arXiv:2304.09148v3 [cs.CV] 2 May 2023

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

Tianrun Chen[1,2+*] Lanyun Zhu[4+] Chaotao Ding[3+] Runlong Cao[3+]

Yan Wang [6] Zejian Li [5] Lingyun Sun [2] Papa Mao[1]

Ying Zang[3*]

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.

Preprint.

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

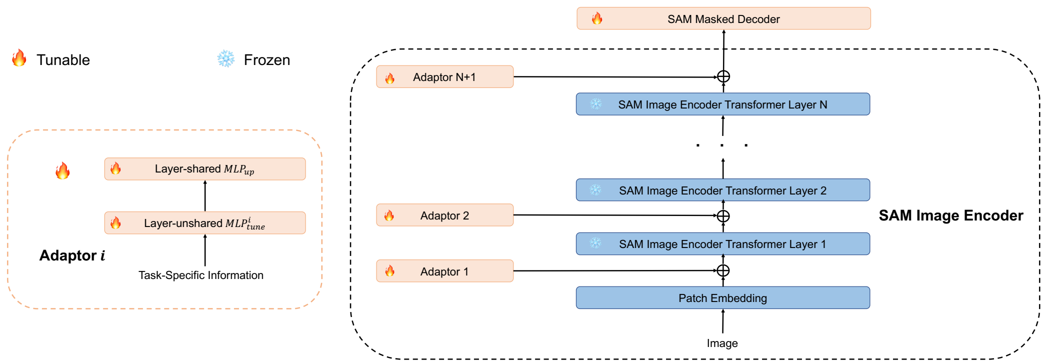

Figure 1: **The architecture of the proposed SAM-Adapter.**

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:

$$P^i = \mathrm{MLP}_{up}\left(\mathrm{GELU}\left(\mathrm{MLP}^i_{tune}\left(F_i\right)\right)\right) \tag{1}$$

in which MLP^i_{tune} are linear layers used to generate task-specific prompts for each Adapter. MLP_{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:

$$F_i = \sum_{1}^{N} w_j F_j \tag{2}$$

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 1250 images (1000 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

Method	CHAMELEON [7]				CAMO [8]				COD10K [6]			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$
SINet[53]	0.869	0.891	0.740	0.440	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
RankNet[54]	0.846	0.913	0.767	0.045	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045
JCOD [55]	0.870	0.924	-	0.039	0.792	0.839	-	0.82	0.800	0.872	-	0.041
PFNet [56]	0.882	**0.942**	0.810	0.330	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
FBNet [57]	0.888	0.939	**0.828**	**0.032**	0.783	0.839	0.702	0.081	0.809	0.889	0.684	0.035
SAM [2]	0.727	0.734	0.639	0.081	0.684	0.687	0.606	0.132	0.783	0.798	0.701	0.050
SAM-Adapter (Ours)	**0.896**	0.919	0.824	0.033	**0.847**	**0.873**	**0.765**	**0.070**	**0.883**	**0.918**	**0.801**	**0.025**

Table 1: Quantitative Result for Camouflage 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_ϕ), 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.

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^i_{tune} has 32 linear layers and MLP^i_{up} 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_α). 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

Method	BER $\downarrow$
Stacked CNN [58]	8.60
BDRAR [59]	2.69
DSC [60]	3.42
DSD [61]	2.17
FDRNet [62]	1.55
SAM [2]	40.51
SAM-Adapter (Ours)	**1.43**

