Multi-modal Segment Anything Model for Camouflaged Scene Segmentation

Guangyu Ren[1*] Hengyan Liu[1*] Michalis Lazarou[2] Tania Stathaki[2]
[1]Xi'an Jiaotong-Liverpool University, China
[2]Imperial College London, United Kingdom
{guangyu.ren, hengyan.liu}@xjtlu.edu.cn, {michalis.lazarou14, t.stathaki}@imperial.ac.uk

Abstract

Camouflaged scenes, where objects blend seamlessly into their environments, pose significant challenges to both human observers and computer vision systems. To address this, we propose a novel framework that leverages off-the-shelf foundation models to generate multi-modal prompts for the Segment Anything Model (SAM), thus eliminating the need for manual prompts and significantly improving overall performance on this downstream task. At first, we generate an image caption using the BLIP model and obtain its text embedding through the use of a text encoder. We then generate a visual embedding through the vision encoder of the BLIP model and use both as inputs to SAM to provide additional semantic information about the image. Finally, we propose a couple of architectural novelties, a) we effectively integrate the multi-modal information in SAM through a multi-level adapter and b) we replace the dense embedding of SAM with the image embedding of its image encoder. Our method achieves new state-of-the-art performance in 11 out of 12 metrics in three benchmark datasets for camouflaged detection. Additionally, our method can be successfully adapted to other tasks such as medical image segmentation performing on par or even outperforming the state-of-the-art methods. Our code is available in https://github.com/ic-qialanqian/Vision-Language-SAM.

1. Introduction

Camouflaged Object Detection (COD) focuses on identifying objects that are visually blended into their surroundings, often sharing similar colors and patterns, which makes them challenging to detect. COD has garnered significant attention in the field of computer vision due to its potential implications and extensive applications in areas such as surveillance systems [15], wildlife monitoring [44], tracking [25, 36] and low-altitude economy (LAE) [9]. However, this task remains a significant challenge for computer vision because of the barely detectable differences between the object and its surroundings, causing traditional segmentation methods to either overlook or misclassify the target objects.

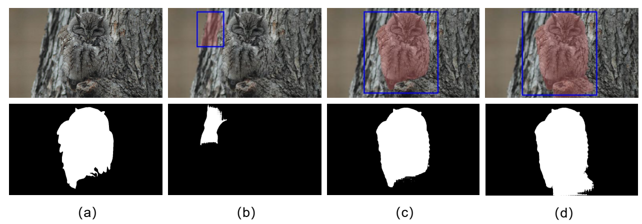

Figure 1. Visualization of segmentation results generated by SAM under different bounding-box prompts. (a) Original image and corresponding ground truth mask. (b) Segmentation result with a bounding-box prompt placed in the background, failing to capture the object. (c) Segmentation result with a properly placed bounding-box prompt, closely matching the ground truth. (d) Segmentation result with a bounding-box prompt slightly larger than in (c), resulting in over-segmentation and mask leakage into the background. This figure highlights the sensitivity of SAM to bounding-box prompt placement and size.

Despite the remarkable performance of deep learning-based methods in various computer vision tasks, they still struggle with Camouflaged Object Detection (COD). A key reason for this underperformance is the limited size of available datasets. Without high-quality datasets, models struggle to learn expressive feature representations, which are essential for accurately distinguishing camouflaged objects from their surroundings. Obtaining more data can be a way to address this issue, nevertheless this can be labor-intensive since it requires humans to annotate correctly numerous images.

Another approach is to leverage large foundation models like the Segment Anything Model (SAM) that have demonstrated adaptability for downstream tasks using sparse prompts (e.g., text, bounding boxes) and dense prompts (e.g., masks) for segmentation. However, even SAM that has been trained on the huge dataset SA-1B struggles with COD for many reasons. One reason is COD is highly complex since background and foreground look very similar. Another problem is a fundamental limitation of SAM, that relies on

[*]Equal contribution.

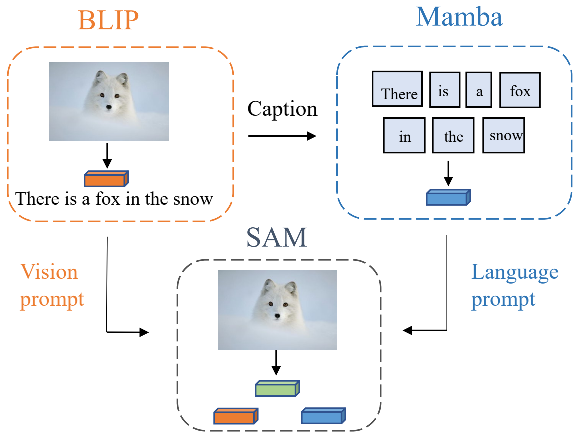

Figure 2. The proposed paradigm for segmenting camouflaged objects. BLIP is used for image captioning to automatically generate descriptions for each input image. A text encoder, Mamba, then extracts text embeddings from the generated descriptions, which are further combined with vision embeddings from BLIP's image encoder to create vision-language prompts for the SAM decoder.

manual user prompts, referred to as sparse prompts, such as a bounding boxes, text and pixel points.

