SAM-PARSER: Fine-Tuning SAM Efficiently by Parameter Space Reconstruction

Zelin Peng*, Zhengqin Xu*, Zhilin Zeng, Xiaokang Yang, Wei Shen†

MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{zelin.peng, fate311, bernardeschi, xkyang, wei.shen}@sjtu.edu.cn

Abstract

Segment Anything Model (SAM) has received remarkable attention as it offers a powerful and versatile solution for object segmentation in images. However, fine-tuning SAM for downstream segmentation tasks under different scenarios remains a challenge, as the varied characteristics of different scenarios naturally requires diverse model parameter spaces. Most existing fine-tuning methods attempt to bridge the gaps among different scenarios by introducing a set of new parameters to modify SAM's original parameter space. Unlike these works, in this paper, we propose fine-tuning SAM efficiently by parameter space reconstruction (SAM-PARSER), which introduces nearly zero trainable parameters during fine-tuning. In SAM-PARSER, we assume that SAM's original parameter space is relatively complete, so that its bases are able to reconstruct the parameter space of a new scenario. We obtain the bases by matrix decomposition, and fine-tuning the coefficients to reconstruct the parameter space tailored to the new scenario by an optimal linear combination of the bases. Experimental results show that SAM-PARSER exhibits superior segmentation performance across various scenarios, while reducing the number of trainable parameters by approximately 290 times compared with current parameter-efficient fine-tuning methods.

Introduction

The recent unveiling of foundation models (Brown et al. 2020; Kirillov et al. 2023; Wang et al. 2023b) has shown unprecedented performance and potential across various domains in artificial intelligence. Among these, Segment Anything Model (SAM) (Kirillov et al. 2023) is one of the most noteworthy foundation models in computer vision, which contains a vast number of parameters and is pre-trained on a large-scale segmentation dataset, i.e., SA-1B. Consequently, SAM exhibits its capability in precise object segmentation across various images (He et al. 2023a; Ji et al. 2023). The effectiveness of SAM has generated significant interest in fine-tuning it for downstream scenarios (Zhang and Liu 2023; Ma and Wang 2023). However, this is a challenging task because the diverse characteristics inherent in different scenarios often necessitate varied model parameter spaces. Furthermore, given the vast number of parameters in SAM, direct fine-tuning of its original parameter space seems less feasible.

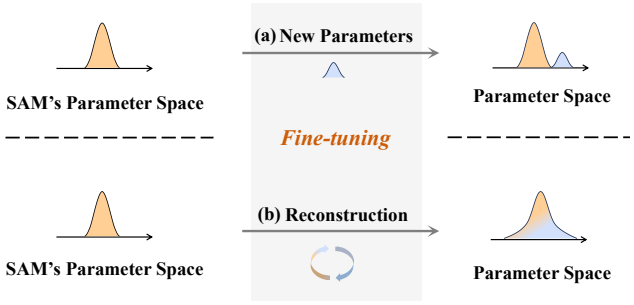

Figure 1: Comparative Overview: Previous Methods vs. SAM-PARSER. (a) Previous parameter-efficient fine-tuning methods adapt SAM to different scenarios via adding new parameters. (b) Our SAM-PARSER adapts SAM to various scenarios through parameter space reconstruction, with nearly zero trainable parameters introduced.

Recent methods tackle these challenges by adopting a parameter-efficient fine-tuning paradigm, inspired by the prominent fine-tuning strategies in Natural Language Processing (NLP), i.e., LoRA (Hu et al. 2022) and Adapter (Houlsby et al. 2019). As shown in Fig. 1, their common idea is to learn the distinct characteristics by introducing a set of new trainable parameters, thereby shifting the original parameter space to satisfy various downstream scenarios, leading to competitive performance for adapting SAM to new scenarios (Ma and Wang 2023; Chen et al. 2023; Wang et al. 2023a). Considering that the original parameter space of SAM is already very huge, we raise a question by the light of nature: Is there any way to adapt SAM to new varied scenarios without introducing new trainable parameters?

In this paper, to offer a solution to this question, we propose fine-tuning SAM efficiently by **par**ameter **s**pac**e** **r**econstruction (SAM-PARSER). Since the original parameter space of SAM is huge, we can assume that its bases are foundational, which are capable of reconstructing the parameter space tailored to varied downstream scenarios. Specifically, we utilize a widely-accepted matrix de-

*These authors contributed equally.
†Corresponding Author.

composition technique, i.e., Singular Value Decomposition (SVD) (Andrews and Patterson 1976), decompose the original parameter space of SAM into the bases and associated coefficients. Then, by fine-tuning these coefficients, we are able to reconstruction of the parameter space tailored for new scenarios through an optimal linear combination of these bases. Consequently, our proposed SAM-PARSER efficiently adapts SAM to different scenarios with nearly zero newly introduced trainable parameters, i.e., a few coefficients, during fine-tuning.

Extensive experiments demonstrate that SAM-PARSER shows competitive performance across three prevalent scenarios in computer vision, including natural image segmentation, remote sensing image segmentation, and medical image segmentation. Additionally, different from many fine-tuning methods that target on the parameters of Transformer layers, we apply our SAM-PARSER solely to the convolutional layers in SAM's encoder. In such a situation, SAM-PARSER only needs to fine-tune 0.5k coefficients, reducing the number of trainable parameters by $\approx 290\times$ compared with current parameter-efficient fine-tuning methods, while achieving superior segmentation performance.

