CamSAM2: Segment Anything Accurately in Camouflaged Videos

Yuli Zhou^{1,2} Guolei Sun^{1*} Yawei Li^{1,3} Yuqian Fu⁴ Luca Benini^{3,5} Ender Konukoglu¹

¹Computer Vision Laboratory, ETH Zurich

²University of Zurich

³Integrated Systems Laboratory, ETH Zurich

⁴INSAIT, Sofia University "St. Kliment Ohridski"

⁵University of Bologna

Abstract

Video camouflaged object segmentation (VCOS), aiming at segmenting camouflaged objects that seamlessly blend into their environment, is a fundamental vision task with various real-world applications. With the release of SAM2, video segmentation has witnessed significant progress. However, SAM2's capability of segmenting camouflaged videos is suboptimal, especially when given simple prompts such as point and box. To address the problem, we propose Camouflaged SAM2 (CamSAM2), which enhances SAM2's ability to handle camouflaged scenes without modifying SAM2's parameters. Specifically, we introduce a decamouflaged token to provide the flexibility of feature adjustment for VCOS. To make full use of fine-grained and highresolution features from the current frame and previous frames, we propose implicit object-aware fusion (IOF) and explicit object-aware fusion (EOF) modules, respectively. Object prototype generation (OPG) is introduced to abstract and memorize object prototypes with informative details using high-quality features from previous frames. Extensive experiments are conducted to validate the effectiveness of our approach. While CamSAM2 only adds negligible learnable parameters to SAM2, it substantially outperforms SAM2 on three VCOS datasets, especially achieving 12.2 mDice gains with click prompt on MoCA-Mask and 19.6 mDice gains with mask prompt on SUN-SEG-Hard, with Hiera-T as the backbone. The code will be available at github.com/zhoustan/CamSAM2.

1. Introduction

Camouflaged object detection (COD) and video camouflaged object segmentation (VCOS) aim to identify objects that blend seamlessly into their surroundings. Unlike standard object segmentation tasks, where objects typi-

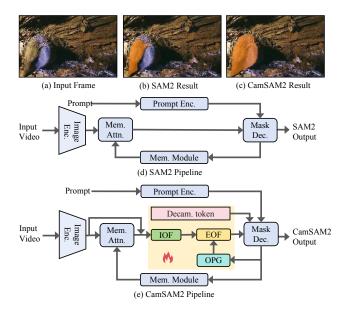


Figure 1. **Illustration of SAM2 and CamSAM2**. (d) SAM2's segmentation of the camouflaged object is suboptimal, primarily because its feature optimization is biased toward natural videos, and its design does not account for the unique challenges inherent to VCOS. (e) CamSAM2 improves SAM2's ability on segmenting and tracking camouflaged objects by introducing a *decamouflaged token*, *IOF* to enhance features with high-resolution features, and *EOF* and *OPG* to further enhance features by exploiting informative object details across time. CamSAM2 only adds a limited number of parameters on SAM2 while keeping all SAM2's parameters fixed and fully inheriting SAM2's zero-shot ability.

cally exhibit clear boundaries and contrast with the background, camouflaged objects are naturally indistinguishable from the background. These tasks have various applications in wildlife monitoring, surveillance, and searchand-rescue operations [40, 44]. COD focuses on detecting camouflaged objects in individual images, while VCOS extends it to video sequences, adding the complexity of modeling temporal information across frames. Despite re-

^{*}Corresponding author.

cent advancements in COD [5, 7, 33, 34, 54, 58, 59] and VCOS [4, 7, 15, 20, 25, 29, 33], the performance remains far from satisfactory compared to standard segmentation tasks.

The recently introduced vision foundation model, Segment Anything Model 2 (SAM2) [39], marks a significant advancement in video segmentation. SAM2 has learned rich and generalizable representations for natural scenes on the SA-1B [19] (11M images, 1B masks) and SA-V [39] (50.9K videos, 35.5M masks) datasets. Therefore, its features are optimized for natural scenes, while SAM2's ability of segmenting camouflaged objects is suboptimal, as in [47, 61]. As shown in Fig. 1, SAM2 segments only part of a camouflaged animal (*hedgehog*) with a single-click prompt, indicating that there is still room for performance improvements in VCOS using SAM2.

This paper aims to develop a model for accurate segmentation in camouflaged videos, requiring both natural image understanding and effective identification of camouflaged objects in complex environments. To achieve this, we identify the following core challenges in adapting SAM2 for VCOS: (1) SAM2 is optimized for natural scenes rather than camouflaged environments. (2) The architecture does not account for the complexities of segmenting and tracking camouflaged objects across time. For VCOS, accurately segmenting camouflaged objects for a frame requires: a) exploiting fine-grained and detailed features from the frame, and b) considering the temporal evolvement of fine-grained features from previous frames. For exploiting temporal information, SAM2 is equipped with a memory module containing a memory encoder and memory bank. However, only low-resolution and coarse features are encoded into the memory, which is suboptimal for accurate VCOS.

