limitation but encounter two main challenges [10]: (i) aligning vision and language representations for rare or long-tailed objects is challenging due to their scarcity, leading to compromised segmentation performance for these objects; and (ii) certain objects are difficult to describe accurately in natural language without specialized knowledge. For example, someone without a background in chemistry may struggle to accurately describe “molecular orbitals”.

To tackle the challenges resulting from long-tailed data shortage and descriptive limitation, image segmentation with visual references (i.e., an annotated reference image that indicates the objects of interest) has become increasingly important. Its ultimate goal is to segment similar objects as indicated in the annotated reference image, regardless of the semantic category of target objects. Recently, various methods [20, 32, 45] have leveraged the exceptional segmentation capabilities of SAM, achieving significant breakthroughs in visual reference segmentation. The main idea of these methods is to generate prompts that direct SAM to predict masks for the target objects. Among these approaches, training-based methods [32] that learn prompt embeddings to guide SAM mask generation have achieved state-of-the-art (SOTA) performance. However, the existing training approaches often direct the prompts toward the boundaries of target regions, resulting in instability and reduced robustness. This robustness challenge is specifically caused by the inherent design of SAM — iden-

tical masks can be generated from different prompts (see Figure 1). This oversight poses a significant challenge for the training-based methods, where stability and generalizability are essential for zero-shot capability.

To address this limitation, we introduce a novel visual reference segmentation method, ProSAM, to enhance the robustness and zero-shot capability of SAM-based visual reference segmentation. Specifically, inspired by the spirit of variational inference [3] in statistics, we propose a variational prompt encoder with reparameterization trick to predict a multivariate prompt distribution in high-dimensional space, such that every prompt sampled from this multivariate prompt distribution can effectively guide SAM to generate a high-quality mask for the target object. During inference, the predicted mean of this multivariate prompt distribution will be utilized to generate the predicted mask with the same visual concept as the reference object. Unlike the existing training-based method [32], which does not favor the prompts closer to the center of the target prompt region, our method encourages the mean of the multivariate prompt distribution to be closer to the center of the target prompt region by injecting the noise into the generated prompts and penalizing the Laplacian during training.

To demonstrate the effectiveness of our method, we conducted extensive experiments on Pascal-5<sup>i</sup> and COCO-20<sup>i</sup> datasets following the same dataset configuration as the SOTA methods [32, 45]. The experimental results demonstrate that our approach consistently outperforms the SOTA method on both datasets. In summary, the contributions of this paper are threefold:

- We identify a commonly overlooked limitation in the SOTA SAM-based visual reference segmentation approach, where prompts are often generated at boundaries of target regions, leading to instability and reduced robustness.
- We propose a probabilistic prompt generation method that leverages variational inference to penalize the prompts that lie in unstable regions, enhancing the robustness of generated prompts.
- Our approach consistently outperforms the SOTA SAM-based visual reference segmentation method on the Pascal-5<sup>i</sup> and COCO-20<sup>i</sup> datasets.

## 2. Related Work

### 2.1. Visual Reference Segmentation

A visual reference is an annotated reference image that represents the object of interest. Segmenting based on visual reference prompt provides a more intuitive and straightforward way to guide the segmentation of the desired object in the target image, regardless of its semantic category. Unlike text prompts, visual reference prompt bypasses the need for cross-modality alignment between text and image, be-

cause it solely relies on visual similarities [10]. This unique strength enhances generalizability in segmenting novel objects that were unseen during training.

With the recent advancements in vision foundation models, several SAM-based methods [20, 32] have achieved significant breakthroughs in visual reference segmentation, by transforming visual references into prompts that SAM can understand. These SAM-based methods can be classified into two categories: training-based approaches [32] and training-free approaches [20, 45]. Notably, VRP-SAM [32], a SAM-based training approach, achieves SOTA performance in this task. However, the existing training-based approaches including VRP-SAM fail to consider the robustness of generated prompts. This oversight motivates us to propose a variational prompt encoder, which has been theoretically and empirically demonstrated to generate more robust prompts.

### 2.2. Variational Inference

To facilitate the robustness of the prompts generated for SAM, we draw inspiration from the principles of variational inference [3] in statistics and propose a SAM-based variational prompt encoder to predict a probabilistic prompt distribution based on the visual reference. In the existing literature, the variational inference has been applied across various tasks (e.g., data generation [13], metric learning [17], person re-identification [41], and semantic segmentation [38]) with different purposes. Specifically, VAE [13] and DVML [17] adopt variational inference to generate more diverse and discriminative samples by modeling the data variance, while DistributionNet [41] and PRCL [38] utilize variational inference to handle the noisy data by estimating the uncertainty of data distribution. Unlike these methods, we leverage variational inference to inject noise into the prompt embeddings with the purpose of penalizing Laplacian during training, such that more robust prompts can be generated to guide SAM. To the best of our knowledge, we are the first SAM-based segmentation method that learns a variational prompt encoder to generate probabilistic prompts, specifically designed to enhance the robustness of SAM mask generation.