Related Work

Foundation Models

Foundation models, initially introduced in the Natural Language Processing (NLP) community (Bommasani et al. 2021; Brown et al. 2020), have marked a significant milestone in the trajectory of artificial intelligence. Among them, GPT (Brown et al. 2020), with its advanced large language model, has drawn a lot of attention as its strong zero-shot generalization to unseen tasks and data. Recent developments have expanded the application of foundation models into the field of computer vision. Numerous vision foundation models (Kirillov et al. 2023; Wang et al. 2023b; Zou et al. 2023) are developed to deal with plenty of vision tasks and data ditributions. Our experiments here are carried out using SAM (Kirillov et al. 2023), a large foundation model trained on a large visual dataset, i.e., SA-1B. Large training data equips SAM with the capability to extract rich semantic features and detailed visual patterns, thereby ensuring its great potential for generalization across a variety of downstream scenarios.

Fine-tuning SAM

SAM consists of an image encoder and an image decoder, where the decoder is much more lightweight than the encoder. Thus, fully fine-tuning the decoder is a default operation in all SAM fine-tuning paradigms. The difference among different SAM fine-tuning paradigms lies in how to fine-tuning the encoder. For simplicity, we state **"fine-tuning SAM" means "fine-tuning SAM's image encoder"** in this paper. Naturally, a fine-tuning baseline is fully fine-tuning the decoder and fixing the encoder (Ma and Wang 2023; Wang et al. 2023c). Then the fine-tuning paradigms for SAM can broadly be organized into two categories: 1) Fully fine-tuning, in which all SAM's parameters are fine-tuned; 2) Parameter-efficient fine-tuning, which freezes SAM's original parameter spaces and focuses on fine-tuning a small number of newly introduced parameters, e.g., LoRA (Hu et al. 2022).

Fully Fine-tuning. A straightforward strategy is to fully fine-tune the entire parameters of SAM (Kirillov et al. 2023), as highlighted in a previous study (Li, Hu, and Yang 2023). However, this approach inevitably demands substantial computational resources, which is less practical.

Parameter-efficient Fine-tuning. These works (Wang et al. 2023a; Chen et al. 2023; Wu et al. 2023; Zhang and Liu 2023; Guan et al. 2023; Dutt et al. 2023; Zhang and Jiao 2023) sought to leverage insights derived from parameter-efficient fine-tuning paradigms in Natural Language Processing (NLP), such as Adapter (Houlsby et al. 2019) and LoRA (Hu et al. 2022). For example, SAMed (Zhang and Liu 2023) proposed the use of LoRA (Hu et al. 2022) to fine-tune SAM by optimizing the introduced rank decomposition matrices integrated into its Transformer blocks. Similarly, SAM-adapter (Wu et al. 2023) employed the adapter technique (Houlsby et al. 2019) to fine-tune SAM. They introduced several trainable adapter layers in each Transformer layer of SAM, while concurrently freezing SAM's original parameter space. Contrary to these methods that adapt SAM to different scenarios by introducing new parameter spaces, our SAM-PARSER directly reconstructs SAM's original parameter space, offering a more straightforward and efficient solution for its fine-tuning.

Parameter Space Decomposition

Parameter space decomposition is a highly prevalent strategy for analyzing the structure of parameter spaces, broadly categorized into two main methods: matrix theory-based (Mijnsbrugge, Ongenae, and Van Hoecke 2021; Cannon, Hanna, and Keppel 2011) and Fast Fourier Transform (FFT) theory-based methods (Li et al. 2022). Matrix theory-based decomposition typically involves analyzing components of the parameter space by techniques like Singular Value Decomposition (SVD) (Mijnsbrugge, Ongenae, and Van Hoecke 2021) or QR decomposition (Panahi, Saeedi, and Arodz 2021). A prominent extension of this decomposition method is tensor theory-based decomposition (Wang et al. 2022), commonly viewed as its higher-order counterpart. For the analysis of tensors, or other high-dimensional matrices, Tucker decomposition (Jie and Deng 2023) is frequently employed. Differently, FFT theory-based decomposition typically entails mapping network weights into the frequency domain to analyze contributions across different frequency components (He et al. 2023b). In this paper, we utilize matrix theory-based decomposition in our SAM-PARSER framework to reconstruct SAM's original parameter space, consistent with established practices in the contemporary parameter-efficient fine-tuning paradigm.

Preliminaries

In this section, we first look more closely at prevalent fine-tuning paradigms. Subsequently, we introduce the architecture of SAM. Ultimately, we discuss why parameter-efficient fine-tuning is favored for SAM and suggest a new perspective for its adaptation.

