

CONVOLUTION MEETS LORA: PARAMETER EFFICIENT FINETUNING FOR SEGMENT ANYTHING MODEL

Zihan Zhong*

Tsinghua University
zhongzh22@mails.tsinghua.edu.cn

Zhiqiang Tang

Amazon Web Services
zqtang@amazon.com

Tong He

Amazon Web Services
htong@amazon.com

Haoyang Fang

Amazon Web Services
haoyfang@amazon.com

Chun Yuan

Tsinghua University
yuanc@sz.tsinghua.edu.cn

ABSTRACT

The Segment Anything Model (SAM) stands as a foundational framework for image segmentation. While it exhibits remarkable zero-shot generalization in typical scenarios, its advantage diminishes when applied to specialized domains like medical imagery and remote sensing. To address this limitation, this paper introduces Conv-LoRA, a simple yet effective parameter-efficient fine-tuning approach. By integrating ultra-lightweight convolutional parameters into Low-Rank Adaptation (LoRA), Conv-LoRA can inject image-related inductive biases into the plain ViT encoder, further reinforcing SAM’s local prior assumption. Notably, Conv-LoRA not only preserves SAM’s extensive segmentation knowledge but also revives its capacity of learning high-level image semantics, which is constrained by SAM’s foreground-background segmentation pretraining. Comprehensive experimentation across diverse benchmarks spanning multiple domains underscores Conv-LoRA’s superiority in adapting SAM to real-world semantic segmentation tasks.¹

1 INTRODUCTION

The AI community have witnessed the explosion development of a series of foundation models in recent years, such as CLIP (Radford et al., 2021), GPT-4 (OpenAI, 2023) and ViT-22B (Dehghani et al., 2023). Recently, Segment Anything (SAM) (Kirillov et al., 2023), a promptable model pretrained on over 1 billion masks and 11 million images, emerged as a foundation model for image segmentation. Despite its impressive zero-shot performance on generic object segmentation, it doesn’t perform well on many real-world segmentation tasks in certain domains (Tang et al., 2023; Ji et al., 2023; Zhou et al., 2023), such as natural images (Borji et al., 2019; Fan et al., 2020a), agriculture (Sriwastwa et al., 2018), remote sensing (Xu et al., 2018) and medical images (Fan et al., 2020b).

Following the pretraining-finetuning paradigm (Dosovitskiy et al., 2020; He et al., 2022; Liu et al., 2021a), it is natural to finetune SAM on downstream tasks to enhance its performance. However, existing works (Zhang & Liu, 2023; Chen et al., 2023; Shaharabany et al., 2023) have failed to either analyze or address certain limitations inherent in SAM. 1) SAM’s image encoder is a plain ViT, which is known to lack of vision-specific inductive biases (Chen et al., 2022) that are useful for dense predictions. 2) SAM’s pretraining is essentially a binary mask prediction task that, where, given one prompt, it separates foreground object from background. The low-level mask prediction pretraining hinders SAM’s ability to capture high-level image semantic information crucial for tasks like multi-class semantic segmentation.

To tackle the above limitations and still retain SAM’s valuable segmentation knowledge acquired during pretraining, we finetune a small set of (extra) model parameters while freezing most of SAM’s pretrained weights, hence parameter efficient finetuning (PEFT). This raises the question: *Can PEFT*

*Work done while interning at Amazon Web Services.

¹Our code is public available at <https://github.com/autogluon/autogluon/tree/master/examples/automm/Conv-LoRA>

RGB Images	GT	VPT	LoRA	Conv-LoRA
------------	----	-----	------	-----------

Figure 1: Comparison of VPT, LoRA, and Conv-LoRA (ours) in binary-class road segmentation (top) and multi-class transparent object segmentation (bottom). Conv-LoRA reinforces image-related local priors, allowing SAM to separate roads from adjacent buildings, while LoRA and VPT struggle in this regard. In the second row, VPT produces a reasonable mask for the bowls but erroneously assigns them to the jar/kettle class (indicated by object color), revealing SAM's limited high-level semantic understanding. Both LoRA and Conv-LoRA rectify this misclassification through retuning SAM's image encoder, with Conv-LoRA delivering a cleaner mask with fewer boundary artifacts.

enhance SAM encoder with image-related local prior and facilitate the acquisition of high-level semantic information?

In this paper, we propose a new PEFT method named Conv-LoRA by diving into Low-Rank Adaptation (LoRA) (Hu et al., 2021). LoRA introduces slim trainable linear projection layers into each transformer layer of SAM's encoder, thereby helping recover its capacity to extract high-level semantic information. Our experiments demonstrate that LoRA surpasses the widely-adopted visual prompt tuning (VPT) (Jia et al., 2022), particularly in the multi-class semantic segmentation tasks. On top of LoRA, Conv-LoRA integrates lightweight convolution layers within its bottleneck structure. Convolution can introduce the image-related local prior (i.e. a pixel exhibits stronger correlation with its neighbors than the distant pixels) (Chen et al., 2022) through the local spatial operations.

Furthermore, it is essential to inject the local prior into the appropriate scale(s) of image features, considering the potential variations in object scales. To this end, Conv-LoRA draws inspiration from the concept of Mixture-of-Experts (MoE) (Shazeer et al., 2017) and incorporates multiple parallel convolutional experts, each specializing in a distinct feature scale. Given that ViT processes image features at a fixed scale, typically downsampling them by a factor of 16 from the original resolution, each expert in Conv-LoRA initially recovers image features at a specific scale, applies convolutional operations, and then reverts the features to the default scale. Compared to ViT-adaptor (Chen et al., 2022) and vision-specific transformers like Swin Transformer (Liu et al., 2021a), Conv-LoRA provides an implicit way to enforce multi-scale local priors, assuming it can leverage image features at the default scale to reconstruct feature information at higher scales. Fortunately, SAM's supervised pretraining, which involves masks of various scales, enables the ViT to acquire knowledge of image features beyond the default scale.

In the spirit of PEFT, we also remove the prompt encoder and add lightweight MLPs in the mask decoder for multi-class prediction. This simple modification has transformed SAM into an end-to-end model that can be retuned on both binary and multi-class semantic segmentation applications. Overall, our contribution can be summarized as follows:

- We present an innovative PEFT technique Conv-LoRA. By incorporating supplementary convolution operations, Conv-LoRA reinforces the local prior of SAM from the perspective of handling the limitation of plain ViT.
- Conv-LoRA uses MoE to model the process of dynamically selecting the proper feature scale to inject the vision-specific inductive biases.
- Our investigations reveal that SAM's pretraining has impeded its ViT encoder's capacity to learn high-level image semantic information. However, LoRA demonstrates the potential to help SAM recover this crucial ability.

- We conduct an extensive benchmark encompassing diverse domains, including natural images, agriculture, remote sensing, and healthcare. Conv-LoRA consistently exhibits superior performance over other PEFT techniques in various downstream tasks.

2 RELATED WORK

Parameter Efficient Fine-Tuning (PEFT). Parameter Efficient Fine-Tuning (PEFT) minimizes computational and storage requirements by selectively fine-tuning a small subset of model parameters, while keeping the majority fixed. PEFT encompasses methods such as adapter-based techniques, selective parameter tuning, prompt-driven fine-tuning, and Low-Rank Adaptation (LoRA) emerging from Natural Language Processing (NLP). In the adapter paradigm (Houlsby et al., 2019; Hu et al., 2021; Sung et al., 2022), compact modules are inserted within transformer layers, and other approaches (Guo et al., 2020; Zaken et al., 2021) involve fine-tuning a small fraction of parameters from pre-trained backbones. Prompt tuning (Lester et al., 2021; Li & Liang, 2021) adds adaptable tokens to input or intermediate sequences, and LoRA (Hu et al., 2021) introduces trainable low-rank matrices into transformer layers for weight updates.

PEFT techniques have also proven effective in the Computer Vision (CV) domain. Visual Prompt Tuning (VPT) (Jia et al., 2022) applies prompt tuning concepts (Lester et al., 2021) to image classification, while Scale and Shift Feature Modulation (SSF) (Lian et al., 2022) uses scale and shift parameters for modulating visual features in image classifiers. Convpass (Jie & Deng, 2022) introduces a convolutional bottleneck to enhance ViT’s performance in image classification. In our study, we focus on developing PEFT for SAM in semantic segmentation tasks, specifically enforcing multi-scale local priors beyond the default scale, distinguishing our approach from Convpass.

