VRP-SAM: SAM with Visual Reference Prompt

Yanpeng Sun[1,2*], Jiahui Chen[2,3*], 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

Abstract

*In this paper, we propose a novel Visual Reference Prompt (VRP) encoder that empowers the Segment Anything Model (SAM) to utilize annotated reference images as prompts for segmentation, creating the VRP-SAM model. In essence, VRP-SAM can utilize annotated reference images to comprehend specific objects and perform segmentation of specific objects in target image. It is note that the VRP encoder can support a variety of annotation formats for reference images, including **point**, **box**, **scribble**, and **mask**. VRP-SAM achieves a breakthrough within the SAM framework by extending its versatility and applicability while preserving SAM's inherent strengths, thus enhancing user-friendliness. To enhance the generalization ability of VRP-SAM, the VRP encoder adopts a meta-learning strategy. To validate the effectiveness of VRP-SAM, we conducted extensive empirical studies on the Pascal and COCO datasets. Remarkably, VRP-SAM achieved state-of-the-art performance in visual reference segmentation with minimal learnable parameters. Furthermore, VRP-SAM demonstrates strong generalization capabilities, allowing it to perform segmentation of unseen objects and enabling cross-domain segmentation. The source code and models will be available at https://github.com/syp2ysy/VRP-SAM*

1. Introduction

In recent, the Segment Anything Model (SAM) [13] has emerged as a foundational visual model for image segmentation. Trained on an extensive datasets comprising billions of labels, SAM has demonstrated remarkable versatility in the realm of universal segmentation. What sets SAM apart is its human-interactive design, allowing segmentation based on user-provided prompts, be they in the form of points, bounding boxes, or coarse masks. This distinctive feature positions SAM as a robust tool that can be adaptable to various tasks and requirements [12, 48].

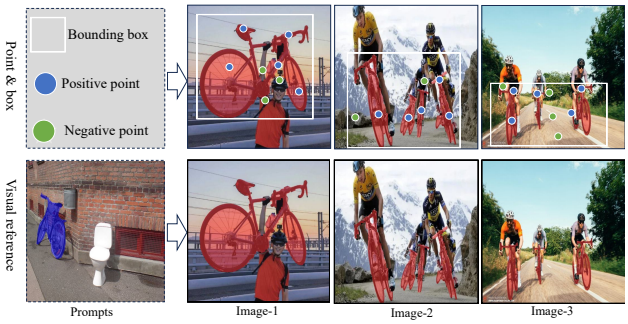

Figure 1. Comparison of SAM's built-in Point and Box prompt modes with Visual Reference Prompt when handling numerous images. The prompts are all provided by users.

However, the existing prompt formats of SAM present significant challenges in practical applications, especially when dealing with complex scenes and numerous images. As shown in Figure 2(a), SAM relies on user-provided prompts (points, boxes, coarse mask) to segment objects in the target image, demanding users to possess a comprehensive understanding of the target objects. In real-world applications, especially in complex scenarios, the level of user familiarity with the target objects can significantly impact the effectiveness of providing specific prompts. Furthermore, variations in the position, size, and quantity of target objects across different images require custom prompts for each image. As illustrated in Figure 1 with the aim of segmenting *'bicycles'*, users need to customize different prompts for each image, significantly impacting efficiency of SAM. Therefore, we propose integrating visual reference prompts to overcome these limitations and enhance adaptability of SAM.

Visual reference prompts is annotated reference images that delineate the objects users expect to segment. As shown in Figure 1, by simply providing a visual reference prompt

*equal contribution
†Corresponding author.

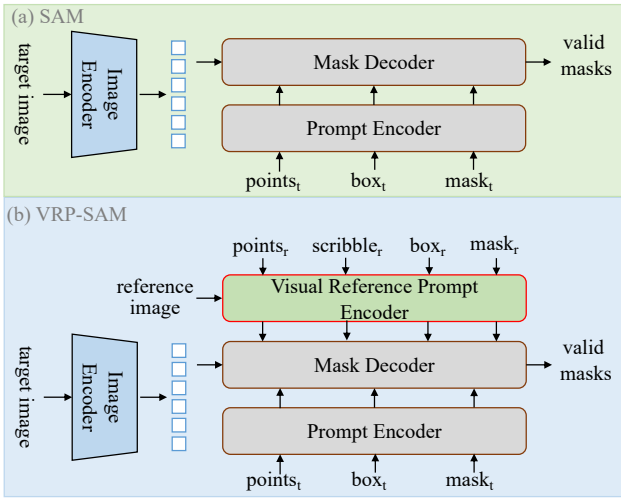

Figure 2. A Comparison between SAM and VRP-SAM. VRP-SAM introduces a visual reference prompt encoder, accepting annotated reference images with **point**, **scribble**, **box**, and **mask** formats, offering a distinct enhancement over SAM.

with *bicycle*, we can segment *bicycle* in different images without requiring users to provide specific prompts for each image. It significantly enhances the efficiency of SAM while reducing reliance on user familiarity with objects. To achieve this goal, some methods [19, 46] incorporate semantic correlation models [9, 25] to establish reference-target correlation [1], obtaining pseudo-masks for the target objects. Following this, a sampling strategy is devised to extract a set of points and bounding boxes from the pseudo-masks, serving as prompts for SAM segment the target image. These methods overlook false positives within the pseudo-mask and exhibit high sensitivity to hyperparameters. Consequently, it heavily relies on the quality of the pseudo-mask and has poor generalization.

