SAM2-ADAPTER: EVALUATING & ADAPTING SEGMENT ANYTHING 2 IN DOWNSTREAM TASKS: CAMOUFLAGE, SHADOW, MEDICAL IMAGE SEGMENTATION, AND MORE

Tianrun Chen [1,2 +*], Ankang Lu [3 +], Lanyun Zhu [4 +], Chaotao Ding [1 +], Chunan Yu [3], Deyi Ji [6],

Zejian Li [5], Lingyun Sun [2], Papa Mao [1] & Ying Zang [3*]

[+] Equal Contribution [*] Corresponding Author
[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 Information Science and Technology, University of Science and Technology of China, P.R. China.
tianrun.chen@zju.edu.cn; 02750@zjhu.edu.cn

ABSTRACT

The advent of large models, also known as foundation models, has significantly transformed the AI research landscape, with models like Segment Anything 2 (SAM2) achieving notable success in diverse image segmentation scenarios. Despite its advancements, SAM encountered limitations in handling some complex low-level segmentation tasks like camouflaged object and medical imaging. This paper introduces SAM2-Adapter, the first adapter designed to overcome the persistent limitations observed in SAM2 and achieve new state-of-the-art (SOTA) results in specific downstream tasks including medical image segmentation, camouflaged (concealed) object detection, and shadow detection. SAM2-Adapter offers generalizability and composability for diverse applications. We present extensive experimental results demonstrating SAM2-Adapter's effectiveness. We show the potential and encourage the research community to leverage the SAM2 model with our SAM2-Adapter for achieving superior segmentation outcomes. We have released our code, pre-trained model, and data processing protocols, which have benefited many researchers in this field.

1 INTRODUCTION

The AI research landscape has been transformed by foundation models trained on vast data Bommasani et al. (2021); Zhu et al. (2024a;c); Chen et al. (2024b). Recently, among the foundation models, Among these, Segment Anything (SAM) Kirillov et al. (2023) stands out as a highly successful image segmentation model with demonstrated success in diverse scenarios. However, in a previously study, researchers found that SAM's performance was limited in some challenging low-level structural segmentation tasks, such as camouflaged object detection and shadow detection. To address this, in 2023, within two weeks of SAM's release, SAM-Adapter is proposed Chen et al. (2023c;b), which aimed to leverage the power of the SAM model to deliver better performance on these challenging downstream tasks. The success of the SAM-Adapter, with its training and evaluation code and checkpoints made publicly available, has already been a valuable resource for many researchers in the community to experiment with and build upon, demonstrating its effectiveness on a variety of downstream tasks.

Now, the research community has pushed the boundaries further with the introduction of an even more capable and versatile successor to SAM, known as Segment Anything 2 (SAM2). Boasting further enhancements in its network architecture and training on an even larger visual corpus, SAM2 has certainly piqued our interest. This naturally leads us to the questions:

- Do the challenges faced by SAM in downstream tasks persist in SAM2?
- Can we replicate the success of SAM-Adapter and leverage SAM2's more powerful pre-trained encoder and decoder to achieve new state-of-the-art (SOTA) results on these tasks?

In this paper, we answer both questions with a resounding "Yes." Our experiments confirm that the challenges SAM encountered in downstream tasks do persist in SAM2, due to the inherent limitations of foundation models—where training data cannot cover the entire corpus and working scenarios vary Bommasani et al. (2021). However, we have devised a solution to address this challenge. By introducing the **SAM2-Adapter**, we've created a multi-adapter configuration that leverages SAM2's enhanced components to achieve new SOTA results in tasks including medical image segmentation, camouflaged object detection, and shadow detection.

This pioneering work is the first attempt to adapt the large pre-trained segmentation model SAM2 to specific downstream tasks and achieve new SOTA performance. SAM2-Adapter builds on the strengths of the original SAM-Adapter while introducing significant advancements.

SAM2-Adapter inherits the core advantages of SAM-Adapter, including:

- **Generalizability**: SAM2-Adapter can be directly applied to customized datasets of various tasks, enhancing performance with minimal additional data. This flexibility ensures that the model can adapt to a wide range of applications, from medical imaging to environmental monitoring.
- **Composability**: SAM2-Adapter supports the easy integration of multiple conditions to fine-tune SAM2, improving task-specific outcomes. This composability allows for the combination of different adaptation strategies to meet the specific requirements of diverse downstream tasks.

SAM2-Adapter enhances these benefits by adapting to SAM2's multi-resolution hierarchical Transformer architecture. By employing multiple adapters working in tandem, SAM2-Adapter effectively leverages SAM2's multi-resolution and hierarchical features for more precise and robust segmentation, which maximizes the potential of the already-powerful SAM2. We perform extensive experiments on multiple tasks and datasets, including ISTD for shadow detection Wang et al. (2018) and COD10K Fan et al. (2020b), CHAMELEON Skurowski et al. (2018), CAMO Le et al. (2019) for camouflaged object detection task, and kvasir-SEG Jha et al. (2020b) for polyp segmentation (medical image segmentation) task. Benefiting from the capability of SAM2 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:

- We are the first to identify and analyze the limitations of the Segment Anything 2 (SAM2) model in specific downstream tasks, continuing our research from SAM.
- Second, we are the first to propose the adaptation approach, **SAM2-Adapter**, to adapt SAM2 to downstream tasks and achieve enhanced performance. This method effectively integrates task-specific knowledge with the general knowledge learned by the large model.
- Third, despite SAM2's backbone being a simple plain model lacking specialized structures tailored for the specific downstream tasks, our extensive experiments demonstrate that SAM2-Adapter achieves SOTA results on challenging segmentation tasks, setting new benchmarks and proving its effectiveness in diverse applications.

