

PERSONALIZE SEGMENT ANYTHING MODEL WITH ONE SHOT

Renrui Zhang^{1,2}, Zhengkai Jiang^{3*}, Ziyu Guo^{2*}, Shilin Yan², Junting Pan¹, Xianzheng Ma²
Hao Dong⁴, Yu Qiao², Peng Gao², Hongsheng Li^{1†}

¹CUHK MMLab ²Shanghai Artificial Intelligence Laboratory

³Institute of Automation, Chinese Academy of Sciences

⁴CFCS, School of CS, Peking University

{zhangrenrui, gaopeng, guoziyu, qiaoyu}@pjlab.org.cn,
kaikaijiang.jzk@gmail.com hsliee.cuhk.edu.hk

ABSTRACT

Driven by large-data pre-training, Segment Anything Model (SAM) has been demonstrated as a powerful promptable framework, revolutionizing the segmentation field. Despite the generality, customizing SAM for specific visual concepts without man-powered prompting is under-explored, e.g., automatically segmenting your pet dog in numerous images. **In this paper, we introduce a training-free Personalization approach for SAM, termed PerSAM.** Given only one-shot data, i.e., a single image with a reference mask, we first obtain a positive-negative location prior for the target concept in new images. Then, aided by target visual semantics, we empower SAM for personalized object segmentation via two proposed techniques: target-guided attention and target-semantic prompting. In this way, we can effectively customize the general-purpose SAM for private use without any training. To further alleviate the ambiguity of segmentation scales, we present an efficient one-shot fine-tuning variant, **PerSAM-F**. Freezing the entire SAM, we introduce a scale-aware fine-tuning to aggregate multi-scale masks, which only tunes **2 parameters** within **10 seconds** for improved performance. To demonstrate our efficacy, we construct a new dataset, PerSeg, for the evaluation of personalized object segmentation, and also test our methods on various one-shot image and video segmentation benchmarks. Besides, we propose to leverage PerSAM to improve DreamBooth for personalized text-to-image synthesis. By mitigating the disturbance of training-set backgrounds, our approach showcases better target appearance generation and higher fidelity to the input text prompt. Code is released at <https://github.com/ZrrSkywalker/Personalize-SAM>.

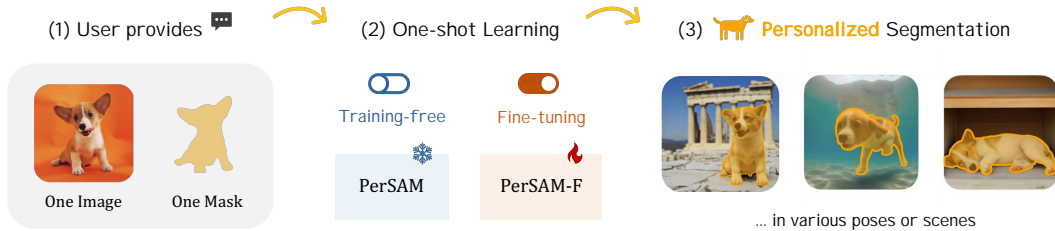


Figure 1: **Personalization of Segment Anything Model.** We customize Segment Anything Model (SAM) (Kirillov et al., 2023) for specific visual concepts, e.g., your pet dog. With only one-shot data, we introduce two efficient solutions: a training-free PerSAM, and a fine-tuning PerSAM-F.

* Equal contribution. † Corresponding author.



Figure 2: **Personalized Segmentation Examples.** Our PerSAM (Left) can segment personal objects in any context with favorable performance, and PerSAM-F (right) further alleviates the ambiguity issue by scale-aware fine-tuning.

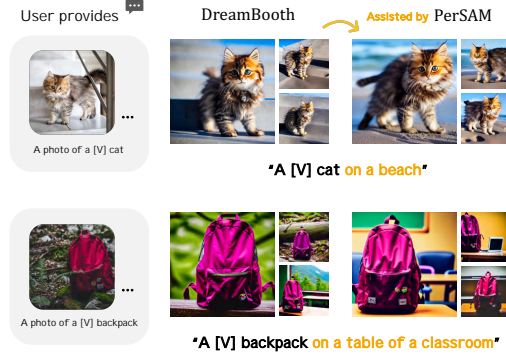


Figure 3: **Improving DreamBooth (Ruiz et al., 2022) with PerSAM.** By mitigating the disturbance of backgrounds during training, our approach can help to achieve higher-quality personalized text-to-image generation.

1 INTRODUCTION

Foundations models in vision (Li et al., 2022; Zou et al., 2023; Wang et al., 2022), language (Brown et al., 2020; Touvron et al., 2023; Radford et al., 2019), and multi-modality (Radford et al., 2021; Jia et al., 2021; Li et al., 2023) have gained unprecedented prevalence, attributed to the availability of large-scale datasets and computational resources. They demonstrate extraordinary generalization capacity in zero-shot scenarios, and display versatile interactivity incorporating human feedback. Inspired by this, Segment Anything (Kirillov et al., 2023) develops a delicate data engine for collecting 11M image-mask data, and subsequently trains a segmentation foundation model, known as SAM. It defines a novel promptable segmentation framework, i.e., taking as input a handcrafted prompt and returning the expected mask, which allows for segmenting any objects in visual contexts.