Toward this end, we propose a straightforward and effective Visual Reference Prompt (VRP) encoder using meta-learning technique, integrated with SAM to create VRP-SAM. It leverages annotated reference images as prompts to segment similar semantic objects in the target image. As illustrated in Figure 2(b), the VRP encoder accepts inputs in various annotation formats, including points, scribbles, boxes, and masks. Specifically, the VRP encoder introduces a semantic-related model to encode reference and target images into the same space. Following meta-learning methods, we first extract prototypes of target objects from annotated information in reference images, enhancing the representation of the target objects in both images. Following meta-learning methods, the prototypes of target objects (users' markers) are first generated from annotated reference images, which aim to highlight such target instances in both reference and target images. Then, we introduce a set of learnable queries to extract the semantic cues of the target objects from attentive enhanced reference features. Then, these queries interact with target images, generating prompt embeddings usable by mask decoder to segment semantically specific objects in the target image. Building upon the foundation of SAM's inherent capabilities, VRP-SAM enhances the model's visual reference segmentation prowess. The introduction of Visual Reference Prompts (VRP) not only diversifies the prompts, enabling the model to swiftly segment objects with identical semantics, but also incorporates a meta-learning mechanism that significantly boosts the generalization of model, particularly in dealing with novel objects and cross-domain scenarios.

To quantitatively assess the generalization capability of VRP-SAM, we follow the dataset configurations commonly used in few-shot segmentation and evaluate our VRP-SAM on Pascal-5^i and COCO-$20i$ datasets. Extensive experiments show that VRP-SAM overcomes the limitations of SAM in prompt format, enabling efficient visual reference segmentation. It is significantly better than sampling-based methods, achieving state-of-the-art results on Pascal-5^i and COCO-$20i$ datasets. Moreover, we present solid evidence of VRP-SAM's superior performance in handling unknown objects and cross-domain scenarios. Our experiments highlight that VRP-SAM achieves the rapid segmentation of a large number of images based on semantics. Furthermore, the incorporation of meta-learning principles significantly enhances its generalization, making it applicable across various scenarios.

2. Related Work

2.1. Application of SAM

Segmentation Anything model (SAM) [13] is a recently introduced category-agnostic interactive segmentation model by Meta. It leverages user-guided instructions for segmentation, such as points, bounding boxes, and coarse masks. Currently, SAM has two primary application approaches. One involves using SAM's segmentation results as prior information to assist downstream tasks. For instance, Inpaint Anything (IA) [43] and Edit Everything [39] leverage SAM's segmentation results for image editing and restoration [42, 47] within the masked regions. Additionally, SEPL [5] employs SAM's segmentation results to enhance pseudo-labels in weakly supervised segmentation tasks [27, 30]. Furthermore, SAM segmentation results serve as prior information in tasks such as crack and volcano crater detection [1, 8].

The second involves guiding SAM's segmentation through various prompt combinations. For example, in the context of visual reference segmentation [31], Matcher [19] samples points and boxes from pseudo-masks, utilizing

[1] SAM, being a category-agnostic model, encounters challenges in adequately capturing reference-target correlations.

SAM to refine these pseudo-masks. TAM [41] is applied in object tracking tasks, where it uses point prompts to initialize object masks and subsequently uses SAM to refine low-quality masks. Additionally, SAMAug [6] introduces a visual point enhancement method tailored for SAM, facilitating automatic image annotation. These examples illustrate SAM's effectiveness across different tasks. However, existing methods are constrained by SAM's existing prompt modalities, and they may struggle when confronted with complex objects and unfamiliar scenes. To break these limitations, we have designed a visual prompt encoder for SAM, expanding its applicability to a wider range of scenarios.

2.2. Visual reference segmentation

Visual reference segmentation [31, 33, 48] aims to guide the segmentation of a target image using a reference image. The goal of this task is to utilize a semantically annotated reference image to instruct the segmentation of objects or regions in the target image that share the same semantics as those in the reference image. In current research, methods can be broadly categorized into two groups: prototype-based and feature-matching-based. Prototype-based methods, such as PFENet [33], PANet [36], and CWT [21], usually focus on distinguishing prototypes with different class-specific features. ASGNet [15], on the other hand, improves the segmentation performance by increasing the number of prototypes. Another feature-matching approach [22, 35] leverages the pixel-level correlations between reference and target images to significantly enhance segmentation performance, as demonstrated by methods like CyCTR [44] and HDMNet [26]. Moreover, modern large-scale vision models [2, 37, 48] have recognized visual reference segmentation as a primary task, given its indispensable role in handling complex objects and unknown scenes. However, it's important to note that SAM [13] does not possess the capability to perform this task, underscoring the necessity of introducing VPR-SAM.