## 3. Preliminary

### 3.1. SAM

SAM is designed to generate a segmentation mask for a given input image  $I$  based on user-specified prompts  $P$ . These prompts can be in various forms, such as points or boxes. Its architecture is composed of three major components: a prompt encoder, an image encoder, and a lightweight mask decoder, denoted by  $f_P^s$ ,  $f_I^s$ , and  $f_M^s$ , respectively. Specifically, the image encoder  $f_I^s$  extracts features from the input image to produce  $F_I^s$ , while the prompt encoder  $f_P^s$  processes  $m$  user-provided prompts to generate

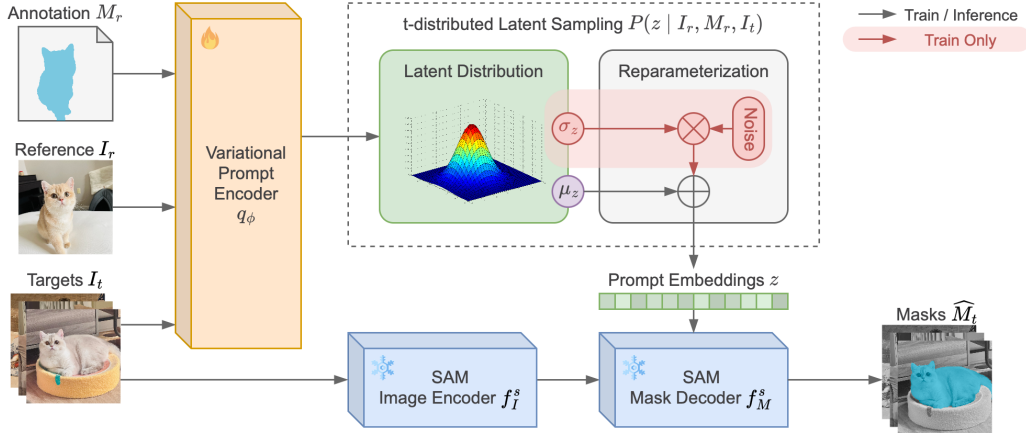


Figure 2. The overview of ProSAM, which segments the target images based on the visual references. Given a pre-trained SAM, a variational prompt encoder is trained to predict a multivariate prompt distribution with reparameterization trick. During the inference, the predicted mean prompt is employed to guide SAM in producing robust mask prediction for the target image.

prompt embeddings  $z$ . This can be expressed as follows:

$$F_I^s = f_I^s(I), \quad z = f_P^s(P), \quad (1)$$

where  $F_I^s \in \mathbb{R}^{h \times w \times c}$  is the image feature map with resolution  $h \times w$  and feature dimension  $c$ , and  $z \in \mathbb{R}^{m \times c}$  represents the prompt embeddings. The encoded image features  $F_I^s$  and prompt embeddings  $z$  are then passed into the decoder  $f_M^s$  to produce the final mask output, represented as:

$$\hat{M} = f_M^s(F_I^s, z), \quad (2)$$

where  $\hat{M}$  represents the final mask predicted by SAM.

### 3.2. Existing Training-Based Method Using SAM

In the existing literature, VRP-SAM [32] is currently the most advanced training-based method, which achieves state-of-the-art performance in this task. Therefore, our method is built on top of VRP-SAM to further enhance the SAM-based visual reference segmentation.

**Visual Reference Prompt Encoder.** In VRP-SAM, the only trainable module is the visual reference prompt encoder, as it utilizes a pre-trained SAM, along with a pre-trained image encoder (e.g., ResNet-50 [7]). This prompt encoder transforms various annotation formats for reference images (e.g., points, boxes, scribbles, and masks) into high-dimensional prompt embeddings that share the same output space as the SAM prompt encoder. There are two major components in their visual reference prompt encoder: feature augments and prompt generator. To be specific, the feature augments leverages a semantic-aware image encoder  $f_I$  to extract the enhanced image features  $F_r^v$  and  $F_t^v$  for reference image  $I_r$  and target image  $I_t$ . These enhanced image features,  $F_r^v$  and  $F_t^v$ , capture the object-specific features extracted from visual annotation  $M_r$  and a

pseudo-mask of target image  $M_t^{\text{pseudo}}$ , respectively. Then, the prompt generator  $f_P^v$  will output a latent prompt  $z$  as follows,

$$z = f_P^v(F_r^v, F_t^v), \quad \text{for } z \in \mathbb{R}^{m \times c}. \quad (3)$$

Lastly, given the generated prompt embeddings  $z$ , a mask prediction  $\hat{M}_t$  is generated as in Equation 2. The predicted mask  $\hat{M}_t$  is expected to encapsulate a visual concept similar to that of the visual reference ( $I_r, M_r$ ). Notably, with  $m$  set to 50 as the default value, the prompt encoder  $f_P^v$  predicts 50 prompt embeddings for each target image, enabling a more comprehensive representation of the visual characteristics of the reference objects.

**Loss Function.** To supervise the learning of its prompt encoder  $f_P^v$ , Binary Cross-Entropy loss and Dice loss are computed between the predicted mask  $\hat{M}_t$  and ground-truth mask  $M_t$  as below,

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\hat{M}_t, M_t) + \mathcal{L}_{\text{Dice}}(\hat{M}_t, M_t). \quad (4)$$

In essence, VRP-SAM focuses solely on mask-level differences, while overlooking the potential for further optimizing the prompt encoder to generate more robust prompts.

## 4. ProSAM

In this paper, our high-level objective is to automatically generate robust prompts to guide SAM in producing high-quality segmentation masks containing the same visual concepts as the visual reference. To this end, we first identify a unique robustness challenge for the SAM-based segmentation method (see Section 4.1). To address this challenge, we propose a variational prompt encoder in Section 4.2 that transforms the visual reference into a multivariate prompt distribution, such that the predicted mean prompt can be

employed to generate high-quality and robust masks during inference. The model training and inference procedures are described in Section 4.3.