Segmentation Models FCN (Long et al., 2015) is a key deep image segmentation model that directly generates pixel-wise segmentation maps from images. U-Net (Ronneberger et al., 2015) employs an encoder-decoder structure with skip connections to preserve fine-grained spatial information. Deeplab (Chen et al., 2017a) integrates atrous (dilated) convolutions for multi-scale context, while PSPNet (Zhao et al., 2017) uses a pyramid pooling module. DANet (Fu et al., 2019), SANet (Zhong et al., 2020), and EMA (Li et al., 2019) utilize attention mechanisms for contextual dependencies. Transformer architectures like PVT (Wang et al., 2021), Swin (Liu et al., 2021b), CvT (Wu et al., 2021), CoaT (Xu et al., 2021), LeViT (Graham et al., 2021), Segformer (Xie et al., 2021a), and PVT v2 (Wang et al., 2022) bring various improvements. SAM (Ji et al., 2023), a recent breakthrough in segmentation, offers a universal approach for segmenting diverse objects and regions in images. Fine-tuning SAM on downstream tasks is recommended due to a lack of high-level semantic information and potential domain bias in the pre-training dataset.

Fine-tuning SAM. Some prior works (Chen et al., 2023; Zhang & Liu, 2023; Wu et al., 2023; Chai et al., 2023; Shaharabany et al., 2023; Hu et al., 2023; Wang et al., 2023) explore fine-tuning SAM for downstream tasks. These methods include tuning SAM’s mask decoder or integrating parameter-efficient tuning methods with SAM’s image encoder. Some of them (e.g., (Chen et al., 2023; Zhang & Liu, 2023; Shaharabany et al., 2023)) provide end-to-end solutions to automate SAM. Our method further addresses the structural limitation of SAM’s image encoder for capturing visual-specific inductive biases by introducing convolution operations. And we unveil that SAM’s pretraining hampers its ViT encoder’s ability to learn high-level semantic information. We also transform SAM into an end-to-end semantic segmentation model with minor architectural adjustments.

Mixture-of-Experts. Mixture-of-Expert (MoE) is designed to expand model capacity while introducing small computational overhead. An MoE layer leverages multiple experts to enhance model capacity, while using the gating network to regulate sparsity for computational savings. Feed-Forward Networks (FFN) are commonly employed as the default choice for experts (Shazeer et al., 2017; Riquelme et al., 2021; Bao et al., 2022; Du et al., 2022; Zhou et al., 2022; Fedus et al., 2022). Some efforts (Zuo et al., 2021; Zhou et al., 2022) focus on more efficient gating mechanisms.

In our work, we utilize the concept of MoE, not aiming at improving it. We compare MoE used in our work with original MoE in three aspects: 1) The original goal of MoE is to expand model capacity without excessively increasing computational overhead, whereas ours is to dynamically inject the local prior into the feature maps of different scales. 2) The structures of experts in MoE are typically the same, whereas ours are not. Each expert specializes in a specific scaling operation

Figure 2: LoRA vs. Conv-LoRA. Both LoRA and Conv-LoRA add an extra trainable encoder-decoder structure parallel to the frozen pre-trained weights. Inside the bottleneck of LoRA, Conv-LoRA inserts lightweight convolution operations managed by MoE with negligible extra parameters.

in our method. 3) While MoE is mostly employed during pre-training, we employ MoE as a part for parameter-efficient tuning on the downstream tasks.

3 METHOD

3.1 CONV-LoRA

LoRA. First, let's briefly recap the design of LoRA (Hu et al., 2021), which uses an encoder-decoder structure to impose a low-rank constraint on the weight updates (Fig. 2 (a)). It freezes the pre-trained model weights and injects small trainable rank decomposition matrices into each layer of the transformer architecture. Specifically, given a pre-trained weight matrix $W_0 \in \mathbb{R}^{b \times a}$, LoRA adds a pair of linear encoder W_e and decoder W_d , i.e., trainable rank decomposition matrices, along its side. W_e and W_d satisfy the low rank constraint $W_e \in \mathbb{R}^{a \times r}$, $W_d \in \mathbb{R}^{b \times r}$, and $r \leq \min(a, b)$. With LoRA, the forward pass changes from $h = W_0x$ to:

$$h = W_0x + W_dW_ex \tag{1}$$

Conv-LoRA aims to incorporate convolution operations between the encoder and decoder components of LoRA (Fig. 2 (b)). On one hand, convolution can inject the image-related local prior, addressing fundamental limitation of the vanilla ViT. On the other hand, the low-rank constraint ensures that the convolution layers remain exceedingly lightweight, preserving the PEFT nature of Conv-LoRA.

A pivotal consideration in designing Conv-LoRA is determining the scale of feature maps at which to introduce the local prior. While the feature maps in ViT are uniform in scale, object masks typically encompass a wide range of scales. Therefore, it is crucial to apply convolution operations at the right scale. To tackle this challenge, we draw inspiration from the concept of Mixture of Experts (MoE) (Shazeer et al., 2017). MoE comprises multiple expert networks and a gating module that dynamically selects which expert(s) to activate during the forward pass (Fig. 3). Adapting this concept to Conv-LoRA, each expert specializes in convolution at a specific scale of feature maps, and a compact gating module learns to dynamically choose the expert(s) based on the input data. Mathematically, with Conv-LoRA, eq. (1) changes to:

$$h = W_0x + W_d \sum_i^{X^0} G(W_ex)_i E_i(W_ex) \tag{2}$$

where $W_0 \in \mathbb{R}^{C_{out} \times C_{in}}$, $W_e \in \mathbb{R}^{r \times C_{in}}$, $W_d \in \mathbb{R}^{C_{out} \times r}$, $x \in \mathbb{R}^{B \times C_{in} \times H \times W}$. B is batch size, $C_{in} = C_{out}$ is the number of input / output channels, and H, W correspond to the height and width. E_i is the i -th expert of all n experts. G is the gating network with only top- k (default 1) values activated. Refer to appendix A for more details of gating.

Inside each expert, three key operations are arranged in sequence: an interpolation function that reconstructs feature maps at a specific scale, a convolutional layer, and a subsequent interpo-

Figure 3: MoE-Conv. It consists of experts and a gating network for dynamic expert selection. Each expert reconstructs feature maps at a specific scale, applies convolution, and returns the feature maps to the default scale. Each expert specializes in one unique feature scale.

lation operation to map the feature maps back to the default feature scale of ViT. Assume expert E_i is in charge of scale s_i , we can formulate it as:

$$E_i(x) = \text{Interpolate}(\text{Conv}_3(\text{Interpolate}(x; s_i)); 1=s) \quad (3)$$

For instance, if $s_i = 4$, expert E_i would initially upscale the feature maps by a factor of 4x, apply the Conv_3 operation, and finally, downscale the feature maps by 4x.

MoE vs. Multi-scale. In contrast to MoE, another method to address diverse scales is employing a multi-scale strategy. This approach utilizes multiple branches to concurrently inject local priors at various scales and aggregates the results. Although seemingly more straightforward, this method comes at a higher computational cost when compared to MoE. The efficiency of MoE stems from its capacity to selectively activate sparse experts, thereby minimizing computational overhead. Given our priority on efficient netuning, we favor MoE as a discerning choice.

3.2 END-TO-END MULTI-CLASS SEGMENTATION WITH SAM

SAM comprises three essential components: an image encoder, a prompt encoder, and a mask decoder. When provided with an image and a prompt, which can take the form of a point, box, mask, or text, the mask decoder generates a mask of the object associated with the given prompt. While this prompt-based approach renders SAM flexible for integration into larger systems, such as interactive segmentation or a combination of detection and subsequent segmentation, it does pose challenges in making SAM an end-to-end model in practical applications. To automate SAM, we freeze the prompt encoder, thus always constant prompt tokens to mask decoder, when netuning it on downstream tasks. Moreover, the original mask decoder is designed to predict binary masks, distinguishing between foreground and background based on the given prompt. To adapt SAM for multi-class semantic segmentation tasks, we introduce a straightforward classification branch (depicted as the red dashed box in Fig. 4) within the mask decoder. This extra branch is responsible for predicting classification scores. Additionally, we apply full netuning to the mask decoder as it is a lightweight module. For more comprehensive information, refer to appendix B.