SAM2-Adapter demonstrates the exceptional ability of the SAM2 model to transfer its knowledge to specific data domains, pushing the boundaries of what is possible in downstream segmentation tasks. We encourage the research community to adopt SAM2 as the backbone in conjunction with our SAM2-Adapter, to achieve even better segmentation results in various research fields and industrial applications. We have released our code, pre-trained model, and data processing protocols, which have benefited many researchers in this field.

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) Long et al. (2015), encoder-decoder structures Ronneberger et al. (2015); Fan et al. (2021); Badrinarayanan et al. (2017); Ji et al. (2024d); Chen et al. (2024a); Ji et al. (2023b; 2022), dilated convolutions Chen et al. (2017; 2018); Liu & Zhu (2021); Zang et al. (2024); Hu et al. (2020), pyramid structures Zhu et al. (2021a); Chen et al. (2017); Zhao et al. (2017); Chen et al. (2018); Zhu et al. (2023a); Fu et al. (2022); Ji et al. (2021), attention modules Zhu et al. (2019; 2024b; 2023b); Ji et al. (2024c; 2023a), and transformers Zheng et al. (2021); Xie et al. (2021); Strudel et al. (2021); Cheng et al. (2022); Zhu et al. (2024a). Recent advancements have improved SAM's performance, such as Ke et al. (2023), which introduces a High-Quality output token and trains the model on fine-grained masks. Other efforts have focused on enhancing SAM's efficiency for broader real-world and mobile use, exemplified by Xiong et al. (2023); Zhang et al. (2023); Zhao et al. (2023). The widespread success of SAM has led to its adoption in various fields, including medical imaging Ma et al. (2024); Deng et al. (2023); Mazurowski et al. (2023); Wu et al. (2023), remote sensing Chen et al. (2023a); Ren et al. (2023); Ji et al. (2024a), motion segmentation Xie et al. (2024); Wang et al. (2021b;a); Feng et al. (2018), and camouflaged object detection Tang et al. (2023). Notably, our previous work SAM-Adapter Chen et al. (2023c;b) tested camouflaged object detection, polyp segmentation, and shadow segmentation, and provide with the first adapter-based method to integrate the SAM's exceptional capability to these downstream tasks.

Adapters. The concept of Adapters was first introduced in the NLP community Houlsby et al. (2019) as a tool to fine-tune a large pre-trained model for each downstream task with a compact and scalable model. In Stickland & Murray (2019), multi-task learning was explored with a single BERT model shared among a few task-specific parameters. In the computer vision community, Li et al. (2022); Ji et al. (2024b; 2019) suggested fine-tuning the ViT Dosovitskiy et al. (2020) for object detection with minimal modifications. Recently, ViT-Adapter Chen et al. (2022) leveraged Adapters to enable a plain ViT to perform various downstream tasks. Liu et al. (2023) 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.

Polyp Segmentation. In recent years, there has been notable progress in polyp segmentation Zhou et al. (2021) due to deep-learning approaches. These techniques employ deep neural networks to derive more discriminative features from endoscopic polyp images. Nonetheless, the use of bounding-box detectors often leads to inaccurate polyp boundary localization. To resolve this, Canny (1986) leveraged fully convolutional networks (FCN) with pre-trained models to identify and segment polyps. Qadir et al. (2021) introduced a technique utilizing Fully Convolutional Neural Networks (FCNNs) to predict 2D Gaussian shapes. Subsequently, the U-Net Kingma & Ba (2017) architecture, featuring a contracting path for context capture and a symmetric expanding path for precise localization, achieved favorable segmentation results. However, these strategies focus primarily on entire polyp regions, neglecting boundary constraints. Therefore, Psi-Net Murugesan et al. (2019) incorporated both region and boundary constraints for polyp segmentation, yet the interplay between regions and boundaries remained underexplored. Mahmud et al. (2021) introduced PolypSegNet, an enhanced encoder-decoder architecture designed for the automated segmentation of polyps in colonoscopy images. To address the issue of non-equivalent images and pixels, Guo et al. (2022) proposed a confidence-aware resampling method for polyp segmentation tasks. Specifically for polyp segmentation, works done by Zhou et al. (2023) and Chen et al. (2023c) present promising results using an unprompted SAM and a domain-adapted SAM respectively. Additionally, Polyp-SAM Li et al. (2023) used SAM for the same task. Roy et al. (2023) evaluated the zero-shot capabilities of SAM on the organ segmentation task.

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 Feng et al. (2013); Pike (2018); Hou & Li (2011); Sengottuvelan et al. (2008) 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.