Specifically, relying on these sparse prompts can be a significant source of error since the size of the bounding box can significantly affect the output of SAM. This is evident in Figure 1, that shows how SAM struggles to segment the object of interest by predicting different segmentation masks for every different bounding box input. For instance, Figure 1(b) demonstrates that if a given bounding box encloses the background, SAM will segment the background instead of the target animal. When the bounding box is slightly enlarged by moving only 10 pixels from Figure 1(c) to (d), which could easily occur in practice, the segmentation mask in (d) shows a visible false positive compared to (c).

In this work we address the challenges of SAM in COD and propose a novel method that leverages SAM to successfully segment COD images. Specifically, we propose to discard the manual prompt inputs completely and utilize multi-modal prompts to guide the segmentation process of SAM so that the object of interest is successfully segmented. We propose a novel way of using text information by utilizing the BLIP model to generate a text caption of the image of interest and embed this caption using a powerful text encoder, the Mamba proposed in [4]. We also use the BLIP model's expressive visual encoder to obtain a vision embedding from the image of interest. Both the vision and language embeddings can provide complementary information that can aid SAM segment the object of interest (Figure 2).

Additionally, we propose two architectural ideas to improve the performance of our method. First, we utilize a multi-level adapter as a way to incorporate the visual information effectively by providing the vision embedding as input in multiple of SAM's transformer layers. The multi-level adapter serves to bridge the gap between the unaligned semantic spaces of vision and text embeddings and SAM's transformer layers. Second, we identify a limitation of SAM, which is the use of a dense embedding that constitutes of a mask embedding that is generated by embedding a user provided input mask using a convolutional network and adding it to the image embedding generated by SAM's image encoder. However, if no input mask is provided, then the masked embedding is replaced with a randomly initialized embedding that can lead to degradation in performance. We propose to completely discard the mask embedding and use only the image embedding from SAM's image encoder, completely alleviating SAM's reliance on manually provided user inputs while also enhancing performance in COD. We refer to our method as *Multi-modal Segment Anything (MM-SAM)*. Our contributions can be summarized as follows:

- We proposed a novel way to exploit text embeddings by using the BLIP model to generate a text caption for the image of interest.
- We are the first, to the best of our knowledge, to utilize both vision and language information to improve SAM's performance in COD.
- We integrate two architectural ideas in SAM that improve SAM's performance, namely a multi-level adapter and a dense embedding relying only on SAM's image embedding.
- We achieve new state of the art performance in 11 out of 12 metrics in 3 widely popular benchmark COD.

2. Related Work

Camouflaged Object Detection Camouflaged object detection (COD) is a challenging task in computer vision that involves identifying objects that closely resemble their background. In early studies, researchers relied on low-level features such as texture, brightness, and color to differentiate the foreground from the background [34, 38]. These features formed the foundational basis for various segmentation techniques. However, the performance of most of these methods heavily depends on accurately annotated samples for training, resulting in significant manual costs and limits in their practical application. Deep learning has been the paradigm of choice in more recent COD works after [3] trained a deep learning model on a new COD dataset achieving remarkable performance in the process.

Currently, the main lines of research that researchers attempt to tackle COD through deep learning is through improvements in network architecture, learning paradigm and supervision level [47]. One study targeting architectural design is [41], where the authors propose a sophisticated architecture using an attention mechanism to fuse multi-level features. Another work extracts frequency-domain features using the discrete cosine transform and fuses them with RGB

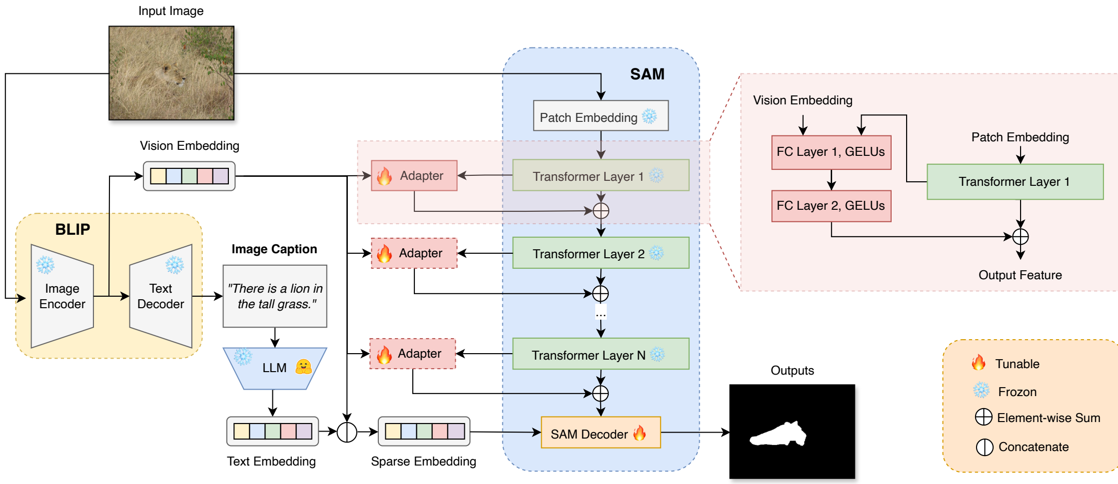

Figure 3. The overall structure of the proposed method. 1) BLIP Image encoder is used to obtain the vision embedding. 2) BLIP's text decoder and the Mamba text encoder are used to obtain the text embedding. 3) The Vision embedding is incorporated into SAM using a multi-level adapter. 4) The text embedding and the vision embedding are concatenated to form sparse embedding. 5) The sparse embedding is concatenated with SAM's image embedding to SAM's decoder. 6) SAM's decoder outputs the predicted segmentation mask.

images via a novel architecture [50].