To tackle the above limitations and fully keep SAM2's ability to process natural videos, we introduce Camouflaged SAM2, dubbed as CamSAM2, equipping SAM2 with the ability to effectively tackle VCOS, as depicted in Fig. 1. CamSAM2 includes a learnable decamouflaged token, which provides flexibility to optimize features for VCOS without modifying SAM2's trained parameters. To exploit the fine-grained features of the frame, we propose the Implicit Object-aware Fusion (IOF) module, which enhances features with implicitly object-aware information. To make use of detailed features from previous frames, we further propose Object Prototype Generation (OPG) to abstract high-quality features within the object region into informative object prototypes through Farthest Point Sampling (FPS) and k-means. Those object prototypes are saved to the memory for easy usage by the Explicit Object-aware Fusion (EOF) module that is designed to integrate explicit object-aware information across the temporal dimension. Our design avoids saving the high-resolution features in the memory and only adds negligible computations to SAM2 while accounting for a large amount of temporal information.

We conduct extensive experiments in §4 to validate the effectiveness of CamSAM2 on three VCOS benchmarks: two camouflaged animal datasets, MoCA-Mask [7] and CAD [3]; and one camouflaged medical dataset, SUN-SEG [17]. Our experiments show that CamSAM2 significantly outperforms SAM2 by achieving improvements of 12.2/13.0 mDice scores with click prompt on MoCA-Mask for Hiera-T/Hiera-S backbones, and 19.6 mDice gains with mask prompt on SUN-SEG-Hard for Hiera-T backbone. When directly evaluating CamSAM2 on CAD without further finetuning, we observe strong zero-shot ability. Since all SAM2's weights remain unchanged, CamSAM2 totally inherits SAM2's capability on segmenting natural videos. In summary, our contributions are three-fold:

- We propose CamSAM2 to equip SAM2 with the ability to segment and track camouflaged objects in videos while keeping SAM2's strong generalizability in natural videos.
- CamSAM2 introduces a decamouflaged token to achieve easy feature adjustments for the VCOS task without affecting SAM2's trained weights. To effectively exploit the crucial fine-grained and high-resolution features from both the current frame and previous frames, we propose IOF, EOF, and OPG modules.
- Our approach clearly outperforms SAM2 and sets new state-of-the-art performance on public VCOS datasets.
 Experiments also show the strong zero-shot ability of CamSAM2 in the domain of VCOS.

2. Related Work

2.1. Camouflaged Object Detection

Camouflaged Scene Understanding (CSU) focuses on interpreting scenes where objects blend closely with their backgrounds, such as natural environments like forests, oceans, and deserts. Early works in this field primarily involved the collection of extensive image and video datasets, such as CAMO [23], COD10K [10], NC4K [26], CAD [3], MoCA-Mask [7], and MoCA-Mask-Pseudo [7], which laid the foundation for CSU.

Traditional Camouflaged Object Detection (COD) methods extract foreground-background features using optical features [2], color, and texture [18]. Deep learning has advanced COD with CNNs and transformers. SINet [10] and SINet-V2 [12] enhance fine-grained cues by applying receptive fields and texture-enhanced modules, while DQNet [45] applies cross-modal detail querying to detect subtle features. Transformer-based models like Camo-Former [54] leverage multi-scale feature extraction with masked separable attention, and WSSCOD [58] employs a frequency transformer and noisy pseudo labels for weak supervision. ZoomNeXt [33] further optimizes multi-scale

extraction via a collaborative pyramid network. These advancements refine COD by integrating sophisticated architectures and diverse supervision strategies.

2.2. Video Camouflaged Object Segmentation

VCOS [20, 25, 49, 57] extends COD to videos, introducing challenges from motion, dynamic backgrounds, and temporal consistency. Former VCOS models tackle these with motion learning, spatial-temporal attention, and advanced segmentation techniques to maintain object coherence across frames. Motion-guided models enhance segmentation by leveraging motion cues. IMEX [14] integrates implicit and explicit motion learning for robust TMNet [57] refines motion features with a transformer-based encoder and neighbor connection decoder. Flow-SAM [50] uses optical flow as input or a prompt, guiding SAM to detect moving camouflaged obiects. Spatial-temporal attention enhances the tracking of camouflaged objects. TSP-SAM [15] and SAM-PM [29] improve SAM's ability to detect subtle movements. Static-Dynamic-Interpretability [20] quantifies static and dynamic information in spatial-temporal models, aiding balanced approaches. Assessing camouflage quality is also essential for VCOS. CAMEVAL [22] introduces scores evaluating background similarity and boundary visibility, refining datasets and improving model robustness. These advancements drive more accurate and effective VCOS systems.