However, SAM inherently loses the capability to segment specific visual concepts. Imagine intending to crop your lovely pet dog in a thick photo album, or find the missing clock from a picture of your bedroom. Utilizing the vanilla SAM would be highly labor-intensive and time-consuming. For each image, you must precisely find the target object within complicated contexts, and then activate SAM with a proper prompt for segmentation. *Considering this, we ask: Can we personalize SAM to automatically segment user-designated visual concepts in a simple and efficient manner?*

To this end, we introduce **PerSAM**, a training-free personalization approach for Segment Anything Model. As shown in Figure 1, our method efficiently customizes SAM using only one-shot data, i.e., a user-provided reference image and a rough mask of the personal concept. Specifically, we first obtain a location confidence map for the target object in the test image by feature similarities, which considers the appearance of every foreground pixel. According to confidence scores, two points are selected as the positive-negative location prior, which are finally encoded as prompt tokens and fed into SAM’s decoder for segmentation. Within the decoder, we propose to inject visual semantics of the target object to unleash SAM’s personalized segmentation power with two techniques:

- **Target-guided Attention.** We guide every token-to-image cross-attention layer in SAM’s decoder by the location confidence map. This explicitly compels the prompt tokens to mainly concentrate on foreground target regions for intensive feature aggregation.
- **Target-semantic Prompting.** To explicitly provide SAM with high-level target semantics, we fuse the original prompt tokens with the embedding of the target object, which provides the low-level positional prompt with additional visual cues for personalized segmentation.

With the aforementioned designs, along with a cascaded post-refinement, PerSAM exhibits favorable personalized segmentation performance for unique subjects in a variety of poses or scenes. Notably, our approach can cope well with scenarios that require segmenting one object among multiple similar ones, simultaneously segmenting several identical objects in the same image, or *tracking different objects along a video*. Nevertheless, as shown in Figure 2, there might be occasional failure cases,

where the object comprises visually distinct subparts or hierarchical structures to be segmented, e.g., the hat on top of a teddy bear, or the head of a robot toy. Such ambiguity casts a challenge for PerSAM in determining the appropriate scale of mask as output, since both the local part and the global shape can be regarded as valid masks by SAM.

To alleviate this issue, we further propose a fine-tuning variant of our approach, **PerSAM-F**. We freeze the entire SAM to preserve its versatile pre-trained knowledge, and only fine-tune **2 parameters** within **10 seconds** on a single A100 GPU. In detail, we enable SAM to produce several potential segmentation results of different mask scales. To adaptively select the best scale for varying objects, we employ a learnable relative weight for each mask scale, and conduct a weighted summation as the final output. By such efficient **scale-aware** training, PerSAM-F avoids over-fitting on the one-shot data and exhibits better segmentation accuracy shown in Figure 2 (Right).

Moreover, we observe that our approach can also assist DreamBooth (Ruiz et al., 2022) to better fine-tune diffusion models for personalized text-to-image generation, as shown in Figure 3. Given a few images containing a specific visual concept, e.g., your pet cat or backpack, DreamBooth learns to convert these images into an identifier [V] in the word embedding space, which, however, can simultaneously include the background information, e.g., stairs or the forest. This would override the newly prompted backgrounds, and disturb the target appearance generation. Therefore, we propose to leverage PerSAM to segment the target object within training images, and only supervise DreamBooth by the foreground area, enabling text-to-image synthesis with higher quality.

We summarize the contributions of our paper as follows:

- **Personalized Object Segmentation.** We first investigate how to customize a general-purpose segmentation model (SAM) into personalized scenarios with minimal expense. To this end, we introduce two efficient and effective methods, along with a new segmentation dataset, PerSeg, for the evaluation of personalized object segmentation.
- **PerSAM and PerSAM-F.** In PerSAM, we propose three training-free techniques to guide SAM by the high-level semantics of target objects. In PerSAM-F, we design a scale-aware fine-tuning with 2 parameters in 10 seconds to well alleviate the mask ambiguity issue.
- Our approach achieves competitive results on various tasks, including the PerSeg benchmark, one-shot part and semantic segmentation, and video object segmentation. In addition, PerSAM can enhance DreamBooth for better personalized text-to-image synthesis.

2 RELATED WORK

Foundation Models. With powerful generalization capacity, pre-trained foundation models can be adapted for various downstream scenarios and attain promising performance. In natural language processing, BERT (Devlin et al., 2018; Lu et al., 2019), GPT series (Brown et al., 2020; OpenAI, 2023; Radford & Narasimhan, 2018; Radford et al., 2019), and LLaMA (Zhang et al., 2023c) have demonstrated remarkable in-context learning abilities, and can be transferred to new tasks by domain-specific prompts. Similarly, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), which conduct contrastive learning on image-text pairs, exhibit exceptional accuracy in zero-shot visual recognition. Painter (Wang et al., 2022) introduces a vision model that unifies network architectures and in-context prompts to accomplish diverse vision tasks, without downstream fine-tuning. CaFo (Zhang et al., 2023d) cascades different foundation models and collaborates their pre-trained knowledge for robust low-data image classification. SAM (Kirillov et al., 2023) presents a foundation model for image segmentation, which is pre-trained by 1 billion masks and conducts prompt-based segmentation. There are some concurrent works extending SAM for high-quality segmentation (Ke et al., 2023), faster inference speed (Zhao et al., 2023; Zhang et al., 2023a), all-purpose matching (Liu et al., 2023), 3D reconstruction (Cen et al., 2023), object tracking (Yang et al., 2023), medical (Ma & Wang, 2023; Huang et al., 2023) image processing. From another perspective, we propose to personalize the segmentation foundation model, i.e., SAM, for specific visual concepts, which adapts a generalist into a specialist with only one shot. Our method can also assist the personalization of text-to-image foundation models, i.e., Stable Diffusion (Rombach et al., 2022) and Imagen (Saharia et al., 2022), which improves the generation quality by segmenting the foreground target objects from the background disturbance.