Other ideas include using multi-task learning such as using an auxiliary loss function to help guide the network such as predicting associated cues in the such as boundary along with the segmentation mask [5] and localizing the camouflaged object simultaneously with segmenting them [30]. Also, different levels of supervision has been explored as a way to alleviate the burden of relying on pixel-wise annotated datasets. Weakly-supervised training regimes have been proposed that use only sparse annotations such as points and scribbles [6] as well as semi-supervised training regimes that use labeled and unlabeled by generating pseudo-labels for the latter and exploiting both to obtain more accurate segmentation predictions [29].

Segment Anything The Segment Anything (SAM) [17] proposed the first foundational model for image segmentation, achieving remarkable zero-shot performance in novel segmentation tasks in a similar way that Large Language models achieved in language tasks. The success of this work inspired new lines of research that attempt to utilize the pre-trained SAM model effectively in novel segmentation tasks. Fine-tuning SAM effectively has been the main research interest in attempting to use SAM successfully in novel situations. [2] uses a simple adapter to incorporate domain-specific information or visual prompts into SAM. In [33], the authors proposed to fine-tune SAM efficiently by parameter space reconstruction. In [21], inspired from the success of adversarial training in NLP, the authors proposed to use adversarial tuning to improve the generalizability of SAM by mapping natural images onto a low-dimensional manifold using a generative model, perturbing their embedding along an adversarial direction within the manifold before generating new adversarial images that are used to fine-tune SAM. Apart from fine-tuning SAM, GenSAM [10] proposed to utilize generic text prompts to automatically generate and utilize visual prompts to aid COD performance.

Language Prompting Language prompting is the concept of providing a large language model with text input, referred to as a prompt, to guide its generated response. It has been shown to help language models achieve impressive improvements in performance in NLP tasks [1]. Different types of prompts have been proposed such as hard prompts [46], which use a pre-designed template to guide model outputs, and soft prompts which use trainable embeddings that are fine-tuned to various tasks [19]. This paradigm shift in NLP inspired the Computer Vision community to research language prompting as a way of improving the performance of computer vision tasks. The seminal work of CLIP [35] trained very robust visual models by training them using image and caption pairs.

Segmentation tasks have also explored language prompts, for example, SAM leverages text prompts provided by a user to improve segmentation performance. More recent works also focused on addressing SAM's limitations through language prompting such as GenSAM [10], which eliminated the need for user prompts by using generic text prompts to guide the segmentation process.

Visual Prompting is inspired by prompt tuning in natural language processing (NLP) [1, 26] but rather than use text as input, visual cues are used as the prompt input. The original SAM model utilized visual prompts such as bounding boxes, point coordinates within the image that indicate areas of interest and masks that are image regions that may have been partially segmented. Along this line of research, [28] proposed focus on visual prompts that are image-specific in contrast to the typical dataset-specific visual prompts. Specifically, the authors proposed to use explicit visual information from every image such as features from frozen patch embeddings and the image's high-frequency components. The work of [2] addressed the limitations of the Segment Anything Model (SAM) in handling specific segmentation tasks like camouflaged object detection, shadow detection, and medical image segmentation by proposing to use a lightweight adapter model to incorporate domain-specific visual prompts into the SAM network.

Our work bears some similarities with GenSAM [10]. While GenSAM uses a generic text prompt followed by a cross-modal chain of thoughts prompting to obtain heatmaps for visual prompts, ours simply leverages BLIP and Mamba to generate image specific text prompts and uses BLIP's image embedding as a visual prompt. Additionally, our work proposes architectural novelties integrated in SAM such as a multi-level adapter and the utilization of an improved dense embedding in contrast to GenSAM that proposes none.

3. Methodology

3.1. Overview of Architecture

We follow the SAM baseline in [2] and directly adopt the original image encoder and mask decoder from the one with ViT Huge (ViT-H) architecture. Figure 3 illustrates the overall structure of the proposed method. We first exploit BLIP for the image captioning task to automatically generate descriptions for each input image. Then an LLM or a text encoder, such as Mamba, is used to extract text embeddings from the generated descriptions, which are further combined with vision embeddings from BLIP's image encoder to create vision-language prompts for the SAM decoder. Additionally, within the SAM image encoder, the vision embeddings from each transformer block are integrated with those from BLIP. This combination is used to train an adapter for fine-tuning SAM's image encoder.

3.2. Vision-Language Prompts

Caption Generation via BLIP BLIP is first used as a caption generator to automatically provide an appropriate description for each camouflaged scene I_i:

$$\text{prompt}_i = \text{BLIP}(I_i), \tag{1}$$

where, the generated prompt_i is a brief sentence that describes and summarizes the primary elements of the given input I_i. Notably, no instruction is provided to specify the template of the generated caption; it relies solely on the input image. For example, in Figure 3, BLIP generates the following caption for the given image: "*There is a **lion** in the **tall grass***.", which highlights the key elements in the image (*i.e.,* lion and tall grass).