4 EXPERIMENTS

Settings. We perform several experiments using SAM on four real-world scenarios, including medical images, natural images, agriculture and remote sensing. We use the batch size of 4 and Adam optimizer with learning rate of 10^{-4} as default, with a weight decay of 10^{-4} . A larger learning rate of $3 \cdot 10^{-4}$ is found useful for the datasets we use in agriculture and remote sensing. The random horizontal flip is applied during training as data augmentation. All the methods are trained for 30 epochs with structure loss (i.e., the combination of weighted IoU loss and binary cross entropy loss) unless otherwise specified. Additionally, our Conv-LoRA follows Shazeer et al. (2017) to introduce extra loss for balancing the utilization among the experts. The weight of the extra loss is set to 1.0 and 2.0 for binary-class and multi-class semantic segmentation respectively. We set the number of experts to be 8 by default, with each expert specializing in a scaling ratio within the

Figure 4: The modified SAM's mask decoder for multi-class semantic segmentation. The classification module (within the red dashed box) is newly added compared to original SAM's mask decoder. N is the number of output mask tokens, K is the number of classes, C is the number of channels, H and W indicate the height and width of the feature map (we omit 'batch size' for simplicity).

continuous range from 1 to 8. And we apply Conv-LoRA to the query, key and value matrices in self-attention layers, same as how LoRA does.

Datasets Our experiments encompass semantic segmentation datasets from various domains, spanning natural images, medical images, agriculture, and remote sensing. In the natural image domain, we explore two specific tasks: camouflaged object segmentation (Fan et al., 2020a; Skurowski et al., 2018; Le et al., 2019) and shadow detection (Vicente et al., 2016). Within medical segmentation, we investigate polyp segmentation (Jha et al., 2020; Bernal et al., 2015; Tajbakhsh et al., 2015; Jha et al., 2017; Silva et al., 2014) and skin lesion segmentation (Codella et al., 2018). For agriculture and remote sensing, we employ the leaf disease segmentation (Rath, 2023) and road segmentation (Mnih, 2013) datasets as representative examples, respectively. We also explore multi-class transparent object segmentation using Trans10K-v1 (Xie et al., 2020) with 3 classes and Trans10K-v2 (Xie et al., 2021b) with 12 fine-grained classes. Further details about each dataset can be found in appendix C.

Baselines We compare our method with the following methods: 1) Fine-tune SAM's mask decoder only. 2) BitFit (Zaken et al., 2021), which only fine-tunes bias terms in the pre-trained model. 3) Adapter (Houlsby et al., 2019), which inserts the trainable bottleneck layers between the transformer layers. 4) SAM-Adapter (Chen et al., 2023), which further tunes the patch embedded features and learns an extra embedding for high-frequency components for low-level semantic segmentation tasks. And it is one of the pioneer works that apply PEFT method to SAM. 5) VPT (Jia et al., 2022), which inserts learnable tokens to hidden states for each transformer layer. 6) LST (Sung et al., 2022), which inserts a trainable side network parallel to the frozen backbone. To control the number of trainable parameters, the side adapter network employs a pre-trained ViT-Tiny model, similar to SAN (Xu et al., 2023). The features from the frozen backbone and the side adapter network are fused at the global attention layers of SAM. 7) SSF (Lian et al., 2022), which inserts learnable scale and shift parameters to modulate visual features during training. 8) LoRA (Hu et al., 2021) inserts trainable bottleneck layers parallel to the frozen linear weight.

To adhere to the page limit, we select representative datasets of different domains to report the main experiment results. Refer to appendix D for the full experiment results. All the experiments for PEFT methods are run for three times to ease the randomness. The average values and the standard error are reported in table 1.

4.1 BINARY-CLASS SEMANTIC SEGMENTATION

In table 1, all PEFT methods consistently outperform fine-tuning the mask decoder alone, underscoring the importance of fine-tuning the image encoder of SAM. Furthermore, Conv-LoRA surpasses other PEFT techniques across diverse datasets from different domains. Compared to other PEFT methods, Conv-LoRA uses lightweight convolution operations to strengthen the vision-specific local prior, which turns out effective in boosting the segmentation performance. The substantial performance gaps between SAM trained from scratch and Conv-LoRA also underscore the considerable assistance provided by SAM's pretraining knowledge in enhancing downstream task performance.

Method	#Params (M) / Ratio (%)	Medical						Natural Images				Agriculture		Remote Sensing	
		Kvasir		CVC-612		ISIC 2017		CAMO		SBU	Leaf		Road		
		S	E	S	E	Jac	Dice	S	E	F1	BER#	IoU	Dice	IoU	Dice
Domain Specific	* / 100%	90.9	94.4	92.6	95.5	80.1	87.5	80.8	85.8	73.1	3.56	62.3	74.1	59.1	73.0
SAM trained from scratch	641.09 / 100%	78.5	82.4	85.9	91.6	73.8	82.5	61.9	67.0	40.5	5.53	52.1	65.5	55.6	71.1
decoder-only	3.51 / 0.55%	86.5	89.5	85.5	89.9	69.7	79.5	78.5	83.1	69.8	14.58	50.8	63.8	48.6	65.1
BitFit	3.96 / 0.62%	90.8	93.8	89.0	91.6	76.4	84.7	86.8	90.7	81.5	3.16	71.4	81.7	60.6	75.2
Adapter	3.92 / 0.61%	91.2	94.0	89.3	92.0	76.7	85.0	87.7	91.3	82.8	2.84	72.1	82.4	61.5	75.9
VPT	4.00 / 0.62%	91.5	94.3	91.0	93.7	76.9	85.1	87.4	91.4	82.1	2.70	73.6	83.8	60.2	74.9
LST	11.49 / 1.77%	89.7	93.3	89.4	92.4	76.4	84.9	83.3	88.0	77.1	3.18	70.2	81.1	60.2	74.9
SAM-Adapter	3.98 / 0.62%	89.6	92.5	89.6	92.4	76.1	84.6	85.6	89.6	79.8	3.14	71.4	82.1	60.6	75.2
SSF	4.42 / 0.69%	91.3	93.9	89.6	91.9	76.6	85.0	87.5	91.4	82.6	3.19	71.5	81.8	61.6	76.0
LoRA	4.00 / 0.62%	91.2	93.8	90.7	92.5	76.6	84.9	88.0	91.9	82.8	2.74	73.7	83.6	62.2	76.5
Conv-LoRA	4.02 / 0.63%	92.0	94.7	91.3	94.0	77.6	85.7	88.3	92.4	84.0	2.54	74.5	84.3	62.6	76.8

Table 1: Results on binary semantic segmentation. #Params (M) / Ratio (%) represents the number of trainable parameters and its proportion relative to the total number of parameters. Domain Specific as a placeholder referring to methods that specifically designed for the tasks. Underlined denotes the better results compared to PEFT methods. See appendix D for more details. Compared to LoRA, Conv-LoRA incurs negligible parameter overhead, but delivers a clear performance boost.

Method	#Params (M) / Ratio (%)	Easy				Hard			
		Acc	mIoU	MAE #	MBER #	Acc	mIoU	MAE #	MBER #
TransLab	42.19 / 100%	95.77	92.23	0.036	3.12	83.04	72.10	0.166	13.30
decoder-only	3.51 / 0.55%	94.68	88.54	0.050	4.24	83.53	68.30	0.186	14.37
VPT	4.00 / 0.62%	98.31	95.73	0.017	1.52	90.42	83.38	0.083	7.21
LoRA	4.00 / 0.62%	98.44	96.26	0.016	1.35	91.94	83.95	0.083	6.35
Conv-LoRA	4.02 / 0.63%	98.63	96.45	0.015	1.27	93.05	84.37	0.075	6.25

Method	# Params (M) / Ratio (%)	Acc	mIoU	Category IoU											
				bg	shelf	jar	freezer	window	door	eyeglass	cup	wall	bowl	bottle	box
TransLab	42.19 / 100%	92.67	69.00	93.90	54.36	64.48	65.14	54.58	57.72	79.85	81.61	72.82	69.63	77.50	56.43
Trans2Seg	56.20 / 100%	94.14	72.15	95.35	53.43	67.82	64.20	59.64	60.56	88.52	86.67	75.99	73.98	82.43	57.17
decoder-only	3.51 / 0.55%	90.66	49.97	93.66	32.75	39.96	35.87	50.70	45.89	57.38	73.16	69.36	54.23	56.58	33.77
VPT	4.00 / 0.62%	94.42	62.81	97.41	29.76	52.82	62.09	55.54	63.61	81.12	83.40	79.61	65.29	72.92	44.77
LoRA	4.00 / 0.62%	94.80	66.01	97.50	42.17	57.82	64.35	53.44	64.08	87.28	85.28	80.43	63.67	77.97	49.56
Conv-LoRA	4.02 / 0.63%	95.07	67.09	97.66	50.51	58.44	51.70	55.69	65.22	85.23	84.84	80.97	72.84	79.83	52.73

Table 2: Results on multi-class semantic segmentation. The tables are the results for three-class and twelve-class semantic segmentation respectively.