2.3. Segment Anything Model 2

SAM2 [39] is a vision foundation model for promotable segmentation across images and videos. Compared to SAM [19] which is limited to image segmentation, SAM2 offers a significant performance leap in video segmentation, producing higher segmentation accuracy while using fewer interactions than previous methods. SAM2 has demonstrated strong capabilities in many tasks, including medical image, video and 3D segmentation [6, 13, 24, 27, 42, 51, 55, 56], video object tracking and segmentation [32, 43], remote sensing [38], 3D mesh and point cloud segmentation [46], image camouflaged object detection and video camouflaged object segmentation [6, 47, 51, 61]. In previous works [47, 61], although SAM2's ability to segment camouflaged videos has surpassed most existing methods, there is still a significant performance gap compared to other VOS tasks, especially when using simple prompts.

3. Method

We propose CamSAM2, equipping SAM2 with the ability to accurately segment camouflaged objects in videos while retaining SAM2's original capabilities. §3.1 briefly reviews the architecture of SAM2. From §3.2 to §3.5, we describe CamSAM2 tailored for VCOS. With fixing SAM2's parameters, CamSAM2 proposes a learnable decamouflaged to-

ken, Implicit and Explicit Object-aware Fusion, and Object Prototype Generation to enhance feature representations, thus leading to improved performance, as shown in Fig. 2. Training and inference strategies are presented in §3.6.

3.1. Preliminaries

SAM2 [39] is a pioneering vision foundation model designed for promptable visual segmentation tasks. Different from SAM [19], SAM2 includes a memory module that stores information about the object from previous frames. It contains an image encoder, memory attention, prompt encoder, mask decoder, memory encoder, and memory bank. For each frame, the image encoder extracts representative visual features, which are then conditioned on the features and predictions of past frames. If a prompt (point, box, or mask) is given, the prompt encoder encodes it into sparse or dense embeddings. Exploiting memory-conditioned features and prompt embeddings, the mask decoder outputs the segmentation mask. The memory encoder then updates the memory bank with the output mask and the unconditioned frame embedding to support the segmentation of subsequent frames. SAM2 is pre-trained on SA-1B [19] and further trained on SA-V [39], achieving strong performance across video and image segmentation tasks. For more details, please refer to [39].

3.2. Decamouflaged Token

Given a video clip containing m frames, we denote all frames as $\{\mathbf{I}_{t-m+1},\cdots,\mathbf{I}_i,\cdots,\mathbf{I}_t\}$ with ground-truth segmentation masks of $\{\mathbf{S}_{t-m+1},\cdots,\mathbf{S}_i,\cdots,\mathbf{S}_t\}$. Especially, \mathbf{I}_t is the current frame for the purpose of easy explanation. We use the image encoder to extract features for all frames, denoted as $\{\mathbf{F}_{t-m+1},\cdots,\mathbf{F}_i,\cdots,\mathbf{F}_t\}$. Here, \mathbf{F}_i can be further represented as $\{\mathbf{F}_i^0,\cdots,\mathbf{F}_i^j,\cdots,\mathbf{F}_i^{L-1}\}$, containing feature maps extracted from L different intermediate layers or blocks, where $\mathbf{F}_i^j \in \mathbb{R}^{c_j \times h_j \times w_j}$, with c_j,h_j , and w_j representing channels, height, and width, respectively.

SAM2's output tokens include an object score (occlusion) token, an IoU token, and mask tokens. To enhance SAM2's ability of segmenting camouflaged objects, we introduce a new learnable decamouflaged token $\mathbf{T} \in \mathbb{R}^{1 \times 256}$, enabling it to optimize features for segmenting camouflaged objects. As depicted in Fig. 2, integrated with SAM2's output tokens, the decamouflaged token undergoes the same layers as output tokens within SAM2's mask decoder: two attention blocks including self-attention, followed by tokento-image (T2I) and image-to-token (I2T) cross attention, and the last T2I cross attention. After this, the output decamouflaged token is denoted as \mathbf{T}' . This token is updated through back-propagated gradients, while SAM2's weights remain frozen. \mathbf{T}' is then passed through an MLP layer to participate in computing CamSAM2's final mask logits,