Vision Embeddings with BLIP Given that BLIP is pretrained on a large-scale dataset and possesses the capability to generate coherent image descriptions, its encoder can serve as a highly efficient feature extractor for generating the vision embedding e_i^{vision} of the given image I_i. The embedding e_i^{vision} contains comprehensive information and will be used to train the model in the latter stages:

$$e_i^{vision} = \text{BLIP}_{encoder}(I_i), \tag{2}$$

where $\text{BLIP}_{encoder}$ represents its image encoder.

Text Embeddings with Mamba To better utilize the generated caption of the input image for model training, an LLM – Mamba is employed to extract the text embedding e_i^{text} from the generated description:

$$e_i^{text} = \text{Mamba}(\text{prompt}_i). \tag{3}$$

Then, the vision embedding e_i^{vision} and the text embedding e_i^{text} are concatenated to form the input token of SAM's decoder:

$$e_i^{sparse} = \text{concat}(e_i^{text}, e_i^{vision}), \tag{4}$$

where e_i^{sparse} denotes the sparse embedding.

3.3. Multi-Level Feature Adapter

One of our key contributions is the development of an adapter to fine-tune the image encoder, which integrates both local-level features from the transformer block of SAM and global image-level features from vision-language models (VLMs), such as BLIP. This differs from previous work like SAM-Adapter [2], which focuses primarily on adapters that incorporate visual prompts and task-specific frequency information. The output embedding of each adapter layer and each transformer block are then added together to produce the final output (see Figure 3). Hence, the proposed adapter layer not only reduces the computational cost of fine-tuning but also allows for better task-specific representation learning by incorporating vision embeddings from BLIP's pretrained image encoder. Moreover, the vision embedding e_i^{vision} of BLIP's image encoder contains image-level feature, while the output embedding $\mathcal{F}_l^{encoder}$ from the l-th transformer block contains pixel-level features. According to [12], the above multi-level adaptation is capable of aligning BLIP's features to the demands of COD contexts. The details of the multi-level feature adapter can be described as:

$$\mathcal{F}_l^{adapter} = \text{Adapter}^l(\text{concat}(\mathcal{F}_l^{encoder}, e_i^{vision})), \tag{5}$$

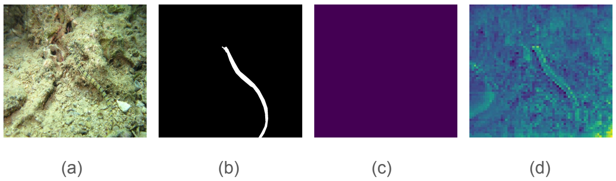

Figure 4. Visualization for the heatmaps of dense embeddings and image embeddings in SAM. (a) RGB image (b) Ground Truth (c) The original dense embedding in SAM (d) The original image embedding of image encoder in SAM.

where $\text{Adapter}^l(\cdot)$ denotes the l-th adapter layer and $\mathcal{F}_l^{adapter}$ is its output feature. To reduce the number of learnable parameters, all adapter layers share the same weights. Finally, $\mathcal{F}_l^{adapter}$ and $\mathcal{F}_l^{encoder}$ are added together to form the input feature $\mathcal{F}_l^*$ for the next transformer block stage as:

$$\mathcal{F}_l^* = \gamma \cdot \mathcal{F}_l^{adapter} + (1-\gamma) \cdot \mathcal{F}_l^{encoder}, \tag{6}$$

where γ is the weight coefficient. By leveraging the proposed multi-level feature adapter, the model can integrate both global and local features, enhancing its ability to discriminate foreground objects in the image.

3.4. Rethinking Dense Embedding in Mask Decoder

In the original SAM framework, the mask decoder is used to map the image embedding $e_i^{encoder}$, the prompt embedding to the final segmentation result. In particular, the prompt embedding consists of a sparse embedding e_i^{sparse} and a dense embedding e_i^{dense}, which are generated by the sparse prompts (*e.g.*, text, bounding boxes, points) and the dense prompts (*e.g.,* masks), respectively. However, when the mask is not accessible, the dense embedding is replaced by a randomly initialized variable and its visualization result is shown in Figure 4c. From the visualization results, it is evident that the embedding initialized by random variables lacks significant semantic information. Instead of adopting a random dense embedding, we propose to use the image embedding $e_i^{encoder}$ generated from the image encoder of SAM since it contains useful semantic information as it is evidently shown in Figure 4d. Therefore, the image embedding $e_i^{encoder}$ is passed through a fully connected layer with an activation function to produce a semantically meaningful dense embedding e_i^{dense}:

$$e_i^{dense} = \text{GELU}(\text{Linear}(e_i^{encoder})), \tag{7}$$

where $\text{GELU}(\cdot)$ is the Gaussian Error Linear Units activation function [8] and $\text{Linear}(\cdot)$ is the fully connected layer. Next, following the original SAM approach, the image embedding $e_i^{encoder}$ is added to the dense embedding e_i^{dense} and then fed into the mask encoder with the sparse embedding e_i^{sparse} to predict the final result $\mathcal{S}_i^{final}$:

$$\mathcal{S}_i^{final} = \text{Decoder}(e_i^{dense} + e_i^{encoder}, e_i^{sparse}). \tag{8}$$

Finally, the model is trained with the linear combination of the focal loss and the dice loss.

