Improving SAM for Camouflaged Object Detection via Dual Stream Adapters

Jiaming Liu, Linghe Kong*, Guihai Chen
School of Computer Science, Shanghai Jiao Tong University
Shanghai, China
{jmliu99, linghe.kong}@sjtu.edu.cn, gchen@cs.sjtu.edu.cn

Abstract

*Segment anything model (SAM) has shown impressive general-purpose segmentation performance on natural images, but its performance on camouflaged object detection (COD) is unsatisfactory. In this paper, we propose SAM-DSA that performs COD for RGB-D inputs via **D**ual **S**tream **A**dapters. While keeping the SAM architecture intact, dual stream adapters are expanded on the image encoder to learn potential complementary information from RGB images and depth images, and fine-tune the mask decoder and its depth-aware replica to perform dual-stream mask prediction. In practice, the dual stream adapters are embedded into the attention block of the image encoder in a parallel manner to facilitate the refinement and correction of the two types of image embeddings. To mitigate channel discrepancies arising from dual stream embeddings that do not directly interact with each other, we augment the association of dual stream embeddings using bidirectional knowledge distillation including a model distiller and a modal distiller. In addition, to predict the masks for RGB and depth attention maps, we integrate the two types of image embeddings which are jointly learned with the prompt embeddings to update the initial prompt, and then feed them into the mask decoders to synchronize the consistency of image embeddings and prompt embeddings. Experimental results on four COD benchmarks show that our SAM-DSA achieves excellent detection performance gains over SAM and achieves state-of-the-art results with a given fine-tuning paradigm.*

1. Introduction

SAM [26] is a visual foundation model (VFM) for promptable image segmentation with strong zero-shot capability and wide generalization. A fact is that SAM's ability to discriminate 2D images depends on the distribution coverage from the training data. Thus, a natural question arises: *can SAM be directly extended to solve the task of segmenting camouflaged images for special scenarios?* Recent research [8, 22] shows that this is not feasible, as camouflaged objects have properties such as low contrast and texture blending to create visual confusion with the background [51].

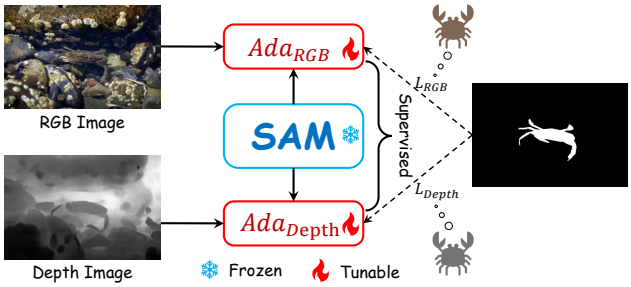

Figure 1. Illustration of the proposed SAM-DSA. The two learnable adapters based on the SAM model are extended to act on paired input streams of RGB images and depth images, respectively. Subject to co-supervision, Ada_{RGB} focuses on perceiving pixel semantics to segment objects and Ada_{Depth} focuses on structural transformation to separate objects from the background.

Based on the above observations, we further argue that fine-tuning is an important step in applying SAM to camouflaged images. We first analyze two key advantages associated with SAM: (1) The training dataset of SAM inherently contains a large number of backgrounds of camouflaged images, and the fixed parameters are sufficient to parse the overall visual environment. (2) After fine-tuning without changing the original weights, SAM still has strong generalization potential as a pre-trained large-scale VFM, which is crucial for effectively deploying models for COD in images with large inter-domain differences.

We then aim to adapt SAM for specific applications by incorporating domain knowledge. Recent studies on parameter-efficient adaptation methods have explored two primary strategies: fine-tuning a small subset of pretrained parameters [44, 46, 66] and integrating lightweight adapter modules [3, 7, 59]. The positive results have been achieved with only a small number of parameter modifications or additions, which motivates us to develop an efficient SAM adapter for the COD task. In contrast to previous research, our desired adapter needs to be applicable to both modalities, as RGB images and depth images can naturally com-

plement obscure regions and merge valid object clues.

It is worth investigating how to design adapters with both modalities in mind — *should we use fine-tuned networks with the same architecture or different ones?* Inspired by the teacher and student networks in DSAM [65], we believe that different modal inputs should have features extracted in different ways in order to perform complementary functions. To this end, we design two adapters to handle RGB-D inputs. Considering the similar semantic and distinct depth differences between objects and backgrounds in the COD, we aim to capture the high-frequency semantic features of the object from the RGB perspective to determine the potential salient regions in the whole image, and then capture the structural geometric features of the object from the depth perspective to highlight the salient object from the background, as shown in Figure 1. Empirically, this semantic-geometric fusion approach proves effective for detecting objects from various camouflaged images.

Specifically, we introduce the RGB adapter and the depth adapter as a combination of linear layer networks with frequency separation mechanisms [53], using down-projection and up-projection linear layers embedded at both ends of the high-frequency feature extractor. The difference is that two frequency separation operations are performed under different specific filtering operators. Depending on the depth of the ViT network [10, 32] and the number of attention blocks in the SAM, the extracted features of RGB images and depth images can be updated multiple times, resulting in the RGB embedding and depth embedding that contains rich details of camouflaged objects. In addition, to further update the initial box prompt, we synchronously use the RGB embedding and the depth embedding to blend with the prompt embedding, and perform channel-wise convolution operations, which makes the dense prompt embeddings close to the representations of the image embeddings on the feature channel. These encoder modules are jointly fine-tuned to transform generic visual perception models into specific models suitable for the COD task.

Moreover, since camouflaged objects have small size, irregular shape, and low contrast, we choose to continue activating two mask decoders to generate mask predictions with refined prompt embeddings and image embeddings. Considering that the source-free depth images may suffer from noise, we adopt inter-modal knowledge distillation, regarding the RGB embeddings and depth embeddings as the teacher and student, respectively. Our intuition is that since RGB embeddings are dense semantic features and depth embeddings are separated structural features, the former are easier to distinguish visually, and thus complementary information between the two modalities is achieved by predicting soft labels of RGB embeddings that can guide the latter. In addition, we argue that the ViT-based PVT network [57, 58] with pyramids is stronger in feature extraction. To take advantage of this advantage, we first map PVT embeddings to the same space as the image embeddings using a bias correction module, and then regard the PVT embeddings and RGB embeddings as teacher and student respectively to perform inter-model knowledge distillation, further enhancing the representation of the RGB embeddings. In brief, our contributions are as follows:

- We propose SAM-DSA to improve SAM into the COD domain. To our knowledge, SAM-DSA is the first to apply dual stream adapters to enhance the performance of SAM on RGB-D-based camouflaged inputs.
- The bidirectional distillation and mixed embedding modules are proposed. The former enriches the representations of image embeddings through modal and model distillations, and the latter refines dense prompt embeddings by fusing hybrid image embeddings.
- On the four COD benchmarks, SAM-DSA is thoroughly compared with existing methods. The results show that our SAM-DSA outperforms state-of-the-art visual foundation models and specific expert models.