Additionally, while Conv-LoRA outperforms certain domain-specific methods that are having more trainable parameters, it may still fall short on specific datasets. It's important to note that Conv-LoRA aims to be a general-purpose PEFT method for adapting SAM to various domains rather than competing with these domain-specific models. Tailoring SAM for specific domain applications with more intricate adjustments might yield superior performance compared to specialized model designs.

4.2 MULTI-CLASS SEMANTIC SEGMENTATION

In table 2, while decoder-only re-tuning approaches achieve comparable segmentation accuracy with domain-specific methods, they exhibit a substantial gap in terms of mIoU metrics. While accuracy measures pixel-level segmentation performance, mIoU takes mask class information into account. We suspect that SAM's image encoder encounters challenges in extracting high-level semantic information that is valuable for classification tasks. Moreover, re-tuning the image encoder using PEFT methods results in a significant boost in mIoU, indicating a restoration of its capability to learn high-level image semantics.

Additionally, we conduct linear probing experiments on SAM's image encoder. Specifically, we freeze SAM's ViT-B encoder and train only a linear head on ImageNet-1K. Similarly, we perform linear probing for the ViT-B pretrained with MAE (He et al., 2022). The results reveal that SAM's image encoder exhibits significantly lower ImageNet-1K accuracy compared to the MAE encoder (54.2% vs. 67.7%). Given that SAM's image encoder is initialized using an MAE encoder, we

Figure 5: Mean attention distance of each attention head, with each dot indicating the mean distance across images for one of the 16 heads at one layer. In contrast to MAE, SAM retains the ability to incorporate local information even in deeper layers.

hypothesize the pre-training focused on low-level foreground-background mask prediction adversely affects the ViT encoder’s ability to capture high-level semantic information for classification.

To be specific, SAM was trained on a dataset that exclusively comprises segmentation masks without explicit semantic information. In theory, to minimize loss, the fundamental objective for its encoder is to project pixels into a metric space where pixels from the same object are in close proximity, while those from distinct objects are distantly positioned. This projection requires an implicit understanding of ‘objectness’, focusing on proximity within an image rather than preserving consistent representations of the same-type object across different images. This introduces a potential challenge in aligning representations with semantics across diverse images.

4.3 ABLATION STUDY

SAM’s local prior assumption. SAM’s local prior is grounded in its extensive segmentation pre-training. Through supervised training on a vast dataset encompassing 1 billion high-quality masks and 11 million images, SAM has honed a robust capability to discern and capture local features within images. Notably, SAM’s encoder retains the ViT architecture, which inherently lacks a dedicated local prior. However, this deficiency is effectively compensated by the significant local prior acquired through segmentation pretraining.

We analyze using the mean attention distance as a metric. In Fig. 5, SAM exhibits numerous heads in the deep blocks with short mean attention distances, suggesting its heightened focus on local information during the later stages of the encoder. In contrast, the MAE pretrained ViT, representing SAM’s initialization, displays consistently long mean attention distances among attention heads in the later stages. Consequently, SAM’s segmentation pretraining induces a transformative shift in ViT’s attention heads, steering them from a global-oriented to a local-oriented configuration. This transformation underscores the efficacy of SAM’s approach in imbuing the model with a distinctive local prior, enhancing its ability to capture fine-grained details within images.

MoE vs. Multi-scale. Conv-LoRA leverages MoE to dynamically inject the local prior into feature maps of different scales (Fig. 6 (a)), as the optimal scale remains unclear. Here we explore the impact of a multi-scale strategy, which fuses the features from all scales simultaneously (Fig. 6 (b)).

We compare performance and training costs in table 3. Dynamic MoE outperforms multi-scale direct addition in terms of performance. This could be attributed to the preference for specific scales’ feature maps based on different inputs. When injecting the local prior into feature maps of all scales simultaneously, the discrepancy of the importance diminishes as the information from critical feature maps is smoothed out. However, with dynamic selection of the top-1 experts for each forward pass, the information from these crucial scales takes precedence and isn’t diluted by other

scales. Dynamic MoE also provides a 1.54x speedup and reduces memory usage by 1.7GB during training. In summary, the comparison underscores the effectiveness and efficiency of MoE.

Method	ISIC 2017				Method	Training Speed (Iter / s)	Training Memory (GB) #
	Jac	T-Jac	Dice	Acc			
Multi-scale	77.4	69.5	85.4	93.7	Multi-scale	0.79	23.4
MoE	77.9	70.3	85.9	93.9	MoE	1.22	21.7

Table 3: MoE vs. Multi-scale, the latter fuses features from all scales simultaneously. The performance and the training cost illustrate the effectiveness and efficiency of dynamic selecting the experts, i.e., injecting the local prior into the feature maps of different scales dynamically.

The 'Optimal' Scale of Feature Map for Different Datasets. While we are unable to definitively determine the optimal scale for introducing the local prior, we could check whether the 'optimal' scale within a given range varies indeed across different datasets.

Specifically, we simply modify Conv-LoRA: use only one expert and a specific scaling ratio. We set the scaling ratio to 1, 2, 4 respectively. The experiments are conducted on Leaf Disease Segmentation dataset and ISIC 2017 dataset.

Scaling Ratio	Leaf			ISIC 2017			
	IoU	Dice	Acc	Jac	T-Jac	Dice	Acc
1	73.6	83.4	95.5	76.8	69.1	85.0	93.7
2	73.8	83.7	95.8	77.3	69.7	85.4	93.8
4	74.0	83.7	95.9	76.9	68.4	85.1	93.7
8	73.2	83.1	96.0	76.9	68.3	85.2	93.6

Table 4: Comparison of the 'optimal' scale of feature map for local prior injection across different datasets. We modify Conv-LoRA to use one expert with a specific scaling ratio.

In table 4, the 'optimal' scale indeed varies across the different datasets. The scaling ratio set to 4 is optimal for Leaf Disease Segmentation dataset, whereas it is set to 2 for ISIC 2017 dataset. These results further confirm our assumption and the necessity for our dynamic local prior injection based on different inputs.

For more ablation experiments, analyses (e.g., further analyses of local prior and MoE) and visualization, refer to appendix E through appendix I.

5 CONCLUSION

Parameter efficient netuning (PEFT) is a popular way when adapting foundation models to various downstream tasks. We present Conv-LoRA, a novel PEFT approach for applying SAM to downstream segmentation applications. Conv-LoRA is simple, generic, and obtains promising results over multiple domains including natural images, agriculture, remote sensing, and healthcare. Moreover, ours shed light on several aspects of SAM: 1) although the large-scale supervised segmentation pretraining can provide image-related local prior knowledge from the data perspective, injecting lightweight convolution operations in the ViT encoder can further boost the exploitation of local prior from another perspective of architecture; 2) the foreground-background segmentation pretraining prevents the image encoder from learning high-level semantic information, which can be alleviated through netuning relatively few parameters in the encoder.

Our efforts primarily focus on developing a general PEFT method for SAM, showing stronger performance than existing PEFT methods in a broad spectrum of benchmarks, other than directly competing with state-of-the-art (SOTA) models in specialized domains. Given that SAM netuned with Conv-LoRA may not yet consistently outperform domain-specific SOTA models, we believe that tailoring the mask decoder and prompt encoder beyond image encoder netuning, and combining Conv-LoRA with other PEFT methods can be promising directions for domain-specific applications.