4. Experiments

4.1. Dataset and Evaluation Metrics

For the task of camouflaged object detection, we utilize the COD10K dataset [3], the CHAMELEON dataset [40], and the CAMO dataset [18] in our experiments, without providing any additional prompts such as bounding boxes, scribbles, points, or full masks. COD10K is the most extensive dataset for camouflaged object detection, comprising 3,040 training samples and 2,026 testing samples. The CHAMELEON dataset contains 76 images sourced from the Internet for testing purposes. The CAMO dataset includes 1,250 images, with 1,000 designated for training and 250 for testing. Adherent to the training protocol outlined in Fan [3] and Chen [2], we use a composite dataset of CAMO and COD10K for training. Performance evaluation is conducted by using the CAMO and COD10K test sets, as well as the entire CHAMELEON dataset. For the task of medical image segmentation, we use ColonDB [42] and Kvasir [16] for polyp image segmentation

Based on the standard evaluation of COD, we utilize common metrics to assess our method: structure-measure (S_α), weighted F-measure (F_β^w), mean E-measure (E_ϕ), and mean absolute error (MAE). The structure-measure (S_α) adopts object perception and region perception to assess the similarity of structural information. The weighted F-measure F_β^w represents the weighted mean F-measure (Fm) metric, which indicates the standard overall performance that is computed as a function of precision and recall. The mean E-measure (E_ϕ) aims at capturing the image-level statistics and local pixel-matching information. The mean absolute error (MAE) metric is commonly used to measure the average pixel-level error, indicating the similarity between prediction masks and ground truth masks. Additionally, we use the Precision-Recall (PR) and F-measure curves to show the overall performance.

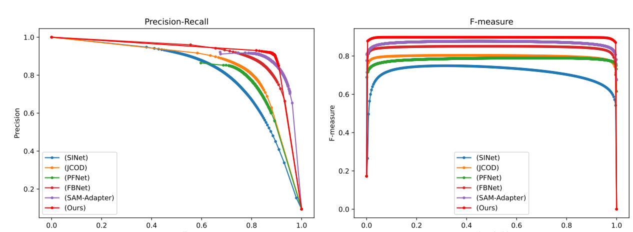

Figure 5. The Precision-Recall (PR) curve and F-measure curve on all testing datasets.

Table 1. Quantitative results for camouflage detection.

Method	CHAMELEON				CAMO				COD10K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$Mae \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$Mae \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$Mae \downarrow$
Task-Generic Prompt Setting												
CLIP-Surgery+SAM [22]	0.689	0.741	0.606	0.147	0.612	0.692	0.520	0.189	0.629	0.698	0.488	0.187
GroundingSAM [37]	0.744	0.776	0.662	0.122	0.707	0.753	0.656	0.157	0.764	0.813	0.670	0.085
GenSAM [10]	0.764	0.807	0.680	0.090	0.719	0.775	0.659	0.113	0.775	0.838	0.681	0.067
ProMaC [11]	0.833	0.790	0.899	0.044	0.767	0.846	0.725	0.090	0.805	0.876	0.716	0.042
Supervised Setting												
SINet [3]	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
JCOD [20]	0.870	0.924	-	0.039	0.792	0.839	-	0.082	0.800	0.872	-	0.041
FBNet [23]	0.888	0.939	0.828	0.032	0.783	0.839	0.702	0.081	0.809	0.889	0.684	0.035
RankNet [31]	0.846	0.913	0.767	0.045	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045
SAM-Adapter [2]	0.896	0.919	0.824	0.033	0.847	0.873	0.765	0.070	0.883	**0.918**	0.801	0.025
PFNet [32]	0.882	0.942	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
FSPNet [14]	-	-	-	-	0.856	0.899	0.799	0.050	0.851	0.895	0.735	0.026
MM-SAM	**0.923**	**0.946**	**0.853**	**0.027**	**0.863**	**0.901**	**0.782**	**0.059**	**0.896**	0.907	**0.808**	**0.023**

