SAM Fails to Segment Anything? – SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More

Tianrun Chen1,2+ Lanyun Zhu4+ Chaotao Ding3+ Runlong Cao3+
Yan Wang6 Zejian Li5 Lingyun Sun2 Papa Mao1

Ying Zang3+

First Online: 14 April, 2023
+ Equal Contribution * Corresponding Author 
{tianrun.chen@zju.edu.cn; 02750@zjhu.edu.cn}

1 KOKONI, Moxin (Huzhou) Tech. Co., LTD, Huzhou, Zhejiang, P.R. China.
2 College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, P.R. China.
3 School of Information Engineering, Huzhou University, Huzhou, Zhejiang, P.R. China.
4 Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore.
5 School of Software Technology, Zhejiang University, Hangzhou, Zhejiang, P.R. China.
6 School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing, P.R. China.

Project Page: http://tianrun-chen.github.io/SAM-Adaptor/

Abstract

The emergence of large models, also known as foundation models, has brought significant advancements to AI research. One such model is Segment Anything (SAM), which is designed for image segmentation tasks. However, as with other foundation models, our experimental findings suggest that SAM may fail or perform poorly in certain segmentation tasks, such as shadow detection and camouflaged object detection (concealed object detection). This study first paves the way for applying the large pre-trained image segmentation model SAM to these downstream tasks, even in situations where SAM performs poorly. Rather than fine-tuning the SAM network, we propose SAM-Adapter, which incorporates domain-specific information or visual prompts into the segmentation network by using simple yet effective adapters. By integrating task-specific knowledge with general knowledge learnt by the large model, SAM-Adapter can significantly elevate the performance of SAM in challenging tasks as shown in extensive experiments. We can even outperform task-specific network models and achieve state-of-the-art performance in the task we tested: camouflaged object detection, shadow detection. We also tested polyp segmentation (medical image segmentation) and achieves better results. We believe our work opens up opportunities for utilizing SAM in downstream tasks, with potential applications in various fields, including medical image processing, agriculture, remote sensing, and more.
1 Introduction

AI research has witnessed a paradigm shift with models trained on vast amounts of data at scale. These models, or known as foundation models, such as BERT, DALL-E, and GPT-3 have shown promising results in many language or vision tasks [1]. Recently, among the foundation models, Segment Anything (SAM) [2] has a distinct position as a generic image segmentation model trained on the large visual corpus [2]. It has been demonstrated that SAM has successful segmentation capabilities in diverse scenarios, which makes it a groundbreaking step toward image segmentation and related fields of computer vision.

However, as computer vision encompasses a broad spectrum of problems, SAM’s incompleteness is evident, which is similar to other foundation models since the training data cannot encompass the entire corpus, and working scenarios are subject to variation [1]. In this study, we first test SAM in some challenging low-level structural segmentation tasks including camouflaged object detection (concealed scenes) and shadow detection, and we find that the SAM model trained on general images cannot perfectly "Segment Anything" in these cases.

As such, a crucial research problem is: How to harness the capabilities acquired by large models from massive corpora and leverage them to benefit downstream tasks?

Here, we introduce the SAM-Adapter, which serves as a solution to the research problem mentioned above. This pioneering work is the first attempt to adapt the large pre-trained image segmentation model SAM to specific downstream tasks with enhanced performance. As its name states, SAM-Adapter is a very simple yet effective adaptation technique that leverages internal knowledge and external control signal. Specifically, it is a lightweight model that can learn alignment with a relatively small amount of data and serves as an additional network to inject task-specific guidance information from the samples of that task. Information is conveyed to the network using visual prompts [3, 4], which has been demonstrated to be efficient and effective in adapting a frozen large foundation model to many downstream tasks with a minimum number of additional trainable parameters.

Specifically, we show that our method is:

• Generalizable: SAM-Adapter can be directly applied to customized datasets of various tasks to enhance performance with the assistance of SAM.
• Composable: It is effortless to combine multiple explicit conditions to fine-tune SAM with multi-condition control.

We perform extensive experiments on multiple tasks and datasets, including ISTD for shadow detection [5] and COD10K [6], CHAMELEON [7], CAMO [8] for camouflaged object detection task, and kvasir-SEG [9] for polyp segmentation (medical image segmentation) task. Benefiting from the capability of SAM and our SAM-Adapter, our method achieves state-of-the-art (SOTA) performance on both tasks. The contributions of this work can be summarized as follows:

• First, we pioneer the analysis of the incompleteness of the Segment Anything (SAM) model as a foundation model and propose a research problem of how to utilize the SAM model to serve downstream tasks.
• Second, we are the first to propose the adaptation approach, SAM-Adapter, to adapt SAM to downstream tasks and achieve enhanced performance. The adapter integrates the task-specific knowledge with general knowledge learnt by the large model. The task-specific knowledge can be flexibly designed.
• Third, despite SAM’s backbone being a simple plain model lacking specialized structures tailored for the two specific downstream tasks, our approach still surpasses existing methods and attains state-of-the-art (SOTA) performance in these downstream tasks.