#### 4.1. Robustness Challenge

Robust prompts are crucial for the SAM model to yield stable and precise final mask predictions. Yet, the importance of prompt robustness has been largely overlooked in the existing literature, as pointed out in Section 3.2. In this section, we pinpoint this critical challenge in existing SAM-based segmentation methods: the failure to account for the robustness of generated prompts.

In practice, there exists a region of prompt embeddings in the high-dimensional space in which every prompt can lead SAM to produce acceptable segmentation masks (see Figure 1). This region is referred to as the target prompt region  $\mathcal{R}_{I_r, M_r, I_t}$ . However, the stability and robustness of different prompts within this region can vary. For less robust prompts, even a small perturbation can lead to significant changes in the masks produced by SAM. This instability is more prevalent when the learned prompts lie near the boundary of the target region. Such unstable prompts in the boundary areas are more likely to be generated if a learnable prompt generator is not explicitly guided to produce prompts toward the center of  $\mathcal{R}_{I_r, M_r, I_t}$  (see Appendix 7.3 for detailed analysis). Theoretically speaking, the loss landscape in the boundary areas typically exhibits sharp gradient variations due to high curvature (quantified by the Hessian  $\nabla^2 \mathcal{L}(z)$ ), given a loss function  $\mathcal{L}$  that solely considers the mask-level differences (e.g., Equation 4).

Ideally, the prompt embedding should reside in a flat region of the loss landscape with low curvature, which is particularly important when generalizing to objects with unseen semantic categories. One straightforward way to enforce this would be to directly regularize the curvature by penalizing the Hessian  $\nabla^2 \mathcal{L}(z)$ . However, incorporating such a curvature-based regularization term directly into the loss function is challenging due to the limitation of the automatic differentiation-based deep learning framework [2], because it requires computing second-order derivatives. Unfortunately, no existing approaches including VRP-SAM attempt to encourage prompt embeddings to lie in flatter regions of the loss landscape. This issue is a challenging but important gap that we aim to bridge in this work.

#### 4.2. Variational Prompt Encoder

To improve the robustness of generated prompts, we propose a simple but effective method to learn a more robust prompt encoder that favors prompts in the flatter regions of the loss landscape without explicitly incorporating a second-order regularization term. To be specific, we introduce a variational prompt encoder, denoted as  $q_\phi(z|I_r, M_r, I_t)$  and parameterized by  $\phi$ , to approximate

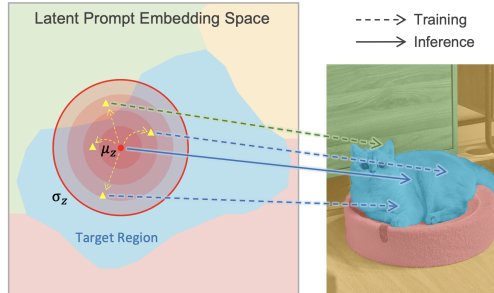


Figure 3. Intuitive illustration of the high-level idea behind the proposed variational prompt encoder. The dashed arrow shows the sampling and prompting procedure during training, while the solid arrow shows the prompting strategy during inference. Note that our generated prompts can be any type of prompt, while this illustration shows a positive point prompt as an example.

the true multivariate prompt distribution  $P(z|I_r, M_r, I_t)$ . The objective is to maximize the likelihood that a sampled prompt embedding  $z \sim q_\phi(z|I_r, M_r, I_t)$  falls within the target prompt region  $\mathcal{R}_{I_r, M_r, I_t}$ . As illustrated in Figure 3, a straightforward intuition behind this framework is to optimize the robustness of expected prompt embedding  $\hat{\mu}_z$  by inducing a margin between  $\hat{\mu}_z$  and the boundary of  $\mathcal{R}_{I_r, M_r, I_t}$  leveraging the standard deviation  $\hat{\sigma}_z$  of the variational prompt distribution  $q_\phi(z|I_r, M_r, I_t)$ . This makes the mean prompt  $\hat{\mu}_z$  less likely to fall outside the target region under small perturbations. Furthermore, based on the theoretical analysis presented in Section 7.1 and Section 7.2 in Appendix, we prove that performing variational optimization leads to an implicit penalty on the curvature of the loss function, thereby encouraging the optimization to favor flatter regions in the loss landscape, which is usually closer to the center of  $\mathcal{R}_{I_r, M_r, I_t}$ . Empirically, our verification study also showcases that our framework effectively pushes prompt embeddings toward the target region center (see Section 5.3.) Therefore, based on both the theoretical analysis and the empirical results, the proposed variational prompt encoder naturally addresses the challenge we presented in Section 4.1.

A straightforward instantiation of  $P(z|I_r, M_r, I_t)$  is assuming  $z$  follows a conditional multivariate Gaussian distribution. However, considering our motivation is to push the mean prompt  $\hat{\mu}_z$  toward the center of the target prompt embedding region, a more heavy-tailed t-distribution would be a better choice due to its statistical properties of having more chance to sample outliers [5]. In other words, when  $\hat{\mu}_z$  is close to the boundary of  $\mathcal{R}_{I_r, M_r, I_t}$ , it is more likely to sample a prompt falling outside the target prompt embedding region (green region in Figure 3) if it follows heavy-tailed t-distribution. Given that the outlier prompt is more likely to result in low-quality mask predictions, the force to push the mean prompt toward the center of  $\mathcal{R}_{I_r, M_r, I_t}$

will be greater with larger gradients. Theoretically, we also show in Section 7.5 that t-distribution results in a larger 4th-order curvature penalty that forces an extra push toward flatter and more stable regions. Inspired by Kim et al. [12], we formulate our variational prompt distribution  $q_\phi(\mathbf{z}|I_r, M_r, I_t)$  with the heavy-tailed property and diagonal covariance as follows,

$$q_\phi(\mathbf{z}|I_r, M_r, I_t) = t\left(\mathbf{z} \middle| \hat{\boldsymbol{\mu}}_{\mathbf{z}}, \frac{\text{diag}(\hat{\boldsymbol{\sigma}}_{\mathbf{z}}^2)}{1 + \nu^{-1}n}, \nu + n\right) \quad (5)$$

where  $\nu$  is a hyper-parameter to control the degree of heavy-tailness and  $n$  is the dimensionality of  $\mathbf{z}$ .