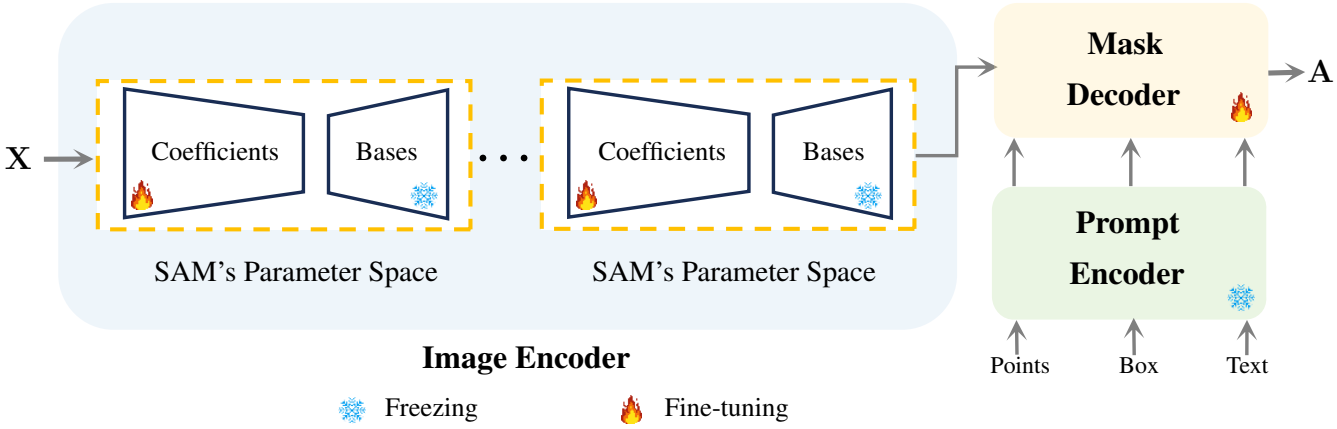

Figure 2: A schematic representation of SAM-PARSER. We reconstruct SAM's parameter spaces for adapting to various scenarios by fine-tuning the coefficients while maintaining the foundational bases are frozen. Concurrently, we fully fine-tuning SAM's mask decoder.

Fine-tuning Paradigm

Fully Fine-tuning. Given a training set $\mathcal{T}$ used for fine-tuning, existing methods train a neural network $\mathcal{F}(\mathbf{X};\mathbf{W})$ to produce a dense prediction map $\mathbf{A}$, where $\mathbf{X}$ is an input image and $\mathbf{W}$ are the trainable parameters of $\mathcal{F}$. Accordingly, the objective function of fully fine-tuning is formulated as:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \ell(\mathbf{A}, \mathbf{Y}), \tag{1}$$

where $\mathbf{Y}$ is a full dense label map, and ℓ is typically chosen as a cross-entropy loss when addressing a segmentation task.

Parameter-efficient Fine-tuning. We provide a brief introduction of two representative parameter-efficient fine-tuning methods in NLP, i.e., Adapter (Houlsby et al. 2019) and LoRA (Hu et al. 2022).

(1) Adapter. Adapter (Houlsby et al. 2019) maintained all the original parameters frozen and integrated a learnable adapter layer between the Multi-Head Attention (MHA) module and the LayerNorm layer in each Transformer layer. This adapter layer, composed of two linear layers and a ReLU activation function, creates an extra parameter space specifically tailored to new scenarios.

(2) LoRA. A prior work (Aghajanyan, Zettlemoyer, and Gupta 2020) illustrated that pre-trained large foundation models often reside on a low intrinsic dimension, allowing them to be projected into a smaller sub-space. Following this, LoRA (Hu et al. 2022) assumed that the change in weights during fine-tuning also has a low intrinsic rank. They introduced a learnable low-rank matrix $\Delta\mathbf{W}$ that works in parallel to an original weighted matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, which is in the MHA module of each transformer layer. $\Delta\mathbf{W}$ is formed through a QR decomposition, denoted by $\Delta\mathbf{W} = \boldsymbol{\beta}\boldsymbol{\alpha}$, where $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\alpha} \in \mathbb{R}^{r\times k}$, and rank $r \ll \min(d, k)$. The size of r determines the size of the newly introduced parameter space, allowing for a more targeted extraction of unique characteristics in new scenarios.

Segment Anything Model (SAM)

SAM (Kirillov et al. 2023) is constructed of an image encoder, with extensive parameters, and a subsequent decoder. The image encoder consists of sequentially connected layers, starting with several transformer layers followed by standard convolutional layers. By training on a large-scale dataset, i.e., SA-1B, this encoder creates a broad parameter space that guides the decoder to achieve precise segmentation results for all objects contained in images. Additionally, to enable user interactions and identify specific objects, SAM is further equipped with a prompt encoder. This specialized encode is proficient in processing both dense (i.e., mask-based) and (i.e., box or point-based) prompts.

Discussion

Compared with fully fine-tuning paradigms, a parameter-efficient fine-tuning paradigm has attracted more attention in the fine-tuning SAM literature (Chen et al. 2023; Wang et al. 2023a; Li, Hu, and Yang 2023). One reason is, as its name suggests, the parameter-efficient fine-tuning paradigm significantly reduces the trainable parameters comparing those of the fully fine-tuning paradigm, thus speeding up the training process. Besides, it is capable of learning unique characteristics of new scenarios by introducing new parameter spaces, making SAM better adapt to varied segmentation tasks.

In analyzing current fine-tuning literature, we find there is a clear trend toward reducing fine-tuning parameters while simultaneously aiming for improved results. Given this trajectory, we question if there exists a way that can match or even surpass the performance of prior paradigms with nearly zero newly introduced trainable parameters.

Methodology

In this section, we introduce our solution that proposes fine-tuning SAM efficiently by **par**ameter **s**pac**e** **r**econstruction (SAM-PARSER). In SAM-PARSER, we assume the bases

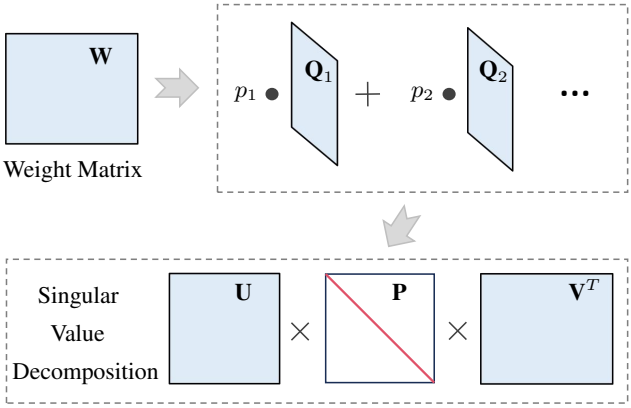

Figure 3: A schematic diagram to illustrate how to decompose weight matrices in SAM's image encoder. First, we decompose the weight matrix **W** into bases and their corresponding coefficients. Following this, we employ singular value decomposition (SVD) (Andrews and Patterson 1976) to further disintegrate the bases **Q** into a pair of bases, i.e., **U** and **V**.