3. Preliminary

In this section, we first review the architecture of Segment Anything Model. Then we formulate the problem setting in this paper.

3.1. SAM Architecture

SAM is an interactive segmentation model composed of an image encoder, a prompt encoder, and a mask decoder. Given a target image I_t and some geometric prompts, SAM first employs a Vision Transformer (ViT) [34] as image encoder to extract image features. Subsequently, the prompt encoder is utilized to generate the prompt embeddings derived from the user-provided points, boxes, or masks. Finally, the mask decoder integrates the image features and prompt embeddings, and generates a high-quality mask in a class-agnostic manner.

To the best of our knowledge, current research based on SAM still relies on geometric prompts for segmentation. Thus far, there have been no efforts to introduce new forms of prompts to enhance SAM. We design a novel visual reference prompt to extend the visual reference segmentation capabilities of SAM.

3.2. Problem Formulation

Visual reference segmentation aims to segment objects in target image that same semantic as the annotated object in reference image. Specifically, let I_t represent the target image and I_r represent the reference image. Given a reference image I_r and its annotation M_r^i, where i denotes the category of the annotated object, we leverage this information to segment all regions in the target image belonging to category i. Depending on the annotation granularity, the reference image have four annotation formats: point, scribble, box, and mask. Point annotation involves providing specific points on the target, scribble requires outlining the target region with arbitrary curves, boxes correspond to bounding boxes around the target, and masks provide pixel-level annotations for the entire target.

4. VRP-SAM

VRP-SAM extends SAM to perform visual reference segmentation without compromising its original functionality. We propose a training-efficient visual reference prompt encoder that, firstly, accommodates various granularities of visual references and, secondly, directly encodes these visual references into prompt embeddings rather than geometric prompts. These prompt embeddings are then fed directly into the mask decoder of SAM, resulting in the generation of the target mask.

As depicted in Figure 3, the VRP Encoder consists of feature augmenter and prompt generator. Next, we will delve into the details of VRP Encoder and the loss function.

4.1. Feature Augmenter

Inspired by meta-learning, the Feature Augmenter separately encodes reference annotations M_r^i into features of both the reference and target images. This process is designed to distinguish between foreground and background representations. To capture semantic correlations between reference and target images, we introduce a semantic-aware image encoder within the VRP encoder, encoding them into the same latent space. To prevent overfitting of the VRP encoder, we freeze the image encoder during the training phase. This ensures a balance between capturing semantic relevance and preventing excessive specialization of the VRP encoder.

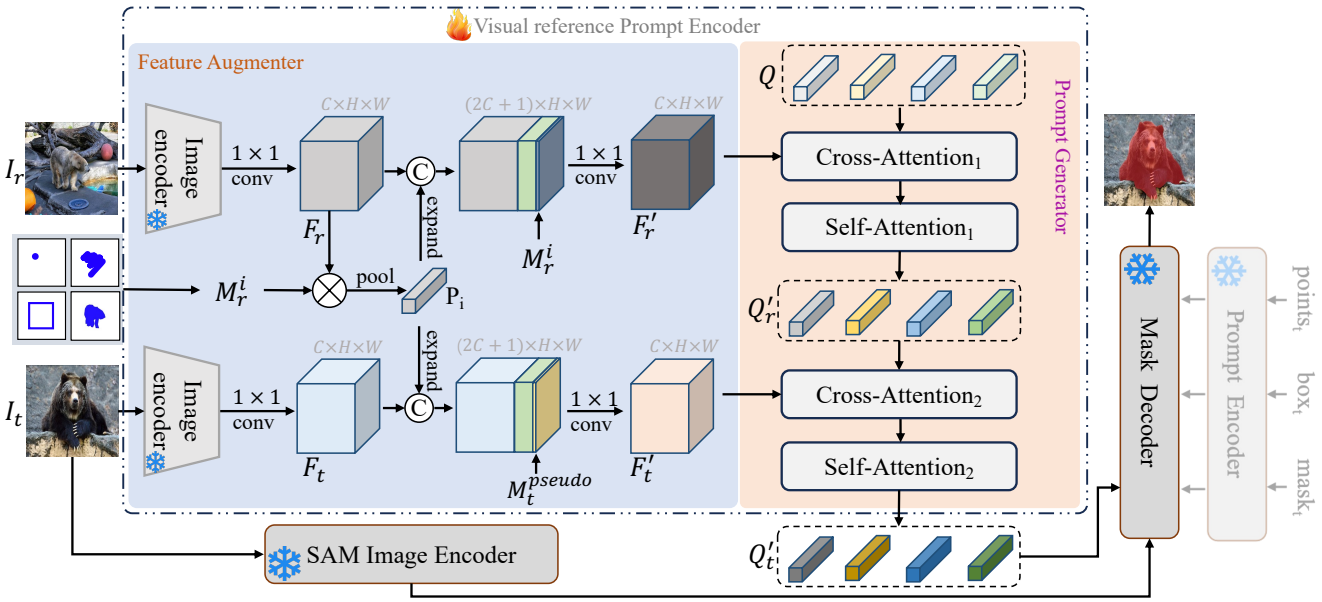

Figure 3. Proposed VRP-SAM framework. Our approach enables SAM to perform visual reference segmentation by extends a VRP encoder. It takes various granularities of visual references as inputs and encodes these visual references into prompt embeddings. Our VRP encoder consists of a feature augmenter and a prompt generator.