To learn a variational prompt distribution via back-propagation, the sampling function for sampling  $\mathbf{z}$  from  $q_\phi(\mathbf{z}|I_r, M_r, I_t)$  must be differentiable. Thus, the reparameterization trick [1] is employed to approximate the sampling process of  $\mathbf{z}$  with a differentiable function. With two independent random variables  $\epsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I)$  and  $\delta \stackrel{\text{iid}}{\sim} \chi^2(\nu + n)$  as the source of randomness, the sampling function  $g_{\mathbf{z}}$  for drawing  $\mathbf{z}$  from  $q_\phi(\mathbf{z}|I_r, M_r, I_t)$  can be formulated as,

$$\mathbf{z} = g_{\mathbf{z}}(\hat{\boldsymbol{\mu}}_{\mathbf{z}}, \hat{\boldsymbol{\sigma}}_{\mathbf{z}}, \delta, \epsilon) \quad (6)$$

$$= \hat{\boldsymbol{\mu}}_{\mathbf{z}} + \frac{1}{\sqrt{\delta}/(\nu + n)} \frac{\hat{\boldsymbol{\sigma}}_{\mathbf{z}}}{\sqrt{1 + \nu^{-1}n}} \odot \epsilon. \quad (7)$$

This formulation enables us to sample  $\mathbf{z} \sim q_\phi(\mathbf{z}|I_r, M_r, I_t)$  as a differentiable function of  $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$  and  $\hat{\boldsymbol{\sigma}}_{\mathbf{z}}$ , predicted by the variational prompt encoder.

### 4.3. Training and Inference

**Training.** During training, our goal is to minimize the expected loss w.r.t. the prompt embedding distribution,

$$\min_{\phi} \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|I_r, M_r, I_t)} [\mathcal{L}(f_M^s(\mathbf{z}, F_{I_t}^s), M_t)] \quad (8)$$

$$= \min_{\phi} \int_{\mathbb{R}^n} q_\phi(\mathbf{z}|I_r, M_r, I_t) \mathcal{L}(f_M^s(\mathbf{z}, F_{I_t}^s), M_t) d\mathbf{z}, \quad (9)$$

where  $F_{I_t}^s = f_I^s(I_t)$ . However, this integral is intractable. Therefore, we employ the Monte Carlo method to approximate it by minimizing the expected loss with  $K$  samples,

$$\begin{aligned} & \min_{\phi} \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|I_r, M_r, I_t)} [\mathcal{L}(f_M^s(\mathbf{z}, F_{I_t}^s), M_t)] \\ &= \min_{\phi} \mathbb{E}_{\epsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I), \delta \stackrel{\text{iid}}{\sim} \chi^2(\nu+n)} [\mathcal{L}(f_M^s(\mathbf{z}, F_{I_t}^s), M_t)] \\ &\approx \min_{\phi} \frac{1}{K} \sum_{k=1}^K \mathcal{L}(f_M^s(g_{\mathbf{z}}(\hat{\boldsymbol{\mu}}_{\mathbf{z}}, \hat{\boldsymbol{\sigma}}_{\mathbf{z}}, \delta_k, \epsilon_k), F_{I_t}^s), M_t). \quad (10) \end{aligned}$$

Note that the loss function can theoretically be any mask-level loss that evaluates the deviation between the predicted mask  $\hat{M}_t$  and  $M_t$ . Following VRP-SAM [32], we adopt the BCE loss and Dice loss to enforce both pixel-wise accuracy and the degrees of overlap (see Equation 4). Regarding the

model architecture of the variational prompt encoder, our approach theoretically can be applicable to any model architecture. In this paper, to ensure a fair comparison with the existing training-based method, we employ an identical model architecture as VRP-SAM prompt encoder (as described in Section 3.2), except adding two linear layers at the end to predict  $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$  and  $\hat{\boldsymbol{\sigma}}_{\mathbf{z}}$  respectively.

**Inference.** During inference, only the predicted mean prompt  $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$  is used to prompt the SAM mask decoder  $f_M^s$  to generate a robust mask prediction  $\hat{M}_t$ . This inference strategy facilitates the robustness of generated prompts for novel objects, leveraging the margin between  $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$  and the boundary of target prompt region  $\mathcal{R}_{I_r, M_r, I_t}$ . More importantly, since we rely solely on the predicted mean prompt for inference, our inference speed and memory usage are on par with that of non-probabilistic methods such as VRP-SAM [32].

## 5. Experiments

### 5.1. Experimental Setup

Table 1. Quantitative comparison with SOTA visual reference segmentation methods based on mIoU. <sup>†</sup> represents the method is based on SAM. For the models trained by in-domain datasets, their results are colored in gray. The colors green and blue indicate the best and second-best results, respectively, among all methods that were not trained with in-domain datasets.