of SAM's original parameter space are foundational, and capable of reconstructing the parameter space adapted to diverse downstream scenarios. The core of the proposed SAM-PARSER lies in fine-tuning the coefficients corresponding to these bases, aiming to achieve the optimal linear combination for reconstructing the parameter spaces tailored to new scenarios. Fig. 2 provides a visual overview of our proposed SAM-PARSER.

Specifically, for a pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, where $d \leq k$, it can be decomposed into bases and their corresponding coefficients, which are formulated as:

$$\mathbf{W} = \sum_{i=1}^{d} p_i \mathbf{Q}_i, \tag{2}$$

where $p_i \in \mathbb{R}$ and $\mathbf{Q}_i \in \mathbb{R}^{d\times k}$ are i^{th} coefficient and base, respectively.

Following previous works (Han et al. 2023; Sun et al. 2022), as shown in Fig. 3, we use a widely-accepted matrix decomposition technique, i.e., Singular Value Decomposition (SVD) (Andrews and Patterson 1976), to derive the bases and their coefficients. In this way, we can rewrite Eq. 2 as:

$$\begin{aligned}\mathbf{W} &= \sum_{i=1}^{d} p_i \mathbf{u}_i (\mathbf{v}_i)^T \\ &= \mathbf{UPV}^T,\end{aligned} \tag{3}$$

where we denote that $\mathbf{u}_i \in \mathbb{R}^d$ and $\mathbf{v}_i \in \mathbb{R}^k$, and then $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_d] \in \mathbb{R}^{d\times d}$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_d] \in \mathbb{R}^{k\times d}$ compose of the bases of **W**. In our proposed SAM-PARSER, we only fine-tune the diagonal matrix

$$\mathbf{P} = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_d \end{bmatrix},$$

which are the coefficients. Accordingly, the objective function of our proposed SAM-PARSER is formulated as:

$$\mathbf{P}^* = \arg\min_{\mathbf{P}} \ell(\mathbf{A}, \mathbf{Y}), \tag{4}$$

as **P** is a diagonal matrix, the number of parameters equals its rank, since the parameter matrix **W** is a full-rank matrix. Unlike previous methods that simply freeze **W** and introduce new parameters, we directly reconstruct **W** by fine-tuning **P**, offering a more direct and efficient way for SAM adaptation across diverse scenarios.

Loss Function. As recommended in previous works (Wang et al. 2023a; Ma and Wang 2023), we incorporate a combination of binary cross-entropy, denoted as ℓ_{ce}, and binary dice loss, represented by ℓ_{dice}, to fine-tune SAM. Then, for multi-instance scenes, we also adopt a focal loss ℓ_{focal} (Lin et al. 2017) for balancing the learning of different instances. Finally, the loss function is derived as:

$$\ell = (1-\lambda)\ell_{ce} + \lambda\ell_{focal} + \ell_{dice}, \tag{5}$$

where $\lambda = 0$ represents a single-instance scene, commonly found in medical image segmentation, e.g., lesion segmentation.

Experiments

In this section, we evaluate our method across a spectrum of downstream segmentation tasks. These tasks fall into three primary categories: (1) Medical image segmentation, (2) Natural image segmentation, and (3) Remote sensing image segmentation. We begin by introducing the datasets utilized, the corresponding evaluation metrics, and implementation details. Following this, we delve into an ablation study of our proposed SAM-PARSER. Finally, we benchmark SAM-PARSER against various other fine-tuning strategies.

Experimental Setup

Datasets. To validate the effectiveness of our proposed methods, we conduct experiments on fine-tuning SAM to five datasets.

(1) CT Abdominal organ (AO) (Ma et al. 2022). Following (Ma and Wang 2023), we randomly split 80% medical images of AO for fine-tuning and 20% for testing.

(2) COCO2017 (COCO) (Lin et al. 2014). We fine-tune SAM by using natural images on the training set and evaluate the models' performance on its validation set.

(3) PASCAL VOC2012 (VOC) (Everingham et al. 2010). We fine-tune SAM by using natural images on its training set and evaluate the models' performance on its validation set.

(4) NWPU VHR-10 (NWPU) (Cheng et al. 2014; Cheng and Han 2016; Cheng, Zhou, and Han 2016). As recommended in (Cheng, Zhou, and Han 2016), we allocate 70% remote sensing images of NWPU for fine-tuning and the remaining 30% for evaluation.

	Trans. la.	Conv. la.	mIoU (%)	F1 (%)
Baseline†	-	-	82.9	90.0
Parameter Space	✓		83.1 (+0.2)	90.2 (+0.2)
	✓	✓	83.3 (+0.4)	90.3 (+0.3)
		✓	**84.2 (+1.3)**	**90.9 (+0.9)**

Table 1: Ablation on selecting which parameter space for reconstruction. "Trans. la.": Transformer layers. "Conv. la.": Convolutional layers. "Baseline†": only fine-tuning SAM's decoder.