2. Related Work

Camouflaged Object Detection (COD) is a challenging task aimed at recognizing objects that blend in with their surroundings, and its applications cover a wide range of fields such as medicine, agriculture, and art [7, 56]. Traditional COD research has relied on low-level features such as texture, brightness, and color, by which foreground and background are distinguished [15, 16, 31, 47, 49, 61]. However, with the advancement of deep neural networks, the development of COD has been significantly boosted. Recent studies have shown that impressive results have been achieved using methods such as mixed-scale semantics [42], iterative refinement [24], and human attention mechanisms [55]. Some methods based on the design of animal hunting models, such as SINetV2 [13] and SLSR [37], as well as methods combining probabilistic representation models with Transformer [63], have also emerged. Some methods [29, 54, 62] extended the RGB COD task into an RGB-D COD task, attempting to utilize depth information to aid in detection. Moreover, Bi *et al.* [2] and Liu *et al.* [34] introduced dynamic allocation mechanisms and multi-scale fusion methods to suppress the interference of inaccurate depth images on COD. Different from auxiliary depth input [60], in our study, we learn data representation from depth images through VFM and interact RGB and depth information in an orderly manner to compensate for the unstable representations of objects in their respective modalities.

Segment Anything Model (SAM) for COD is not consistently effective, mainly due to the large domain differences between the training data of SAM and domain-specific data, resulting in their lack of sufficient domain-specific semantic guidance. This flaw also occurs in the upgraded SAM 2

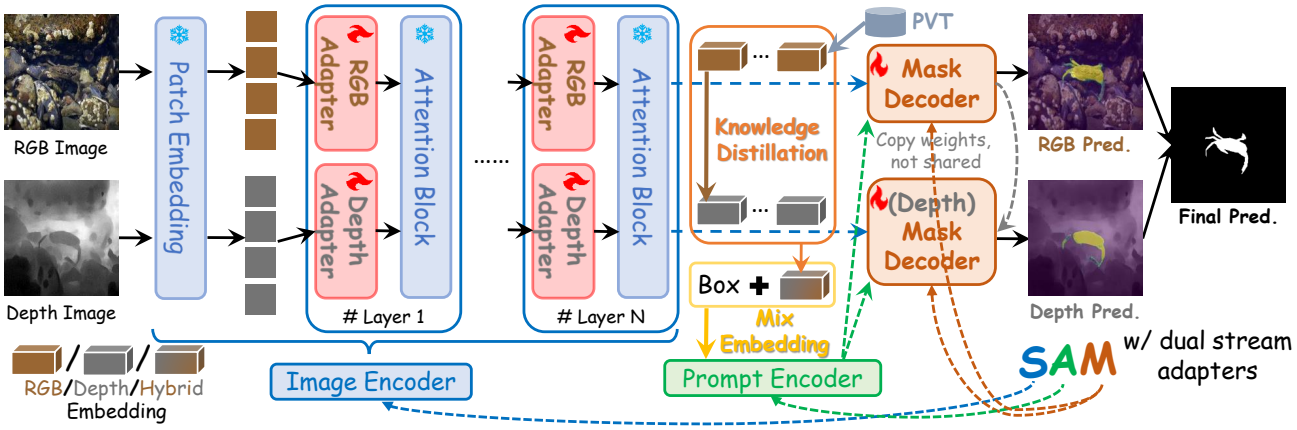

Figure 2. Overall pipeline of our SAM-DSA. The dual stream images are fed into SAM in parallel to extract image features that are fine-tuned by the respective adapters. The knowledge distiller is used to address the differences caused by the dual-stream features being decoded without direct interaction. The initialized box prompt is mixed with the image features to generate a more refined dense prompt embedding. Finally, the prediction results of the two classes of feature maps are weighted summed to obtain the final detection results.

[48] with a memory, making it difficult to generalize historical information from COD videos. For this reason, medical imaging (MI) researchers have developed models such as MedSAM-(2) [38, 72], Med-SA [59], and MA-SAM [3] for tumor and myocardial segmentation. COD researchers have proposed models such as SAM(2)-Adapter [6, 7], TSP-SAM [20], and DSAM [65] for camouflage and concealed scene recognition. SAM-COD [4, 5] learns prior knowledge from a expert model trained on COD data via a semantic matcher and proposes an adaptive knowledge distillation strategy to ensure reliable representations, integrates three types of labels for weakly-supervised COD. Anyway, introducing domain-specific knowledge into SAM is an effective measure to improve performance and robustness on specific downstream tasks.

Attributed to its origin, it is a feasible approach to fine-tune large pre-trained models using scalable adapters. Recently, ViT-Adapter [9] utilized adapters to enable ordinary ViTs to perform a variety of downstream tasks. Liu *et al.* [33] introduced an explicit visual prompting (EVP) technique that merged explicit visual cues into adapters. Unlike previous usage scenarios and approaches, we delve deeper into this idea by introducing dual stream adapters for RGB-D inputs to fully utilize the complementary nature of the two types of images when fine-tuning the SAM in order to facilitate the stable performance of SAM on the COD task.

3. Method

Figure 2 shows the overall pipeline of the proposed SAM-DSA. The dual stream adapters are embedded in the frozen SAM model to learn specific prior knowledge. Sequentially, we first briefly outline the deconstruction of the original SAM, and then introduce the technical implementation details of the modifications to the image encoder, prompt encoder, and mask decoder respectively.

3.1. SAM Overview

SAM [26] consists of an image encoder, a prompt encoder, and a mask decoder. The image encoder converts raw images into patched image embeddings using the Vision Transformer (ViT) network [10]. The prompt encoder encodes prompts (points, boxes, masks, etc.) into prompt embeddings. The mask decoder includes prompt self-attention blocks and bidirectional cross-attention blocks (prompt-to-image and reverse attentions). After applying the attention blocks, the feature maps are up-sampled and converted to segmentation masks through a fully connected layer.

In this work, we use the weights of the pre-trained SAM to initialize the weights of the three feature processors of our method, and fine-tune the mask decoder during training. According to DSAM [65], we input perturbed ground-truth box prompts to the original mask decoder of the SAM.