Data	Methods	Label Type	F-0	F-1	F-2	F-3	Means
COCO-20 <sup>i</sup>	Painter[35]		31.2	35.3	33.5	32.4	33.1
	SegGPT[36]		56.3	57.4	58.9	51.7	56.1
	PerSAM <sup>†</sup> [45]	mask	23.1	23.6	22.0	23.4	23.0
	PerSAM-F <sup>†</sup> [45]		22.3	24.0	23.4	24.1	23.5
	Matcher <sup>†</sup> [20]		52.7	53.5	52.6	52.1	52.7
	VRP-SAM <sup>†</sup> [32]	point	32.03	39.36	46.44	40.52	39.85
		scribble	44.83	48.22	51.61	47.66	48.08
		box	44.63	49.2	56.56	49.34	49.93
		mask	47.02	54.24	59.91	51.95	53.28
	ProSAM <sup>†</sup>	point	33.32	40.23	47.82	41.2	40.64
scribble		47.37	48.97	53.44	48.39	49.54	
box		45.39	50.01	57.92	50.41	50.93	
mask		48.74	55.55	60.72	53.49	54.63	
PASCAL-5 <sup>i</sup>	VRP-SAM <sup>†</sup> [32]	point	63.69	70.95	63.22	54.53	63.35
		scribble	70.04	74.67	65.93	59.12	67.44
		box	71.3	75.98	65.95	61.27	68.75
		mask	74.01	76.77	69.46	64.34	71.14
	ProSAM <sup>†</sup>	point	64.71	72.11	63.89	55.64	64.08
	scribble	71.16	75.69	66.31	60.93	68.52	
	box	72.38	76.81	67.07	62.76	69.76	
	mask	75.26	77.57	70.09	65.22	72.04	

**Datasets.** To evaluate the effectiveness and generalizability of ProSAM, we conducted comprehensive experiments on Pascal-5<sup>i</sup> [29] and COCO-20<sup>i</sup> [25] under the same few-shot setting as the existing visual reference segmentation methods [16, 32, 44]. In this setting, these two datasets are divided into 4 folds. In each fold, Pascal-5<sup>i</sup> in-

Table 2. Quantitative comparison with one-shot segmentation methods based on mIOU. The green and blue colors indicate the best and second-best results, respectively.

Methods	Image Encoder	Learnable Params	COCO-20 <sup>i</sup>					PASCAL-5 <sup>i</sup>				
			F-0	F-1	F-2	F-3	Mean	F-0	F-1	F-2	F-3	Mean
PFENet [33]	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 [23]		2.6M	36.3	43.1	38.7	38.7	39.2	64.3	70.7	60.3	60.5	64.0
CyCTR [42]		15.4M	38.9	43.0	39.6	39.8	40.3	65.7	71.0	59.5	59.7	64.0
SSP [6]		8.7M	35.5	39.6	37.9	36.7	37.4	60.5	67.8	66.4	51.0	61.4
NTRENet [19]		19.9M	36.8	42.6	39.9	37.9	39.3	65.4	72.3	59.4	59.8	64.2
DPCN [18]		-	42.0	47.0	43.3	39.7	43.0	65.7	71.6	69.1	60.6	66.7
VAT [9]		3.2M	39.0	43.8	42.6	39.7	41.3	67.6	72.0	62.3	60.1	65.5
BAM [15]		4.9M	39.4	49.9	46.2	45.2	45.2	69.0	73.6	67.6	61.1	67.8
HDMNet [27]		4.2M	43.8	<b>55.3</b>	<b>51.6</b>	<b>49.4</b>	<b>50.0</b>	71.0	<b>75.4</b>	<b>68.9</b>	62.1	<b>69.4</b>
ProSAM		1.73M	<b>48.74</b>	<b>55.55</b>	<b>60.72</b>	<b>53.49</b>	<b>54.63</b>	<b>75.26</b>	<b>77.57</b>	<b>70.09</b>	<b>65.22</b>	<b>72.04</b>
DCAMA [30]	Swin-B	47.7M	<b>49.5</b>	52.7	52.8	48.7	<b>50.9</b>	<b>72.2</b>	73.8	64.3	<b>67.1</b>	69.3

cludes 15 base classes for training and 5 novel classes for testing, while COCO-20<sup>i</sup> has 60 base classes for training and 20 novel classes for testing. Therefore, the robustness and generalizability of each visual reference segmentation method can be fully assessed under this setting. Following VRP-SAM [32], 1,000 pairs of visual reference and target images are randomly selected to evaluate our testing performance for each fold.

**Implementation Details.** To ensure a strictly fair comparison with VRP-SAM, the SOTA method in the visual reference segmentation, we ensure that all the experimental settings (e.g., random seed, LR scheduler, optimizer) and hyper-parameters (e.g., number of layers, prompt embedding dimensions, number of prompts) are identical to VRP-SAM. Specifically, the model architecture of the variational prompt encoder is similar to VRP-SAM (see Section 8 in Appendix for more details), except two linear layers have been added at the end of the prompt encoder to predict the mean and log standard deviation of the multivariate prompt distribution. In other words, our method is easy to implement with only a few lines of code, but can effectively boost the robustness of generated prompts. During training, the number of Monte Carlo samples  $K$  has been set to 10, with the degrees of freedom  $\nu$  of 5. Also, we employ the AdamW [21] optimizer with weight decay of 1e-6 and the cosine annealing learning rate scheduler with warm restart [22] after 15 epochs, and the initial learning rate of 1e-4. The model is trained with 100 epochs with a fixed random seed of 321. For the choice of image encoder utilized by the variational prompt encoder, ResNet-50 [7] is adopted following VRP-SAM. In addition, our variational prompt encoder predicts 50 multivariate prompt distributions per target image and only uses 50 mean prompts during inference, which guarantees the same number of prompts have been used in the experimental study for both VRP-SAM and ProSAM. In terms of the visual annotation types, we follow the same procedure as VRP-SAM [32] and SEEM [46] to automatically generate points, scribbles, and boxes based on the mask annotations. Lastly, the mean intersection over

union (mIOU) is adopted to evaluate our segmentation performance across all datasets. All experiments on COCO-20<sup>i</sup> were conducted on 4 RTX 4090 GPUs with a batch size of 2 per GPU, whereas the experiments on Pascal-5<sup>i</sup> were conducted on 1 H100 GPU with a batch size of 2. On a single H100 GPU with batch size 2, ProSAM takes 0.19s per training batch and 0.12s per inference batch.