	Bases	Coefficients	mIoU (%)	Params (K)
Baseline†	-	-	82.9	3963.2
Component		✓	84.2 (+1.3)	3963.7 (+**0.5**)
	✓		84.9 (+2.0)	4852.4 (+888.7)
	✓	✓	**85.3 (+2.4)**	4852.9 (+889.2)

Table 2: Ablation on fine-tuning bases or coefficients. "Baseline†": only fine-tuning SAM's decoder. Fine-tuning both the bases and coefficients equates to a full fine-tuning strategy, representing an upper bound of our SAM-PARSER.

(5) WHU building extraction (WHU) (Ji, Wei, and Lu 2018). It is tailored for remote sensing image segmentation. We fine-tune SAM on the training set and evaluate its performance on the validation set.

Evaluation Metrics. In line with previous studies (Ma and Wang 2023; Wang et al. 2023a), we utilize the Dice Similarity Coefficient (DSC) for evaluating medical image segmentation. For both natural and remote sensing image segmentation, we adopt mean intersection-over-union (mIoU) and the F1 score (F1) as our evaluation metrics, as recommended by (Zheng et al. 2022).

Implementation Details. In all of our experiments, we employ the ViT-Base version of SAM (Kirillov et al. 2023) as our backbone, integrating a box prompt for its prompt encoder input. In line with MedSAM (Ma and Wang 2023), we apply a random perturbation to each bounding box, varying between 0 and 20 pixels. Our training employs the Adam optimizer (Kingma and Ba 2014). For medical image segmentation, the initial learning rate is set to 10^{-6}, and the weight decay is 5×10^{-4} with one image per mini-batch. The number of fine-tuning epochs is set to 25. For natural and remote sensing image segmentation, we follow SonarSAM (Wang et al. 2023a), the initial learning rate is set to 1.5×10^{-4}, and the weight decay is 5×10^{-5} with one image per mini-batch. The number of fine-tuning epochs is set to 10.

Ablation Study

SAM's whole parameter space is huge, where trainable parameters are mainly from transformer layers and convolutional layers. Previous fine-tuning methods predominantly target on the Transformer layers to introduce new parameters. Here, we delve into the parameters sub-spaces formed by the parameters from both the transformer layers and the convolutional layers and conduct ablation studies to determine reconstructing which parameter space/sub-space of SAM offers better performance. We also conduct an ablation study to verify our choice to fine-tune the coefficients rather than bases in SAM-PARSER. All results are reported on the NWPU dataset.

Select which parameter space for reconstruction? SAM's encoder consists of sequentially connected layers, starting with several transformer layers followed by standard convolutional layers. In this study, we approach the reconstruction of SAM's original parameter space through the reconstruction of its parameter sub-spaces, i.e., the sub-space formed by the parameters from 1) the transformer layers, 2) convolutional layers, and 3) both of them. As shown in Table. 1, only using parameters from transformer layers or using those from layers of both two types to form the parameter space for reconstruction leads to obvious performance drops, 0.9% mIoU and 0.7% mIoU, respectively. This indicates that the convolutional layers represent a more informative and crucial parameter space for reconstruction. Notably, when performing our method on the parameter space formed by the parameters only from the convolutional layers, we observe that the required trainable parameter size is only 0.5k while leading to superior performance.

Fine-tuning bases or coefficients? We ablate this experiment to validate the influence of fine-tuning different components in the original parameter space of SAM, i.e., coefficients and bases. As shown in Table. 2, when compared with the strategies of solely fine-tuning the bases or fully fine-tuning both coefficients and bases (the latter representing the upper bound of our SAM-PARSER), our SAM-PARSER achieves a dramatic reduction in trainable parameters by approximately 1700 times. Surprisingly, the performance drop is marginal at only 0.7% and 1.1%, respectively. This demonstrates that fine-tuning coefficients offers an efficient way for adapting SAM to new scenarios.

Main Results

In this section, we compare our approach against several prevailing fine-tuning techniques for SAM. These include: (1) Exclusively fully fine-tuning SAM's decoder, which is used as our baseline, (2) Leveraging LoRA (Hu et al. 2022) for fine-tuning SAM's image encoder and fully fine-tuning SAM's decoder, and (3) Utilizing Adapter (Houlsby et al. 2019) for fine-tuning SAM's image encoder and fully fine-tuning SAM's decoder.

Quantitative Results. The quantitative results for fine-tuning SAM are shown in Tab 3.

(1) Natural Image Segmentation. Table 3 gives the evaluation results comparing to other fine-tuning methods in natural image segmentation. On PASCAL VOC2012 *val* set (Everingham et al. 2010), our proposed method outperforms Adapter (Houlsby et al. 2019) by 0.7% and 0.6% in terms of mIoU and F1, respectively. But on COCO2017 *val* set (Lin et al. 2014), our proposed method is worse than LoRA (Hu et al. 2022) by 0.2% mIoU, a negative case we observe. We infer that for large natural image datasets, there is a significant overlap between their parameter spaces and SAM's original parameter space. Therefore, the gains from parameter space reconstruction are limited.