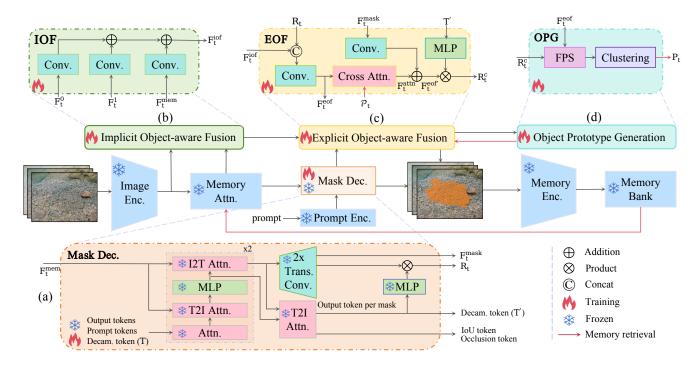


Figure 2. **Overall architecture of CamSAM2**. CamSAM2 effectively captures and segments camouflaged objects by leveraging implicit and explicit object-aware information from the current or previous frames. It includes the following key components: (a) the decamouflaged token, which extends SAM2's token structure to learn features suitable for segmenting camouflaged objects; (b) an *IOF* module to enrich memory-conditioned features with implicitly object-aware high-resolution features; (c) an *EOF* module to aggregate explicit object-aware features; and (d) an *OPG* module, generating informative object prototypes, which guides cross-attention in EOF. These components work together to preserve fine details, enhance segmentation quality, and track camouflaged objects across time.

which will be explained in §3.4.

3.3. Implicit Object-aware Fusion

Early-layer features from the image encoder capture high-resolution details, such as edges and textures, essential for distinguishing subtle differences between the camouflaged object and the background. In contrast, deeper layers focus on high-level semantic information. In SAM2, memory-conditioned features are computed by only conditioning high-level semantic features on the memory, without using detailed features from early layers. These early-layer features are implicit object-aware, as features for background and non-relevant objects also exist with similar magnitude. To this end, we propose an IOF module that fuses these implicit object-aware features with memory-conditioned features.

For SAM2, three feature maps from the Hiera image encoder [41] are extracted for each frame, i.e., L=3. We have \mathbf{F}_t^0 , \mathbf{F}_t^1 , and \mathbf{F}_t^2 for the current frame \mathbf{I}_t . We denote the memory-conditioned feature as \mathbf{F}_t^{mem} , encoded by the memory-attention module on \mathbf{F}_t^2 , as in [39]. The high-resolution features \mathbf{F}_t^0 and \mathbf{F}_t^1 are fused with \mathbf{F}_t^{mem} via compression modules and point-wise addition to create a refined feature representation $\mathbf{F}_t^{iof} \in \mathbb{R}^{c_0 \times h_0 \times w_0}$, where a compression module $C(\cdot)$ consists of two convolutional

layers, followed by an upsampling layer, to create compact representations. This process is given by:

$$\mathbf{F}_t^{iof} = C_0(\mathbf{F}_t^0) + C_1(\mathbf{F}_t^1) + C_2(\mathbf{F}_t^{mem}). \tag{1}$$

3.4. Explicit Object-aware Fusion

After obtaining \mathbf{F}_t^{iof} , we further refine it by EOF, which exploits *explicit* object-aware information from the current frame and previous frames, through employing object mask logits and object prototypes (see §3.5). We have three steps to fuse informative features. *First*, feature \mathbf{F}_t^{iof} , with shape $\mathbb{R}^{c_0 \times h_0 \times w_0}$ from the previous IOF module, is directly concatenated with SAM2's mask logits \mathbf{R}_t , which has shape $\mathbb{R}^{1 \times h_0 \times w_0}$. This concatenated feature is then processed through a convolutional layer to reduce the channels back to c_0 , resulting in output with the original shape $\mathbb{R}^{c_0 \times h_0 \times w_0}$, denoted as:

$$\mathbf{F}_{t}^{eof} = \operatorname{Conv}\left(\left[\mathbf{F}_{t}^{iof}; \mathbf{R}_{t}\right]\right).$$
 (2)

Second, \mathbf{F}_t^{eof} goes through a cross-attention layer. Prototypes generated from previous frames, representing clustered camouflaged features, serve as informative priors to help distinguish the camouflaged object from its background. A cross-attention mechanism takes \mathbf{F}_t^{eof} as a

query, and leverages these prototypes as keys and values, effectively exploiting the information within the object prototypes to refine \mathbf{F}_t^{eof} . Formally, we update \mathbf{F}_t^{eof} by conducting cross-attention with prototypes $\mathcal{P}_t = \{\mathbf{P}_0, \mathbf{P}_1, \dots, \mathbf{P}_{t-1}\}$ from previous frames, given by:

$$\mathbf{F}_{t}^{attn} = \operatorname{Attn}(\mathbf{F}_{t}^{eof}, \mathcal{P}_{t}, \mathcal{P}_{t}). \tag{3}$$