To the best of our knowledge, this work pioneers to demonstrate the exceptional ability of SAM to transfer to other specific data domains with remarkable accuracy. While we only tested it on a few datasets, we expect SAM-Adapter can serve as an effective and adaptable tool for various downstream segmentation tasks in different fields, including medical and agriculture. This study will usher in a new era of utilizing large pre-trained image models in diverse research fields and industrial applications.
2 Related Work

Semantic Segmentation. In recent years, semantic segmentation has made significant progress, primarily due to the remarkable advancements in deep-learning-based methods such as fully convolutional networks (FCN) [10], encoder-decoder structures [11, 12, 13, 14, 15], dilated convolutions [16, 17, 18, 19, 20], pyramid structures [21, 18, 22, 19, 23], attention modules [24, 25, 26], and transformers [27, 28, 29, 30]. Building upon previous research, Segment Anything (SAM) [2] introduces a large ViT-based model trained on a large visual corpus. This work aims to leverage the SAM to solve specific downstream image segmentation tasks.

Adapters. The concept of Adapters was first introduced in the NLP community [31] as a tool to fine-tune a large pre-trained model for each downstream task with a compact and scalable model. In [32], multi-task learning was explored with a single BERT model shared among a few task-specific parameters. In the computer vision community, [33] suggested fine-tuning the ViT [34] for object detection with minimal modifications. Recently, ViT-Adapter [35] leveraged Adapters to enable a plain ViT to perform various downstream tasks. [4] introduce an Explicit Visual Prompting (EVP) technique that can incorporate explicit visual cues to the Adapter. However, no prior work has tried to apply Adapters to leverage pretrained image segmentation model SAM trained at large image corpus. Here, we mitigate the research gap.

Camouflaged Object Detection (COD). Camouflaged object detection, or concealed object detection is a challenging but useful task that identifies objects blend in with their surroundings. COD has wide applications in medicine, agriculture, and art. Initially, researches of camouflage detection relied on low-level features like texture, brightness, and color [36, 37, 38, 39] to distinguish foreground from background. It is worth noting that some of these prior knowledge is critical in identifying the objects, and is used to guide the neural network in this paper.

Le et al. [8] first proposed an end-to-end network consisting of a classification and a segmentation branch. Recent advances in deep learning-based methods have shown a superior ability to detect complex camouflaged objects [6, 40, 41]. In this work, we leverage the advanced neural network backbone (a foundation model – SAM) with the input of task-specific prior knowledge to achieve the state-of-the-art (SOTA) performance.

Shadow Detection. Shadows can occur when an object surface is not directly exposed to light. They offer hints on light source direction and scene illumination that can aid scene comprehension [42, 43]. They can also negatively impact the performance of computer vision tasks [44, 45]. Early method use hand-crafted heuristic cues like chromacity, intensity and texture [46, 43, 47]. Deep learning approaches leverage the knowledge learnt from data and use delicately designed neural network structure to capture the information (e.g. learned attention modules) [48, 49, 50]. This work leverage the heuristic priors with large neural network models to achieve the state-of-the-art (SOTA) performance.

3 Method

3.1 Using SAM as the Backbone

As previously illustrated, the goal of the SAM-Adapter is to leverage the knowledge learned from the SAM. Therefore, we use SAM as the backbone of the segmentation network. The image encoder of SAM is a ViT-H/16 model with 14x14 windowed attention and four equally-spaced global attention blocks. We keep the weight of pretrained image encoder frozen. We also leverage the mask decoder of the SAM, which consists of a modified transformer decoder block followed by a dynamic mask prediction head. We use the pretrained SAM’s weight to initialize the weight of the mask decoder of our approach and tune the mask decoder during training. We input no prompts into the original mask decoder of SAM.

3.2 Adapters

Next, the task-specific knowledge $F_i$ is learned and injected into the network via Adapters. We employ the concept of prompting, which utilizes the fact that foundation models have been trained on
large-scale datasets. Using appropriate prompts to introduce task-specific knowledge [4] can enhance the model’s generalization ability on downstream tasks, especially when annotated data is scarce.