Method	AO	COCO		VOC		NWPU		WHU	
	DSC (%)	mIoU (%)	F1 (%)	mIoU (%)	F1 (%)	mIoU (%)	F1 (%)	mIoU (%)	F1 (%)
Baseline	66.4 ± 1.1	67.2 ± 0.2	76.3 ± 0.1	72.3 ± 0.4	81.7 ± 0.5	71.6 ± 0.6	81.4 ± 0.5	66.7 ± 0.2	76.0 ± 0.3
Baseline†	89.7 ± 1.2	74.4 ± 0.2	82.7 ± 0.3	78.7 ± 0.5	88.3 ± 0.2	82.9 ± 0.6	90.0 ± 0.4	81.1 ± 0.2	87.5 ± 0.1
FacT (Jie and Deng 2023)	89.6 ± 0.8	74.8 ± 0.3	82.8 ± 0.3	79.1 ± 0.3	88.4 ± 0.2	83.4 ± 0.1	90.1 ± 0.1	81.4 ± 0.2	87.6 ± 0.3
LoRA (Hu et al. 2022)	90.1 ± 0.6	**75.3** ± 0.1	83.1 ± 0.1	79.5 ± 0.0	89.0 ± 0.2	83.9 ± 0.1	90.7 ± 0.2	**81.9** ± 0.1	88.2 ± 0.1
Adapter (Houlsby et al. 2019)	89.4 ± 0.8	74.9 ± 0.1	83.1 ± 0.1	78.9 ± 0.0	88.6 ± 0.0	82.0 ± 0.2	89.6 ± 0.5	81.0 ± 0.3	87.5 ± 0.4
SAM-PARSER (Ours)	**90.9** ± 0.3	75.0 ± 0.1	**83.2** ± 0.0	**79.6** ± 0.1	**89.2** ± 0.1	**84.2** ± 0.2	**90.9** ± 0.1	81.8 ± 0.1	**88.4** ± 0.1

Table 3: Segment anything model (SAM) fine-tuned on five datasets. "AO": CT Abdominal organ *test* set (Ma et al. 2022) for medical image segmentation. "COCO": COCO2017 *val* set (Lin et al. 2014) for natural image segmentation. "VOC": PASCAL VOC2012 *val* set (Everingham et al. 2010) from natural image segmentation. "NWPU": NWPU VHR-10 *val* set (Cheng et al. 2014; Cheng and Han 2016; Cheng, Zhou, and Han 2016) for remote sensing image segmentation. "WHU": WHU building extraction *val* set (Ji, Wei, and Lu 2018) for remote sensing image segmentation. "Baseline": without any form of fine-tuning. "Baseline†": only fine-tuning SAM's decoder. These results are evaluated with three different seeds.

Method	Params (K)	Time (Fps)
Baseline†	3963.2	6.0
FacT (Jie and Deng 2023)	3977.5 (+14.3)	3.3 (−2.7)
LoRA (Hu et al. 2022)	4080.2 (+144)	3.4 (−2.6)
Adapter (Houlsby et al. 2019)	4314.7 (+351.5)	2.9 (−3.1)
SAM-PARSER (Ours)	**3963.7** (+**0.5**)	**5.8** (−**0.2**)

Table 4: Computation efficiency analysis for different fine-tuning methods. "Baseline†": only fine-tuning SAM's decoder.

Method	mIoU (%)	F1 (%)
Baseline	71.4	82.7
Baseline†	81.2	89.4
FacT (Jie and Deng 2023)	81.7 (+0.5)	89.7 (+0.3)
LoRA (Hu et al. 2022)	**82.4 (+1.2)**	**90.1 (+0.7)**
Adapter (Houlsby et al. 2019)	81.8 (+0.6)	89.9 (+0.5)
SAM-PARSER (Ours)	81.6 (+0.4)	89.6 (+0.2)

Table 5: Extended experiments. Fine-tuning SAM on the *val* set of SSDD dataset (Zhang et al. 2021). "Baseline": without any form of fine-tuning. "Baseline†": only fine-tuning SAM's decoder.

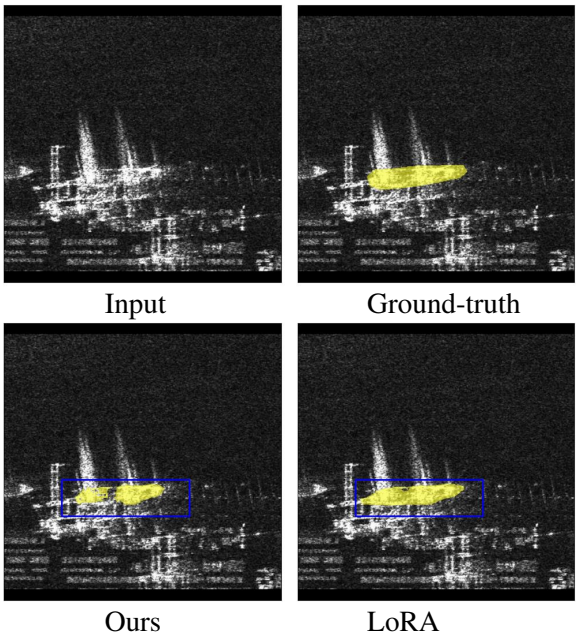

Figure 4: **A failure case on SSDD *val* set.**

(2) Remote Sensing Image Segmentation. In Table 3, we compare our approach with others in remote sensing image segmentation. On both NWPU VHR-10 *val* set and WHU building extraction *val* set, our approach with little computing overhead boosts the baseline by a clear margin of 1.3% mIoU and 0.8% mIoU, respectively.

(3) Medical Image Segmentation. Table 3 reports the comparison results on CT Abdominal organ test set (Ma et al. 2022). It can be clearly seen that fine-tuning SAM with our proposed method outperforms the widely-used fine-tuning method LoRA (Hu et al. 2022) by 0.8% DSC (from 90.3% to 91.1%). It demonstrates the effectiveness of our presented strategy for mitigating interfering information and aligning with the parameter space specific to medical image segmentation tasks.