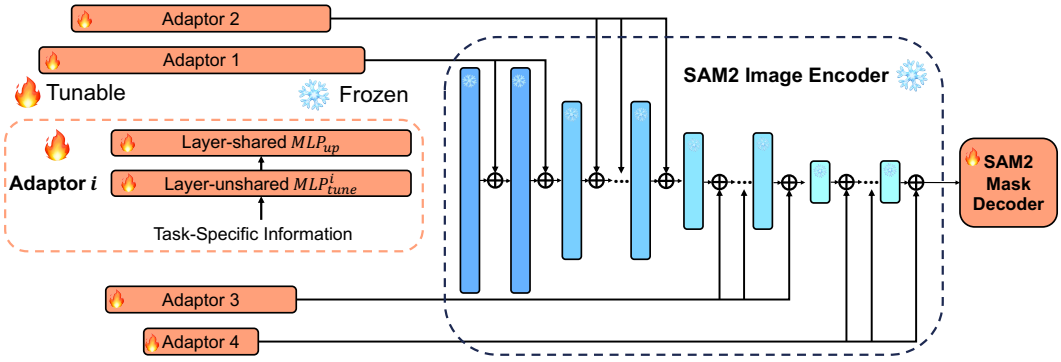

Figure 1: **The architecture of the proposed SAM-Adapter.**

Le et al.Le et al. (2019) 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 Fan et al. (2020b); Mei et al. (2021b); Lin et al. (2023b). In this work, we leverage the advanced neural network backbone (a foundation model – SAM2) 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 Karsch et al. (2011); Lalonde et al. (2012). They can also negatively impact the performance of computer vision tasks Nadimi & Bhanu (2004); Cucchiara et al. (2003). Early method use hand-crafted heuristic cues like chromacity, intensity and texture Huang et al. (2011); Lalonde et al. (2012); Zhu et al. (2010). 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) Le et al. (2018); Cun et al. (2020); Zhu et al. (2018b). 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 2 AS THE BACKBONE

The core of our SAM2-Adapter is built upon the powerful image encoder and mask decoder components of the SAM2 model. Specifically, we leverage the MAE pre-trained Hiera image encoder from SAM2, keeping its weights frozen to preserve the rich visual representations it has learned from pretraining on large-scale datasets. Additionally, we utilize the mask decoder module from the original SAM2 model, initializing its weights with the pretrained SAM2 parameters and then fine-tuning it during the training of our adapter. We do not provide any additional prompts as input to the original SAM2 mask decoder.

We next learn and inject task-specific knowledge F^i into the network via Adapters. We employ the concept of prompting, which utilizes the fact that foundation models like SAM2 have been trained on large-scale datasets. Using appropriate prompts to introduce task-specific knowledge Liu et al. (2023) can enhance the model's generalization ability on downstream tasks, especially when annotated data is scarce.

The architecture of the proposed SAM2-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 Liu et al. (2023). It is worth noting that the different from SAMKirillov et al. (2023), the image encoder of SAM2 has four stages with hierarchical resolutions. Therefore, we initialized four different adapter and insert the four adapter in different layers of each stage. In each stage, the weight of the adapter is shared. Specifically, each of 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 fea-

tures. P^i refers to the output prompt that is attached to each transformer layer of SAM model. GELU is the GELU activation function Hendrycks & Gimpel (2016). The information F^i can be chosen to be in various forms.

3.2 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:

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

4 Experiments

4.1 Tasks and Datasets

In our experiments, we selected two challenging low-level structural segmentation tasks and one medical imaging task to evaluate the performance of the SAM2-Adapter: camouflaged object detection and shadow detection, and polyp segmentation.

For the camouflaged object detection task, we utilized three prominent datasets: COD10K Fan et al. (2020b), CHAMELEON Skurowski et al. (2018), and CAMO Le et al. (2019). 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. The CAMO dataset consists of 1,250 images, with 1,000 for training and 250 for testing. Following the training protocol in Fan et al. (2020b), we used the combined dataset of CAMO and the training set of COD10K for model training. For evaluation, we used the test sets of CAMO and COD10K, as well as the entire CHAMELEON dataset. For the shadow detection task, we employed the ISTD dataset Wang et al. (2018), which contains 1,330 training images and 540 test images.For polyp segmentation (medical image segmentation), we use the kvasir-SEG dataset Jha et al. (2020b). The train-test split followed the settings of the Medico multimedia task at MediaEval 2020: Automatic Polyp Segmentation Jha et al. (2020a).

For evaluation metrics, we followed the protocol in Liu et al. (2023) and used commonly-used metrics such as S-measure (S_m), mean E-measure (E_ϕ), and MAE for the camouflaged object detection task. For the shadow detection task, we used the balance error rate (BER) metric. For the polyp segmentation task, we used mean Dice score (mDice) and mean Intersection-over-Union (mIoU) as the evaluation measures.

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 Liu et al. (2023), 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 one linear layer 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 hiera-large version of SAM2. 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 20 epochs. The experiments are implemented using PyTorch on three NVIDIA Tesla A100 GPUs.