Third, the attention-refined feature \mathbf{F}_t^{attn} is combined with the upscaled mask feature \mathbf{F}_t^{mask} from SAM2 mask decoder. The upscaled mask feature is first processed through a convolutional layer and then fused with \mathbf{F}_t^{attn} via point-wise addition, as follows:

$$\mathbf{F}_{t}^{eof\prime} = \mathbf{F}_{t}^{attn} + \text{Conv}(\mathbf{F}_{t}^{mask}). \tag{4}$$

Finally, we calculate the mask logits \mathbf{R}_t^c of CamSAM2, by processing the output decamouflaged token \mathbf{T}' through an MLP layer, and then performing point-wise product with the $\mathbf{F}_t^{eof'}$, as shown below:

$$\mathbf{R}_{t}^{c} = \text{MLP}(\mathbf{T}') \cdot \mathbf{F}_{t}^{eof'}. \tag{5}$$

This approach incorporates both implicit and explicit camouflaged information, which can enhance mask generation for more accurate segmentation for the VCOS task.

3.5. Object Prototype Generation

To effectively represent the camouflaged features within the mask (object) region, we employ Farthest Point Sampling (FPS) [31] to identify k points within the predicted mask region, which act as cluster centers. This approach ensures that the sampled points are well-distributed throughout the mask, capturing diverse and important characteristics of the camouflaged object. Then, we group all pixels in the predicted mask region into k clusters by conducting one-iteration k-means, using the sampled k points as initial centers. The prototype of each cluster is represented as the mean of the spatial features of the points in the cluster. This prototype generation process is denoted as \mathcal{F}_p , as shown in:

$$\mathbf{P}_t = \{ P_t^i \mid 1 \le i \le k \} = \mathcal{F}_p(\mathbf{F}_t^{eof}, \mathbf{R}_t^c), \tag{6}$$

where \mathbf{P}_t represents the camouflaged object prototypes extracted from high-resolution and detailed features for the frame \mathbf{I}_t . The prototypes are concatenated and then saved in the memory, which will be used by EOF (§3.4) when segmenting the subsequent frames.

3.6. Training and Inference Strategies

Training Strategies. We simulate interactive prompting of the model in the training process, prompting on the first frame of the sampled sequence. Following the training strategy of SAM2, we use three types of prompts (mask, bounding box, 1-click point of foreground) for training, with the probability of 0.5, 0.25, and 0.25, respectively.

To train the model, we use a combined loss of binary cross-entropy (BCE) and dice loss for mask predictions across the entire video. This loss applies to both SAM2's mask logits \mathbf{R}_i^c , compared with the ground-truth mask \mathbf{S}_i of frame \mathbf{I}_i , as follows:

$$\mathcal{L}_{C} = \sum_{i=t-m+1}^{t} \left[\mathcal{L}_{BCE}(\mathbf{R}_{i}, \mathbf{S}_{i}) + \mathcal{L}_{BCE}(\mathbf{R}_{i}^{c}, \mathbf{S}_{i}) \right],$$

$$\mathcal{L}_{D} = \sum_{i=t-m+1}^{t} \left[\mathcal{L}_{Dice}(\mathbf{R}_{i}, \mathbf{S}_{i}) + \mathcal{L}_{Dice}(\mathbf{R}_{i}^{c}, \mathbf{S}_{i}) \right],$$

$$\mathcal{L} = \mathcal{L}_{C} + \mathcal{L}_{D},$$
(7)

where \mathcal{L} is the final loss for our approach, summing the BCE loss \mathcal{L}_C and the dice loss \mathcal{L}_D .

Inference. During inference, we provide a prompt at the first frame of a video, following [29]. Our final output is the average of the logits of SAM2 and CamSAM2 masks for the error correction.

4. Experiments

4.1. Experimental Setup

Our experiments are conducted on three video datasets: two popular camouflaged animal datasets, MoCA-Mask [7] and CAD [3], and one camouflaged medical dataset, SUN-SEG [17]. The pioneering Moving Camouflaged Animals dataset (MoCA) [21] comprises 37K frames from 141 YouTube video sequences. The dataset MoCA-Mask is reorganized from the MoCA, containing 71 video sequences with 19,313 frames for training and 16 video sequences with 3,626 frames for testing, respectively, with pixel-wise ground-truth masks on every five frames. It also generates a MoCA-Mask-Pseudo dataset, which contains pseudo masks for unlabeled frames with a bidirectional optical-flow-based consistency check strategy. The Camouflaged Animal Dataset (CAD) includes 9 short videos in total that have 181 hand-labeled masks on every five frames. SUN-SEG is the largest benchmark for video polyp segmentation, derived from SUN-database [30]. It consists of a training set with 112 clips (19,544 frames) and two test sets: SUN-SEG-Easy, containing 119 clips (17,070 frames), and SUN-SEG-Hard, comprising 54 clips (12,522 frames).