Qualitative results. Here, we visualize some representative example segmentation results of our method against prevailing fine-tuning methods, e.g., LoRA (Hu et al. 2022) in three datasets. As shown in Fig 5, we observe that our approach is able to diverse scenarios and produce more accurate results.

Model Efficiency. Our fine-tuning method brings performance improvements with nearly zero additional computational overhead. To validate this, we show the statistics of Params, i.e., the number of network trainable parameters, and Fps, i.e., training speed, in Table 4. It is clearly shown that the complexity of our method is significantly smaller than that of other fine-tuning methods. For example, the increase in Params in our method is **290** times less than that

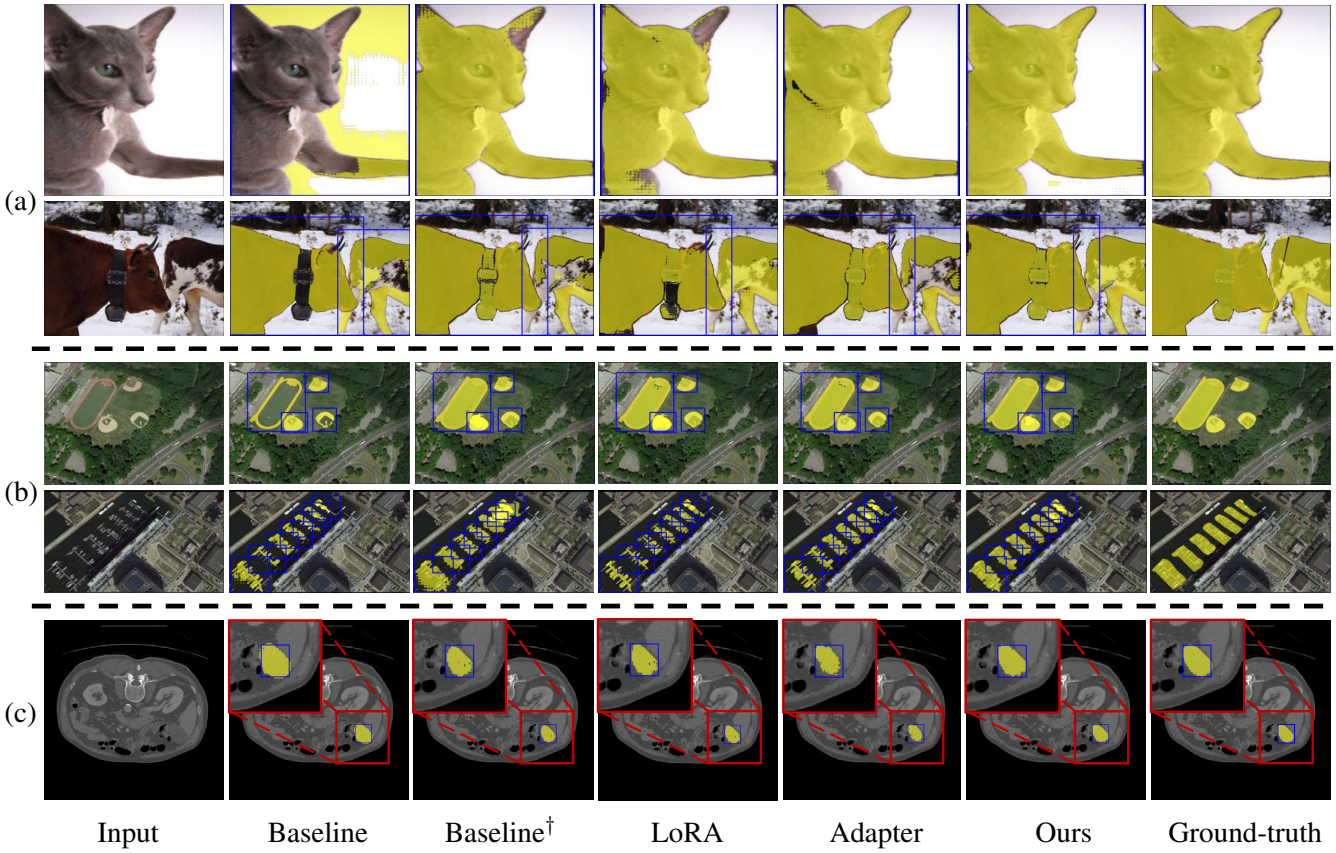

Figure 5: Qualitative segmentation results on three scenarios, i.e., (a) natural image segmentation on PASCAL VOC2012 *val* set (Everingham et al. 2010), (b) remote sensing image segmentation on NWPU VHR-10 *val* set (Cheng, Zhou, and Han 2016), and (c) medical image segmentation on CT Abdominal organ *test* set. "Baseline": without any form of fine-tuning. "Baseline†": only fine-tuning SAM's decoder.

in LoRA (Hu et al. 2022). This significant reduction in overhead stems from that our SAM-PARSER emphasizes reconstructing the coefficients in SAM's original parameter space, rather than creating an entirely new sub-space.

Analysis of a failure case. Our method is based on an assumption that the original parameter space of SAM is complete enough to cover new scenarios. However, this might be invalid when the image space of a new scenario is extremely different. To investigate this, we carry out extended experiments using the SSDD dataset (Zhang et al. 2021), dedicated to radar image segmentation—**a stark departure from natural image segmentation.** As shown in Fig 4, compared with LoRA, our SAM-PARSER achieves inferior segmentation results. According to the experimental results in Table 5, we can observe that our approach still outperforms the baseline by a mIoU of 0.4%, but falls behind LoRA (Hu et al. 2022) and Adapter (Houlsby et al. 2019). This demonstrates that in scenarios with distinctive characteristics, our proposed SAM-PARSER faces challenges as their parameter space cannot be reconstructed solely by the bases from the SAM's orignal parameter space. As a potential solution, incorporating existing parameter-efficient methods, like LoRA, to expand the bases of the original parameter space, could further boost our method to reconstruct the parameter space of these scenarios.