Method	CHAMELEON Skurowski et al. (2018)				CAMO Le et al. (2019)				COD10K Fan et al. (2020b)			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	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 $\downarrow$
SINetFan et al. (2020a)	0.869	0.891	0.740	0.440	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
RankNetLv et al. (2021)	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 Li et al. (2021)	0.870	0.924	-	0.039	0.792	0.839	-	0.82	0.800	0.872	-	0.041
PFNet Mei et al. (2021a)	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 Lin et al. (2023a)	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 Kirillov et al. (2023)	0.727	0.734	0.639	0.081	0.684	0.687	0.606	0.132	0.783	0.798	0.701	0.050
SAM2 Ravi et al. (2024)	0.359	0.375	0.115	0.357	0.350	0.411	0.079	0.311	0.429	0.505	0.115	0.218
SAM-Adapter Chen et al. (2023c;b)	0.896	0.919	0.824	0.033	0.847	0.873	0.765	0.070	0.883	0.918	0.801	0.025
SAM2-Adapter (Ours)	**0.915**	**0.955**	**0.889**	**0.018**	**0.855**	**0.909**	**0.810**	**0.051**	**0.899**	**0.950**	**0.850**	**0.018**

Table 1: Quantitative Segmentation Result Comparison for Camouflaged Object Detection

4.3 Experiments for Camouflaged Object Detection

We first evaluated SAM on the challenging task of camouflaged object detection, where foreground objects often blend with visually similar background patterns. Our experiments revealed that SAM did not perform well in this task. As shown in Figure 2, SAM failed to detect several concealed objects. This was further confirmed by the quantitative results presented in Table 1, where SAM's performance was significantly lower than existing state-of-the-art methods across all evaluated metrics, while SAM2, on its own, had the lowest performance, which fails to produce any meaningful results.

In contrast, Figure 3 clearly demonstrates that by introducing the SAM2-Adapter, our method significantly elevates the model's performance. 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 methods.

Furthermore, the SAM2-Adapter set a new SOTA performance. Visualized results show that SAM2-Adapter segments more precisely without adding extra false information, further demonstrating the robustness and accuracy of our approach.

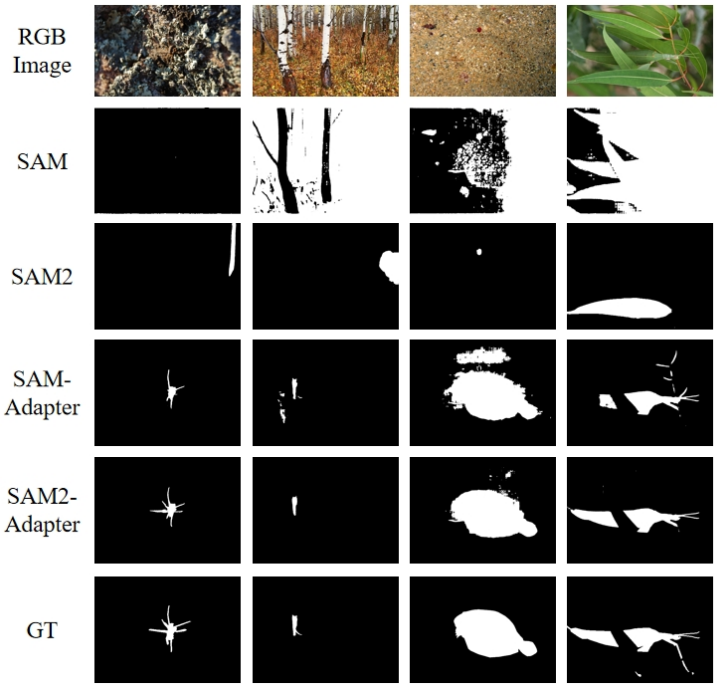

Figure 2: **Shadow Detection Visualization** As shown in the figure, SAM often fails to detect animals that are visually camouflaged within their natural environments and can sometimes produce irrelevant results. SAM2 also struggles with similar issues and produce non-meaningful outcomes. However, by incorporating SAM-Adapter, our approach significantly improves object segmentation performance. Furthermore, SAM2-Adapter demonstrates even better performance than SAM-Adapter. The samples depicted are from the CHAMELEON dataset.