ACKNOWLEDGEMENTS

We thank all the anonymous reviewers for their helpful comments. This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012), Shenzhen Key Laboratory (ZDSYS20210623092001004), and Beijing Key Lab of Networked Multimedia.

REFERENCES

- Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlm: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems* 35:32897–32912, 2022.
- Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, CV Jawahar, and Manohar Paluri. Improved road connectivity by joint learning of orientation and segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2019:10385–10393, 2019.
- Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilafino. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics* 48:99–111, 2015.
- Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.
- Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media* 5:117–150, 2019.
- Shurong Chai, Rahul Kumar Jain, Shiyu Teng, Jiaqing Liu, Yinhao Li, Tomoko Tateyama, and Yen-wei Chen. Ladder re-tuning approach for sam integrating complementary networks. *arXiv preprint arXiv:2306.12737*, 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* 36(4):834–848, 2017a.
- Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017b.
- Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?—sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.
- Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense prediction. *arXiv preprint arXiv:2205.08534*, 2022.
- Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* 34:17864–17875, 2021.
- Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2022:1290–1299, 2022.
- Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic) 2018. *IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 168–172. IEEE, 2018.

- Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. *International Conference on Machine Learning* pp. 7480–7512. PMLR, 2023.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*2020.
- Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning* pp. 5547–5569. PMLR, 2022.
- Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 2777–2787, 2020a.
- Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pronet: Parallel reverse attention network for polyp segmentation. *International conference on medical image computing and computer-assisted intervention* pp. 263–273. Springer, 2020b.
- Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (10):6024–6042, 2022. doi: 10.1109/TPAMI.2021.3085766.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research* (1): 5232–5270, 2022.
- Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 3146–3154, 2019.
- Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve Jegou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 12259–12269, 2021.
- Demi Guo, Alexander M Rush, and Yoon Kim. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*2020.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 16000–16009, 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.
- Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv preprint arXiv:2306.13731*2023.
- Debesh Jha, Pia H Smedsrud, Michael A Riegler, Håvard D Johansen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II* 26, pp. 451–462. Springer, 2020.

- Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05752*, 2023.
- Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.
- Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022.
- 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.
- Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding* 184:45–56, 2019.
- Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9167–9176, 2019.
- Xiang Lisa Li and Percy Liang. Pre-tuning: Optimizing continuous prompts for generalization. *arXiv preprint arXiv:2101.00190*, 2021.
- Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, pp. 109–123, 2022.
- Weicong Liang, Yuhui Yuan, Henghui Ding, Xiao Luo, Weihong Lin, Ding Jia, Zheng Zhang, Chao Zhang, and Han Hu. Expediting large-scale vision transformer for dense prediction without re-tuning. *Advances in Neural Information Processing Systems*, pp. 35462–35477, 2022.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021a.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- Volodymyr Mnih. Machine learning for aerial image labeling. University of Toronto (Canada), 2013.
- OpenAI. Gpt-4 technical report, 2023.
- Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.
- 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. *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, pp. 12116–12128, 2021.

- Sovit Ranjan Rath. Leaf disease segmentation dataset, 2023. URL <https://www.kaggle.com/datasets/sovitath/leaf-disease-segmentation-with-trainvalid-split>.
- Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, Andr Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18 pp. 234–241. Springer, 2015.
- Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014.
- Przemysław Skurowski, Hassan Abdulameer, J B aszczyk, Tomasz Depta, Adam Kornacki, and P Kozie . Animal camou age analysis: Chameleon database. *Published manuscript* 2(6):7, 2018.
- Apurva Sriwastwa, Shikha Prakash, Swati Swarit, Khushboo Kumari, Sitanshu Sekhar Sahu, et al. Detection of pests using color based image segmentation. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* 393–396. IEEE, 2018.
- Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.
- Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging* 35(2):630–644, 2015.
- Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camou aged object detection. *arXiv preprint arXiv:2304.04709*, 2023.
- David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdal, Aaron Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering* 8(1):1–10, 2017.
- Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI* 14 pp. 816–832. Springer, 2016.
- An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation. *arXiv preprint arXiv:2308.07156*, 2023.

- Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 568–578, 2021.
- Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* 8(3):415–424, 2022.
- Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Pvt: Introducing convolutions to vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision* pp. 22–31, 2021.
- Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.
- Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pp. 696–711. Springer, 2020.
- Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Seg-former: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34:12077–12090, 2021a.
- Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. *arXiv preprint arXiv:2101.08461*, 2021b.
- Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 2945–2954, 2023.
- Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 9981–9990, 2021.
- Yongyang Xu, Zhong Xie, Yaxing Feng, and Zhanlong Chen. Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sensing* 10(9):1461, 2018.
- Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bit t: Simple parameter-efficient re-tuning for transformer-based masked language models. *arXiv preprint arXiv:2106.10192*, 2021.
- Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.
- Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* pp. 14–24. Springer, 2021.
- 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.
- Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 13065–13074, 2020.
- Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*, 2023.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson WH Lau. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4702–4711, 2021.

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. Taming sparsely activated transformer with stochastic experts preprint arXiv:2110.04260 2021.

A GATING NETWORK OF CONV-LORA

In order to calculate the gating scores $G(x) \in \mathbb{R}^{B \times k}$, i.e., the scores for B samples in a batch assigned to k experts, we need to calculate the values $H(x) \in \mathbb{R}^{B \times n}$ of n experts first. As the size of input x is $B \times r \times H \times W$ (we use x to represent $W \times x$ in eq. (2) for simplicity), we apply global average pooling (denoted as 'AvgPool') followed by a reshaping operation (denoted as 'Reshape'), to obtain $x_h \in \mathbb{R}^{B \times r}$. Specifically, the size of k is changed to $B \times r \times 1 \times 1$ after 'AvgPool', and then changed to $B \times r$ after 'Reshape'. Then, following (Shazeer et al., 2017), we apply the trainable gating weight $W_g \in \mathbb{R}^{r \times n}$, and the noise term $W_{noise} \in \mathbb{R}^{r \times n}$ to calculate $H(x)$:

$$\begin{aligned} x_h &= \text{Reshape}(\text{AvgPool}(x); (B; r)) \\ H(x)_i &= (x_h \cdot W_g)_i + \text{StandardNormal}(\text{Softplus}(x_h \cdot W_{noise})_i) \end{aligned} \quad (4)$$

We keep only the top k values (denoted as 'KeepTopK') based on $H(x)_i$ of each expert E_i , setting the rest to 1, whose corresponding gate values is equal to 0 after a 'Softmax' operation. If $G(x)_i$ is 0, we need not compute $E_i(x)$:

$$\begin{aligned} G(x) &= \text{Softmax}(\text{KeepTopK}(H(x); k)) \\ \text{where } \text{KeepTopK}(v; k)_i &= \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v. \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \quad (5)$$

B MASK DECODER FOR MULTI-CLASS SEGMENTATION

SAM's mask decoder follows the design in mask-based segmentation, wherein the image is grouped into N (i.e., the number of output tokens in Fig. 4) regions represented by binary masks. However, unlike other mask-based methods such as Maskformer (Cheng et al., 2021) and Mask2former (Cheng et al., 2022), SAM does not incorporate a classification module; it solely identifies the foreground, lacking the ability for multi-class segmentation. What we need to do is straightforward: incorporate a classification branch to obtain class predictions for the corresponding masks.

During training, we need to match the set of mask predictions and the set of ground truth segments. We follow the design in Mask2Former (Cheng et al., 2022), which proposes a memory-friendly way for bipartite matching between the predictions and ground truths. It efficiently alleviates the memory demands in semantic segmentation tasks that necessitate high-resolution mask prediction, particularly for SAM involving the processing of high-resolution input images (1024 × 1024). For the masks that are not allocated to any ground truths, an extra 'no object' category is introduced to ensure the one-to-one matching. Hence, the classification branch should produce $K + 1$ class predictions, assuming there are originally K categories. For semantic inference, we can exclude the 'no object' category and perform a straightforward matrix multiplication between the masks and the classification predictions to obtain pixel-wise predictions.

Following Mask2Former, we use binary cross-entropy loss and dice loss L_{dice} for mask loss, cross entropy loss L_{cls} for classification predictions:

$$L_{\text{multi-class}} = L_{\text{mask}} + L_{\text{cls}} + L_{\text{MoE}} \quad (6)$$

where $L_{\text{mask}} = L_{\text{ce}} + L_{\text{dice}}$.