Implementation Details. The proposed CamSAM2 is implemented with PyTorch [35]. CamSAM2 is initialized with the parameters of SAM2. We freeze all parameters used in SAM2 and initialize other parameters randomly. We set betas = (0.9, 0.999) for the optimizer Adam and use the initial learning rate of 1e-3 with the StepLR of 10. We train CamSAM2 on 4 NVIDIA RTX 4090 GPUs for 15 epochs. For camouflaged animal segmentation, we train the model using the MoCA-Mask-Pseudo training set and eval-

Model	Backbone	Params (M)	Prompt	$S_m \uparrow$	$F^{\omega}_{\beta}\uparrow$	MAE ↓	$F_{\beta}\uparrow$	$E_m \uparrow$	mDice ↑	mIoU ↑
EGNet [60]	ResNet-50	111.7	-	54.7	11.0	3.5	13.6	57.4	14.3	9.6
BASNet [37]	ResNet-50	87.1	-	56.1	15.4	4.2	17.3	59.8	19.0	13.7
CPD [48]	ResNet-50	47.9	-	56.1	12.1	4.1	15.2	61.3	16.2	11.3
PraNet [11]	ResNet-50	32.6	-	61.4	26.6	3.0	29.6	67.4	31.1	23.4
SINet [10]	ResNet-50	48.9	-	59.8	23.1	2.8	25.6	69.9	27.7	20.2
SINet-V2 [12]	Res2Net-50	27.0	-	58.8	20.4	3.1	22.9	64.2	24.5	18.0
PNS-Net [16]	ResNet-50	142.9	-	52.6	5.9	3.5	8.4	53.0	8.4	5.4
RCRNet [52]	ResNet-50	53.8	-	55.5	13.8	3.3	15.9	52.7	17.1	11.6
MG [53]	VGG	4.8	-	53.0	16.8	6.7	19.5	56.1	18.1	12.7
SLT-Net-LT [7]	PVTv2-B5	82.3	-	63.1	31.1	2.7	33.1	75.9	36.0	27.2
ZoomNeXt [33]	PVTv2-B5	84.8	-	73.4	47.6	1.0	49.7	73.6	49.7	42.2
SAM2 [39]	Hiera-T	38.9	1-click	68.2	50.7	7.7	52.5	73.6	52.1	44.8
CamSAM2	Hiera-T	39.4	1-click	75.2	61.7	7.3	63.7	82.0	64.3	54.6
SAM2 [39]	Hiera-T	38.9	box	81.5	69.9	0.6	70.9	89.4	72.7	62.3
CamSAM2	Hiera-T	39.4	box	82.9	72.4	0.6	73.2	94.2	75.5	64.8
SAM-PM [29]	ViT-L	303.0	mask	72.8	56.7	0.9	-	81.3	59.4	50.2
SAM2 [39]	Hiera-T	38.9	mask	84.7	76.0	0.4	76.9	91.9	77.1	67.9
CamSAM2	Hiera-T	39.4	mask	86.2	78. 7	0.4	79.6	96.2	80.2	70.5

Table 1. Comparisons between our method and existing approaches on MoCA-Mask. CamSAM2 outperforms the existing method by achieving new state-of-the-art performance. The results of all these methods (excluding SAM2) are from corresponding publications. The best results are shown in **bold**. ↑: the higher the better, ↓: the lower the better.

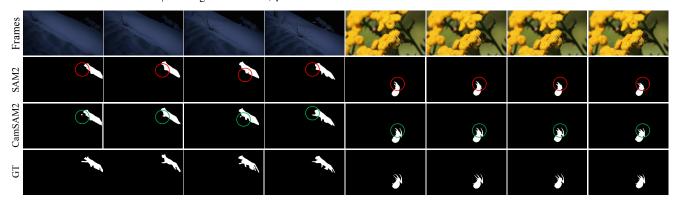


Figure 3. Qualitative comparisons between SAM2 and CamSAM2 using 1-click prompt with the Hiera-T backbone on two MoCA-Mask clips. From *top* to *bottom*: the input frames, SAM2's results, CamSAM2's results, and ground-truth masks. CamSAM2 demonstrates improved accuracy in segmenting camouflaged objects, especially in complex backgrounds, as shown by the circles. *Best viewed in color*:

uate it on the MoCA-Mask test set and CAD. During inference for SAM2 and CamSAM2, we apply the 1-click, box, and mask prompts only on the first frame of each video. For camouflaged polyp segmentation, we train the model using the SUN-SEG training set and perform inference using mask prompt on the first frame of each video on the SUN-SEG-Easy and SUN-SEG-Hard test sets.