## 5.2. Quantitative Evaluation

To assess the effectiveness of ProSAM, we compare our method against the existing visual reference segmentation methods via mIOU. Specifically, for both Pascal-5<sup>i</sup> and COCO-20<sup>i</sup>, ProSAM is trained and tested for each fold separately to evaluate the models on the *unseen classes only*. Unfortunately, we are unable to reproduce the VRP-SAM results they reported in their paper, possibly due to the different hardware we had. Therefore, in order to ensure the fairness of our quantitative comparison against VRP-SAM, we reported the experimental results of both VRP-SAM and our method under completely identical experimental settings as mentioned in Section 5.1. The only difference is the hyper-parameters introduced by learning multivariate prompt t-distributions (i.e., Monte Carlo Samples  $K$  and degrees of freedom  $\nu$ ), because those are not applicable to VRP-SAM. For qualitative evaluations, please refer to Section 11 in the Appendix.

### Quantitative Comparison with FM-based Methods.

Powered by the recent advancement of vision foundation models, several visual reference segmentation methods achieved a performance breakthrough [20, 32, 35, 36, 45]. According to Table 1, our method with mask annotation achieves the best performance among all SAM-based methods, and surprisingly obtains a comparable result as SegGPT that is trained and tested on the same set of classes within each fold. Compared with Matcher [20], which employs a larger image encoder DINOv2 ViT-L, we can still achieve a superior performance using a more lightweight, non-ViT backbone, ResNet-50. Compared with VRP-SAM, our variational prompt encoder enables us to consistently

outperform VRP-SAM on both datasets, no matter which type of visual reference has been used. Also, the comparison over the confusion matrix against VRP-SAM (see Section 10.1 in Appendix) showcases that ProSAM improves the balanced performance by reducing both the false negatives and false positives.

#### Quantitative Comparison with Few-Shot Methods.

To comprehensively evaluate our method, we also compare ProSAM with the SOTA few-shot segmentation methods. As shown in Table 2, we achieve SOTA performance on both datasets with the least number of learnable parameters. Note that all the experimental results are computed on *novel classes only*, which demonstrate the robustness and generalizability of our variational prompt encoder. Especially compared with DCAMA [30], which has 47.7M learnable parameters, we can still outperform it with only 1.73M learnable parameters.

**Generalizability Study Under Domain Shift.** The generalization capability of ProSAM under domain shift is critical, as it highlights our performance in scenarios where there is a substantial difference between the domains of the training data and testing data. Following the previous works [23, 32, 33], we conducted the generalization study under the domain shift from COCO-20<sup>i</sup> to PASCAL-5<sup>i</sup>. Specifically, the models are trained on COCO-20<sup>i</sup> and exclusively tested on novel classes from PASCAL-5<sup>i</sup> that were not included in COCO-20<sup>i</sup> during training. As shown in Table 3, ProSAM with ResNet-50 is able to outperform all other methods, even though their performance on mIOU is already sufficiently high. This result demonstrates the extraordinary generalization capability of ProSAM under the significant domain difference between training and testing.

Table 3. The generalization evaluation under the domain shift from COCO-20<sup>i</sup> to PASCAL-5<sup>i</sup>. For all methods, mask annotations are employed as the visual reference. The average mIOU across 4 folds is reported as the evaluation metric.

Methods	Image Encoder	Mean mIOU
ProSAM	ResNet-50	<b>77.65</b>
VRP-SAM [32]		76.44
RPMM [40]		49.6
PFENet [33]	ResNet-50	61.1
RePRI [4]		63.2
VAT-HM [24]		65.1
HSNet [23]	ResNet-101	64.1
DGPNNet [11]	ResNet-101	70.1
FP-Trans [43]	DeiT-B/16	69.7

### 5.3. Verification Study

As described in Section 4.2, the primary motivation behind the proposed variational prompt encoder is to generate more robust prompts that lie away from the boundaries of the target prompt region  $\mathcal{R}_{I_r, M_r, I_t}$ . To empirically validate this, we conducted two complementary studies to assess the robustness of prompts generated by ProSAM. Due to the page limit, additional verification studies are included

in Appendix Section 9.

**Robustness to Noise Perturbation.** If our generated prompts truly stay away from the boundaries of the target prompt region, then a small perturbation applied in the latent prompt embedding space should result in minimal degradation in mask quality. To evaluate this, we injected Gaussian noise into the prompt embeddings of both ProSAM and VRP-SAM during inference and analyzed their robustness. As shown in Figure 4, when the standard deviation of Gaussian noise is set to 1.2, the mIOU of VRP-SAM degrades by approximately 40%, while ProSAM only shows a 20% drop. Overall, VRP-SAM exhibits a much greater sensitivity to noise, suggesting that its prompts tend to lie near the boundaries of target prompt regions, whereas ProSAM prompts remain significantly more stable under perturbations, providing evidence that they are located closer to the center of the target prompt region.