The architecture of the proposed SAM-Adapter is illustrated in Figure 1. We aim to keep the design of the adapter to be simple and efficient. Therefore, we choose to use an adapter that consists of only two MLPs and an activate function within two MLPs [4]. Specifically, the adapter takes the information $F^i$ and obtains the prompt $P^i$:

$$ P^i = \text{MLP}_{\text{up}} \left( \text{GELU} \left( \text{MLP}_{\text{tune}}^i(F^i) \right) \right) $$

in which $\text{MLP}_{\text{tune}}^i$ are linear layers used to generate task-specific prompts for each Adapter. $\text{MLP}_{\text{up}}$ is an up-projection layer shared across all Adapters that adjusts the dimensions of transformer features. $P^i$ refers to the output prompt that is attached to each transformer layer of SAM model. GELU is the GELU activation function [51]. The information $F^i$ can be chosen to be in various forms.

### 3.3 Input Task-Specific Information

It is worth noting that the information $F^i$ can be in various forms depending on the task and flexibly designed. For example, it can be extracted from the given samples of the specific dataset of the task in some form, such as texture or frequency information, or some hand-crafted rules. Moreover, the $F^i$ can be in a composition form consisting multiple guidance information:

$$ F^i = \sum_{j=1}^{N} w_j F^j $$

in which $F^j$ can be one specific type of knowledge/features and $w^j$ is an adjustable weight to control the composed strength.

### 4 Experiments

#### 4.1 Tasks and Datasets

We select two challenging low-level structural segmentation task for SAM – camouflaged object detection and shadow detection. For camouflaged object detection, we choose COD10K dataset [6], CHAMELEON dataset [7], and CAMO dataset [8] in our experiment. COD10K is the largest dataset for camouflaged object detection containing 3,040 training and 2,026 testing samples. CHAMELEON includes 76 images collected from the Internet for testing. CAMO dataset consists of 1,250 images (1,000 images for the training set and 250 images for the testing set). Following the training protocol in [6], we use combined dataset of CAMO and COD10K (the training set of camouflaged images) for training, and use the test set of CAMO, COD10K and the entire CHAMELEON dataset for performance evaluation. For shadow detection, we use ISTD dataset [5], which contains 1,330 training images and 540 test images. We choose kvasir-SEG [9] for polyp segmentation (medical image segmentation) task, and the train-test split follows the settings in Medico multimedia task at
### 4.2 Implementation Details

In the experiment, we choose two types of visual knowledge, patch embedding $F_{pe}$ and high-frequency components $F_{hfc}$, following the same setting in [4], which has been demonstrated effective in various of vision tasks. $w^j$ is set to 1. Therefore, the $F_i$ is derived by $F_i = F_{hfc} + F_{pe}$.

The $MLP_{\text{tune}}^i$ has 32 linear layers and $MLP_{\text{up}}^i$ is one linear layer that maps the output from GELU activation to the number of inputs of the transformer layer. We use ViT-H version of SAM. Balanced BCE loss is used for shadow detection. BCE loss and IOU loss are used for camouflaged object detection and polyp segmentation. AdamW optimizer is used for all the experiments. The initial learning rate is set to 2e-4. Cosine decay is applied to the learning rate. The training of camouflaged object segmentation is performed for 20 epochs. Shadow segmentation is trained for 90 epochs. Polyp segmentation is trained for 120 epochs. The experiments are implemented using PyTorch on four NVIDIA Tesla A100 GPUs.

### 4.3 Experimental Result for Camouflaged Object Detection

We first evaluate SAM in camouflaged object detection task, which is a very challenging task as foreground objects are often with visual similar patterns to the background. Our experiments revealed that SAM did not perform well in this task. As shown in Figure 2, SAM failed to detect some concealed objects. This can be further confirmed by the quantitative results presented in Table 1. In fact, SAM’s performance was significantly lower than the existing state-of-the-art methods in all metrics evaluated.

In Figure 2, it can be found clearly that by introducing the SAM-Adapter, our method significantly elevates the performance of the model (+17.9% in $S_{\alpha}$). Our approach successfully identifies concealed objects, as evidenced by clear visual results. Quantitative results also show that our method outperforms the existing state-of-the-art method.

### 4.4 Experimental Result for Shadow Detection

We also evaluated SAM on the task of shadow detection. However, as depicted in Figure 4, SAM struggled to differentiate between the shadow and the background information with parts missing or mistakenly added.