Evaluation Metrics. We adopt seven evaluation metrics to measure the quality of predicted pixel-wise masks: S-measure (S_m) [8], F-measure (F_β) [1], weighted F-measure (F_β^ω) [28], mean absolute error (MAE) [36], E-measure (E_m) [9], mean Dice (mDice), and mean IoU (mIoU).

4.2. Experimental Results

Results on MoCA-Mask. Tab. 1 compares the performance of three promptable methods. CamSAM2 clearly

outperforms SAM-PM and SAM2; even with the 1-click prompt, CamSAM2 still outperforms SAM-PM, which uses mask prompt on the first frame. The promptable methods clearly outperform other models on the MoCA-Mask, highlighting the strength of integrating prompt-based strategies for VCOS and demonstrating the potential of promptable methods to excel in scenarios where camouflaged objects are particularly challenging to segment and track.

Tab. 2 presents a detailed comparison between SAM2 and CamSAM2 on the MoCA-Mask dataset, across different prompt types (1-click, box, and mask) with Hiera-T and Hiera-S backbones. CamSAM2 demonstrates consistent improvements over SAM2 across all prompt types and backbones. With a 1-click prompt, it achieves mDice/mIoU gains of 12.2/9.8 (Hiera-T) and 13.0/12.0 (Hiera-S), demonstrating its effectiveness in segmenting camouflaged objects with minimal input. For a box prompt, CamSAM2 im-

Model	Prompt	mDice ↑	mIoU ↑				
Hiera-T							
SAM2 CamSAM2	Click	52.1 64.3 (+12.2)	44.8 54.6 (+9.8)				
SAM2 CamSAM2	Box	72.7 75.5 (+2.8)	62.3 64.8 (+2.5)				
SAM2 CamSAM2	Mask	77.1 80.2 (+3.1)	67.9 70.5 (+2.6)				
Hiera-S							
SAM2 CamSAM2	Click	54.9 67.9 (+13.0)	46.7 58.7 (+12.0)				
SAM2 CamSAM2	Box	73.8 75.5 (+1.7)	63.8 65.4 (+1.6)				
SAM2 CamSAM2	Mask	80.3 81.3 (+1.0)	70.7 71.6 (+0.9)				

Table 2. **Detailed comparisons between CamSAM2 and SAM2 on MoCA-Mask.** CamSAM2 *consistently* outperforms SAM2 for all considered prompt types and backbones. Improvements of CamSAM2 over SAM2 are shown in dark green.

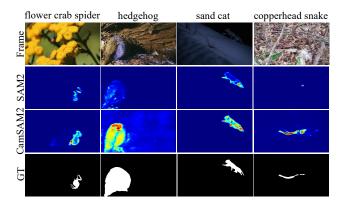


Figure 4. Visualization of the attention maps obtained from SAM2 and CamSAM2 using 1-click point prompt with the Hiera-T backbone. The attention map is extracted from the last token-to-image cross-attention in the mask decoder. From *top* to *bottom*: the input frames, attention map of SAM2's mask token and the image embedding, attention map of the decamouflaged token and the image embedding, and ground-truth masks. The higher attention regions are indicated by warmer colors.

proves mIoU by 2.5 (Hiera-T) and 1.6 (Hiera-S). With a mask prompt, it achieves 2.6 and 0.9 mIoU gains for Hiera-T and Hiera-S, respectively.

Fig. 3 shows qualitative results for two video clips of MoCA-Mask from SAM2 and CamSAM2 using 1-click prompts with the Hiera-T as the backbone. Fig. 4 presents a comparative visualization of the attention maps generated by SAM2 and CamSAM2 for various objects. The attention maps are extracted from the last token-to-image cross attention layer in the mask decoder. The token serves as

the query, and the image embedding serves as the key and value. CamSAM2 demonstrates superior attention precision over SAM2, with a larger activated region and stronger activations around target objects, validating the effectiveness of our proposed methods in enhancing VCOS quality.

Notably, CamSAM2 introduces only a marginal increase in parameters (0.5M). Despite this minimal increase, CamSAM2 achieves significant improvements while keeping all of SAM2's parameters unchanged, fully inheriting SAM2's capability for segmenting and tracking common objects in natural scenes.

Results on CAD. We evaluate the zero-shot performance of CamSAM2 and SAM2 on the CAD dataset using Hiera-T and Hiera-S backbones with point and box prompts, as shown in Tab. 3. CamSAM2 demonstrates notable improvements over SAM2, particularly in the 1-click prompt setting, where it gains 3.4 and 1.8 in mDice and mIoU, highlighting CamSAM2's enhanced capability when only minimal guidance is available. In the box prompt setting, CamSAM2 also shows clear gains, with an increase of 1.8 and 2.5 in mDice and mIoU. These observations indicate that CamSAM2 outperforms SAM2 in zero-shot scenarios, underscoring its effectiveness and suitability for practical segmentation tasks that require minimal user input.