Conclusion

We proposed fine-tuning SAM efficiently by parameter space reconstruction, called SAM-PARSER. In SAM-PARSER, we assume that SAM's initial parameter space is relatively complete, which allows us to use its bases for reconstruction of parameter space tailored to varied downstream scenarios. To achieve this, we employed SVD technique to decompose the original parameter space into the bases and associated coefficients. By fine-tuning these coefficients, we can achieve the optimal linear combination for reconstructing the parameter space of a new scenario. Extensive experiments have demonstrated our superior performance, while adding nearly zero trainable parameters.

Acknowledgments

This work was supported by NSFC 62322604, 62176159, Natural Science Foundation of Shanghai 21ZR1432200, and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

References

Aghajanyan, A.; Zettlemoyer, L.; and Gupta, S. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Andrews, H.; and Patterson, C. 1976. Singular value decomposition (SVD) image coding. *IEEE transactions on Communications*, 24(4): 425–432.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cannon, K.; Hanna, C.; and Keppel, D. 2011. Efficiently enclosing the compact binary parameter space by singular-value decomposition. *Physical Review D*, 84(8): 084003.

Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Zhang, S.; Wang, Y.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2023. SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, and More. *arXiv preprint arXiv:2304.09148*.

Cheng, G.; and Han, J. 2016. A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117: 11–28.

Cheng, G.; Han, J.; Zhou, P.; and Guo, L. 2014. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98: 119–132.

Cheng, G.; Zhou, P.; and Han, J. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12): 7405–7415.

Dutt, R.; Ericsson, L.; Sanchez, P.; Tsaftaris, S. A.; and Hospedales, T. 2023. Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity. *arXiv preprint arXiv:2305.08252*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Guan, Z.; Hu, M.; Zhou, Z.; Zhang, J.; Li, S.; and Liu, N. 2023. Badsam: Exploring security vulnerabilities of sam via backdoor attacks. *arXiv preprint arXiv:2305.03289*.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*.

He, S.; Bao, R.; Li, J.; Grant, P. E.; and Ou, Y. 2023a. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*.

He, Z.; Yang, M.; Feng, M.; Yin, J.; Wang, X.; Leng, J.; and Lin, Z. 2023b. Fourier Transformer: Fast Long Range Modeling by Removing Sequence Redundancy with FFT Operator. *arXiv preprint arXiv:2305.15099*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ji, G.-P.; Fan, D.-P.; Xu, P.; Cheng, M.-M.; Zhou, B.; and Van Gool, L. 2023. SAM Struggles in Concealed Scenes–Empirical Study on" Segment Anything". *arXiv preprint arXiv:2304.06022*.

Ji, S.; Wei, S.; and Lu, M. 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1): 574–586.

Jie, S.; and Deng, Z.-H. 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, 1060–1068.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Li, Y.; Hu, M.; and Yang, X. 2023. Polyp-sam: Transfer sam for polyp segmentation. *arXiv preprint arXiv:2305.00293*.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2022. Fourier neural operator for parametric partial differential equations. In *The Eleventh International Conference on Learning Representations*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Ma, J.; and Wang, B. 2023. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*.

Ma, J.; Zhang, Y.; Gu, S.; Zhu, C.; Ge, C.; Zhang, Y.; An, X.; Wang, C.; Wang, Q.; Liu, X.; Cao, S.; Zhang, Q.; Liu, S.; Wang, Y.; Li, Y.; He, J.; and Yang, X. 2022. AbdomenCT-1K: Is Abdominal Organ Segmentation a Solved Problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6695–6714.

Mijnsbrugge, D. V.; Ongenae, F.; and Van Hoecke, S. 2021. Parameter Efficient Neural Networks With Singular Value Decomposed Kernels. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11.

Panahi, A.; Saeedi, S.; and Arodz, T. 2021. Shapeshifter: a parameter-efficient transformer using factorized reshaped matrices. *Advances in Neural Information Processing Systems*, 34: 1337–1350.

Sun, Y.; Chen, Q.; He, X.; Wang, J.; Feng, H.; Han, J.; Ding, E.; Cheng, J.; Li, Z.; and Wang, J. 2022. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. *Advances in Neural Information Processing Systems*, 35: 37484–37496.

Wang, D.; Zheng, Y.; Lian, H.; and Li, G. 2022. High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, 117(539): 1338–1356.

Wang, L.; Ye, X.; Zhu, L.; Wu, W.; Zhang, J.; Xing, H.; and Hu, C. 2023a. When SAM Meets Sonar Images. *arXiv preprint arXiv:2306.14109*.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Wang, Y.; Zhou, W.; Mao, Y.; and Li, H. 2023c. Detect Any Shadow: Segment Anything for Video Shadow Detection. *arXiv preprint arXiv:2305.16698*.

Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; and Arbel, T. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. 2021. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sensing*, 13(18): 3690.

Zhang, Y.; and Jiao, R. 2023. How Segment Anything Model (SAM) Boost Medical Image Segmentation? *arXiv preprint arXiv:2305.03678*.

Zheng, H.; Gong, M.; Liu, T.; Jiang, F.; Zhan, T.; Lu, D.; and Zhang, M. 2022. HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images. *Pattern Recognition*, 129: 108717.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.