3.2. Dual Stram Adapter

For the input RGB image and the paired source-free depth image [62], we propose dual stream adapters for extracting high-frequency features, as shown in Figure 3. While keeping the both adapters simple and efficient, we further introduce different high-frequency wavelet filters for different adapters. Specifically, we choose to use an adapter consisting of two linear layers and activation functions, and the final adapter output $\bar{\boldsymbol{X}}_{Ada}$ is a combination of the input embedding $\boldsymbol{X}$ and transformed embedding by the adapter.

$$\bar{\boldsymbol{X}}_{Ada} = \boldsymbol{X} + \boldsymbol{L}_{up}(\boldsymbol{\sigma}(\boldsymbol{L}_{down}(\boldsymbol{X}))), \tag{1}$$

where $\boldsymbol{L}_{down}(*)$ is a downward projected linear layer, $\boldsymbol{\sigma}()$ is a nonlinear activation function ReLU [41], and $\boldsymbol{L}_{up}()$ is an upward projected linear layer. Taking the RGB or depth image embedding $\boldsymbol{X}$ as input, the computation within the adapter module obtains the image embedding with high-frequency details through the residual structure.

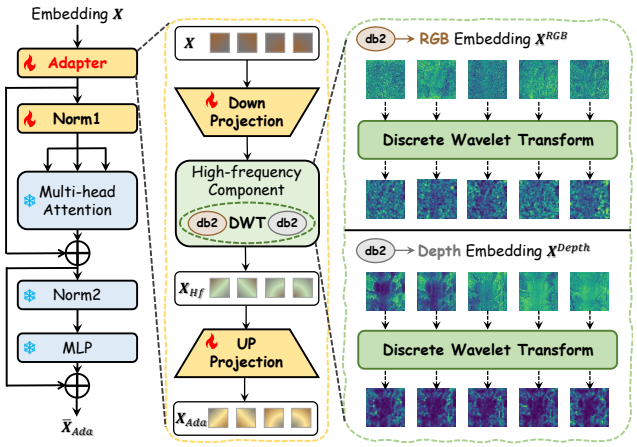

Figure 3. Illustration of the proposed adapter. The patched embedding of the image is used as input and our dual-stream adapter extracts the image embedding with high-frequency details by feature projection transform and discrete wavelet transform. Note that the residual structure is omitted for ease of display.

DWT. In contrast to the previous adapters that only perform downward-upward projection for single RGB inputs [7], our adapter is equipped with the Discrete Wavelet Transform (DWT) for extracting high-frequency features. DWT [1] is able to efficiently separate different frequency components of an image through multi-scale analysis, and process low-frequency and high-frequency information separately to capture subtle high-frequency features, which is particularly applicable to the changes in details of camouflaged objects, such as changes in texture and contouring. In practice, we synchronously set the same wavelet `db2` for RGB images and depth images, and merge the high-frequency subbands in the horizontal, vertical, and diagonal directions. In the case of RGB embedding,

$$\boldsymbol{Lf}_{RGB}, \boldsymbol{Hf}_{RGB} = DWT_{\texttt{haar}}(\boldsymbol{X}_{RGB}), \quad (2)$$

$$\boldsymbol{LH}_{RGB}, \boldsymbol{HL}_{RGB}, \boldsymbol{HH}_{RGB} = \boldsymbol{Hf}_{RGB}[0|1|2], \quad (3)$$

$$\boldsymbol{X}_{Hf}^{RGB} = \sqrt{\boldsymbol{LH}_{RGB}^2 + \boldsymbol{HL}_{RGB}^2 + \boldsymbol{HH}_{RGB}^2}, \quad (4)$$

where $\boldsymbol{LH}$, $\boldsymbol{HL}$, and $\boldsymbol{HH}$ are the representations of the high frequency components in the three directions, respectively, and the L2 paradigm is used to synthesize the high-frequency information. By such calculation, a comprehensive high-frequency feature map can be obtained, covering the detailed variations of the image in different directions.

Normalization. In addition to the dual stream adapters, we also employ an activation strategy for the normalization layer `Norm1` close to the tunable adapter, which allows our SAM-DSA quickly adapt to the new data distribution. At the same time, in order to prevent the pre-trained features from changing dramatically during fine-tuning, `Norm2` behind the multi-head attention layer remains frozen.

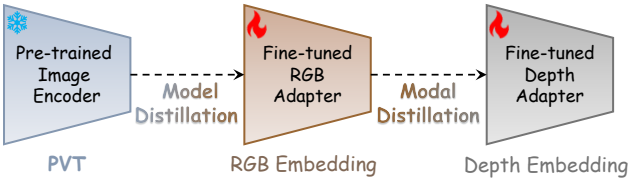

Figure 4. Illustration of the proposed bidirectional knowledge distillation. The model distillation from the pre-trained image encoder to the fine-tuned RGB adapter, and modal distillation from the RGB adapter to the depth adapter are executed sequentially.

3.3. Bidirectional Knowledge Distillation

As illustrated in Figure 4, we design two knowledge distillation processes: one is the model distillation between the pre-trained encoder and the fine-tuned adapter, and the other is the modal distillation between the RGB adapter and the depth adapter. Based on the proposed bidirectional knowledge distillation (KD), the fine-tuned model not only inherits more advanced representations from the pre-trained model, but also overcomes inter-modal differences and suppresses possible noise from depth images.

Model distillation. Specifically, we utilize a pre-trained specialized model (*i.e.* PVTv2 [58]) as a teacher and pass its learned knowledge to the foundation fine-tuning model (*i.e.* our adapter) as a student via knowledge distillation. The pre-trained model typically have stronger feature extraction capabilities and thus can provide richer learning signals to the fine-tuned model. In this way, the fine-tuning model is able to inherit the knowledge from the pre-trained model and further optimize for a specific task.

$$\boldsymbol{X}_{PVT}^{RGB} = BC(PVT(\boldsymbol{I})), \quad (5)$$

$$\boldsymbol{L}_{KD-model} = \boldsymbol{KL}(\boldsymbol{X}_{PVT}^{RGB}, \bar{\boldsymbol{X}}_{Ada}^{RGB}), \quad (6)$$

where $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ is an RGB image, PVT is the frozen pre-trained model and BC is the trainable bias-calibrated network $BC(\boldsymbol{X}) = \alpha * \boldsymbol{X} + \beta$ that maps $\boldsymbol{X}$ to the dimensions of $\bar{\boldsymbol{X}}_{Ada}^{RGB}$. $\boldsymbol{KL}(Teacher, Student)$ is the Kullback-Leibler divergence [27], which is used to measure the difference between the teacher model and the student model.