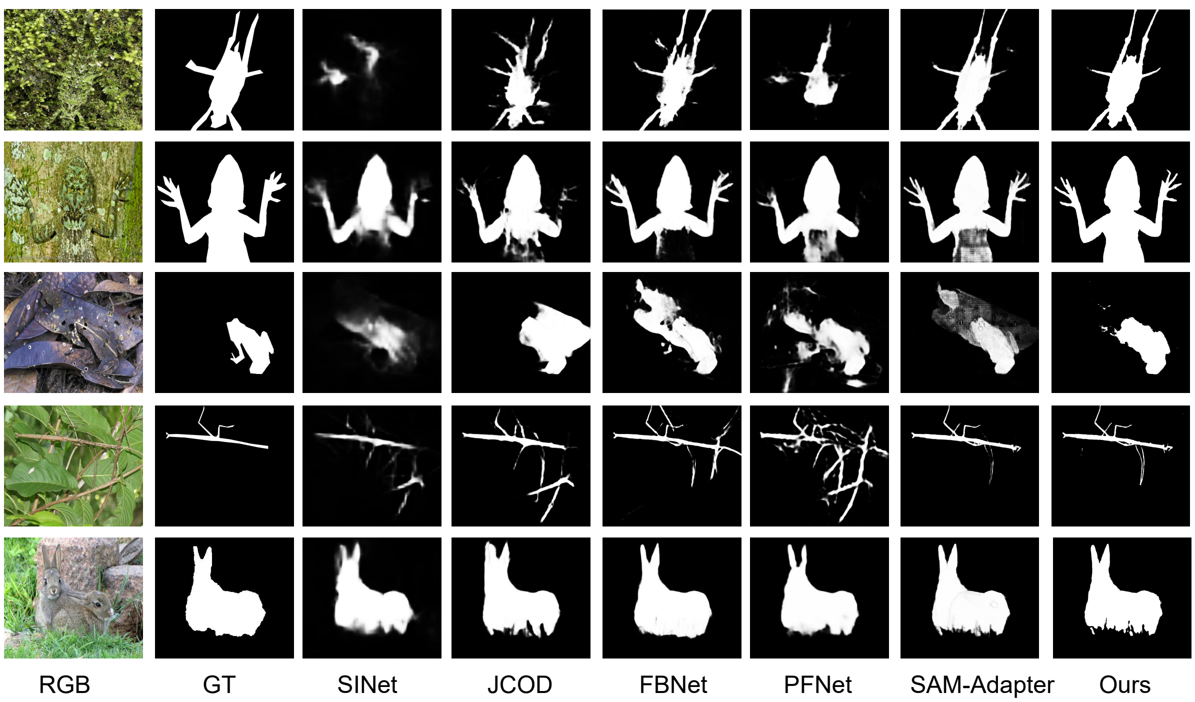

Figure 6. Comparisons with other state-of-the-art methods.

4.2. Implementation Details

We use the SAM model from Chen et al. [2] as our baseline. For all experiments, we utilize the AdamW optimizer. The initial learning rate is set to 2e-4, and a cosine decay schedule is applied. All training and testing images are resized to 1024×1024 resolution, and the model is trained with a batch size of 1. The camouflaged object segmentation task is trained for 20 epochs using PyTorch on a single NVIDIA RTX 3090 GPU with 24GB of memory.

4.3. Comparisons with State-of-the-art Methods

Comparisons with current state-of-the-art methods, namely SINet [3], RankNet [31], JCOD [20], PFNet [32], FBNet [23], SAM [17], SCOD [7], SAM-Adapter [2] and GenSAM [10] are presented in Table 1. To provide a comprehensive assessment of each method's efficacy, this comparative study evaluates multiple algorithms for camouflage detection across three benchmark datasets: CHAMELEON, CAMO, and COD10K.

The proposed method consistently outperforms existing

Table 2. Ablation studies on SAM.

Method's Variant					CHAMELEON				CAMO				COD10K			
SAM	Language	Vision	Dense	Adapter	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$
✓					0.836	0.872	0.666	0.065	0.790	0.839	0.619	0.107	0.807	0.801	0.606	0.054
✓	✓				0.845	0.881	0.706	0.054	0.784	0.848	0.630	0.101	0.818	0.833	0.651	0.046
✓	✓	✓			0.854	0.898	0.725	0.048	0.811	0.863	0.676	0.085	0.821	0.818	0.642	0.042
✓	✓	✓	✓		0.872	0.915	0.769	0.043	0.822	0.882	0.712	0.075	0.839	0.871	0.706	0.034
✓	✓	✓	✓	✓	0.923	0.946	0.853	0.027	0.863	0.901	0.782	0.059	0.896	0.907	0.808	0.023

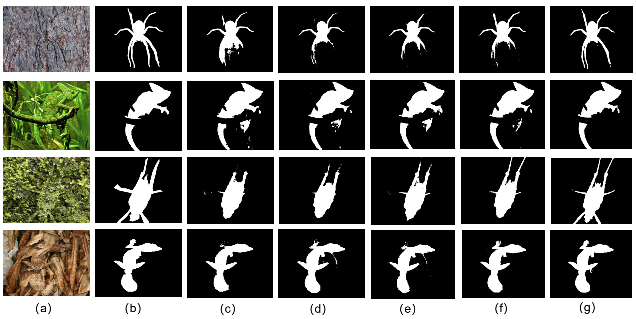

Figure 7. Qualitative results on ablation studies. Each column refers to a proposed method and they are gradually added to the SAM baseline from left to right. (a) RGB images. (b) Ground truth. (c) SAM baseline. (d) Language prompts. (e) Vision prompts. (f) Dense embedding. (g) Adapter.

algorithms across all datasets and metrics, demonstrating its robustness and accuracy in camouflage detection. Significant improvements in S_α, E_ϕ, F_β^w, and MAE highlight the ability of the method to maintain structural integrity, alignment, and precision while minimizing errors. Figure 5 presents the PR curve and F-measure curve on all testing sets, clearly indicating that our method (red line) achieves superior performance.