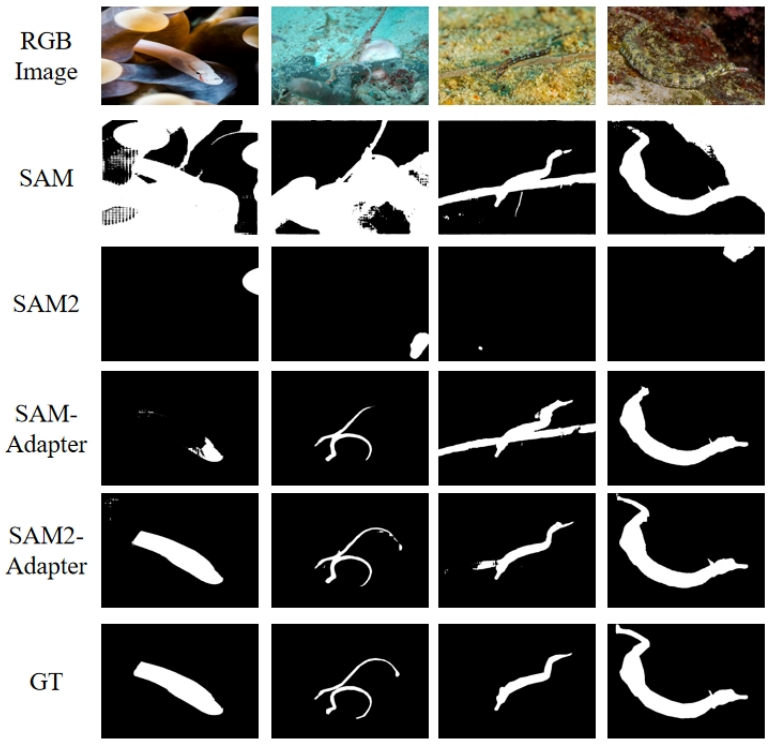

Figure 3: **Visualization for Camouflaged Image Segmentation in COD-10K dataset** As shown in the figure, SAM struggles to detect animals that are visually camouflaged within their natural environments and can sometimes produce results that lack meaningful segmentation. SAM2 also faces similar challenges, often resulting in no output or false results. However, by incorporating SAM2-Adapter, our method significantly improves object segmentation performance, surpassing SAM-Adapter. For other dataset, please refer to *More Results* section.

Method	BER ↓
Stacked CNN Vicente et al. (2016)	8.60
BDRAR Zhu et al. (2018a)	2.69
DSC Hu et al. (2018)	3.42
DSD Zheng et al. (2019)	2.17
FDRNet Zhu et al. (2021b)	1.55
SAM Kirillov et al. (2023)	40.51
SAM2 Ravi et al. (2024)	50.81
SAM-Adapter	**1.43**
SAM2-Adapter (Ours)	**1.43**

Table 2: Result for Shadow Detection

4.4 Experiments for Shadow Detection

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

Similarly, SAM2 also struggled with the ”shadow” concept without proper prompting, failing to produce meaningful results. In our study, we compared various methods for shadow detection and found that SAM’s performance was significantly poorer than existing methods. However, by integrating the SAM-Adapter, we achieved a substantial improvement in performance. The SAM-Adapter enhanced the detection of shadow regions, making them more clearly identifiable. Furthermore, SAM2-Adapter worked just as effectively as SAM-Adapter, delivering comparable results. Our findings were validated through quantitative analysis, and Table 2 demonstrates the significant performance boost provided by the SAM-Adapter and matched by the SAM2-Adapter for shadow detection.

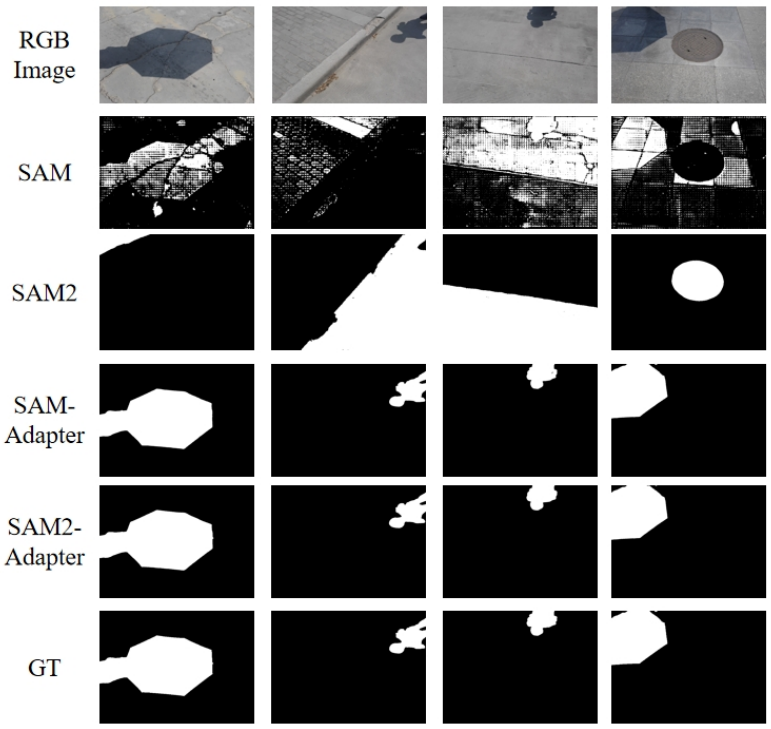

Figure 4: **Shadow Detection Visualized.** Both SAM and SAM2 have no understanding about the "shadow" concept without proper prompting. They produce meaningless results. SAM-Adapter and SAM2-Adapter perform equally well in shadow detection tasks.

Method	mDice ↑	mIoU ↑
UNet Ronneberger et al. (2015)	0.821	0.756
UNet++ Zhou et al. (2018)	0.824	0.753
SFA Fang et al. (2019)	0.725	0.619
SAM Kirillov et al. (2023)	0.778	0.707
SAM2 Ravi et al. (2024)	0.200	0.029
SAM-Adapter	0.850	0.776
SAM2-Adapter (Ours)	**0.873**	**0.806**

Table 3: Quantitative Result for Polyp Segmentation

4.5 Experiments for Polyp Segmentation

We illustrate the application of SAM2-Adapter in the context of medical image segmentation, specifically focusing on polyp segmentation. Polyps, which have the potential to become malignant, are identified during colonoscopy and removed through polypectomy. Accurate and swift detection and removal of polyps are crucial in preventing colorectal cancer, a leading cause of cancer-related deaths globally.