C DATASET DETAILS

Polyp Segmentation (medical images) Polyp Segmentation, the task to identify abnormal growths known as polyps within gastrointestinal endoscopic images, plays a critical role in early colorectal cancer diagnosis and treatment planning, and it presents a formidable challenge due to the considerable diversity in polyp shapes and sizes. We choose two polyp segmentation datasets: Kvasir (Jha

et al., 2020) and CVC-ClinicDB/CVC-612 (Bernal et al., 2015). Kvasir contains 1000 images and CVC-ClinicDB, also called CVC-612, includes 612 open-access images. Fan et al. (2020b) divides the images into a 9:1 ratio for training and testing. Additionally, we randomly divide a validation set comprising 20% of the images from the training set, for validation during training. Furthermore, we also use images from CVC-ColonDB (Tajbakhsh et al., 2015), EndoScope-IVZ (Suzuki et al., 2017) and ETIS (Silva et al., 2014) for testing following the setting in (Fan et al., 2020b). We consistently train all methods for 30 epochs, unless otherwise specified for specific datasets.

Skin Lesion Segmentation (medical images) Skin Lesion Segmentation involves the segmentation of various types of skin lesions within medical images, serving a crucial role in the early diagnosis and treatment of skin disorders, notably skin cancer. However, this task is particularly challenging due to ambiguous boundaries and color variations. We choose ISIC 2017 (Codella et al., 2018) for skin lesion segmentation. ISIC 2017 provides 2000 images for training, 150 images for validation and 600 images for testing.

Camouflaged Object Segmentation (natural images) Camouflaged Object Segmentation focuses on identifying objects that are concealed within complex or visually cluttered backgrounds, which is more challenging compared to traditional object segmentation. We choose three camouflaged object detection datasets: COD10K (Fan et al., 2020a), CHAMELEON (Skurkowski et al., 2018), and CAMO (Le et al., 2019). COD10K contains 3040 training and 206 testing samples. CHAMELEON includes 76 images collected from the Internet for testing. And CAMO provides 1000 images for training and 250 for testing. Following (Fan et al., 2020a), we train on the combined dataset consisting of the training images from COD10K and CAMO for 20 epochs, and test on the three datasets. Additionally, we randomly split 10% of the images from the training set for validation.

Shadow Detection (natural images) Shadow Detection focuses on the recognition of shadow regions within a scene and can facilitate the estimation of lighting conditions or the removal of shadows. We choose SBU (Vicente et al., 2016), which is the largest annotated shadow dataset. SBU contains 4085 and 638 images for training and testing. We randomly split 10% of the images from the training set for validation and we train the methods for 10 epochs with balanced binary cross entropy loss.

Leaf Segmentation (agriculture) Leaf segmentation involves the identification of individual plant leaves within agricultural images and plays a crucial role in advancing automation for plant diseases control and high quality food production. We choose a Leaf Disease Segmentation dataset (Rath, 2023), which contains 498 images for training and 90 images for testing. We randomly split the training images into 80% for training, 20% for validation.

Road Segmentation (remote sensing) Road segmentation detects road or street regions within images or video frames, and is crucial for applications in autonomous driving, traffic analysis, and urban planning. We choose Massachusetts Roads Dataset (Mnih, 2013), which contains 1107 images for training, 13 images for validation and 48 images for testing. And we train the methods for 20 epochs.

Multiclass Semantic Segmentation (natural images) We choose Trans10K-v1 (Xie et al., 2020) and Trans10K-v2 (Xie et al., 2021b) dataset for multi-class transparent object segmentation. Trans10K-v1 dataset contains 10428 images, with background as one category and two more categories of transparent objects: Transparent Things (e.g., cups, bottles) and Stuff (e.g., windows). Trans10K-v2 dataset is based on Trans10K-v1 dataset, with more fine-grained categories annotations. The dataset contains background plus two main categories divided into 11 fine-grained categories: 1) Transparent Things containing cup, bottle, jar, bowl and eyeglass. 2) Transparent Stuff containing windows, shelf, box, freezer, glass walls and glass doors. In respect to fine-grained categories and high diversity, Trans10K-v2 is more challenging than Trans10K-v1. All the datasets use 5000, 1000 and 4428 images for training, validation and testing, respectively.

D FULL EXPERIMENT RESULTS

Here are the full experiment results for binary semantic segmentation. Our Conv-LoRA demonstrates superiority across diverse datasets from various domains compared to other PEFT methods. Noted that models indicated *italics* in the following tables are specifically designed for the corresponding tasks. We re-run the experiments on all the datasets using the model code provided by

the authors. We choose methods that are either widely established or relatively novel within their respective domains. With the exception of the Leaf Segmentation dataset, which is sourced from Kaggle, we utilize the author's provided sample code to conduct our experiments.

Polyp Segmentation

Method	# Params (M) / Ratio(%)	Kvasir				CVC-612			
		S *	E *	F ¹ *	MAE #	S *	E *	F ¹ *	MAE #
PraNet (Fan et al., 2020b)	32.55 / 100%	90.9	94.4	88.7	2.9	92.6	95.5	88.7	1.0
decoder-only	3.51 / 0.55%	86.5	89.5	77.9	5.1	85.5	89.9	74.7	3.0
BitFit	3.96 / 0.62%	90.8	93.8	87.4	3.2	89.0	91.6	81.7	2.3
Adapter	3.92 / 0.61%	91.2	94.0	88.2	3.1	89.3	92.0	82.5	2.2
VPT	4.00 / 0.62%	91.5	94.3	90.0	2.8	91.0	93.7	84.8	2.1
LST	11.49 / 1.77%	89.7	93.3	86.9	3.7	89.4	92.4	83.7	2.4
SAM-Adapter	3.98 / 0.62%	89.6	92.5	86.9	3.6	89.6	92.4	82.2	2.2
SSF	4.42 / 0.69%	91.3	93.9	88.3	3.0	89.6	91.9	82.0	2.2
LoRA	4.00 / 0.62%	91.2	93.8	88.4	3.1	90.7	92.5	84.5	2.2
Conv-LoRA	4.02 / 0.63%	92.0	94.7	89.7	2.6	91.3	94.0	85.5	1.9

Method	# Params (M) / Ratio(%)	CVC-ColonDB				ETIS				CVC-T			
		S *	E *	F ¹ *	MAE #	S *	E *	F ¹ *	MAE #	S *	E *	F ¹ *	MAE #
PraNet (Fan et al., 2020b)	32.55 / 100%	82.0	84.5	70.9	4.0	79.3	80.6	58.2	2.3	93.9	97.1	85.2	0.8
decoder-only	3.51 / 0.55%	76.7	80.7	59.5	5.2	67.9	71.4	41.0	7.4	86.4	88.4	67.5	2.3
BitFit	3.96 / 0.62%	83.8	86.8	72.7	3.9	84.7	87.1	67.4	1.7	91.5	94.2	81.3	1.4
Adapter	3.92 / 0.61%	83.6	86.3	71.7	3.6	85.3	86.9	67.0	1.8	92.9	94.6	84.1	1.2
VPT	4.00 / 0.62%	83.9	87.3	72.9	3.5	86.3	88.0	69.4	1.8	94.6	97.7	85.2	0.6
LST	11.49 / 1.77%	82.5	86.6	72.2	4.3	81.5	84.3	62.2	3.4	92.0	93.7	82.5	1.2
SAM-Adapter	3.98 / 0.62%	83.1	86.3	70.8	3.8	83.2	85.3	63.5	2.2	92.1	94.2	81.8	1.2
SSF	4.42 / 0.69%	83.9	86.9	72.1	3.9	84.7	87.4	66.7	1.7	92.1	93.9	83.6	1.4
LoRA	4.00 / 0.62%	84.4	87.2	73.5	4.1	85.5	86.5	68.4	1.8	93.5	95.9	85.7	1.3
Conv-LoRA	4.02 / 0.63%	84.7	88.0	75.3	3.4	87.1	88.8	71.6	1.5	93.7	96.5	87.2	0.9

Table 5: Quantitative results for Polyp Segmentation.