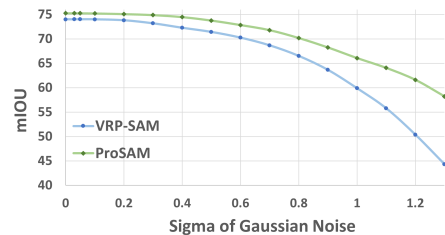


Figure 4. Gaussian noise perturbation on ProSAM prompts and VRP-SAM prompts in the latent space. Both ProSAM model and the VRP-SAM model are trained on PASCAL-5<sup>i</sup> F-0, and the noise perturbation is injected during inference.

**Proximity to Center Prompts.** We further assess prompt robustness by measuring the similarity between the predicted prompts and center point prompts in the ground-truth masks. Since the center prompt of the target prompt regions is unobservable, we approximate it using the latent prompt embedding of the center point prompt in the ground-truth masks because those are the most robust prompts we can obtain. Specifically, we identified the center pixels in the ground-truth masks and fed them into the SAM prompt encoder to get their corresponding prompt embeddings. Then, for each fold in PASCAL-5<sup>i</sup>, we computed the cosine similarity between the predicted prompts and the embeddings of the center point prompts. As shown in Table 4, ProSAM prompts consistently exhibit higher cosine similarity to the center point prompts than VRP-SAM prompts. This observation further supports our hypothesis that ProSAM generates prompts more aligned with the central and robust regions of the target prompt region. Additionally, we evaluated the statistical relationship between cosine similarity and segmentation quality by computing the Pearson correlation coefficient between cosine similarity and mIOU. A positive correlation coefficient of 0.365 suggests that prompts closer to the center point prompt are

more likely to yield high-quality segmentation results.

Table 4. The average cosine similarity between predicted prompts and the embeddings of center point prompts in ground-truth mask.

Cosine Similarity to Center Prompts	Methods	
	VRP-SAM	ProSAM
F-0	0.0072	<b>0.0122</b>
F-1	0.0127	<b>0.0291</b>
F-2	-0.0040	<b>0.0341</b>
F-3	0.0106	<b>0.0420</b>

## 5.4. Ablation Study

To thoroughly evaluate the effectiveness of different components of ProSAM, we conduct ablation studies on three different aspects: formulations of prompt distribution, inference strategies, and choices of image encoder. Due to the page limit, our comparison against VRP-SAM with the same number of learnable parameters and more choices of image encoder, is presented in Appendix Section 10.

**Formulations of Prompt Distribution.** As discussed in Section 4.2, a straightforward approach to modeling the prompt distribution  $P(z|I_r, M_r, I_t)$  is to assume it follows a multivariate Gaussian distribution. Consequently, we perform an ablation study under this assumption, applying the reparameterization trick inspired by VAE [13]. As shown in Table 5, with different  $K$  and  $\nu$ , ProSAM with Gaussian prompt distributions consistently under-perform compared to t-distributions. This result highlights that the heavy-tailed nature of the t-distribution encourages a more robust mean prompt by enforcing a larger 4th-order curvature penalty, which in turn significantly enhances our testing performance on novel classes.

Table 5. Ablation study on Gaussian prompt distribution.  $K$  denotes the number of Monte-Carlo samples, and  $\nu$  represents the degrees of freedom of multivariate t-distribution.

Method	Parameters		PASCAL-5 <sup>i</sup> Means	Method	Parameters		PASCAL-5 <sup>i</sup> Means
	K	$\nu$			K	$\nu$	
ProSAM	10	5	<b>72.04</b>	ProSAM w/ Gaussian	10	-	71.41
	10	3	71.55		15	-	71.01
	15	5	71.42				
	15	3	71.02				

**Choices of Image Encoder.** Image encoder, as part of the prompt encoder, plays an important role in generating accurate prompt embeddings. A more powerful image encoder can generate image embeddings with more accurate semantics, hence leading to better prompt embeddings. To evaluate the generalizability of our method, we use a powerful self-supervised image encoder, the pre-trained DINOv2 [26] ViT-B to extract visual features. As shown in Table 6, we can see that DINOv2 indeed can significantly improve the mIOU for both VRP-SAM and our method, compared with ResNet-50. More importantly, we achieve a higher mIOU compared to VRP-SAM when both meth-

ods utilize DINOv2 as the image encoder. This result highlights the strong generalization capability of our variational prompt encoder across different image encoders.

Table 6. Ablation study on different image encoders.

Methods	Image Encoder	COCO-20 <sup>i</sup>				
		F-0	F-1	F-2	F-3	Mean
VRP-SAM	ResNet-50	47.02	54.24	59.91	51.95	53.28
	DINOv2 ViT-B/14	53.68	59.74	60.24	58.96	58.15
ProSAM	ResNet-50	48.74	55.55	60.72	53.49	54.63
	DINOv2 ViT-B/14	<b>54.49</b>	<b>60.57</b>	<b>61.81</b>	<b>59.8</b>	<b>59.16</b>

**Inference Strategies.** During inference, the predicted mean prompt  $\hat{\mu}_z$  is employed to generate the predicted mask, since the robustness of  $\hat{\mu}_z$  can be guaranteed due to the existence of margin between  $\hat{\mu}_z$  and the boundary of target prompt region  $\mathcal{R}_{I_r, M_r, I_t}$ . However, it is still interesting to see the effectiveness of utilizing randomly sampled prompts from the learned prompt distribution during inference. Therefore, we experimented with 5 different inference strategies, in which we sample  $K$  prompts and merge their corresponding  $K$  masks in 5 different ways. Specifically, they merge either the logit masks or the binary masks given a threshold of 0.5. From Table 7, all the inference strategies with different ways of merging  $K$  masks achieve comparable results to using only the mask prompted by the  $\hat{\mu}_z$ . This indicates that our learned multivariate prompt distribution spans over the target prompt region.