Large Models in Segmentation. As a fundamental task in computer vision, segmentation (Long et al., 2015; Jiang et al., 2022; Zhao et al., 2017; Xu et al., 2021; Jiang et al., 2023; Lin et al., 2022) requires a pixel-level comprehension of a image. Various segmentation-related tasks have been explored, such as semantic segmentation, classifying each pixel into a predefined set of classes (Badrinath et al., 2017; Chen et al., 2017; Zheng et al., 2021; Cheng et al., 2022; Xie et al., 2021; Song et al., 2020); instance segmentation, focusing on the identification of individual object instances (He et al., 2017; Wang et al., 2020; Tian et al., 2020); panoptic segmentation, assigning both class labels and instance identification (Kirillov et al., 2019; Li et al., 2019); and interactive segmentation, involving human intervention for refinement (Hao et al., 2021; Chen et al., 2021). Recently, inspired by language foundation models (Zhang et al., 2023c; Brown et al., 2020), several concurrent works have proposed large-scale vision models for image segmentation. They are pre-trained by extensive mask data and exhibit strong generalization capabilities on numerous image distributions. Segment Anything Model (SAM) (Kirillov et al., 2023) utilizes a data engine with model-in-the-loop annotation to learn a promptable segmentation framework, which generalizes to downstream scenarios in a zero-shot manner. Painter (Wang et al., 2022) and SegGPT (Wang et al., 2023) introduce a robust in-context learning paradigm and can segment any images by a given image-mask prompt. SEEM (Zou et al., 2023) further presents a general segmentation model prompted by multi-modal references, e.g., language and audio, incorporating versatile semantic knowledge. In this study, we introduce a new task termed personalized object segmentation, and annotate a new dataset PerSeg for evaluation. Instead of developing large segmentation models, our goal is to personalize them to segment user-provided objects in any poses or scenes. We propose two approaches, PerSAM and PerSAM-F, which efficiently customize SAM for personalized segmentation.

Parameter-efficient Fine-tuning. Directly tuning the entire foundation models on downstream tasks can be computationally expensive and memory-intensive, posing challenges for resource-constrained applications. To address this issue, recent works have focused on developing parameter-efficient methods (Sung et al., 2022; He et al., 2022; Rebuffi et al., 2017; Qin & Eisner, 2021) to freeze the weights of foundation models and append small-scale modules for fine-tuning. Prompt tuning (Lester et al., 2021; Zhou et al., 2022; Jia et al., 2022; Liu et al., 2021) suggests using learnable soft prompts alongside frozen models to perform specific downstream tasks, achieving more competitive performance with scale and robust domain transfer compared to full model tuning. Low-Rank Adaption (LoRA) (Hu et al., 2021; Cuenca & Paul, 2023; Zhang et al., 2023b; Hedegaard et al., 2022) injects trainable rank decomposition matrices concurrently to each pre-trained weight, which significantly reduces the number of learnable parameters required for downstream tasks. Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020; Lin et al., 2020; Chen et al., 2022) are designed to be inserted between layers of the original transformer, introducing lightweight MLPs for feature transformation. Different from existing works, we adopt a more efficient adaption method delicately designed for SAM, i.e., the scale-aware fine-tuning of PerSAM-F with only 2 parameters and 10 seconds. This effectively avoids the over-fitting issue on one-shot data, and alleviates the ambiguity of segmentation scale with superior performance.

3 METHOD

In Section 3.1, we first briefly revisit Segment Anything Model (SAM) (Kirillov et al., 2023), and introduce the task definition for personalized object segmentation. Then, we illustrate the methodology of our PerSAM and PerSAM-F in Section 3.2 and 3.3, respectively. Finally, we utilize our approach to assist DreamBooth (Ruiz et al., 2022) for better text-to-image generation in Section 3.4.

3.1 PERSONALIZED OBJECT SEGMENTATION

A Revisit of Segment Anything. SAM consists of three components, a prompt encoder, an image encoder, and a lightweight mask decoder, respectively denoted as Enc_P , Enc_I , and Dec_M . As a promptable framework, SAM takes as input an image I , and a set of prompts P , which can be a point, a box, or a coarse mask. Specifically, SAM first utilizes Enc_I to obtain the input image feature, and adopts Enc_P to encode the human-given prompts of a length k into prompt tokens as

$$F_I = Enc_I(I), \quad T_P = Enc_P(P), \quad (1)$$

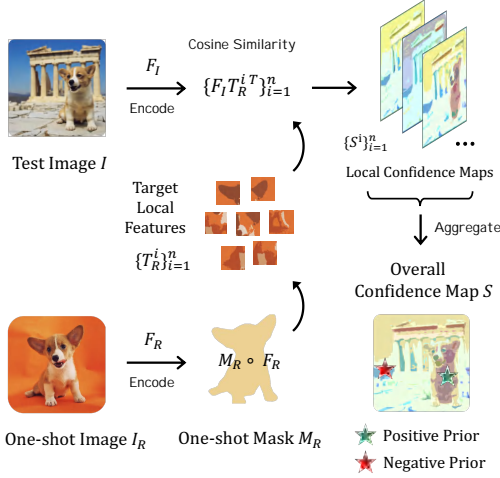


Figure 4: **Positive-negative Location Prior.** We calculate a location confidence map for the target object in new test image by the appearance of all local parts. Then, we select the location prior as the point prompt for PerSAM.