Modal distillation. While $\boldsymbol{X}_{Hf}^{RGB}$ guided by a specialized model is rich in semantic information, $\boldsymbol{X}_{Hf}^{Depth}$ consists of deep structural information that usually contains certain noise and does not always highlight the camouflaged object [36]. Therefore, we continue to perform modal distillation by using RGB embeddings as teachers to instruct depth embeddings as students to learn the semantic knowledge.

$$\boldsymbol{L}_{KD-modal} = \boldsymbol{KL}(\bar{\boldsymbol{X}}_{Ada}^{RGB}, \bar{\boldsymbol{X}}_{Ada}^{Depth}). \quad (7)$$

As a result, we utilize the KD metrics $\boldsymbol{L}_{KD-model}$ and $\boldsymbol{L}_{KD-modal}$ of the bidirectional distillation as loss items during the training phase. This strategy can enhance the representation of the modified visual foundation model to improve robustness and accuracy in the COD task.

3.4. Mixed Prompt Embedding

During image encoding, SAM causes loss of detail information due to 16×16 downsampling. To solve this issue, in addition to using dense PVT embedding to guide image embedding, we additionally use dual stream embeddings to compensate for the dense prompt embedding.

Specifically, first the box prompt embedding is processed by a deepwise convolutional layer; then the expert and adapter embeddings are concatenated to generate the new hybrid embedding; finally the dense prompt embedding is updated by another deepwise convolutional layer. This process can effectively fuse the dual stream image and prompt embeddings to enhance the representation capability of new dense prompt embedding. Mathematically,

$$\boldsymbol{X}_{Hybrid} = cat[\boldsymbol{X}_{PVT}^{RGB}, \bar{\boldsymbol{X}}_{Ada}^{RGB} + \bar{\boldsymbol{X}}_{Ada}^{Depth}], \tag{8}$$

$$\boldsymbol{X}_{box}^{dense} = DWConv(cat[\boldsymbol{X}_{box}^{dense}, \boldsymbol{X}_{Hybrid}]), \tag{9}$$

where $cat[*,*]$ represents the embedding concatenation operation. Here, $DWConv$ is used for high-frequency filtering in the diagonal direction, as we observe that it applies to the summation of dual stream embeddings.

3.5. Training and Inference

In the training phase, we optimize our model using two losses. On the one hand, DiceCELoss is employed as a loss function for dual stream predictions, which computes the weighted sum of the dice loss and the cross-entropy loss. On the other hand, a bidirectional distillation loss is employed to utilize the in-channel knowledge to achieve soft alignment between the model and the modal. In summary, the loss function of our SAM-DSA is as follows.

$$\boldsymbol{L}_{DiceCE} = DiceCE(\boldsymbol{Y}^{RGB}|\boldsymbol{Y}^{Depth}, \boldsymbol{Y}^{GT}) \tag{10}$$

$$\boldsymbol{L} = \lambda \boldsymbol{L}_{DiceCE} + (1-\lambda)\boldsymbol{L}_{KD}, \tag{11}$$

where $\boldsymbol{Y}^{GT}$ denotes the binarized ground truth and is used twice to supervises dual stream predictions $\boldsymbol{Y}^{RGB}$ and $\boldsymbol{Y}^{Depth}$ respectively. Moreover, λ is a hyperparameter used to balance the prediction loss and distillation loss of SAM-DSA, and the default value is 0.9.

In order to obtain the final detection result during inference, we weight the RGB prediction $\boldsymbol{Y}^{RGB}$ and depth prediction $\boldsymbol{Y}^{Depth}$. In the end, pixels with a mean value greater than 0.5 are considered to belong to the foreground pixel, otherwise they are considered to be background pixels.

$$\boldsymbol{Y} = (0.5 * \boldsymbol{Y}^{RGB} + 0.5 * \boldsymbol{Y}^{Depth}) > 0.5. \tag{12}$$

Considering the real application where depth images may not be available or of poor quality, we can use only the RGB stream loss and prediction as a result of our SAM-DSA. The construction and results of this setting will be discussed in the subsequent experimental analysis.

4. Experiments

4.1. Experimental Settings

Datasets. The experiments are conducted on four datasets, namely CHAMELEON [50], CAMO [28], COD10K [13] and NC4K [37]. CHAMELEON contains 76 images collected from the Internet for testing. CAMO contains 1250 images randomly divided into 1000 training and 250 test sets. COD10K contains 5066 images, of which 3040 are training and 2026 are test sets. NC4K contains 4121 images, all of which are used as test sets. Following protocol from [13], our study uses a dataset consisting of the training set of COD10K and CAMO, which contains 3040 images and 1000 images, respectively, and the rest of the images of COD10K and CAMO and the entire CHAMELEON and NC4K images are used as the test dataset.

Evaluation metrics. The six evaluation metrics in the COD field are used, including the structural metric S_α [11], the adaptive/maximum E-measure $\alpha E/E^x$ [12], the weighted/maximum F-measure F^ω/F^x [39], and average absolute error M. Where, S_α measures the structural similarity between the predicted results and the actual segmented region. E-measure considers the structure and texture difference of two images. F-measure is the harmonic mean of precision and recall. M is the average absolute difference between the predicted value and the true value.

Implementation details. We implement our SAM-DSA in Pytorch [45]. The pre-trained models consist of the specialized weight PVTv2 [58] and VFM weight SAM-B [26]. We resize the inputs of all images to 1024x1024 by bilinear interpolation, train on a NVIDIA 3090 GPU with the batch size of 1, and use the AdamW optimizer [35] with an initialized learning rate of 1×10^{-4} for 100 epochs.

4.2. Comparison Results

We compare our SAM-DSA with 20 RGB stream-based, and 8 RGB-D stream-based state-of-the-art (SOTA) COD methods which are converted from the salient object detection (SOD) methods CDINet [68], DCF [23], CMINet [69], SPNet [71], DCMF [54], SPSN [29], PopNet [62]. In particular, we focus on the performance of VFM methods with the SAM weight, including SAM [26], SAM-Adapter [7], MedSAM [38] and DSAM [65].

Among them, although MedSAM [38] was initially applied in the field of medical image processing, it belongs to the enhanced SAM, and its application in COD also significantly improves the performance. In addition, we abandon the SAM-Adapter [7] using SAM-H due to the conditions. For fairness, prediction results are provided directly from their papers or generated by their already trained models.

Quantitative evaluation. Table 1 shows that the proposed SAM-DSA outperforms other SOTA methods on most of the evaluation metrics in both settings. On the one hand,

Table 1. Quantitative results of different COD methods on four datasets, *i.e.*, CAMO [28], CHAMELEON [50], COD10K [13] and NC4K [37]. Red/blue/orange indicates the 1st/2nd/3rd best result for the current setting. † represents the SAM-based methods.