While numerous deep learning approaches have been developed for polyp identification, and the pre-trained SAM model shows promise in identifying some polyps, its performance can be significantly improved with our SAM-Adapter approach. However, without proper prompting, the SAM2 model fails to produce meaningful results. Our SAM2-Adapter addresses this issue and outperforms the original SAM-Adapter. The results of our study, presented in Table 3 and the visualization results in Figure 6, underscore the effectiveness of SAM2-Adapter in improving the accuracy and reliability of polyp detection.

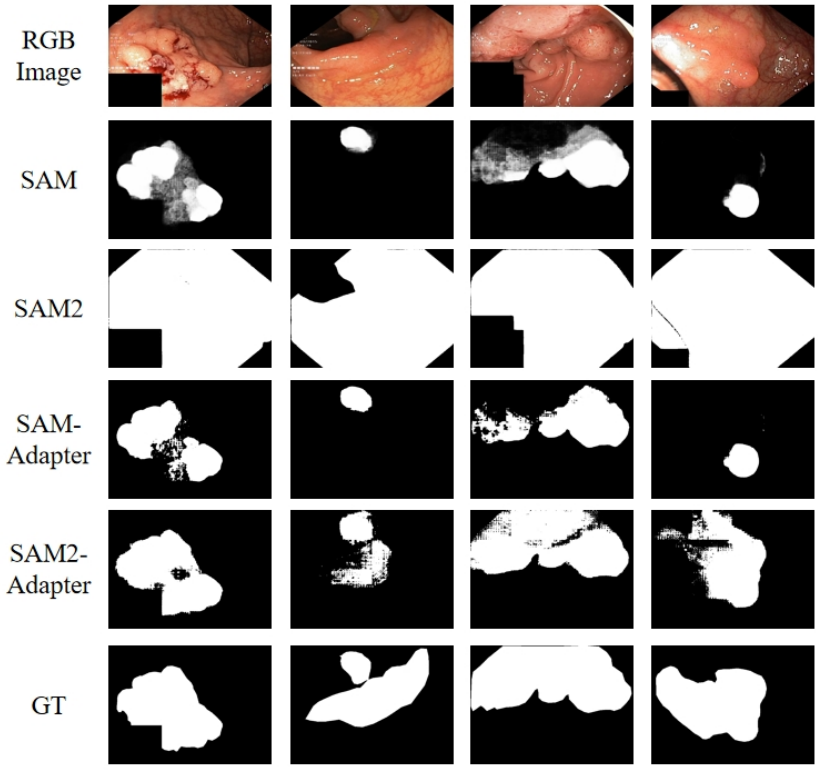

Figure 5: **Visualization of Polyp Segmentation Results.** As illustrated in the figure, although SAM can identify some polyp structures in the image, the result is not accurate. Without proper prompting, SAM 2 failed to deliver meaningful polyp segmentation results. By using SAM2-Adapter, our approach significantly outperform SAM-Adapter with more accurate (and complete) segmentation results.

5 Conclusion and Future Work

In this paper, we introduced SAM2-Adapter, a novel adaptation method designed to leverage the advanced capabilities of the Segment Anything 2 (SAM2) model for specific downstream segmentation tasks. SAM2-Adapter utilizes a multi-adapter configuration that is specifically tailored to SAM2's multi-resolution hierarchical Transformer architecture. This approach effectively addresses the limitations encountered with SAM, enabling the achievement of new state-of-the-art (SOTA) performance in challenging segmentation tasks such as camouflaged object detection, shadow detection, and polyp segmentation.

Our experiments demonstrate that SAM2-Adapter not only retains the beneficial features of its predecessor, including generalizability and composability but also enhances these capabilities by integrating seamlessly with SAM2's advanced architecture. This integration allows SAM2-Adapter to outperform previous methods and set new benchmarks across various datasets and tasks.

The continued presence of challenges from SAM in SAM2 highlights the inherent complexities of applying foundation models to diverse real-world scenarios. Nevertheless, SAM2-Adapter effectively addresses these issues, showcasing its potential as a robust tool for high-quality segmentation in a range of applications.

We encourage researchers and engineers to adopt SAM2 as the backbone for their segmentation tasks, coupled with SAM2-Adapter, to realize improved performance and advance the field of image segmentation. Our work not only extends the capabilities of SAM2 but also paves the way for future innovations in adapting large pre-trained models for specialized applications.

REFERENCES

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

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.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model, 2023a. URL https://arxiv.org/abs/2306.16269.

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

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)*, pp. 801–818, 2018.

Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3367–3375, 2023b.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, 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, medical image segmentation, and more, 2023c. URL https://arxiv.org/abs/2304.09148.

Tianrun Chen, Chaotao Ding, Lanyun Zhu, Tao Xu, Deyi Ji, Ying Zang, and Zejian Li. xlstm-unet can be an effective 2d\& 3d medical image segmentation backbone with vision-lstm (vil) better than its mamba counterpart. *arXiv preprint arXiv:2407.01530*, 2024a.