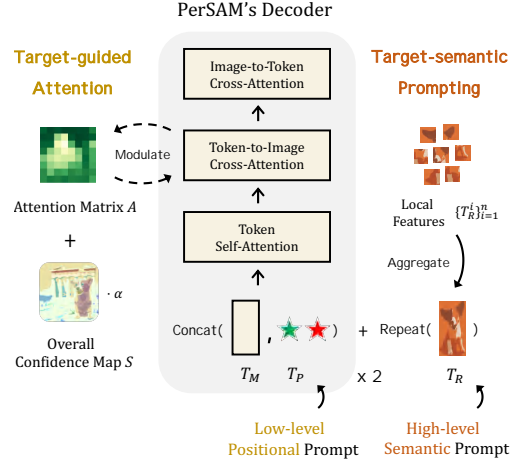


Figure 5: **Target-guided Attention (Left) & Target-semantic Prompting (Right).** To inject SAM with target semantics, we explicitly guide the cross-attention layers, and propose additional prompting with high-level cues.

where $F_I \in \mathbb{R}^{h \times w \times c}$ and $T_P \in \mathbb{R}^{k \times c}$, with h, w denoting the resolution of the image feature map and c denoting the feature dimension. After that, the encoded image and prompts are fed into the decoder Dec_M for attention-based feature interaction. **SAM constructs the input tokens of the decoder by concatenating several learnable mask tokens T_M as prefixes to the prompt tokens T_P . These mask tokens are responsible for generating the mask output, formulated as**

$$M = \text{Dec}_M(F_I, \text{Concat}(T_M, T_P)), \quad (2)$$

where M denotes the final segmentation mask predicted by SAM.

Task Definition. Although SAM is generalized enough for any object by prompting, it lacks the ability to automatically segment specific subject instances. Considering this, we define a new task for personalized object segmentation. The user provides only a single reference image, and a mask indicating the target visual concept. The given mask can either be an accurate segmentation, or a rough sketch drawn on-the-fly. Our goal is to customize SAM to segment the designated object within new images or videos, without additional human prompting. For evaluation, we annotate a new dataset for personalized segmentation, named PerSeg. The raw images are collected from the works for subject-driven diffusion models (Gal et al., 2022; Ruiz et al., 2022; Kumari et al., 2022), containing various categories of visual concepts in different poses or scenes. In this paper, we propose two efficient solutions for this task, which we specifically illustrate as follows.

3.2 TRAINING-FREE PERSAM

Location Confidence Map. Conditioned on the user-provided image I_R and mask M_R , PerSAM first obtains a confidence map that indicates the location of the target object in the new test image I . As shown in Figure 4, we apply an image encoder to extract the visual features of both I_R and I . The encoder can be SAM’s frozen backbone or other pre-trained vision models, for which we adopt SAM’s image encoder Enc_I by default. We formulate the process as

$$F_I = \text{Enc}_I(I), \quad F_R = \text{Enc}_I(I_R), \quad (3)$$

where $F_I, F_R \in \mathbb{R}^{h \times w \times c}$. Then, we utilize the reference mask $M_R \in \mathbb{R}^{h \times w \times 1}$ to crop the features of foreground pixels within the visual concept from F_R , resulting in a set of n local features as

$$\{T_R^i\}_{i=1}^n = M_R \circ F_R, \quad (4)$$

where $T_R^i \in \mathbb{R}^{1 \times c}$ and \circ denotes spatial-wise multiplication. After this, we calculate n confidence maps for each foreground pixel i by the **cosine similarity between T_R^i and test image feature F_I** as

$$\{S^i\}_{i=1}^n = \{F_I T_R^i\}_{i=1}^n, \quad \text{where } S^i \in \mathbb{R}^{h \times w}. \quad (5)$$

Note that F_I and T_R^i have been pixel-wisely L2-normalized. Each S^i represents the distribution probability for a different local part of object in the test image, such as the head, the body, or the paws of a dog. On top of this, we adopt an average pooling to aggregate all n local maps to obtain the overall confidence map of the target object as

$$S = \frac{1}{n} \sum_{i=1}^n S^i \in \mathbb{R}^{h \times w}. \quad (6)$$

By incorporating the confidences of every foreground pixel, S can take the visual appearance of different object parts into consideration, and acquire a relatively comprehensive location estimation.

Positive-negative Location Prior. To provide PerSAM with a location prior on the test image, we select two points with the highest and lowest confidence values in S , denoted as P_h and P_l , respectively. The former represents the most likely center position of the target object, while the latter inversely indicates the background. Then, they are regarded as the positive and negative point prompts, and fed into the prompt encoder as

$$T_P = \text{Enc}_P(P_h, P_l) \in \mathbb{R}^{2 \times c}, \quad (7)$$

which denote the prompt tokens for SAM’s decoder. In this way, SAM would tend to segment the contiguous region surrounding the positive point, while discarding the negative one’s on the image.