It is important to note that the superior segmentation capabilities of the proposed method are not solely dependent on SAM in COD. As shown in Table 1, SAM struggles with COD due to the lack of prompt guidance. This issue can be mitigated by equipping SAM with either a vision or language prompt, such as those used in SAM-Adapter or GenSAM. These prompts provide additional semantic guidance, enhancing performance relative to SAM alone. However, even with these improvements, they still fall short of the performance achieved by our proposed method.

We further visualize testing examples in diverse scenarios in Figure 6. The top four rows show animals well camouflaged and embedded in the background. Our approach successfully segments challenging samples, whereas other methods still struggle in certain situations. The final row illustrates a simpler case, where the rabbits stand out against the background. In that case, all methods obtain accurate masks of the rabbits. Notably, while the ground truth overlooks the overlapping grass and rabbits, JCOD, SAM-Adapter, and

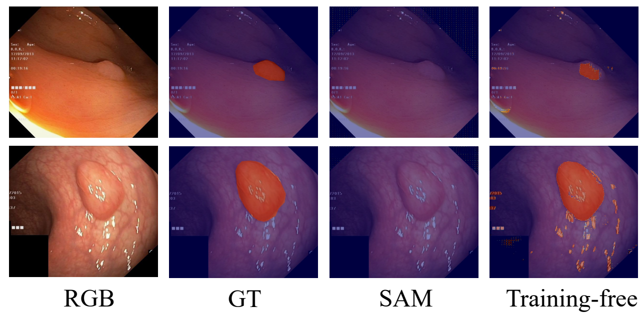

Figure 8. Visualization of training-free mode for polyp segmentation

our method still manage to segment the overlapping grass.

4.4. Ablation Study

We conducted an ablation study to validate our method's effectiveness. Results, detailed in Table 2, show that adding each component—language prompts, vision module, dense connections, and adapter—progressively improve model performance for COD. The baseline (first row) represents SAM applied directly, without prompts.

Initially, we examined vision-language prompts applied to SAM, which significantly improved segmentation prediction, particularly enhancing E_ϕ and F_β^w scores. Adding a vision module further improved metrics, indicating that multi-modal information benefits the model. Dense connections notably enhanced F_β^w and reduced MAE, indicating greater prediction precision. Finally, the adapter mechanism, added to the full model (SAM + Language + Vision + Dense), yielded the highest scores, emphasizing its importance in integrating features from different modules.

Figure 7 illustrates qualitative ablation results, showing segmented masks under various conditions. While SAM alone roughly segments objects, it produces many false positives and false negatives. Adding language prompts reduces these errors by providing semantic guidance. Further improvements come from adding vision prompts, though some issues remain. The final configuration (SAM + Language + Vision + Dense + Adapter) maximally utilizes multi-scale and multi-modal information, resulting in clearer object segmentation and elimination of false negatives.

Table 3. Results for medical image segmentation under training-free setting. Training-free mode means that our model directly incorporates embeddings from foundation models without the proposed adapter for fine-tuning. Zero-shot indicates that our entire model including the proposed adapter is fine-tuned on COD dataset.

Method	Venue	CVC-ColonDB				Kvasir			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$Mae \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$Mae \downarrow$
GPT4V+SAM [17]	Arxiv23	0.242	0.246	0.051	0.578	0.253	0.236	0.128	0.614
LLaVA1.5+SAM [17, 24]	NeruIPS23	0.357	0.355	0.194	0.491	0.403	0.400	0.293	0.479
X-Decoder [51]	CVPR23	0.331	0.327	0.095	0.462	0.384	0.371	0.202	0.449
SEEM [52]	NeruIPS23	0.284	0.280	0.085	0.570	0.367	0.337	0.215	0.520
GroundingSAM [17, 27]	ICCV23	0.206	0.195	0.071	0.711	0.468	0.521	0.353	0.387
GenSAM [10]	AAAI24	0.379	0.494	0.059	0.244	0.487	0.619	0.210	0.172
ProMaC [11]	NeruIPS24	0.530	**0.583**	**0.243**	**0.176**	0.573	0.726	0.394	0.166
MM-SAM(Training-Free)	Ours	0.451	0.527	0.107	0.317	0.475	0.637	0.159	0.235
MM-SAM(Zero-Shot)	Ours	**0.565**	0.520	0.220	0.185	**0.740**	**0.756**	**0.535**	**0.134**

Table 4. Zero-shot on saliency and blur detection.

Method	DUTS		CUHK	
	$E_\phi \uparrow$	$Mae \downarrow$	$mIoU \uparrow$	$F_\beta \uparrow$
SAMed [48]	0.764	0.104	0.554	0.717
SEEM [52]	0.599	0.326	0.608	0.675
Painter [45]	0.811	0.113	0.186	0.273
PerSAM [49]	0.641	0.257	0.558	0.716
AlignSAM [13]	0.782	0.082	0.685	**0.769**
MM-SAM	**0.877**	**0.058**	**0.698**	0.746

4.5. Training-free Medical Image Segmentation

We compare our model with other state-of-the-art (SOTA) training-free methods on medical image segmentation. The proposed method directly incorporates embeddings from foundation models without using the proposed adapter for fine-tuning on any dataset. Table 3 shows that without any fine-tuning, our method achieves competitive performance with other state-of-the-art methods. Figure 8 illustrates that foundation models can directly provide multi-modal prompts to SAM for medical image segmentation. Additionally, using our model fine-tuned on COD dataset we achieve impressive performance on polyp segmentation, demonstrating the generalizability of our paradigm.