The Feature Augmenter is illustrated in Figure 3. Initially, leveraging a semantic-aware image encoder (e.g. ResNet-50), we encode I_r and I_t separately, resulting in the feature map $F_r \in \mathbf{R}^{C\times H\times W}$ and $F_t \in \mathbf{R}^{C\times H\times W}$. Subsequently, we extract the prototype feature P_i corresponding to class i from F_r using the mask M_r^i. This process can be summarized as follows:

$$P_i = MaskAvgPool(F_r, M_r^i) \tag{1}$$

where M_r^i represents one of following annotation formats: **point**, **scribble**, **box**, or **mask**. To enhance the context information about class i, we concatenate the prototype features and mask with F_r and F_t. F_r is concatenated with mask m_i, and F_t is concatenated with pseudo-mask m_i^{pseudo}. m_i^{pseudo} is obtained using a common training-free approach, and detailed descriptions are provided in the Appendix. Subsequently, we employ a shared 1×1 convolution layer to reduce dimensionality of enhanced features.This process can be summarized as follows:

$$F_r^{'} = Conv(concat(F_r, P_i, m_i)) \tag{2}$$

$$F_t^{'} = Conv(concat(F_t, P_i, m_i^{\text{pseudo}})) \tag{3}$$

Ultimately, we obtain enhanced image features $F_r^{'} \in \mathbf{R}^{C\times H\times W}$ and $F_t^{'} \in \mathbf{R}^{C\times H\times W}$, which have enhanced context information for the category i. Thus, the enhanced features comprise foreground representations for class i and background representations for the other classes. Subsequently, we feed the enhanced features into the Prompt Generator to obtain a set of visual reference prompts.

4.2. Prompt Generator

The purpose of the Prompt Generator is to obtain a set of visual reference prompts embedding for the SAM mask decoder. As depicted in Figure 3, the process commences with the introduction of a set of learnable queries denoted as $Q \in \mathbf{R}^{N\times C}$, where N indicates the number of visual reference prompts. These queries first engage with the reference features $F_r^{'}$ to obtain the category-specific information through cross-attention and self-attention layer:

$$Q_r^{'} = \text{SelfAttn}_1(\text{CrossAttn}_1(Q, F_r^{'})) \tag{4}$$

Here, we obtain a set of queries, $Q_r^{'}$, possessing knowledge about the object to be segmented. Subsequently, we employ cross-attention to interact these queries with target image feature to obtain the foreground information in target image. Following this, a self-attention layer is used to update the queries, generating a set prompts $Q_t^{'}$ that align with the representation of SAM :

$$Q_t^{'} = \text{SelfAttn}_2(\text{CrossAttn}_2(Q_r^{'}, F_t^{'})) \tag{5}$$

The final $Q_t^{'}$ serve as the visual reference prompt embeddings for SAM, equipped with the capability to guide the segmentation of the foreground in the target image. By inputting this set of visual reference prompt embeddings into the mask decoder, the mask $M_t^i \in \mathbf{R}^{1\times H\times W}$ for category i in the target image can be obtained.

4.3. Loss Function

We employ Binary Cross-Entropy (BCE) loss and Dice loss to supervise the learning of the Visual Reference Prompt

Table 1. Compare with other foundation models. Results of one-shot semantic segmentation on COCO-20^i.Gray indicates the model is trained by in-domain datasets. † indicates the method using SAM.

Methods	Label type	F-0	F-1	F-2	F-3	Means
Painter [37]		31.2	35.3	33.5	32.4	33.1
SegGPT [38]		56.3	57.4	58.9	51.7	56.1
PerSAM† [46]	*mask.*	23.1	23.6	22.0	23.4	23.0
PerSAM-F† [46]		22.3	24.0	23.4	24.1	23.5
Matcher† [19]		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

Encoder. The BCE loss ensures pixel-wise accuracy, while the Dice loss provides additional context for pixel-level segmentation. Therefore, the total loss of VRP-SAM is:

$$L_{\text{total}} = \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]}_{\text{BCE Loss}} + \underbrace{1 - \frac{2\sum_{i=1}^{N}(p_i \cdot y_i)}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} y_i^2}}_{\text{Dice Loss}} \tag{6}$$

where N represents the total number of pixels, y_i is the ground truth label for pixel i, and p_i is the predicted probability of pixel i belonging to the object. By combining these two losses, we comprehensively consider both accuracy and contextual effects, thus guiding the Visual Reference Prompt Encoder more effectively in generating precise segmentation results.