Target-guided Attention. Although the positive-negative point prompt has been obtained, we further propose a more explicit semantic guidance to the cross-attention operation in SAM’s decoder, which concentrates the feature aggregation within foreground target regions. As shown in Figure 5, the overall confidence map S in Equation 6 can clearly indicate the rough region of the target visual concept in the test image (hotter colors indicate higher scores). Based on such a property, we utilize S to guide the attention map in every token-to-image cross-attention layer of the decoder. Specifically, we denote every attention map after the softmax function as $A \in \mathbb{R}^{h \times w}$, and then modulate its attention distribution by

$$A^g = \text{softmax}\left(A + \alpha \cdot \text{softmax}(S)\right), \quad (8)$$

where α denotes a balancing factor. With the attention bias, the mask and prompt tokens are compelled to capture more visual semantics associated with the target subject, other than the unimportant background area. This contributes to more effective feature aggregation in attention mechanisms, and enhances the final segmentation accuracy of PerSAM in a training-free manner.

Target-semantic Prompting. The vanilla SAM only receives prompts with low-level positional information, such as the coordinate of a point or a box. To provide SAM’s decoder with more high-level cues, we propose to utilize the visual feature of the target concept as an additional high-level semantic prompting. We first obtain the global embedding T_R of the object in the reference image by both average pooling between different local features as

$$T_R = \frac{1}{n} \sum_{i=1}^n T_R^i \in \mathbb{R}^{1 \times c}. \quad (9)$$

Then, we element-wisely add T_R to all the input tokens of the test image in Equation 2, before feeding them into the decoder block, which is shown in Figure 5 as

$$T^g = \text{Repeat}(T_R) + \text{Concat}(T_M, T_P), \quad (10)$$

where T^g denotes the input token guided by target semantics for the decoder Dec_M , and the Repeat operation duplicates the target visual embedding. Aided by the simple token incorporation, PerSAM is not only prompted by low-level location points, but also high-level target visual cues.

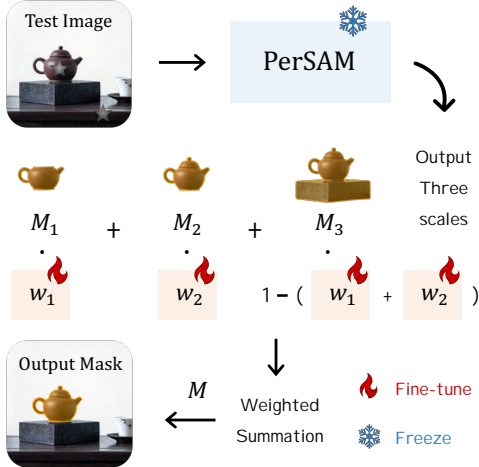


Figure 6: **The Scale-aware Fine-tuning in PerSAM-F.** To alleviate the scale ambiguity, PerSAM-F adopts two learnable weights for adaptively aggregating three-scale masks.

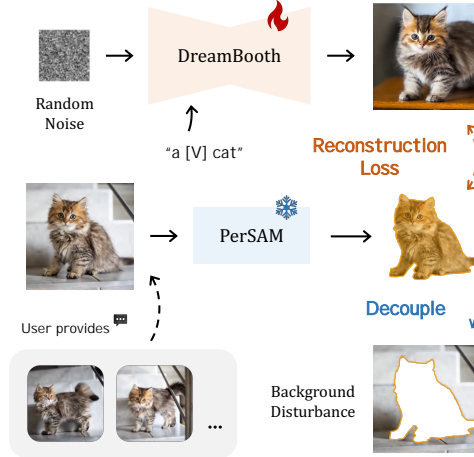


Figure 7: **PerSAM-assisted DreamBooth.** We utilize PerSAM to decouple the target objects from the background for improving the generation of DreamBooth.

Cascaded Post-refinement. Via the above techniques, we obtain an initial segmentation mask on the test image from SAM’s decoder, which however, might include rough edges and isolated background noises. **For further refinement, we iteratively feed the mask back into the decoder Dec_M for a two-step post-processing.** In the first step, we prompt the decoder by the currently predicted mask along with the previous positive-negative point prompt. For the second step, we acquire the bounding box enclosing the mask from the first step, and prompt the decoder additionally with this box for more accurate object localization. As we only iterate the lightweight decoder without the large-scale image encoder, the post-processing is efficient and only costs an extra 2% latency.

3.3 FINE-TUNING OF PERSAM-F

Ambiguity of Segmentation Scales. The training-free PerSAM can tackle most cases with satisfactory segmentation accuracy. However, some target objects contain hierarchical structures, which leads to the ambiguity of mask scales. As shown in Figure 6, the teapot on top of a platform is comprised of two parts: a lid and a body. If the positive point prompt (denoted by a green pentagram) is located at the body, while the negative prompt (denoted by a red pentagram) does not exclude the platform in a similar color, PerSAM would be misled for segmentation. Such an issue is also discussed in SAM (Kirillov et al., 2023), where it proposes an alternative to simultaneously generate multiple masks of three scales, corresponding to the whole, part, and subpart of an object. Then, the user is required to manually select one mask out of three, which is effective but consumes extra manpower. In contrast, our personalized task aims to customize SAM for automatic object segmentation without the need for human prompting. This motivates us to further develop a scale-aware version of PerSAM by parameter-efficient fine-tuning.

Scale-aware Fine-tuning. For adaptive segmentation with the appropriate scale, we introduce a fine-tuning variant, PerSAM-F. Unlike the training-free model only producing one mask, PerSAM-F first follows PerSAM to obtain the location prior, and refers to SAM’s original solution to output three-scale masks, denoted as M_1 , M_2 , and M_3 , respectively. On top of this, we adopt two learnable mask weights, w_1 , w_2 , and calculate the final mask output by a weighted summation as

$$M = w_1 \cdot M_1 + w_2 \cdot M_2 + (1 - w_1 - w_2) \cdot M_3, \quad (11)$$

where w_1, w_2 are both initialized as $1/3$. To learn the optimal weights, we conduct one-shot fine-tuning on the reference image, and regard the given mask as the ground truth. Note that, we freeze the entire SAM model to preserve its pre-trained knowledge, and only fine-tune the **2 parameters** of w_1, w_2 within **10 seconds** on a single A100 GPU. In this way, our PerSAM-F efficiently learns the scale-aware semantics of objects, and adaptively outputs the best segmentation scale for different concepts, improving the generalization capacity of PerSAM.