Method	CAMO (250)						CHAMELEON (76)						COD10K (2026)						NC4K (4121)					
	$M\downarrow$	$F^x\uparrow$	$F^\omega\uparrow$	$S_m\uparrow$	$E^x\uparrow$	$\alpha E\uparrow$	$M\downarrow$	$F^x\uparrow$	$F^\omega\uparrow$	$S_m\uparrow$	$E^x\uparrow$	$\alpha E\uparrow$	$M\downarrow$	$F^x\uparrow$	$F^\omega\uparrow$	$S_m\uparrow$	$E^x\uparrow$	$\alpha E\uparrow$	$M\downarrow$	$F^x\uparrow$	$F^\omega\uparrow$	$S_m\uparrow$	$E^x\uparrow$	$\alpha E\uparrow$
RGB-based COD Methods																								
SINet [14]	.099	.762	.606	.751	.790	.825	.044	.845	.740	.868	.908	.938	.051	.708	.551	.771	.832	.867	.058	.804	.723	.808	.873	.883
SLSR [37]	.080	.791	.696	.787	.843	.855	.046	-	.794	.842	-	.896	.037	.756	.673	.804	.854	.882	.048	.836	.766	.839	.898	.902
MGL-R [67]	.088	.791	.719	.775	.820	.848	.031	.868	.828	.893	.932	.923	.035	.767	.686	.813	.874	.865	.053	.828	.762	.832	.876	.867
PFNet [40]	.085	.793	.695	.782	.845	.852	.033	.859	.810	.882	.927	.942	.040	.747	.660	.800	.880	.868	.053	.820	.745	.829	.891	.894
UJSC [30]	.072	.812	.728	.800	.861	.853	.030	-	.848	.894	-	.943	.035	.761	.684	.808	.886	.891	.047	.838	.771	.841	.900	.907
C2FNet [52]	.079	.802	.719	.796	.856	.864	.032	.871	.828	.888	.936	.932	.036	.764	.686	.813	.894	.886	.049	.831	.762	.838	.898	.901
UGTR [63]	.086	.800	.695	.783	.829	.859	.031	.862	.810	.887	.926	.921	.036	.769	.660	.816	.873	.850	.052	.831	.745	.839	.884	.889
SegMAR [24]	.080	.799	.742	.794	.857	.872	.032	.871	.835	.887	.935	.950	.039	.750	.724	.799	.876	.895	.050	.828	.781	.836	.893	.905
ZoomNet [42]	.074	.818	.752	.801	.858	.883	.033	.829	.845	.859	.915	.952	.034	.771	.729	.808	.872	.893	.045	.841	.784	.843	.893	.907
SINetv2 [13]	.070	-	.743	.820	.882	.875	.030	-	.816	.888	-	.942	.037	-	.680	.815	.887	.863	.048	-	.770	.847	.903	.898
DGNet [21]	.057	-	.769	.839	.915	.901	.029	-	.816	.890	-	.934	.033	-	.693	.822	.911	.877	.042	-	.784	.857	.922	.907
FSPNet [19]	.050	-	.799	.856	.928	.899	-	-	-	-	-	-	.026	-	.735	.851	.930	.895	.035	-	.816	.879	.937	.915
PRNet [18]	.050	-	.855	.872	-	.922	.020	-	.881	.914	-	.960	.022	-	.810	.874	-	.937	.031	-	.865	.891	-	.933
CamoFormer-R [64]	.066	-	.756	.817	-	.884	.024	-	.843	.900	-	.949	.029	-	.730	.838	-	.898	.024	-	.793	.857	-	.915
CamoFocus-R [25]	.071	-	.752	.812	-	.873	.027	-	.849	.898	-	.953	.033	-	.719	.825	-	.903	.043	-	.788	.847	-	.910
ZoomNeXt-R [43]	.069	-	.760	.822	-	.885	.020	-	.864	.912	-	.969	.026	-	.758	.855	-	.926	.038	-	.808	.869	-	.925
SAM† [26]	.132	-	.606	.684	-	.687	.081	-	.639	.727	-	.734	.049	-	.701	.783	-	.798	.078	-	.696	.767	-	.776
SAM-Adapter† [7]	.070	-	.765	.847	-	.873	.033	-	.824	.896	-	.919	.025	-	.801	.883	-	.918	-	-	-	-	-	-
MedSAM† [38]	.065	-	.779	.820	-	.904	.036	-	.813	.868	-	.936	.033	-	.751	.841	-	.917	.041	-	.821	.866	-	.929
COMPrompter† [70]	.054	-	.819	.853	-	.919	.030	-	.830	.884	-	.946	.026	-	.779	.861	-	.933	.036	-	.840	.880	-	.935
SAM-DSA†	.047	.876	.839	.866	.948	.946	.031	.881	.841	.888	.965	.956	.023	.866	.817	.881	.969	.942	.032	.892	.857	.889	.963	.954
RGB-D-based COD Methods																								
CDINet [68]	.100	.638	-	.732	.766	-	.036	.787	-	.879	.903	-	.044	.610	-	.778	.821	-	.067	.697	-	.793	.830	-
DCF [23]	.089	.724	-	.749	.834	-	.037	.821	-	.850	.923	-	.040	.685	-	.766	.864	-	.061	.765	-	.791	.878	-
CMINet [69]	.087	.798	-	.782	.827	-	.032	.881	-	.891	.930	-	.039	.768	-	.811	.868	-	.053	.832	-	.839	.888	-
SPNet [71]	.083	.807	-	.783	.831	-	.033	.872	-	.888	.930	-	.037	.776	-	.808	.869	-	.054	.828	-	.825	.874	-
DCMF [54]	.115	.737	-	.728	.757	-	.059	.807	-	.830	.853	-	.063	.679	-	.748	.776	-	.077	.782	-	.794	.820	-
SPSN [29]	.084	.782	-	.773	.829	-	.032	.866	-	.887	.932	-	.042	.727	-	.789	.854	-	.059	.803	-	.813	.867	-
PopNet [62]	.073	.821		.806	.869		.022	.893	-	.910	.962	-	.031	.789	-	.827	.897	-	.043	.852	-	.852	.908	-
DSAM† [65]	.061	.834	.794	.832	.920	.920	.042	.824	.784	.854	.935	.925	.033	.807	.758	.846	.931	.912	.040	.862	.828	.871	.940	.936
SAM-DSA†	.044	.884	.849	.875	.953	.952	.029	.887	.850	.893	.969	.961	.022	.873	.827	.887	.972	.948	.029	.898	.866	.896	.972	.959

when our SAM-DSA is oriented only to the RGB stream, it outperforms the existing methods substantially, bringing an overall improvement of 2%-6% on four datasets and breaking through the bottleneck of existing technologies. Moreover, we notice that compared to recent CNN-based or Transformer-based COD methods (e.g., ZoomNet [42], SegMaR [24], and CamoFormer [64]), although they employ multi-stage or multi resolution training and inference and other strategies to enhance model performance which add additional computational burden, our SAM-DSA constructed based on the VFM model (SAM-B) still outperforms them in most benchmark tests. On the other hand, when SAM-DSA is oriented to the RGB-D stream, *i.e.*, with the aid of source-free depth images, it is more capable of mining the detailed information of camouflaged objects.