5. Experiments

5.1. Setting

Datasets: To validate segmentation performance and generalization capability of VRP-SAM, we conducted extensive experiments following the few-shot setting [14, 31, 44] on COCO-20^i [24] and PASCAL-5^i [29] datasets. Specifically, we organized all classes from both datasets into 4 folds. For each fold, PASCAL-5^i [29] comprises 15 base classes for training and 5 novel classes for testing, while COCO-20^i [24] includes 60 training base classes and 20 testing novel classes. To assess performance of model, we randomly sampled 1000 reference-target pairs in each fold. In each fold, As the mentioned datasets lack labels for point, scribble, and box annotations, we followed the SEEM [48] to generate these annotation labels by simulating user inputs randomly based on the reference ground truth masks.

Implementation details: In visual reference prompt encoder, we use VGG-16 [3] and ResNet-50 [9] as the image encoder and initialize it with ImageNet [28] pre-trained weights. We employed the AdamW optimizer [20] along with a cosine learning rate decay strategy for training VRP-SAM. Specifically, on the COCO-20^i dataset, we conducted 50 epochs of training with an initial learning rate of 1e-4 and a batch size of 8. For the PASCAL-5^i dataset, the model was trained for 100 epochs with an initial learning rate of 2e-4 and a batch size of 8. In VRP, the number of queries is set to 50 by default, and the input image size of all experiments needs to be adjusted to 512×512. Following SEEM [48], during training, we obtain annotations for points, scribbles, and boxes based on mask annotations. We provide detailed descriptions of this process in the Appendix. To ensure a fair comparison, VRP-SAM is exclusively compared to previous works [16, 26, 37] using visual reference prompts based on the the mean intersection over union (mIoU).

5.2. Comparison with the State-of-the-art

Comparison with other foundation models. Leveraging the foundation model enables few-shot segmentation. We compared various approaches utilizing Painter [37], SegGPT [38], and SAM as foundation models for few-shot segmentation, providing evaluation metrics on the COCO-20^i dataset. As shown in Table 1, VRP-SAM achieves 53.9% mean mIoU without training on novel classes, achieving comparable with SegGPT. Note that the training data of SegGPT include all classes in COCO. Furthermore, our VRP-SAM outperforms other SAM-base methods including Matcher [19], which employs a DINOv2 [25] pretrained ViT-L [34] as image encoder.

Comparison with few-shot methods. To validate the effectiveness of VRP-SAM, we compared it with state-of-the-art few-shot segmentors on the novel set of COCO-20^i and PASCAL-5^i datasets. To ensure a fair comparison, VRP-SAM is trained and tested separately in each fold, ensuring no overlap in classes between the training and test sets. The results in Table 2 demonstrate that VRP-SAM achieves state-of-the-art results on COCO-20^i and PASCAL-5^i, with mIoU scores of 53.9 and 71.9, respectively. Notably, when using VGG-16 as the image encoder for VRP, VRP-SAM achieves mIoU scores of 48.0 and 68.7 on COCO-20^i and PASCAL-5^i datasets. It indicates that VRP-SAM achieves optimal performance on the novel set with only 1.6M learnable parameters, highlighting its powerful generalization capability.

5.3. Comparison with Geometric Prompts

In this paper, we design Visual Reference Prompts to represent target information. Different from the Geometric Prompts (GP) provided by SAM, our VRP is more flexible and robust. We conducted experiments comparing GP and VRP to validate the superiority of our method (see Table 3).

Table 2. Performance of one-shot semantic segmentation on COCO-20^i and PASCAL-5^i. The red and blue colors respectively represent the optimal and suboptimal results.

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 [33]	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 [26]		4.2M	40.7	50.6	48.2	44.0	45.9	64.8	71.4	67.7	56.4	65.1
VRP-SAM		1.6M	43.6	51.7	50.0	46.5	48.0	70.0	74.7	68.3	61.9	68.7
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 [22]		2.6M	36.3	43.1	38.7	38.7	39.2	64.3	70.7	60.3	60.5	64.0
CyCTR [44]		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 [18]		19.9M	36.8	42.6	39.9	37.9	39.3	65.4	72.3	59.4	59.8	64.2
DPCN [17]		-	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 [26]		4.2M	43.8	55.3	51.6	49.4	50.0	71.0	75.4	68.9	62.1	69.4
VRP-SAM		1.6M	48.1	55.8	60.0	51.6	53.9	73.9	78.3	70.6	65.0	71.9
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

Table 3. Comparison with geometric prompts. Geometric prompts are randomly sampled from the pseudo-mask. † indicates the carefully designed sampling strategy proposed in [19].