Table 1: **Personalized Object Segmentation on the PerSeg Dataset.** We compare the overall mIoU, bIoU, and learnable parameters for different methods (Bar et al., 2022; Wang et al., 2022; 2023; Zou et al., 2023), along with the mIoU for 10 objects in PerSeg. ‘*’ denotes works concurrent to ours.

Method	mIoU	bIoU	Param.	Can	Barn	Clock	Cat	Back-pack	Teddy Bear	Duck Toy	Thin Bird	Red Cartoon	Robot Toy
Painter	56.4	42.0	354M	19.1	3.2	42.9	94.1	88.1	93.0	33.3	20.9	98.2	65.0
VP	65.9	25.5	383M	61.2	58.6	59.2	76.6	66.7	79.8	89.9	67.4	81.0	72.4
SEEM*	87.1	55.7	341M	65.4	82.5	72.4	91.1	94.1	95.2	98.0	71.3	97.0	95.8
SegGPT*	94.3	76.5	354M	96.6	63.8	92.6	94.1	94.4	93.7	97.2	92.6	97.3	96.2
PerSAM	89.3	71.7	0	96.2	38.9	96.2	90.70	95.39	94.6	97.3	93.7	97.0	60.6
PerSAM-F	95.3	77.9	2	96.7	97.5	96.1	92.3	95.5	95.2	97.3	94.0	97.1	96.7

Table 2: **Video Object Segmentation** on DAVIS 2017 val (Pont-Tuset et al., 2017). We utilize gray color to denote the methods involving in-domain training.

Method	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
AGSS	67.4	64.9	69.9
AFB-URR	74.6	73.0	76.1
Painter	34.6	28.5	40.8
SEEM	58.9	55.0	62.8
SegGPT	75.6	72.5	78.6
PerSAM	66.9	71.3	75.1
PerSAM-F	76.1	74.9	79.7

Table 3: **One-shot Semantic and Part Segmentation** on FSS-1000 (Li et al., 2020), LVIS-92ⁱ (Gupta et al., 2019), PASCAL-Part (Morabia et al., 2020), and PACO-Part (Ramanathan et al., 2023). We report the mIoU scores and utilize gray color to denote the methods involving in-domain training.

Method	One-shot Semantic Seg.		One-shot Part Seg.	
	FSS-1000	LVIS-92 ⁱ	PASCAL-Part	PACO-Part
HSNet	86.5	17.4	32.4	22.6
VAT	90.3	18.5	33.6	23.5
Painter	61.7	10.5	30.4	14.1
SegGPT	85.6	18.6	-	-
PerSAM	81.6	15.6	32.5	22.5
PerSAM-F	86.3	18.4	32.9	22.7

3.4 PERSAM-ASSISTED DREAMBOOTH

For personalized text-to-image synthesis, DreamBooth (Ruiz et al., 2022) fine-tunes a pre-trained diffusion model by the given 3~5 photos of a specific object, i.g., a pet cat. It learns to generate the cat referred to by a text prompt, “a [V] cat”, and calculates the loss over the entire reconstructed images. This This would inject the redundant background information in the training images into the identifier [V]. Therefore, as shown in Figure 7, we introduce our strategy to alleviate the disturbance of backgrounds in DreamBooth. Given an object mask for any of the few-shot images, we leverage our PerSAM to segment all the foreground targets, and discard the gradient back-propagation for pixels belonging to the background area. Then, the Stable Diffusion is only fine-tuned to memorize the visual appearances of the target object. With no supervision imposed on the background, our PerSAM-assisted DreamBooth can not only synthesize the target object with better visual correspondence, but also increase the diversity of the new backgrounds guided by the input text prompt.

4 EXPERIMENT

We first evaluate our approach for personalized segmentation on PerSeg in Section 4.1, along with various existing one-shot segmentation benchmarks in Section 4.2. Then, we illustrate the effectiveness of our PerSAM-assisted DreamBooth in Section 4.3. Finally, we conduct several ablation studies to investigate our designs on PerSeg in Section 4.4.

4.1 PERSONALIZED EVALUATION

PerSeg Dataset. To test the personalization capacity, we construct a new segmentation dataset, termed PerSeg. The raw images are collected from the training data of subject-driven diffusion works (Ruiz et al., 2022; Gal et al., 2022; Kumari et al., 2022). PerSeg contains 40 objects of various categories in total, including daily necessities, animals, and buildings. In different poses or scenes, each object is associated with 5~7 images and masks, where we fix one image-mask pair as the user-provided one-shot data. The mIoU and bIoU (Cheng et al., 2021) are adopted for evaluation. Please refer to the Appendix for implementation details and an enlarged data scale of PerSeg.