Qualitative evaluation. We perform qualitative visualization results of SAM-DSA on eight scenes selected from four datasets to compare with three methods using SAM, as show in Figure 5. We analyze four aspects: object size, edge complexity, perceived missing parts, and multiple objects. It can be found that our SAM-DSA can effectively balance contours and details, and can achieve clear segmentation of edges for objects with high edge complexity. In addition, SAM-DSA can correctly segment highly camouflaged parts and extra objects (the first and sixth rows) that are easily overlooked, on the contrary, other methods may exhibit omission or over-segmentation. In short, SAM techniques themselves lack a comprehensive understanding of the foreground and background, making them misclassify certain regions or miss some objects when dealing with complex conditions. In contrast, our SAM-DSA corrects this information to a great extent with the dual stream adapters.

Visualization evaluation. To show more clearly the impact of our dual-stream adapters on the dual-stream inputs, we visualize their respective feature maps and detection results after binarization (corresponding to Figure 5), and the final detection results, as shown in Figure 6. It can be seen that the dual-stream adapters are able to capture their respective details from different graphs, which complement each other to obtain predictions that ultimately optimize both. More results can be found in the supplementary material.

4.3. Ablation Study

To demonstrate the impact of each module from SAM-DSA, we conduct ablation studies by selectively adding and removing modules from the baseline. Consequently, we

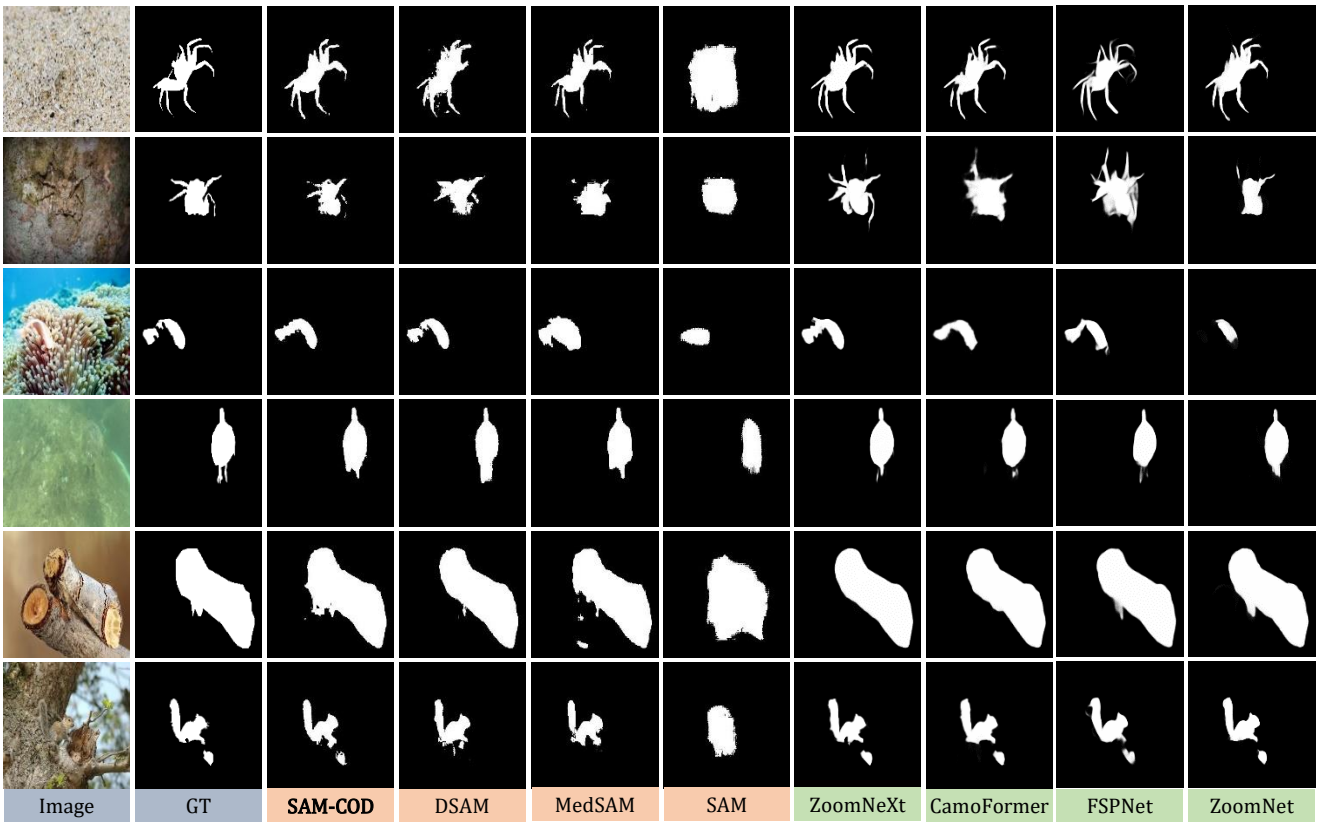

Figure 5. Comparison of our SAM-DSA and other methods in the COD task. We are mainly concerned with those of SAM-based methods.

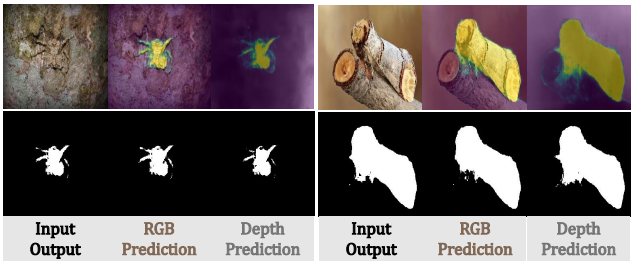

Figure 6. Comparison of the dual-stream adapter for dual-stream inputs. The final prediction result is optimized by collecting common high-confidence regions from the dual stream predictions.

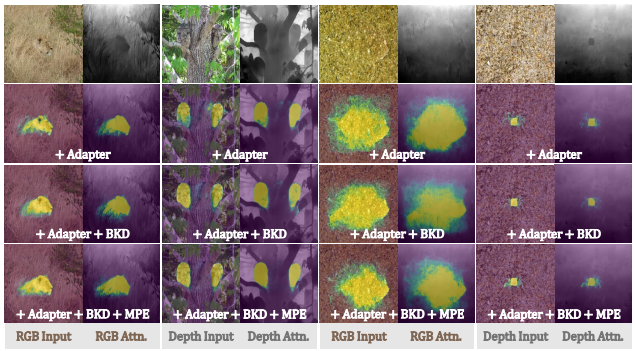

Figure 7. Attentional visualization of dual stream inputs with Adapter, BKD, and MPE modules, respectively. The components used for ablation study increase cumulatively from top to bottom.

consider replacing existing components and discover more experimental phenomena to confirm the reliability of the our method, and the results are shown in Table 2.