Skin Lesion Segmentation

Method	# Params (M) / Ratio(%)	ISIC 2017			
		Jac ^o	T-Jac ^o	Dice ^o	Acc ^o
Transfuse (Zhang et al., 2021)	26.30 / 100%	80.1	74.5	87.5	94.7
decoder-only	3.51 / 0.55%	69.7	56.6	79.5	91.2
BitFit	3.96 / 0.62%	76.4	68.2	84.7	93.5
Adapter	3.92 / 0.61%	76.7	68.0	85.0	93.6
VPT	4.00 / 0.62%	76.9	68.4	85.1	93.7
LST	11.49 / 1.77%	76.4	66.7	84.9	93.5
SAM-Adapter	3.98 / 0.62%	76.1	67.2	84.6	93.4
SSF	4.42 / 0.69%	76.6	68.3	85.0	93.6
LoRA	4.00 / 0.62%	76.6	68.6	84.9	93.6
Conv-LoRA	4.02 / 0.63%	77.6	69.6	85.7	93.9

Table 6: Quantitative results for Skin Lesion Segmentation.

Camou aged Object Segmentation

Method	# Params (M) / Ratio(%)	CHAMELEON				CAMO				COD10K			
		S *	E *	F ¹ *	MAE #	S *	E *	F ¹ *	MAE #	S *	E *	F ¹ *	MAE #
SINet-v2 (Fan et al., 2022)	26.98 / 100%	89.2	94.0	79.9	3.1	80.8	85.8	73.1	7.7	81.2	88.3	65.9	3.7
decoder-only	3.51 / 0.55%	87.3	90.5	79.2	3.7	78.5	83.1	69.8	8.7	82.8	87.8	70.3	3.7
BitFit	3.96 / 0.62%	93.2	96.1	88.7	1.8	86.8	90.7	81.5	5.3	91.1	95.1	85.5	1.8
Adapter	3.92 / 0.61%	93.9	96.6	89.9	1.7	87.7	91.3	82.8	5.0	91.5	95.3	86.4	1.7
VPT	4.00 / 0.62%	93.2	96.1	88.9	1.9	87.4	91.4	82.1	5.0	91.1	94.9	85.3	1.8
LST	11.49 / 1.77%	92.0	95.3	87.2	2.3	83.3	88.0	77.1	6.9	88.4	93.2	80.8	2.3
SAM-Adapter	3.98 / 0.62%	92.7	95.9	87.7	2.0	85.6	89.6	79.8	5.9	90.1	94.1	83.7	2.0
SSF	4.42 / 0.69%	94.0	97.0	90.1	1.5	87.5	91.4	82.6	5.0	91.3	95.1	86.2	1.7
LoRA	4.00 / 0.62%	93.8	96.7	89.8	1.6	88.0	91.9	82.8	4.8	91.5	95.2	86.4	1.7
Conv-LoRA	4.02 / 0.63%	94.1	96.9	90.6	1.6	88.3	92.4	84.0	4.5	91.6	95.5	86.8	1.6

Table 7: Quantitative results for Camou age Detection.

Shadow Detection

Method	# Params (M) / / Ratio(%)	SBU BER #
FDRNet(Zhu et al., 2021)	10.77 / 100%	3.56
decoder-only	3.51 / 0.55%	14.58
BitFit	3.96 / 0.62%	3.16 <small>0.128</small>
Adapter	3.92 / 0.61%	2.84 <small>0.093</small>
VPT	4.00 / 0.62%	2.70 <small>0.055</small>
LST	11.49 / 1.77%	3.18 <small>0.012</small>
SAM-Adapter	3.98 / 0.62%	3.14 <small>0.063</small>
SSF	4.42 / 0.69%	3.19 <small>0.046</small>
LoRA	4.00 / 0.62%	2.74 <small>0.079</small>
Conv-LoRA	4.02 / 0.63%	2.54 <small>0.081</small>

Table 8: Quantitative results for Shadow Detection.

Leaf Segmentation

Method	# Params (M) / Ratio (%)	IoU "	Leaf Dice "	Acc "
DeepLabv3 (Chen et al., 2017b)	41.99 / 100%	62.3	74.1	94.0
decoder-only	3.51 / 0.55%	50.8	63.8	89.6
BitFit	3.96 / 0.62%	71.4 <small>1.15</small>	81.7 <small>1.01</small>	95.5 <small>0.30</small>
Adapter	3.92 / 0.61%	72.1 <small>0.47</small>	82.4 <small>0.36</small>	95.3 <small>0.18</small>
VPT	4.00 / 0.62%	73.6 <small>0.26</small>	83.8 <small>0.26</small>	95.9 <small>0.14</small>
LST	11.49 / 1.77%	70.2 <small>0.87</small>	81.1 <small>0.82</small>	95.3 <small>0.35</small>
SAM-Adapter	3.98 / 0.62%	71.4 <small>0.20</small>	82.1 <small>0.10</small>	95.3 <small>0.05</small>
SSF	4.42 / 0.69%	71.5 <small>0.63</small>	81.8 <small>0.44</small>	95.5 <small>0.26</small>
LoRA	4.00 / 0.62%	73.7 <small>0.20</small>	83.6 <small>0.13</small>	95.7 <small>0.07</small>
Conv-LoRA	4.02 / 0.63%	74.5 <small>0.39</small>	84.3 <small>0.34</small>	96.0 <small>0.07</small>

Table 9: Quantitative results for Leaf Segmentation.

Road Segmentation

Method	# Params (M) / Ratio (%)	IoU "	Road Dice "	Acc "
LinkNet34MTL(Batra et al., 2019)	22.00 / 100%	59.1	73.0	97.7
decoder-only	3.51 / 0.55%	48.6	65.1	96.4
BitFit	3.96 / 0.62%	60.6 <small>0.15</small>	75.2 <small>0.11</small>	97.5 <small>0.02</small>
Adapter	3.92 / 0.61%	61.5 <small>0.11</small>	75.9 <small>0.12</small>	97.6 <small>0.01</small>
VPT	4.00 / 0.62%	60.2 <small>1.87</small>	74.9 <small>1.50</small>	97.4 <small>0.22</small>
LST	11.49 / 1.77%	60.2 <small>0.26</small>	74.9 <small>0.22</small>	97.5 <small>0.01</small>
SAM-Adapter	3.98 / 0.62%	60.6 <small>0.06</small>	75.2 <small>0.04</small>	97.5 <small>0.01</small>
SSF	4.42 / 0.69%	61.6 <small>0.03</small>	76.0 <small>0.02</small>	97.6 <small>0.01</small>
LoRA	4.00 / 0.62%	62.2 <small>0.21</small>	76.5 <small>0.18</small>	97.6 <small>0.02</small>
Conv-LoRA	4.02 / 0.63%	62.6 <small>0.36</small>	76.8 <small>0.27</small>	97.7 <small>0.05</small>

Table 10: Quantitative results for Road Segmentation.

E ADDITIONAL ABLATION

The Rank of LoRA. The performance of SAM for twelve-class segmentation could be improved by introducing more extra parameters (table 11). However, this may disobey the rule of 'parameter efficiency' of PEFT. Exploring the design of a more efficient way for introducing 'classification prior' for SAM is a worthwhile endeavor for the future.

LoRA	#Params (M)	Test	
		Acc "	mIoU "
r = 3	4.00	94.80	66.01
r = 6	4.49	95.15	66.24
r = 12	5.48	95.23	66.69
r = 24	7.44	95.14	67.02

Table 11: The performance trend when increasing the rank of LoRA for twelve-class segmentation.

Figure 6: Comparison between (a) MoE, which uses the gating network to dynamically select the experts, and (b) Multi-scale, which selects all the experts simultaneously.

Combined with Other PEFT Methods. We also evaluate the combination of Conv-LoRA and VPT on Poly Segmentation in table 12. This combination achieves superior or competitive performance, demonstrating the potential of introducing Conv-LoRA in addition to other PEFT methods. As a future work, such a combination might motivate techniques that further reduce the number of trainable parameters while ensuring the enhanced performance.

Method	# Params / Ratio(%)	CVC-ColonDB				ETIS			
		S "	E "	F! "	MAE #	S "	E "	F! "	MAE #
VPT	4.00 / 0.62%	85.0	88.4	75.0	3.6	86.5	87.4	70.1	1.9
Conv-LoRA	4.03 / 0.63%	85.1	88.8	75.7	3.4	86.8	88.5	70.2	1.7
Conv-LoRA+VPT	4.23 / 0.66 %	86.0	89.0	76.8	3.5	87.7	88.1	70.3	1.7

Table 12: Performance of combining Conv-LoRA and VPT on Polyp Segmentation.