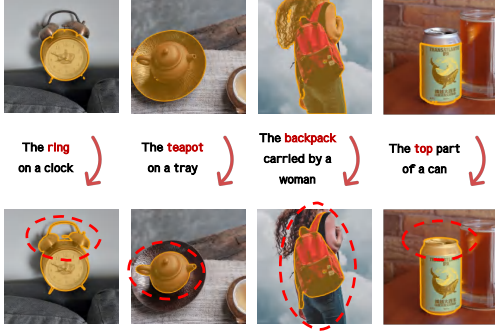


Figure 8: **Visualization of PerSAM-F’s Improvement.** Our scale-aware fine-tuning can well alleviate the scale ambiguity of PerSAM.

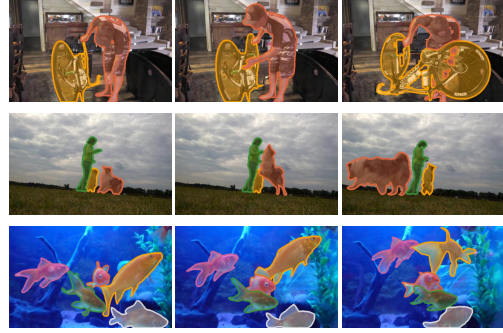


Figure 9: **Visualization of Video Object Segmentation.** Our approach performs well for segmenting multiple objects in a video.

Performance. In Table 1, we observe the fine-tuned PerSAM-F achieves the best results, which effectively enhances PerSAM by +2.7% and +5.9% overall mIoU and bIoU. We show more visualization of PerSAM-F’s improvement in Figure 8. Visual Prompting (VP) (Bar et al., 2022), Painter (Wang et al., 2022), SEEM (Zou et al., 2023), and SegGPT (Wang et al., 2023) are in-context learners that can also segment objects according to the given one-shot prompt data. As shown, the training-free PerSAM can already achieve better performance than Painter, VP, and SEEM with different margins. By the efficient 2-parameter fine-tuning, our PerSAM-F further surpasses the powerful SegGPT by +2.4% and +4.1% overall mIoU and bIoU. Different from their motivations to develop segmentation generalists, our method is specially designed for personalized object segmentation, and exhibits much more efficiency in both time and computational resources.

4.2 EXISTING SEGMENTATION BENCHMARKS

Video Object Segmentation. Given the first-frame image and object masks, our PerSAM and PerSAM-F achieve competitive object segmentation and tracking performance on the validation set of DAVIS 2017 (Pont-Tuset et al., 2017). As shown in Table 2, compared to methods without video training, the training-free PerSAM largely surpasses Painter by +32.3% $\mathcal{J}\&\mathcal{F}$ score, and our PerSAM-F can achieve +0.5% better performance than SegGPT. Notably, our one-shot fine-tuning approach can outperform methods (Lin et al., 2019; Liang et al., 2020) fully trained by extensive video data. The results fully illustrate our strong generalization ability for temporal video data and complex scenarios, which contain multiple similar or occluded objects, as visualized in Figure 9.

One-shot Semantic and Part Segmentation. In Table 3, we evaluate our approach for one-shot image segmentation respectively on four datasets, FSS-1000 (Li et al., 2020), LVIS-92ⁱ (Gupta et al., 2019), PASCAL-Part (Morabia et al., 2020), and PACO-Part (Ramanathan et al., 2023), where we follow Matcher (Liu et al., 2023) for data pre-processing and evaluation. As shown, our PerSAM-F attains consistently better results than Painter, and performs comparably to SegGPT. For models (Min et al., 2021; Hong et al., 2022) with in-domain training, our approach can achieve higher scores than HSNet. The experiments well demonstrate that, our proposed approach is not limited to object-level segmentation, but also works for category-wise and part-wise personalization of SAM.

4.3 PERSAM-ASSISTED DREAMBOOTH

We follow all the hyperparameters in DreamBooth (Ruiz et al., 2022) to fine-tune a pre-trained Stable Diffusion (Rombach et al., 2022) for personalized image synthesis. In addition to Figure 3, we visualize more examples of PerSAM-assisted DreamBooth in Figure 10. For the dog lying on a grey sofa, the “jungle” and “snow” by DreamBooth are still the sofa with green and white decorations. Assisted by PerSAM-F, the newly-generated background is totally decoupled with the sofa and well corresponds to the textual prompt. For the barn in front of the mountains, our approach also alleviates the background disturbance to correctly generate the “forest” and “blue sky”.

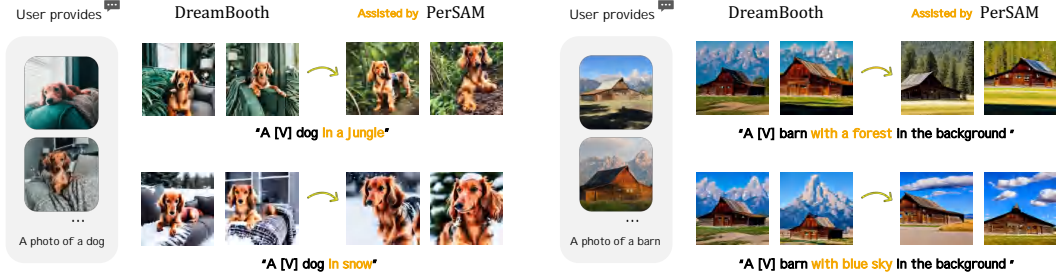


Figure 10: **Visualization of PerSAM-guided DreamBooth.** The improved DreamBooth (Ruiz et al., 2022) can better preserve the diversity for synthesizing various contexts in new images.

Table 4: **Ablation of Main Components** in our proposed method.