Table 7. Ablation study on different inference strategies. Their mIOU on F-0 of PASCAL-5<sup>i</sup> is presented.

Methods	Inference Strategies	mIOU
ProSAM	mean prompt only	<b>75.26</b>
ProSAM	max of $K$ logit masks	74.69
	mean of $K$ logit masks	75.16
	max of $K$ binary masks	74.67
	mean of $K$ binary masks	74.92
	majority vote of $K$ binary masks	75.05

## 6. Conclusion

This paper presents ProSAM, a novel probabilistic prompt generation method that significantly enhances the robustness and zero-shot segmentation capability of SAM-based visual reference segmentation methods. By introducing a variational prompt encoder to learn a multivariate prompt distribution, we address the shortcomings of less robust prompts in existing approaches and thus consistently generate stable, high-quality masks. Through comprehensive experiments, ProSAM demonstrates superior performance over SOTA methods on the Pascal-5<sup>i</sup> and COCO-20<sup>i</sup> datasets, highlighting its potential as a reliable and effective solution for visual reference segmentation. Our findings underscore the importance of probabilistic prompt generation approaches in prompt-based segmentation and pave the way for future research in this domain.

## References

- [1] Najmeh Abiri and Mattias Ohlsson. Variational auto-encoders with student's t-prior. *arXiv preprint arXiv:2004.02581*, 2020. 5
- [2] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018. 4
- [3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 2
- [4] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proc. CVPR*, pages 13979–13988, 2021. 7
- [5] Peng Ding. On the conditional distribution of the multivariate t distribution. *The American Statistician*, 70(3):293–295, 2016. 4
- [6] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *Proc. ECCV*, pages 701–719, 2022. 6
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 3, 6
- [8] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. CLIP-S4: Language-guided self-supervised semantic segmentation. In *Proc. CVPR*, pages 11207–11216, 2023. 1
- [9] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *Proc. ECCV*, pages 108–126, 2022. 6
- [10] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-Rex2: Towards generic object detection via text-visual prompt synergy. In *Proc. ECCV*, pages 38–57, 2024. 1, 2
- [11] Joakim Johnander, Johan Edstedt, Michael Felsberg, Fahad Shahbaz Khan, and Martin Danelljan. Dense Gaussian processes for few-shot segmentation. In *Proc. ECCV*, pages 217–234. Springer, 2022. 7
- [12] Juno Kim, Jaehyuk Kwon, Mincheol Cho, Hyunjong Lee, and Joong-Ho Won.  $t^3$ -variational autoencoder: Learning heavy-tailed data with student's t and power divergence. In *Proc. ICLR*, 2024. 5
- [13] Diederik P Kingma. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 8
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proc. ICCV*, pages 4015–4026, 2023. 1
- [15] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proc. CVPR*, pages 8057–8067, 2022. 6
- [16] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proc. CVPR*, pages 8057–8067, 2022. 5
- [17] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou. Deep variational metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 689–704, 2018. 2
- [18] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proc. CVPR*, pages 11553–11562, 2022. 6
- [19] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *Proc. CVPR*, pages 11573–11582, 2022. 6
- [20] 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. 1, 2, 5, 6
- [21] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6
- [23] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proc. ICCV*, pages 6941–6952, 2021. 6, 7
- [24] Seonghyeon Moon, Samuel S Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Muhammad Haris Khan, and Mubbasir Kapadia. HM: Hybrid masking for few-shot segmentation. In *Proc. ECCV*, pages 506–523, 2022. 7
- [25] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proc. ICCV*, pages 622–631, 2019. 5
- [26] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8, 6
- [27] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proc. CVPR*, pages 23641–23651, 2023. 6
- [28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1
- [29] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 5
- [30] 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 *Proc. ECCV*, pages 151–168, 2022. 6, 7
- [31] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [32] Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. VRP-SAM: Sam with visual reference prompt. In *Proc. CVPR*, pages 23565–23574, 2024. 1, 2, 3, 5, 6, 7
- [33] 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. 6, 7
- [34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 4
- [35] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proc. CVPR*, pages 6830–6839, 2023. 5, 6
- [36] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *Proc. ICCV*, pages 1130–1140, 2023. 5, 6
- [37] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, and Liu Ren. USE: Universal segment embeddings for open-vocabulary image segmentation. In *Proc. CVPR*, pages 4187–4196, 2024. 1
- [38] Haoyu Xie, Changqi Wang, Mingkai Zheng, Minjing Dong, Shan You, Chong Fu, and Chang Xu. Boosting semi-supervised semantic segmentation with probabilistic representations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2938–2946, 2023. 2
- [39] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proc. CVPR*, pages 2945–2954, 2023. 1
- [40] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *Proc. ECCV*, pages 763–778, 2020. 7
- [41] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 552–561, 2019. 2
- [42] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Proc. NeurIPS*, 34:21984–21996, 2021. 6
- [43] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Proc. NeurIPS*, 34:21984–21996, 2021. 7
- [44] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Proc. NeurIPS*, 34:21984–21996, 2021. 5
- [45] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1, 2, 5, 6
- [46] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Proc. NeurIPS*, 36, 2024. 6