MoE-Conv vs. Blocks with Various Kernel sizes. We further compare the design of our MoE-Conv and convolutional blocks with various kernel sizes. Convolutional blocks with various kernel sizes are commonly used in multi-scale feature fusion. Here we follow the Inception module proposed by (Szegedy et al., 2015), which consists of multiple kernel sizes with 1, 3, 3, 5, 5. To ensure a fair relative comparison, we opt for four experts in MoE-Conv to align with the branch number of the Inception structure. We conduct experiments on Leaf Dataset and ISIC 2017 Dataset twice for evaluation.

Method	# Params / Ratio(%)	Leaf			ISIC 2017			
		IoU "	Dice "	Acc "	Jac "	T-Jac "	Dice "	Acc "
Inception	4.01 / 0.63%	72.1	82.3	95.0	76.9	68.5	84.9	93.3
MoE-Conv	4.01 / 0.63%	74.0	83.8	95.6	77.4	69.5	85.5	93.9

Table 13: MoE-Conv vs. Blocks with Various Kernel sizes (indicated as 'Inception').

In table 13, MoE-Conv outperforms the design of convolutional blocks with various kernel sizes. We attribute this to convolutions with various kernel sizes only operate on the default feature scale. Maybe a larger kernel size can inject some approximate local prior in features of a smaller scale, but it probably can't bring local prior into features of larger scales than the default. The features in ViT are downscaled 16x from the original image, so injecting priors in larger scale features are generally more useful, as evidenced in table 4, where the optimal scale is usually larger than the default scale.

Computational Cost. In table 14, we compare the training / inference speed and per epoch training time with the ISIC 2017 dataset and a single V100 GPU. While Conv-LoRA is slower in speed than others, it achieves robust performance gains across various semantic segmentation tasks (table 1). We find that the main cost comes from the upscale and downscale operations. A possible future direction is exploring how to inject local prior without explicitly scaling up and down features. We also notice that there are some other orthogonal works, like token merging (Liang et al., 2022; Bolya et al., 2022), that may accelerate Conv-LoRA.

Method	Training		Inference
	iter/s	min/epoch	iter/s
BitFit	1.82	18.3	4.62
Adapter	1.73	19.3	4.38
VPT	1.41	23.6	4.39
LoRA	1.64	20.3	4.62
Conv-LoRA	1.22	27.3	3.64

Table 14: Computational cost comparison.

F LOCAL PRIOR ANALYSIS

We elaborate on 'SAM's local prior assumption' mentioned in our abstract in section 4.3. Additionally, as our Conv-LoRA is designed to reinforce existing inductive biases within the features, we investigate the presence of these biases in the features.

We use two metrics: 1) Mean attention distance, which reflects the extent of local or global information that a self-attention layer is aggregating (Dosovitskiy et al., 2020; Raghu et al., 2021). We calculate the attention distance for each attention head, defined as the average distance between the position of the query patch and the locations to which it attends, weighted by the attention weights. Then we calculate the average attention distance for each layer by computing the mean across 500 randomly sampled images from downstream semantic segmentation tasks. 2) Relative log amplitudes (Park & Kim, 2022), which reflects whether the model tends to reduce or amplify high-frequency signals (e.g., edges, textures) in the feature map. We compute the relative log amplitudes of the Fourier-transformed feature map in each layer and average them across layers. Then we calculate the mean of the relative log amplitudes over 100 randomly sampled segmentation images.

Inductive biases of the features. Extensive pre-training on large-scale datasets has been demonstrated to equip ViTs with inductive biases (Dosovitskiy et al., 2020; Raghu et al., 2021). SAM's large-scale segmentation pretraining further amplifies these inductive biases within the features.

To evaluate the inductive biases of the features, we conduct comprehensive comparisons among randomly initialized ViT, MAE pretrained ViT, and SAM ViT. We use mean attention distance and relative log amplitudes for evaluation. In fig. 7, as the training progresses from randomly initialized ViT to MAE ViT and ultimately to SAM ViT, we observe a consistent trend: the mean attention distance decreases, while the high-frequency signals in the features increase, indicating a focus on local information. These findings further confirm the presence of inductive biases in features following large-scale pre-training. Furthermore, the findings affirm the rationality of Conv-LoRA in reinforcing the inherent inductive biases within the features.

G MOE ANALYSIS

What does MoE learn? MoE learns to dynamically select an appropriate scale for injecting local priors based on input features. To illuminate its functionality, we conduct an analysis by

Figure 7: Evaluation of inductive biases within features. **Left:** Mean attention distance average over all attention heads. **Right:** Relative log amplitudes of Fourier transformed feature maps. Compared to randomly initialized ViT, SAM and MAE could learn the inductive biases from large-scale data pre-training. And SAM could further amplify these biases within the features.

tracking the frequency of each expert's selection across different datasets during inference. The detailed results are presented in Fig. 8.

Notably, distinct datasets exhibit preferences for different experts. For instance, in Leaf Segmentation, MoE tends to favor experts with upsampling ratios of 3 and 4, while on the ISIC 2017 dataset, it tends to select the expert with an upsampling ratio of 2. Connecting these insights with the optimal scale ablation results in table 4, it becomes evident that MoE selects the expert adaptively for each sample, and consequently, the distributions of MoE selection across datasets reflect their different data distributions. This observation supports that MoE behaves adaptively and effectively with respect to each sample's property. This adaptability reinforces the significance of MoE in tailoring its selection to the unique characteristics of diverse datasets, enhancing its effectiveness in local prior injection.

Figure 8: The number of times each expert is selected during inference for on Leaf Dataset and ISIC 2017 Dataset. Distinct datasets exhibit preferences for different experts.

H LOW-DATA REGIME EXPERIMENTS

Convolutional layers could introduce inductive biases, thereby enhancing data efficiency during fine-tuning. Therefore, we undertake experiments in a low-data regime to validate whether Conv-LoRA can indeed improve data efficiency in semantic segmentation.

Specifically, we conduct experiments under the few-shot setting, wherein acquiring data for downstream tasks is challenging, and only a limited number of training samples per task are available. Our experiments are performed on the Trans10K-v2 dataset, which is used for twelve-class transparent object segmentation. We randomly select the training samples, to meet the settings of 1, 2, 4, 8, and 16 shots (i.e., the number of labeled training examples per class), and the experiments are run for 100 epochs.

Method	# Params / Ratio(%)	Shot	mIoU "
LoRA	4.00 / 0.62%	1/2/4/8/16	13.32/18.01/22.70/25.89/38.11
Conv-LoRA	4.02 / 0.63%	1/2/4/8/16	14.53/19.81/23.05/26.02/38.63

Table 15: Results of few-shot learning on multi-class semantic segmentation. 'Shot' indicates the number of labeled training examples per class.

In table 15, the performance improvement of Conv-LoRA compared to LoRA, is particularly pronounced in an extremely low-data setting (e.g., 1-shot). The result demonstrates that the reduction of inductive biases in Conv-LoRA contribute to improved data efficiency.

I MORE VISUALIZATION RESULTS

Conv-LoRA vs. LoRA Feature. We visualize the feature maps from the fine-tuned SAM's image encoder, which incorporates LoRA and our Conv-LoRA respectively. In fig. 9, the image features from SAM's image encoder equipped with Conv-LoRA could provide more fine-grained information, e.g., slim edges, which is beneficial to later mask prediction. This further demonstrates the effectiveness of reinforcing the image-related local priors with network architecture.

Figure 9: Feature map visualization for binary-class and multi-class semantic segmentation. Compared to LoRA, applying Conv-LoRA to SAM's image encoder could capture more fine-grained details.

Mask prediction. We also provide more visualization results for the mask prediction across various datasets when applying different PEFT methods (VPT, LoRA and Conv-LoRA). This further confirms the superiority of our Conv-LoRA.

RGB Image GT VPT LoRA Conv-LoRA

Figure 10: Visualization Results on Skin Lesion Segmentation.

RGB Image GT VPT LoRA Conv-LoRA

Figure 11: Visualization Results on Camou aged Object Segmentation.