Table 2: Quantitative Result - Shadow Detection

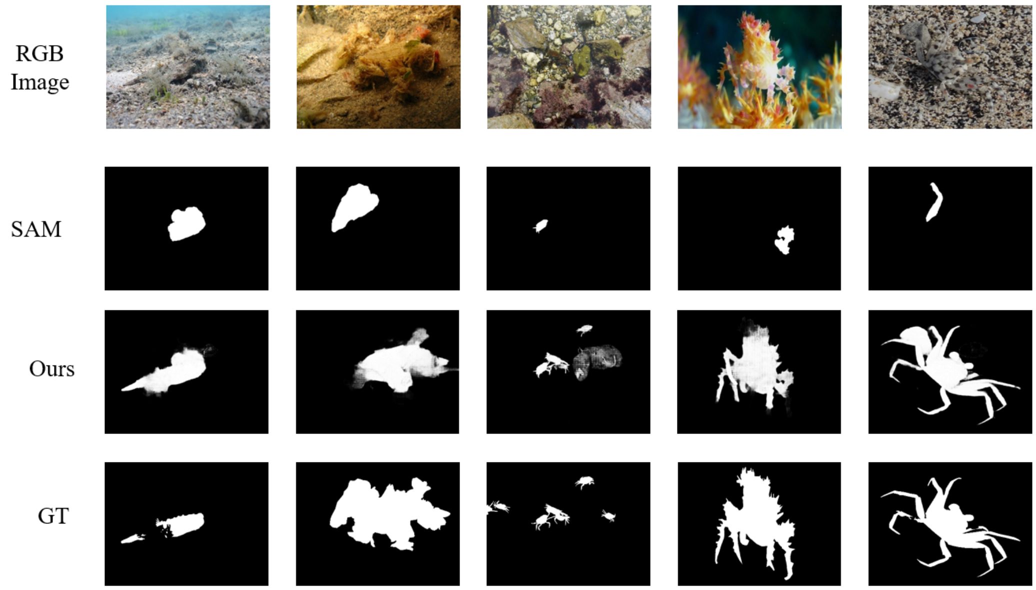

Figure 2: **The Visualization Results of Camouflaged Image Segmentation.** As illustrated in the figure, the SAM failed to perceive those animals that are visually 'hidden'/concealed in their natural surroundings. By using SAM-Adapter, our approach can significantly elevate the performance of object segmentation with SAM. The samples are from the COD-10K dataset, for other dataset, please refer to *More Results* section.

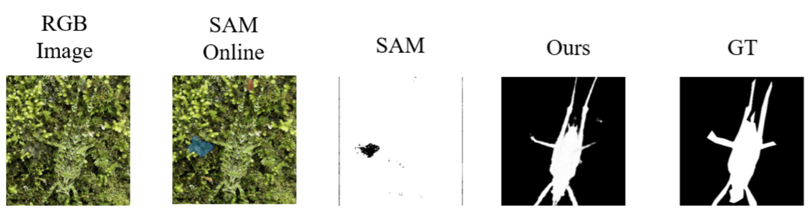

Figure 3: **The Visualization Results of Camouflaged Image Segmentation with Different Prompting Approach of SAM.** The difference of this evaluation approach is that we use the SAM with input point prompts sampled in a unified manner across the image (the *everything* mode that produce multiple masks 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.

our approach. The SAM-Adapter was able to improve the detection of shadow regions, making them more clearly identifiable. Our results were verified through quantitative analysis, and Table 2 demonstrates the performance boost that was brought about by the SAM-Adapter for shadow detection.

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.

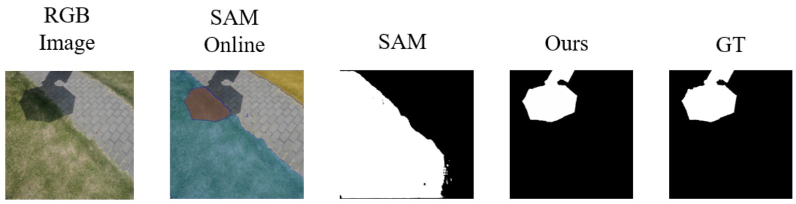

Figure 4: **Shadow Detection with Different Prompting Approach of SAM.** We use SAM with input point prompts sampled in a unified manner across the image (*SAM online* in the figure), and a box of a whole image (*SAM* in the figure). SAM cannot fully identify the shadow in different prompting modes. By using SAM-Adapter, our approach elevate the performance with SAM.