Tianrun Chen, Chunan Yu, Jing Li, Jianqi Zhang, Lanyun Zhu, Deyi Ji, Yong Zhang, Ying Zang, Zejian Li, and Lingyun Sun. Reasoning3d–grounding and reasoning in 3d: Fine-grained zero-shot open-vocabulary 3d reasoning part segmentation via large vision-language models. *arXiv preprint arXiv:2405.19326*, 2024b.

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

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

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.

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, pp. 10680–10687, 2020.

Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang, and Yuankai Huo. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging, 2023. URL https://arxiv.org/abs/2304.04155.

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.

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*, pp. 2777–2787, 2020a.

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*, pp. 2777–2787, 2020b.

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*, pp. 9716–9725, 2021.

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*, pp. 302–310. Springer, 2019.

Weitao Feng, Deyi Ji, Yiru Wang, Shuorong Chang, Hansheng Ren, and Weihao Gan. Challenges on large scale surveillance video analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 69–76, 2018.

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*, pp. 93–96, 2013.

Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022.

Xiaoqing Guo, Zhen Chen, Jun Liu, and Yixuan Yuan. Non-equivalent images and pixels: Confidence-aware resampling with meta-learning mixup for polyp segmentation. *Medical image analysis*, 78:102394, 2022.

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

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.

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.

Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2020.

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*, pp. 7454–7462, 2018.

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*, pp. 898–905. IEEE, 2011.

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*, 2020a.

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*, pp. 451–462. Springer, 2020b.

Deyi Ji, Hongtao Lu, and Tongzhen Zhang. End to end multi-scale convolutional neural network for crowd counting. In *Eleventh international conference on machine vision*, volume 11041, pp. 761–766. SPIE, 2019.

Deyi Ji, Haoran Wang, Hanzhe Hu, Weihao Gan, Wei Wu, and Junjie Yan. Context-aware graph convolution network for target re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1646–1654, 2021.

Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16876–16885, 2022.

Deyi Ji, Feng Zhao, and Hongtao Lu. Guided patch-grouping wavelet transformer with spatial congruence for ultra-high resolution segmentation. *International Joint Conference on Artificial Intelligence*, pp. 920–928, 2023a.

Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23621–23630, 2023b.

Deyi Ji, Siqi Gao, Mingyuan Tao, Hongtao Lu, and Feng Zhao. Changenet: Multi-temporal asymmetric change detection dataset. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2725–2729. IEEE, 2024a.

Deyi Ji, Siqi Gao, Lanyun Zhu, Qi Zhu, Yiru Zhao, Peng Xu, Hongtao Lu, Feng Zhao, and Jieping Ye. View-centric multi-object tracking with homographic matching in moving uav. *arXiv preprint arXiv:2403.10830*, 2024b.

Deyi Ji, Wenwei Jin, Hongtao Lu, and Feng Zhao. Pptformer: Pseudo multi-perspective transformer for uav segmentation. *International Joint Conference on Artificial Intelligence*, 2024c.

Deyi Ji, Feng Zhao, Lanyun Zhu, Wenwei Jin, Hongtao Lu, and Jieping Ye. Discrete latent perspective learning for segmentation and detection. In *Forty-first International Conference on Machine Learning*, 2024d.

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.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality, 2023. URL `https://arxiv.org/abs/2306.01567`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL `https://arxiv.org/abs/1412.6980`.

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.

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.

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)*, pp. 662–678, 2018.

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.

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*, pp. 10071–10081, 2021.

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*, pp. 280–296. Springer, 2022.

Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. Polyp-sam: Transfer sam for polyp segmentation, 2023. URL https://arxiv.org/abs/2305.00293.

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, 2023a.

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, 2023b.

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

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

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

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*, pp. 11591–11601, 2021.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL http://dx.doi.org/10.1038/s41467-024-44824-z.

Tanvir Mahmud, Bishmoy Paul, and Shaikh Anowarul Fattah. Polypsegnet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. *Computers in biology and medicine*, 128:104119, 2021.

Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918, October 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102918. URL http://dx.doi.org/10.1016/j.media.2023.102918.

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*, pp. 8772–8781, 2021a.

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*, pp. 8772–8781, 2021b.

Balamurali Murugesan, Kaushik Sarveswaran, Sharath M Shankaranarayana, Keerthi Ram, and Mohanasankar Sivaprakasam. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation, 2019. URL https://arxiv.org/abs/1902.04099.

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.

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

Hemin Ali Qadir, Younghak Shin, Johannes Solhusvik, Jacob Bergsland, Lars Aabakken, and Ilangko Balasingham. Toward real-time polyp detection using fully cnns for 2d gaussian shapes prediction. *Medical Image Analysis*, 68:101897, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Simiao Ren, Francesco Luzi, Saad Lahrichi, Kaleb Kassaw, Leslie M. Collins, Kyle Bradbury, and Jordan M. Malof. Segment anything, from space?, 2023. URL https://arxiv.org/abs/2304.13000.

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*, pp. 234–241. Springer, 2015.

Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model, 2023. URL https://arxiv.org/abs/2304.05396.

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*, pp. 6–10. IEEE, 2008.

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

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*, pp. 5986–5995. PMLR, 2019.

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*, pp. 7262–7272, 2021.

Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection, 2023. URL https://arxiv.org/abs/2304.04709.

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*, pp. 816–832. Springer, 2016.

Haoran Wang, Licheng Jiao, Fang Liu, Lingling Li, Xu Liu, Deyi Ji, and Weihao Gan. Ipgn: Interactiveness proposal graph network for human-object interaction detection. *IEEE Transactions on Image Processing*, 30:6583–6593, 2021a.

Haoran Wang, Licheng Jiao, Fang Liu, Lingling Li, Xu Liu, Deyi Ji, and Weihao Gan. Learning social spatio-temporal relation graph in the wild and a video benchmark. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):2951–2964, 2021b.

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*, pp. 1788–1797, 2018.

Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation, 2023. URL https://arxiv.org/abs/2304.12620.

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.

Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is sam (and flow), 2024. URL https://arxiv.org/abs/2404.12389.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything, 2023. URL https://arxiv.org/abs/2312.00863.

Ying Zang, Chenglong Fu, Runlong Cao, Didi Zhu, Min Zhang, Wenjun Hu, Lanyun Zhu, and Tianrun Chen. Resmatch: Referring expression segmentation in a semi-supervised manner. *arXiv preprint arXiv:2402.05589*, 2024.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023. URL https://arxiv.org/abs/2306.14289.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. URL https://arxiv.org/abs/2306.12156.

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*, pp. 5167–5176, 2019.

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

Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps?, 2023. URL https://arxiv.org/abs/2304.07583.

Yuan Zhou, Haiyang Wang, Shuwei Huo, and Boyu Wang. Full-attention based neural architecture search using context auto-regression, 2021. URL https://arxiv.org/abs/2111.07139.

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*, pp. 3–11. Springer, 2018.

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*, pp. 223–230. IEEE, 2010.

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*, pp. 12537–12546, 2021a.

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

Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Learning gabor texture features for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1621–1631, 2023b.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3065–3075, 2024a.

Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Addressing background context bias in few-shot segmentation through iterative modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3370–3379, 2024b.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024c.

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)*, pp. 121–136, 2018a.

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)*, pp. 121–136, 2018b.

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*, pp. 4702–4711, 2021b.

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*, pp. 593–602, 2019.

A Appendix

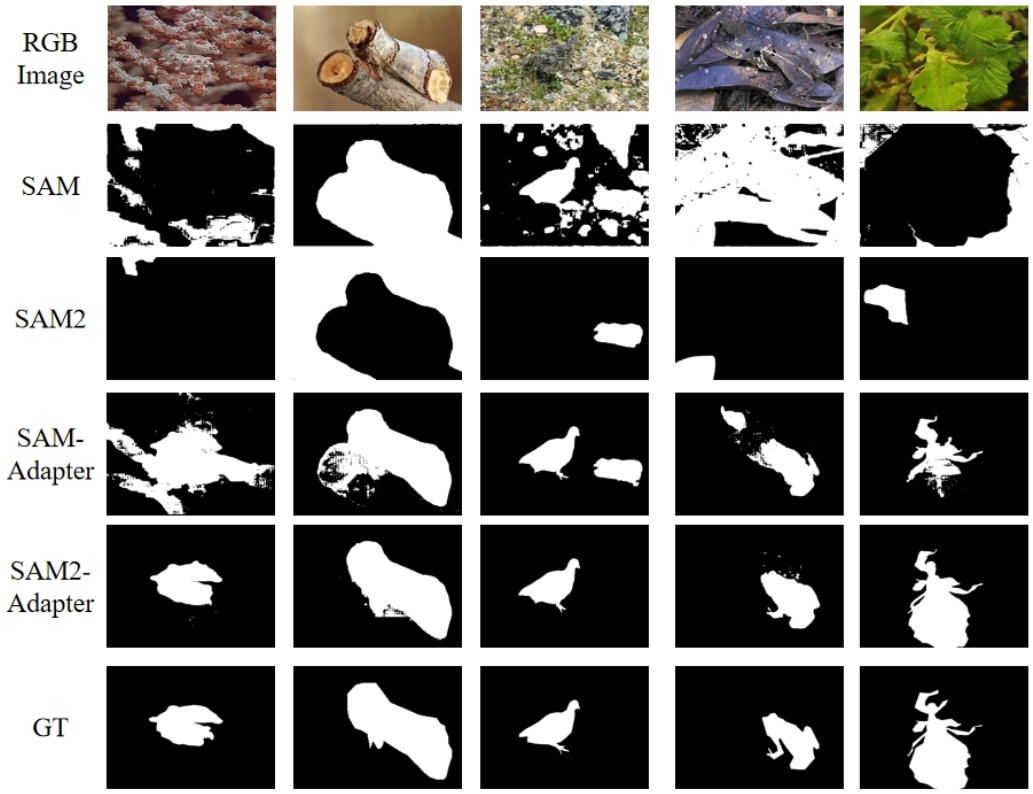

Figure 6: **Camouflaged Segmentation of CAMO dataset.** The SAM and SAM 2 failed to perceive those animals that are visually 'hidden'/concealed in their natural surroundings. By using SAM2-Adapter, our approach can significantly elevate the performance of object segmentation with SAM.