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.* † [19]	52.7
VRP-SAM	ResNet-50	*point.*	38.4
		scribble.	47.3
		box.	49.7
		mask.	53.9
	DINOv2-B/14	*mask.*	60.4

In the GP-SAM experiments, we initially used an Image Encoder to generate a pseudo-mask for the target image following [33], and subsequently obtained points or bounding boxes from the pseudo-mask [2] as geometric prompts. Experimental results consistently demonstrate the ongoing superiority of our VRP over the GP approach. We visualize the segmentation results of GP-SAM and our VRP-SAM in Figure 4. Visualization results indicate that the GP approach is prone to generating false-positive prompts, significantly impacting segmentation performance. In contrast, our VRP effectively avoids such issues.

5.4. Generalization Evaluation

Domain shift: Next, we assess the effectiveness of VRP-SAM in a domain shift scenario, which necessitates significant domain differences between training and testing sets. Following prior work [22, 32, 33], we trained on COCO-20^i and tested on PASCAL-5^i, where the classes in training set do not overlap with those in test set. The results in Table 4, representing the average across four folds, clearly demonstrate the superior performance of VRP-SAM in domain shift scenario. Notably, VRP-SAM with scribble annotations outperforms FP-Trans [45] by 2.8%, while VRP-SAM with mask annotations surpasses DGPNet [11] by 5.8%. This affirms the robust generalization capability of VRP-SAM and its effectiveness in domain transfer scenarios.

Visualization: To assess the generalization capability of VRP-SAM across diverse image styles, we curated a collection of target images from web, spanning various gen-

Figure 4. The visualization results of VRP-SAM and GP-SAM.

[2]For point prompts, we randomly sample 5 points from the pseudo-mask. For box prompt, we adopt the bounding box of the pseudo-mask.

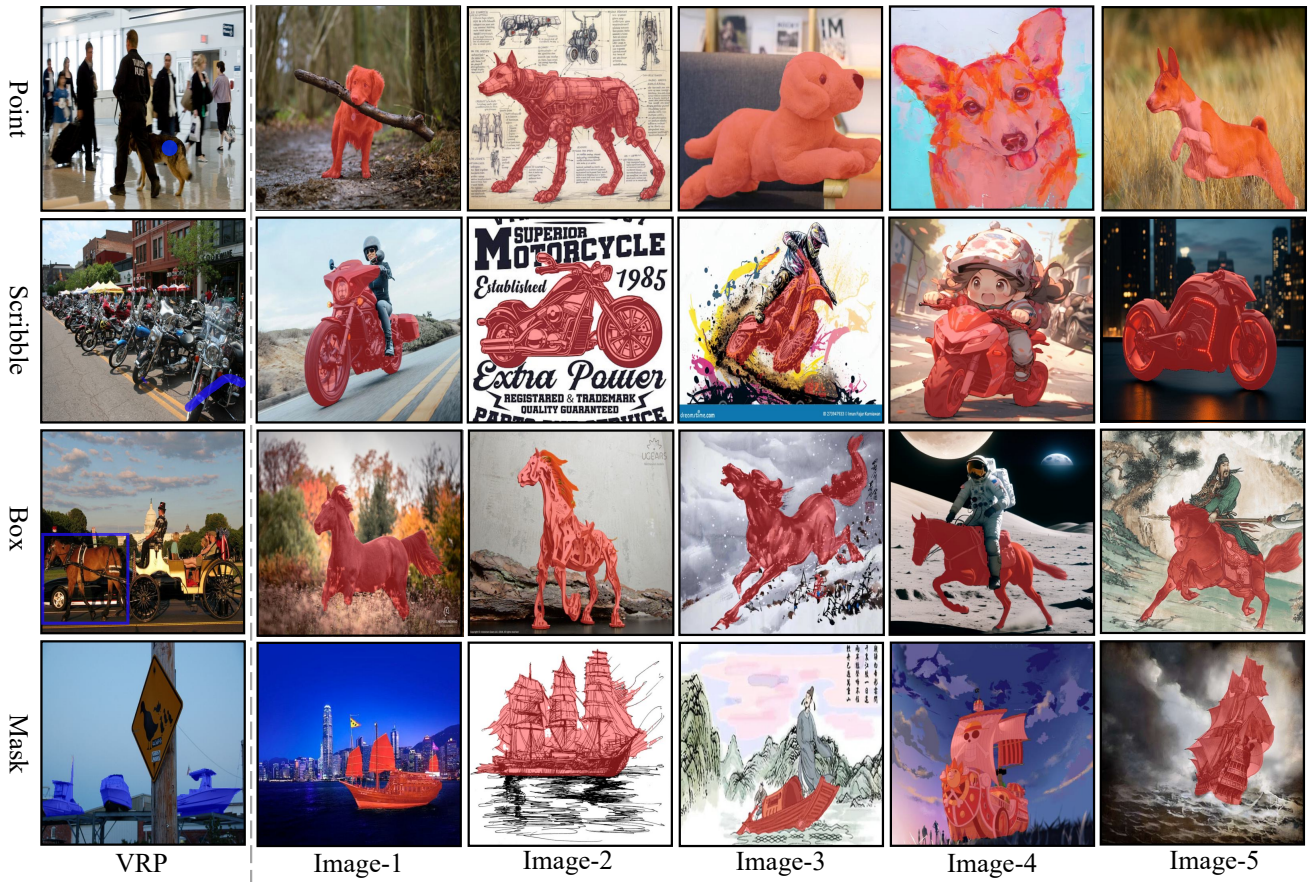

Figure 5. Qualitative results of VRP-SAM across diverse image styles is presented. The target images were sourced from the internet.