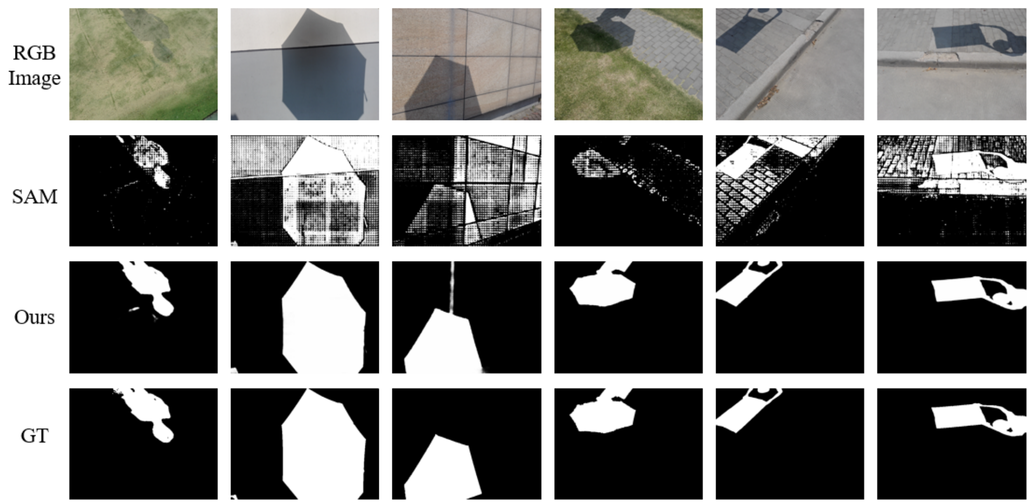

Figure 5: **The Visualization Results of Shadow Detection.** As illustrated in the figure, the SAM failed to distinguish the shadow and the background object. The *SAM* is used with the box prompt with the size of a whole image as the input and no input point prompts. By using SAM-adaptor, our approach can significantly elevate the performance of object segmentation with SAM.

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.

t

Method	mDice ↑	mIoU ↑
UNet [11]	0.821	0.756
UNet++ [63]	0.824	0.753
SFA [64]	0.725	0.619
SAM [2]	0.778	0.707
SAM-Adapter (Ours)	**0.850**	**0.776**

Table 3: Quantitative Result for Polyp Segmentation

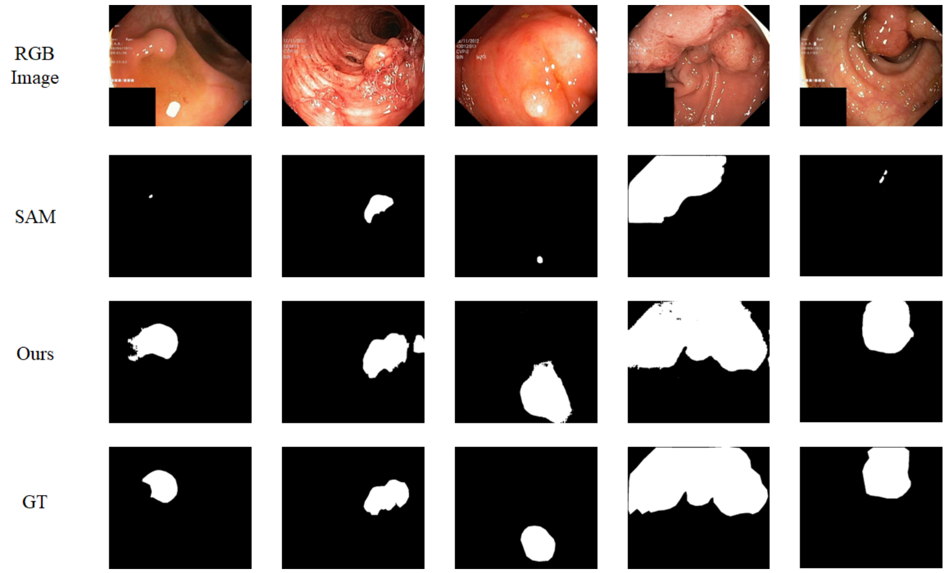

Figure 6: **The Visualization Result of Polyp Segmentation.** As illustrated in the figure, although SAM can identify some polyp structures in the image, the result is not accurate. By using SAM-Adapter, our approach elevate the performance with SAM.

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

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] 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.

[3] Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, and Shanghang Zhang. Exploring sparse visual prompt for cross-domain semantic segmentation. *arXiv preprint arXiv:2303.09792*, 2023.

[4] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. *arXiv preprint arXiv:2303.10883*, 2023.

[5] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018.

[6] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020.

[7] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018.

[8] 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.

[9] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[12] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, 2018.

[13] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9725, 2021.

[14] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[15] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *European Conference on Computer Vision*, pages 775–793. Springer, 2020.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[18] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[20] Zhikang Liu and Lanyun Zhu. Label-guided attention distillation for lane segmentation. *Neurocomputing*, 438:312–322, 2021.

[21] Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, and Junjie Yan. Learning statistical texture for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12537–12546, 2021.

[22] 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*, pages 2881–2890, 2017.

[23] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. *arXiv preprint arXiv:2304.05015*, 2023.

[24] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6798–6807, 2019.

[25] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[26] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 593–602, 2019.

[27] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[28] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[29] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[30] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[31] 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*, pages 2790–2799. PMLR, 2019.

[32] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019.

[33] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022.

[34] 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.

[35] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[36] Xue Feng, Cui Guoying, and Song Wei. Camouflage texture evaluation using saliency map. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 93–96, 2013.