Component ablation. Our SAM-DSA consists of three main components: the core dual-stream adapter (Adapter) module, and the bidirectional knowledge distillation (BKD) and mixed prompt embedding (MPE) modules to better adapt to dual-stream representation learning. The effect of Adapter on the baseline is obvious, with an average improvement of about 8% on the positive evaluation criteria on COD10K. Note that BKD and MPE have a more discriminative effect on the learning of dual-stream inputs, especially in the details of the fusion of the target and the environment. To show this phenomenon intuitively, we show the mask prediction of their application in Figure 7. It can be found that the combination of BKD and MPE can make the adapter more accurate in perceiving camouflaged objects.

Adapter ablation. We develop a series of detailed discussions of the adapter, including adapted features added incrementally to frozen blocks of attention (LoRA [17] designed for image vision), without discrete wavelet transforms, and using the adapter only for RGB inputs. From the results, the adapter in the form of LoRA is not able to perceive the camouflaged object effectively with limited incremental cues. The features learned by the dual-stream adapter without

Table 2. Ablation study and performance comparison of each module on four benchmark datasets.

Ablation settings	CAMO (250)				CHAMELEON (76)				COD10K (2026)				NC4K (4121)			
	$M\downarrow$	$F^{\omega}\uparrow$	$S_m\uparrow$	$\alpha E\uparrow$	$M\downarrow$	$F^{\omega}\uparrow$	$S_m\uparrow$	$\alpha E\uparrow$	$M\downarrow$	$F^{\omega}\uparrow$	$S_m\uparrow$	$\alpha E\uparrow$	$M\downarrow$	$F^{\omega}\uparrow$	$S_m\uparrow$	$\alpha E\uparrow$
Ablation study																
Baseline (B)	.073	.759	.803	.901	.054	.729	.815	.915	.037	.735	.829	.908	.047	.804	.854	.925
B + Adapter	.056	.814	.849	.935	.034	.827	.875	.953	.026	.794	.865	.938	.035	.846	.881	.951
B + Adapter + BKD	.051	.823	.863	.945	.032	.834	.882	.949	.023	.811	.870	.940	.032	.854	.887	.953
B + Adapter + BKD + MPE	.044	.849	.875	.952	.029	.850	.893	.961	.022	.827	.887	.948	.029	.866	.896	.959
Adapter structure																
w/ LoRA form	.056	.815	.847	.936	.034	.819	.870	.955	.029	.769	.848	.933	.041	.816	.860	.941
w/o DWT	.050	.834	.860	.944	.031	.842	.885	.962	.025	.803	.868	.944	.036	.840	.875	.951
RGB-stream adapter	.047	.839	.866	.946	.031	.841	.888	.956	.023	.817	.881	.942	.032	.857	.889	.954
Dual-stream adapter	.044	.849	.875	.952	.029	.850	.893	.961	.022	.827	.887	.948	.029	.866	.896	.959
Knowledge distillation																
$\boldsymbol{L}_{KD-model}(\boldsymbol{L}_{PVT\rightarrow RGB})$	.048	.838	.865	.945	.031	.844	.888	.960	.024	.812	.875	.947	.035	.842	.878	.951
$\boldsymbol{L}_{KD-modal}(\boldsymbol{L}_{RGB\rightarrow Depth})$	.048	.840	.867	.946	.030	.846	.889	.961	.024	.815	.876	.948	.035	.843	.879	.953
$\boldsymbol{L}_{(PVT\rightarrow Depth)}$ + $\boldsymbol{L}_{(Depth\rightarrow RGB)}$	.047	.842	.868	.948	.029	.847	.890	.962	.023	.818	.880	.948	.033	.848	.882	.957
$\boldsymbol{L}_{(PVT\rightarrow RGB)}$ + $\boldsymbol{L}_{(RGB\rightarrow Depth)}$	.044	.849	.875	.952	.029	.850	.893	.961	.022	.827	.887	.948	.029	.866	.896	.959
Prompt embedding																
$\bar{\boldsymbol{X}}_{Ada}^{RGB}$ *or* $\bar{\boldsymbol{X}}_{Ada}^{Depth}$	.055	.811	.848	.934	.037	.810	.865	.947	.029	.771	.852	.930	.040	.817	.863	.941
$\bar{\boldsymbol{X}}_{Ada}^{RGB} + \bar{\boldsymbol{X}}_{Ada}^{Depth}$	.053	.819	.851	.939	.036	.814	.866	.950	.029	.776	.853	.932	.038	.829	.868	.945
$cat[\boldsymbol{X}_{Ada}^{RGB}, \bar{\boldsymbol{X}}_{Ada}^{Depth}]$	.049	.838	.864	.947	.033	.841	.884	.961	.024	.812	.873	.948	.035	.843	.878	.952
$cat[\boldsymbol{X}_{PVT}^{RGB}, \bar{\boldsymbol{X}}_{Ada}^{RGB} + \bar{\boldsymbol{X}}_{Ada}^{Depth}]$	.044	.849	.875	.952	.029	.850	.893	.961	.022	.827	.887	.948	.029	.866	.896	.959

DWT have certain defects. Moreover, Norm1 and Norm2 have less impact on the final prediction results, which shows that our adapter do not depend on the normalized structure of the VFM network. Last but not list, for the RGB input only, which is also a realistic condition as discussed before, the detection performance is slightly degraded without the complement of depth information, but it also outperforms numerous advanced methods with the same input setting.

Distillation ablation. For the ablation study of knowledge distillation, we first study the effect of single knowledge distillation, that is, the knowledge transfer occurs in $\boldsymbol{L}_{KD-model}$ between models and $\boldsymbol{L}_{KD-modal}$ between modalities. From the results, we find that traditional single knowledge distillation cannot adapt to our dual-stream adapter. Due to the lack of expert knowledge prior, only $\boldsymbol{L}_{KD-modal}$ performs the worst. Then we try to use expert embedding (PVT) to guide the adapter embedding (Depth), and then establish a relationship with embedding (RGB). This order of BKD is different from ours, which we analyze: as PVT is essentially trained on a large number of RGB images, the dense PVT embedding first guides the regional depth embedding, which has a slight gap in results.