res such as natural landscapes, artworks, and complex environments. Visual Reference Prompts (VRPs) were selected from the COCO dataset to guide the segmentation of these target images. The segmentation results on different image styles are presented in Figure 5. It demonstrate that VRP-SAM adeptly adapts to varied image styles, accurately delineating target objects. Notably, VRP-SAM excels in handling ink-style images, showcasing refined segmentation performance. This compellingly underscores the robustness and effectiveness of VRP-SAM when confronted with images of unknown styles.

5.5. Ablation Study

To validate the effectiveness of VRP-SAM, we conducted extensive ablation studies on PASCAL-5^i. ResNet-50 was chosen as the image encoder for VRP to ensure experimental consistency. These experiments aimed to thoroughly investigate the performance of VRP-SAM under various conditions.

Loss: To assess the impact of Binary Cross-Entropy (BCE) and Dice losses on VRP-SAM, experiments were conducted on the PASCAL-5^i dataset. Results in Table 5 indicate comparable performance when using BCE or Dice loss individually. Notably, the optimal performance for VRP-SAM is achieved when both losses are combined. It emphasizes the importance of BCE and Dice losses in guiding VRP-SAM to produce more robust and accurate masks. BCE loss ensures precise pixel-level classification, while Dice loss facilitates accurate spatial localization of objects. Simultaneously employing BCE and Dice losses in VRP-SAM maximizes their complementary advantages.

Table 4. Evaluation (Mean IoU (%)) under the domain shift from COCO-20^i to PASCAL-5^i.

Method	Image encoder	Label type	Mean
RPMM [40]	ResNet-50	*mask.*	49.6
PFENet [33]			61.1
RePRI [4]			63.2
VAT-HM [23]			65.1
HSNet [22]	ResNet-101	*mask.*	64.1
DGPNet [11]			70.1
FP-Trans [45]	DeiT-B/16	*mask.*	69.7
VRP-SAM	ResNet-50	*point.*	63.5
		scribble.	72.5
		box.	72.3
		mask.	75.9

The number of query: In Figure 6, we conducted a comprehensive analysis of the impact of varying query quantities on VRP-SAM's performance. We observed

Table 5. Ablation study on different loss function in VRP-SAM.

Method	Label type	Bce loss	Dice loss	Means
VRP-SAM	*mask.*	✓	✗	70.1
		✗	✓	70.3
		✓	✓	**71.9**

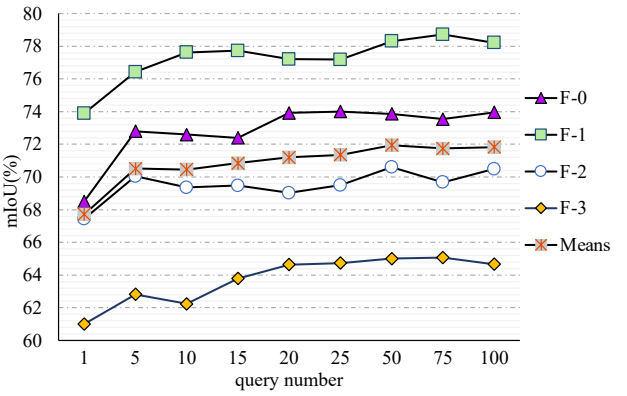

Figure 6. Ablation study on VRP-SAM with different query numbers. The x-axis shows the number of queries, and the y-axis represents model performance.

a positive correlation between increased query numbers and improved segmentation quality. However, once the query count surpassed 50, the performance gain started to diminish. This phenomenon suggests that 50 queries provide ample effective guidance for the model, and further increases in query count do not yield significant improvements, leading to performance saturation. To maintain high performance while minimizing model learnable parameters, we set the query quantity in VRP-SAM to 50. This decision optimally balances guidance information and model efficiency.

Initialization of query: In Table 6, we compared various query initialization strategies, such as random initialization, foreground prototype (FP), background prototype (BP), and a hybrid prototype with half foreground and half background (half-FP & half-BP). Surprisingly, random initialization outperformed all other strategies, showcasing superior performance. This unexpected outcome can be attributed to the intricate segmentation task and the dataset's object diversity. Random initialization enables the model to explore a broader range of states, avoiding potential local minima. In contrast, prototype-based initializations, whether foreground, background, or a combination, might introduce biases, limiting adaptability to diverse object characteristics.