4.6. Zero-shot Segmentation on Other Tasks

We compare our proposed method with other state-of-the-art approaches based on foundation models for saliency detection and blur detection tasks. Using the same evaluation metrics as in AlignSAM [13], we directly apply our method on DUTS [43] and CUHK [39] datasets without further fine-tuning. Table 4 illustrates the superior performance of our method compared to other publicly available SOTA results. These results demonstrate that our framework maintains robust generalization ability on diverse tasks.

Table 5. Results on different text encoders.

Method	CAMO				COD10K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta \uparrow$	$Mae \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta \uparrow$	$Mae \downarrow$
SAM	0.790	0.839	0.619	0.107	0.807	0.801	0.606	0.054
SAM+CLIP	0.779	0.848	0.629	0.101	0.812	0.816	0.631	0.047
SAM+Bert	0.782	0.854	0.634	0.099	0.812	0.823	0.631	0.045
SAM+Llama	0.774	0.852	0.620	0.102	0.806	0.826	0.626	0.047
SAM+Mamba	0.784	0.848	0.630	0.101	0.818	0.833	0.651	0.046

4.7. Text Encoder Analysis

In Table 5, we conduct experiments on different text encoders to demonstrate the improvement of language prompts. The SAM baseline includes only the image encoder and mask decoder. Compared to CLIP, which is used in the original SAM for text prompts, Mamba show more robust performance on COD fine-tuning. However, various LLMs provide similar improvements, likely due to the simple captions generated by BLIP.

5. Conclusion

In this paper, we propose a versatile framework for camouflaged object detection (COD) that leverages off-the-shelf foundation models. To the best of our knowledge, this is the first framework to employ both vision-language models (VLMs) and large language models (LLMs) for the COD task. Our method abandons the commonly used prompt encoder in SAM and incorporates accessible and stable prompts, namely vision and language prompts, directly into the mask decoder and the proposed multi-modal adapter for fine-tuning. We further investigate the architecture of SAM and mitigate the dense embedding issue due to the lack of mask prompts in practice. Extensive experiments demonstrate the superiority of our method over other state-of-the-art methods across three datasets.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[2] 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*, pages 3367–3375, 2023. 3, 4, 5, 6

[3] 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*, 2020. 2, 5, 6

[4] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[5] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3

[6] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[7] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 781–789, 2023. 6

[8] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[9] Bintao Hu, Hengyan Liu, Jianbo Du, Miguel López-Benítez, Celimuge Wu, Xiaoli Chu, and Dusit Niyato. Madqn-enhanced computation offloading and resource allocation for 6g low-altitude economy vehicular networks. *IEEE Transactions on Cognitive Communications and Networking*, 2025. 1

[10] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12511–12518, 2024. 3, 4, 6, 8

[11] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. *arXiv preprint arXiv:2408.15205*, 2024. 6, 8

[12] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024. 4

[13] Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3205–3215, 2024. 8

[14] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5557–5566, 2023. 6

[15] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 1

[16] 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. 5

[17] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3, 6, 8

[18] 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. 5

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 3

[20] 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. 6

[21] Bo Li, Haoke Xiao, and Lv Tang. Asam: Boosting segment anything model with adversarial tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3699–3710, 2024. 3

[22] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162, 2025. 6

[23] 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. 6

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 8

[25] Hengyan Liu, Wenzhang Zhang, Tianhong Dai, Longfei Yin, and Guangyu Ren. Multimodal frequeny spectrum fusion schema for rgb-t image semantic segmentation. In *2024 33rd*

International Conference on Computer Communications and Networks (ICCCN), pages 1–6. IEEE, 2024. 1

[26] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 8

[28] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023. 4

[29] Yang Liu, Cong qing Wang, and Yong jun Zhou. Camouflaged people detection based on a semi-supervised search identification network. *Defence Technology*, 21:176–183, 2023. 3

[30] 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*, 2021. 3

[31] 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. 6

[32] 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. 6

[33] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 3

[34] Thomas W Pike. Quantifying camouflage and conspicuousness using visual salience. *Methods in Ecology and Evolution*, 9(8):1883–1895, 2018. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[36] Guangyu Ren, Yinxiao Yu, Hengyan Liu, and Tania Stathaki. Dynamic knowledge distillation with noise elimination for rgb-d salient object detection. *Sensors*, 22(16):6188, 2022. 1

[37] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 6

[38] 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. 2

[39] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014. 8

[40] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 5

[41] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 1025–1031. International Joint Conferences on Artificial Intelligence, 2021. 2

[42] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 5

[43] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 8

[44] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR 2011*, pages 3401–3408. IEEE, 2011. 1

[45] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 8

[46] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[47] Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. A survey of camouflaged object detection and beyond. *arXiv preprint arXiv:2408.14562*, 2024. 2

[48] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 8

[49] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 8

[50] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[51] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu

Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 8

[52] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 8