Prompt ablation. We utilize the prompt embedding from SAM and use sparse box embeddings in the initial stage. Then to make the prompt embeddings dense pixel-aware, we use point embeddings generated in the image encoder mixed with the prompt embeddings. The first option is to use separate adapter embeddings for different inputs, which turns out to be the worst configuration due to the presence of learning bias in the mixing process. Then we use the common feature summation operation, using the dual-stream embedding as mixed pixel features. It can be found that the embedding summation yields better results, as this operation directly integrates the representations of the two types of inputs. We then introduce the expert embedding PVT, which yields moderate gains when simply concatenated with the RGB embedding. Finally, we achieve the best results with our default configuration after adding depth embeddings to the image representation.

5. Conclusion

This paper proposes an improved visual foundation model SAM-DSA for the camouflaged object detection task. By extending the dual stream adapters, introducing the bidirectional knowledge distillation and mixed prompt embedding, the proposed SAM-DSA significantly improves the performance of the segmentation performance on RGB-D input. Experimental results show that SAM-DSA can effectively utilize the complementary information of RGB and depth images, provide an efficient solution for camouflaged object detection. This approach further demonstrates that significant performance improvements can be achieved by designing modules to suit specific tasks while maintaining the integrity of the infrastructure.

Future work. Our future work will be dedicated to further optimizing SAM-DSA to handle more complex multimodal inputs, such as combining thermal imaging data or spectral data to improve the applicability of the model. In addition, the lightweight SAM and adapter can be explored to reduce computational complexity and make it more efficient in real-time vision applications.

References

[1] Silvia Maria Alessio and Silvia Maria Alessio. Discrete wavelet transform (dwt). *Digital signal processing and spectral analysis for scientists: concepts and applications*, pages 645–714, 2016. 4

[2] Hongbo Bi, Yuyu Tong, Jiayuan Zhang, Cong Zhang, Jinghui Tong, and Wei Jin. Depth alignment interaction network for camouflaged object detection. *Multimedia Systems*, 30(1):51, 2024. 2

[3] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024. 1, 3

[4] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod+: Sam-guided unified framework for weakly-supervised camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[5] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *Proceedings of the European Conference on Computer Vision*, pages 315–331, 2025. 3

[6] Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024. 3

[7] 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. 1, 2, 3, 4, 5, 6

[8] 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. *arXiv preprint arXiv:2304.09148*, 2023. 1

[9] 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. 3

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[11] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 5

[12] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 698–704, 2018. 5

[13] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021. 2, 5, 6

[14] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2777–2787, 2020. 6

[15] Xue Feng, Cui Guoying, and Song Wei. Camouflage texture evaluation using saliency map. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 93–96, 2013. 2

[16] 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. 2

[17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022. 7

[18] Xihang Hu, Xiaoli Zhang, Fasheng Wang, Jing Sun, and Fuming Sun. Efficient camouflaged object detection network based on global localization perception and local guidance refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5452–5465, 2024. 6

[19] 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

[20] Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. Endow sam with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19058–19067, 2024. 3

[21] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023. 6

[22] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications, 2024. 1

[23] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9471–9481, 2021. 5, 6

[24] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4713–4722, 2022. 2, 6

[25] Abbas Khan, Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, Giulia De Masi, and Fakhri Karray. Camofocus: Enhancing camouflage object detection with split-feature focal modulation and context refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1434–1443, 2024. 6

[26] 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. 1, 3, 5, 6

[27] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 4

[28] 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, 6

[29] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *Proceedings of the European Conference on Computer Vision*, pages 630–647. Springer, 2022. 2, 5, 6

[30] 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

[31] Jiaming Liu, Linghe Kong, Jiajie Yan, and Guihai Chen. Mesh-aligned 3d gaussian splatting for multi-resolution anti-aliasing rendering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[32] Jiaming Liu, Yue Wu, Maoguo Gong, Zhixiao Liu, Qiguang Miao, and Wenping Ma. Inter-modal masked autoencoder for self-supervised learning on point clouds. *IEEE Transactions on Multimedia*, 26:3897–3908, 2023. 2

[33] 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. 3

[34] Xinran Liu, Lin Qi, Yuxuan Song, and Qi Wen. Depth awakens: A depth-perceptual attention fusion network for rgb-d camouflaged object detection. *Image and Vision Computing*, 143:104924, 2024. 2

[35] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017. 5

[36] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Nick Barnes, and Deng-Ping Fan. Toward deeper understanding of camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3462–3476, 2023. 4

[37] 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. 2, 5, 6

[38] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3, 5, 6

[39] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014. 5

[40] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8772–8781, 2021. 6

[41] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, pages 807–814, 2010. 3

[42] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2160–2170, 2022. 2, 6

[43] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6

[44] Jay N Paranjape, Nithin Gopalakrishnan Nair, Shameema Sikder, S Swaroop Vedula, and Vishal M Patel. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, pages 187–201, 2024. 1

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[46] 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*, volume 38, pages 4515–4523, 2024. 1

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

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[49] P Sengottuvelan, Amitabh Wahi, and A Shanmugam. Performance of decamouflaging through exploratory image analysis. In *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology*, pages 6–10, 2008. 2

[50] 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, 6

[51] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516):423–427, 2009. 1

[52] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for cam-

ouflaged object detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1025–1031, 2021. 6

[53] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *Proceedings of the European Conference on Computer Vision*, pages 343–360, 2024. 2

[54] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022. 2, 5, 6

[55] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8565–8574, 2021. 2

[56] Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, and Feng Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17201–17211, 2024. 2

[57] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2, 4, 5

[59] 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. *arXiv preprint arXiv:2304.12620*, 2023. 1, 3

[60] Yue Wu, Jiaming Liu, Maoguo Gong, Peiran Gong, Xiaolong Fan, A Kai Qin, Qiguang Miao, and Wenping Ma. Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. *IEEE Transactions on Multimedia*, 26:1626–1638, 2023. 2

[61] Yue Wu, Jiaming Liu, Maoguo Gong, Qiguang Miao, Wenping Ma, and Cai Xu. Joint semantic segmentation using representations of lidar point clouds and camera images. *Information Fusion*, 108:102370, 2024. 2

[62] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1032–1042, 2023. 2, 3, 5, 6

[63] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4146–4155, 2021. 2, 6

[64] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6

[65] Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. Exploring deeper! segment anything model with depth perception for camouflaged object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4322–4330, 2024. 2, 3, 5, 6

[66] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 1

[67] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12997–13007, 2021. 6

[68] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for rgb-d salient object detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 2094–2102, 2021. 5, 6

[69] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4338–4347, 2021. 5, 6

[70] Xiaoqin Zhang, Zhenni Yu, Li Zhao, Deng-Ping Fan, and Guobao Xiao. Comprompter: reconceptualized segment anything model with multiprompt network for camouflaged object detection. *Science China Information Sciences*, 68(1):112104, 2025. 6

[71] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4681–4691, 2021. 5, 6

[72] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024. 3