Variant	mIoU	Gain
Positive Prior	69.1	-
+ Negative Prior	72.5	+3.4
+ Post-refinement	83.9	+11.4
+ Guided Attention	85.8	+1.9
+ Semantic Prompt	89.3	+3.5
+ Scale Tuning	95.3	+6.0

Table 5: **Ablation of Different Fine-tuning Methods.**

Method	Param.	mIoU
PerSAM	0	89.32
Prompt Tuning	12K	76.5
Adapter	196K	78.3
LoRA	293K	90.0
3 Mask Weights	3	92.9
PerSAM-F	2	95.3

Table 6: **Ablation of using Box-image as Reference.**

Method	Mask	Box
Painter	56.4	42.0
VP	65.9	38.1
SEEM	87.1	64.9
SegGPT	94.3	36.0
PerSAM	89.3	88.1
PerSAM-F	95.3	94.9

4.4 ABLATION STUDY

Main Components. In Table 4, we investigate our different components by starting from a baseline that only adopts the positive location prior. Then, we add the negative point prompt and cascaded post-refinement, enhancing +3.6% and +11.4% mIoU, respectively. On top of that, we introduce the high-level target semantics into SAM’s decoder for attention guidance and semantic prompting. The resulting +1.9% and +3.5% improvements fully indicate their significance. Finally, via the efficient scale-aware fine-tuning, PerSAM-F boosts the score by +6.0%, demonstrating superior accuracy.

Different Fine-tuning Methods. In Table 5, we experiment with other parameter-efficient fine-tuning (PEFT) methods for PerSAM-F, i.e., prompt tuning (Liu et al., 2021), Adapter (Houlsby et al., 2019), and LoRA (Hu et al., 2021). We freeze the entire SAM, and only tune the PEFT modules injected into every transformer block in PerSAM’s decoder. As shown, the prompt tuning and Adapter would over-fit the one-shot data and severely degrade the accuracy. Instead, our scale-aware fine-tuning can best improve the performance of PerSAM, while tuning the least learnable parameters.

Using Box-image as Reference. Requiring an accurate mask as one-shot data might be too strict for some users. In Table 6, we relax the input restrictions to a bounding box designating the expected object. For our method, we can regard the box as a prompt and utilize off-the-shelf SAM to generate the one-shot mask. Therefore, the box reference only leads to a marginal performance drop in PerSAM and PerSAM-F, but severely influences other methods.

5 CONCLUSION

In this paper, we propose to personalize Segment Anything Model (SAM) for specific visual concepts with only one-shot data. Firstly, we introduce PerSAM, which injects high-level target semantics into SAM with training-free techniques. On top of this, we present a scale-aware fine-tuning variant, PerSAM-F. With only 2 learnable parameters, PerSAM-F effectively alleviates the ambiguity of mask scales and achieves leading performance on various benchmarks. Besides, we also verify the efficacy of our approach to assist DreamBooth in fine-tuning better text-to-image diffusion models. We hope our work may expand the applicability of SAM to a wider range of scenarios.

REFERENCES

- Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.
- Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. *arXiv preprint arXiv:2304.12308*, 2023.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.
- Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7345–7354, 2021.
- Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.
- Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.
- Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.
- Pedro Cuenca and Sayak Paul. Using lora for efficient stable diffusion fine-tuning. <https://huggingface.co/blog/lora>, January 2023.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019.
- Yuying Hao, Yi Liu, Zewu Wu, Lin Han, Yizhou Chen, Guowei Chen, Lutao Chu, Shiyu Tang, Zhiliang Yu, Zeyu Chen, et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1551–1560, 2021.
- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=0RDcd5Axok>.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- Lukas Hedegaard, Aman Alok, Juby Jose, and Alexandros Iosifidis. Structured pruning adapters. *arXiv preprint arXiv:2211.10155*, 2022.

-
- 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*, pp. 108–126. Springer, 2022.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.
- Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2022.
- Zhengkai Jiang, Zhangxuan Gu, Jinlong Peng, Hang Zhou, Liang Liu, Yabiao Wang, Ying Tai, Chengjie Wang, and Liqing Zhang. Stc: spatio-temporal contrastive learning for video instance segmentation. In *European Conference on Computer Vision Workshops*, pp. 539–556. Springer, 2023.
- 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.
- Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.
- 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.
- Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.
- Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. *arXiv preprint arXiv:2211.09808*, 2022.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2869–2878, 2020.

-
- Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7026–7035, 2019.
- Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33: 3430–3441, 2020.
- Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3949–3957, 2019.
- Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*, 2020.
- Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pp. 388–404. Springer, 2022.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.
- 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.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13–23, 2019.
- Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.
- Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6941–6952, 2021.
- Keval Morabia, Jatin Arora, and Tara Vijaykumar. Attention-based joint detection of object and semantic part. *arXiv preprint arXiv:2007.02419*, 2020.
- OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.
- Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapter-fusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.
- Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.
- Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021.
- Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

-
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7141–7151, 2023.
- Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in Neural information processing systems*, 30, 2017.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Xiangyu Zhang, Hongbin Sun, Jian Sun, and Nanning Zheng. Rethinking learnable tree filter for generic feature transform. *Advances in Neural Information Processing Systems*, 33:3991–4002, 2020.
- Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vi-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022.
- Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision*, pp. 282–298. Springer, 2020.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.
- Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. *arXiv preprint arXiv:2212.02499*, 2022.
- Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.
- Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.
- Mutian Xu, Junhao Zhang, Zhipeng Zhou, Mingye Xu, Xiaojuan Qi, and Yu Qiao. Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3056–3064, 2021.
- 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.

-
- Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.
- Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023b.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023c.
- Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*, 2023d.
- Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.
- Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.
- Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
- 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.