Number of VRPs: In our investigation of the influence of number of visual reference prompts on segmentation results, we explored utilization of few visual reference prompts. This approach involves sending several visual reference prompts to the VRP encoder, generating multiple prompt embeddings. These embeddings are concatenated and forwarded to the mask decoder, resulting in the final mask. The result in Table 7, we observed a substantial enhancement in segmentation performance with an increase in the number of visual reference prompts. The results emphasize positive influence of incorporating multiple visual reference prompts, enhancing the model's ability to accurately capture intricate details and nuances associated with target object. The diverse prompts contribute enriched contextual information, facilitating a more comprehensive understanding of the object's characteristics and leading to generation of precise and nuanced segmentation masks.

Table 6. Ablation study of different query initialization methods on VRP-SAM.

Method	Label type	Image encoder	Mean IoU
VRP-SAM	*mask.*	FP	68.2
		BP	62.6
		half-FP & half-BP	67.4
		random	**71.9**

Table 7. Ablation Study on Few Visual Reference prompts for VRP-SAM.

Method	Label type	1-VRP	5-VRP
VRP-SAM	*point.*	62.9	64.1 (+1.2)
	scribble.	66.8	68.4 (+1.6)
	box.	69.4	70.5 (+1.1)
	mask.	71.9	72.9 (+1.0)

6. Conclusion

In this paper, we present VRP-SAM, an innovative extension of the SAM framework achieved through integration of a Visual Reference Prompt (VRP) encoder. This addition empowers SAM to leverage visual reference prompts for guided segmentation. The core methodology involves encoding annotated reference images through the VRP encoder, which then interacts with the target image to generate meaningful prompts for segmentation within the SAM framework. The seamless fusion of VRP encoder with SAM, resulting in VRP-SAM, enhances the model's generality and adaptability. VRP-SAM overcomes limitations posed by SAM's existing prompt formats, especially in complex scenarios and large datasets. The introduced visual reference prompts, including point, box, scribble, and mask annotations, offer a flexible solution, expanding the model's applicability. Extensive empirical studies conducted showcase VRP-SAM's state-of-the-art performance in visual reference segmentation with minimal learnable parameters. Notably, VRP-SAM demonstrates robust generalization capabilities, excelling in segmentation tasks for novel objects and cross-domain scenarios.

Acknowledge This work was partially supported by the National Key Research and Development Program of China under Grant 2022ZD0118802 and the National Natural Science Foundation of China (Grant No. U20B2064 and U21B2043).

References

[1] Mohsen Ahmadi, Ahmad Gholizadeh Lonbar, Abbas Sharifi, Ali Tarlani Beris, Mohammadsadegh Nouri, and Amir Sharifzadeh Javidi. Application of segment anything model for civil infrastructure defect assessment. *arXiv preprint arXiv:2304.12600*, 2023.

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems*, pages 25005–25017, 2022.

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

[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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021.

[5] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *arXiv preprint arXiv:2305.05803*, 2023.

[6] Haixing Dai, Chong Ma, Zhengliang Liu, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, Dajiang Zhu, Wei Liu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023.

[7] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 701–719, 2022.

[8] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. Deep learning universal crater detection using segment anything model (sam). *arXiv preprint arXiv:2304.07764*, 2023.

[9] 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.

[10] 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.

[11] Joakim Johnander, Johan Edstedt, Michael Felsberg, Fahad Shahbaz Khan, and Martin Danelljan. Dense gaussian processes for few-shot segmentation. In *European Conference on Computer Vision*, pages 217–234, 2022.

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv:2306.01567*, 2023.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[14] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *IEEE conference on computer vision and pattern recognition*, pages 8057–8067, 2022.

[15] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021.

[16] Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. Ctnet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9904–9917, 2021.

[17] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2022.

[18] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11573–11582, 2022.

[19] 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.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[21] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *IEEE International Conference on Computer Vision*, pages 8741–8750, 2021.

[22] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *IEEE international conference on computer vision*, pages 6941–6952, 2021.

[23] 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 *European Conference on Computer Vision*, pages 506–523, 2022.

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

[25] 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.

[26] 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.

[27] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

[30] Yanpeng Sun and Zechao Li. Ssa: Semantic structure aware inference for weakly pixel-wise dense predictions without cost. *arXiv preprint arXiv:2111.03392*, 2021.

[31] 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.

[32] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023.

[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.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[35] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

[36] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *IEEE international conference on computer vision*, pages 9197–9206, 2019.

[37] 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.

[38] 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.

[39] Defeng Xie, Ruichen Wang, Jian Ma, Chen Chen, Haonan Lu, Dong Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*, 2023.

[40] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 763–778, 2020.

[41] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

[42] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything models. *arXiv preprint arXiv:2306.04121*, 2023.

[43] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[44] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. In *Advances in Neural Information Processing Systems*, pages 21984–21996, 2021.

[45] 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.

[46] 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.

[47] 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.

[48] 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.