In our study, we evaluated various methods for shadow detection and found that our results were significantly poorer than existing methods. However, by integrating the SAM-Adapter, we were able to significantly improve the performance of

<table>
<thead>
<tr>
<th>Method</th>
<th>BER ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stacked CNN [58]</td>
<td>8.80</td>
</tr>
<tr>
<td>BDRAR [59]</td>
<td>2.69</td>
</tr>
<tr>
<td>DSC [60]</td>
<td>3.42</td>
</tr>
<tr>
<td>DSD [61]</td>
<td>2.17</td>
</tr>
<tr>
<td>FDRNet [62]</td>
<td>1.55</td>
</tr>
<tr>
<td>SAM [2]</td>
<td>40.51</td>
</tr>
<tr>
<td>SAM-Adapter (Ours)</td>
<td>1.43</td>
</tr>
</tbody>
</table>

Table 2: Quantitative Result - Shadow Detection

---

mediaeval 2020: Automatic polyp segmentation [52]. For evaluation metrics, we follow the protocol in [4] and use commonly-used S-measure ($S_m$), mean E-measure ($E_{\phi}$), and MAE for evaluation of camouflaged object detection. We use balance error rate (BER) for shadow detection. We use For SAM, We use the official implementation and tried different prompting approaches.
Figure 2: The Visualization Results of Camouflaged Image Segmentation. As illustrated in the figure, the SAM failed to perceive those animals that are visually ‘hidden’/concealed in their natural surroundings. By using SAM-Adapter, our approach can significantly elevate the performance of object segmentation with SAM. The samples are from the COD-10K dataset, for other dataset, please refer to More Results section.

Figure 3: The Visualization Results of Camouflaged Image Segmentation with Different Prompting Approach of SAM. The difference of this evaluation approach is that we use the SAM with input point prompts sampled in a unified manner across the image (the everything mode that produce multiple masks of the SAM online demo, denoted SAM online in the figure), and no input points but a mask box with the size of the image as the prompt, denoted SAM. It can be found that in different prompting mode, SAM cannot fully identify the object. By using SAM-Adapter, our approach can significantly elevate the performance of object segmentation with SAM.

4.5 Experimental Result for Polyp Segmentation

We showcase an example of using SAM-Adapter in medical image segmentation. We use the example of polyp segmentation. Polyps, which can become malignant, are identified during colonoscopy and removed through polypectomy. Accurate and speedy detection and removal of polyps are critical in preventing colorectal cancer, which is a leading cause of cancer-related deaths worldwide.
Numerous deep learning approaches have been developed for identifying polyps, and while pre-trained SAM is capable of identifying some polyps, we have found that its performance can be significantly improved with our SAM-Adapter approach. The results of our study, as illustrated in Table 3 and the visualization results in Figure 6, demonstrate the effectiveness of the SAM-Adapter in enhancing the identification of polyps.

<table>
<thead>
<tr>
<th>Method</th>
<th>mDice ↑</th>
<th>mIoU ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet [11]</td>
<td>0.821</td>
<td>0.756</td>
</tr>
<tr>
<td>UNet++ [63]</td>
<td>0.824</td>
<td>0.753</td>
</tr>
<tr>
<td>SFA [64]</td>
<td>0.725</td>
<td>0.619</td>
</tr>
<tr>
<td>SAM [2]</td>
<td>0.778</td>
<td>0.707</td>
</tr>
<tr>
<td>SAM-Adapter (Ours)</td>
<td>0.850</td>
<td>0.776</td>
</tr>
</tbody>
</table>

Table 3: Quantitative Result for Polyp Segmentation
5 Conclusion

In this work, we first extend the Segment Anything (SAM) model and apply it to some downstream tasks. Our experiments reveal that, like other foundational models, SAM is not effective in some vision tasks, for example, dealing with concealed objects. Therefore, we propose the SAM-Adapter, which utilizes SAM as the backbone and injects customized information into the network through simple yet effective Adapters to enhance performance in specific tasks. We evaluate our approach in camouflaged object detection and shadow detection tasks and demonstrate that the SAM-Adapter not only significantly improves SAM’s performance but also achieves state-of-the-art (SOTA) results. Our approach is also capable of enhancing the performance of medical image segmentation, as we show in our polyp segmentation task. We anticipate that this work will pave the way for applying SAM in downstream tasks and will have significant impacts in various image segmentation and computer vision fields.

6 Future Work

This study showcases the effectiveness and versatility of using adapters and large foundation models. Moving forward, we plan to extend the SAM-Adapter to tackle even more challenging image segmentation tasks and broaden its application to other fields. We also anticipate the development of more specialized designs tailored to specific tasks.

References


7 More Results
Figure 7: The Visualization Results of Camouflaged Image Segmentation of CAMO dataset. As illustrated in the figure, the SAM failed to perceive those animals that are visually ‘hidden’/concealed in their natural surroundings. By using SAM-Adapter, our approach can significantly elevate the performance of object segmentation with SAM.