[37] Thomas W Pike. Quantifying camouflage and conspicuousness using visual salience. *Methods in Ecology and Evolution*, 9(8):1883–1895, 2018.

[38] Jianqin Yin Yanbin Han Wendi Hou and Jinping Li. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Engineering*, 15:2201–2205, 2011.

[39] P Sengottuvelan, Amitabh Wahi, and A Shanmugam. Performance of decamouflaging through exploratory image analysis. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 6–10. IEEE, 2008.

[40] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021.

[41] Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson WH Lau. Frequency-aware camouflaged object detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–16, 2023.

[42] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011.

[43] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, 98:123–145, 2012.

[44] Sohail Nadimi and Bir Bhanu. Physical models for moving shadow and object detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 26(8):1079–1087, 2004.

[45] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 25(10):1337–1342, 2003.

[46] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *2011 international conference on computer vision*, pages 898–905. IEEE, 2011.

[47] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *2010 IEEE Computer Society conference on computer vision and pattern recognition*, pages 223–230. IEEE, 2010.

[48] Hieu Le, Tomas F Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+ d net: Training a shadow detector with adversarial shadow attenuation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 662–678, 2018.

[49] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10680–10687, 2020.

[50] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.

[51] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[52] Debesh Jha, Steven A Hicks, Krister Emanuelsen, Håvard Johansen, Dag Johansen, Thomas de Lange, Michael A Riegler, and Pål Halvorsen. Medico multimedia task at mediaeval 2020: Automatic polyp segmentation. *arXiv preprint arXiv:2012.15244*, 2020.

[53] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020.

[54] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11601, 2021.

[55] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10071–10081, 2021.

[56] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021.

[57] Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson WH Lau. Frequency-aware camouflaged object detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–16, 2023.

[58] 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. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 816–832. Springer, 2016.

[59] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.

[60] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7454–7462, 2018.

[61] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2019.

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

[63] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[64] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 302–310. Springer, 2019.

7 More Results

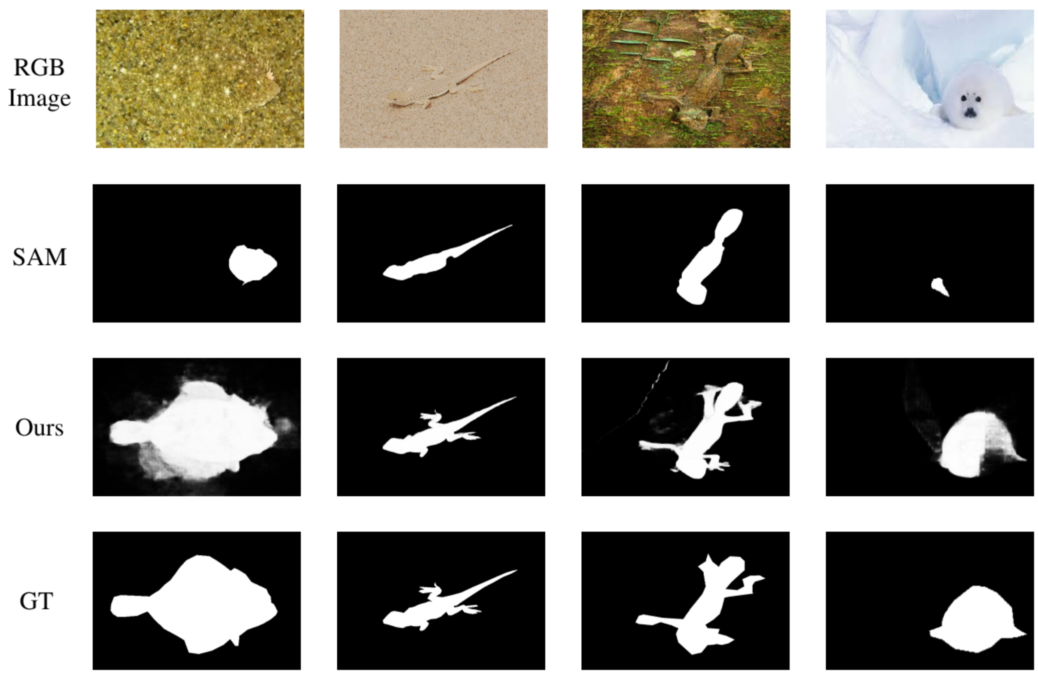

Figure 7: **The Visualization Results of Camouflaged Image Segmentation of CAMO dataset.** As illustrated in the figure, the SAM failed to perceive those animals that are visually 'hidden'/concealed in their natural surroundings. By using SAM-Adapter, our approach can significantly elevate the performance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