Results on SUN-SEG. Tab. 4 presents the performance comparison on the SUN-SEG dataset, showing that Cam-SAM2 consistently outperforms SAM2 across all metrics on both SUN-SEG-Easy and SUN-SEG-Hard test sets. Notably, CamSAM2 achieves substantial improvements. Specifically, mDice improves by 10.7 on SUN-SEG-Easy, rising from 73.6 to 84.3, and by 19.6 on SUN-SEG-Hard, increasing from 61.0 to 80.6, demonstrating its effectiveness in segmenting camouflaged polyps. Based on the results above, our method significantly enhances SAM2 across different camouflaged scenarios, demonstrating its effectiveness and broad applicability in VCOS.

4.3. Ablation Studies

To understand the impact of individual components in Cam-SAM2, we conducted ablation studies on the MoCA-Mask test set using the Hiera-T backbone and a 1-click point prompt. The goal is to measure the contributions of key components, including the decamouflaged token, IOF, EOF, and OPG, on segmentation performance. Additionally, we evaluated the effects of different distance metrics and prototype numbers in the OPG process. We analyze the results below.

Impact of Key Components. As shown in Tab. 5, each main component in CamSAM2 clearly contributes to its high performance. Starting from baseline (SAM2), adding the decamouflaged token alone improves mDice from 52.1 to 54.9 and mIoU from 44.8 to 47.0. Adding IOF further

Model	Backbone	Params (M)	Prompt	$S_m \uparrow$	$F^{\omega}_{\beta}\uparrow$	$MAE\downarrow$	$F_{\beta} \uparrow$	$E_m \uparrow$	mDice ↑	mIoU ↑
EGNet [60]	ResNet-50	111.7	-	61.9	29.8	4.4	35.0	66.6	32.4	24.3
BASNet [37]	ResNet-50	87.1	-	63.9	34.9	5.4	39.4	77.3	39.3	29.3
CPD [48]	ResNet-50	47.9	-	62.2	28.9	4.9	35.7	66.7	33.0	23.9
PraNet [11]	ResNet-50	32.6	-	62.9	35.2	4.2	39.7	76.3	37.8	29.0
SINet [10]	ResNet-50	48.9	-	63.6	34.6	4.1	39.5	77.5	38.1	28.3
SINet-V2 [12]	Res2Net-50	27.0	-	65.3	38.2	3.9	43.2	76.2	41.3	31.8
PNS-Net [16]	ResNet-50	142.9	-	65.5	32.5	4.8	41.7	67.3	38.4	29.0
RCRNet [52]	ResNet-50	53.8	-	62.7	28.7	4.8	32.8	66.6	30.9	22.9
MG [53]	VGG	4.8	-	59.4	33.6	5.9	37.5	69.2	36.8	26.8
SLT-Net-LT [7]	PVTv2-B5	82.3	-	69.6	48.1	3.0	52.4	84.5	49.3	40.2
ZoomNeXt [33]	PVTv2-B5	84.8	-	75.7	59.3	2.0	63.1	86.5	59.9	51.0
SAM2 [39]	Hiera-T	38.9	1-click	75.7	58.3	3.3	62.2	81.4	59.2	48.9
CamSAM2	Hiera-T	39.4	1-click	77.1	62.2	3.2	68.1	83.9	62.6	50.7
SAM2 [39]	Hiera-T	38.9	box	85.4	77.3	1.7	79.5	95.1	77.8	66.7
CamSAM2	Hiera-T	39.4	box	87.2	79.5	1.3	81.4	96.3	79.6	69.2

Table 3. Comparisons between our method and existing approaches on CAD. Our approach clearly outperforms existing methods.

Model	$S_m \uparrow$	$F^{\omega}_{\beta}\uparrow$	$E_m \uparrow$	mDice ↑			
SUN-SEG-Easy							
SAM2 [39] CamSAM2	83.4 88.3	71.6 82.6	83.0 93.4	73.6 84.3			
SUN-SEG-Hard							
SAM2 [39] CamSAM2	75.5 86.4	58.4 78.2	73.4 91.2	61.0 80.6			

Table 4. Comparisons between CamSAM2 and SAM2 on SUN-SEG-Easy and SUN-SEG-Hard.

Decam. Token	IOF	EOF	OPG	mDice ↑	mIoU ↑
				52.1	44.8
\checkmark				54.9	47.0
\checkmark	\checkmark			55.2	47.5
\checkmark	\checkmark	\checkmark		55.9	47.9
\checkmark	\checkmark	\checkmark	\checkmark	64.3	54.6

Table 5. **Ablation study on the effectiveness of main components of CamSAM2.** It shows the effectiveness of each key component of CamSAM2.

raises mDice to 55.2 and mIoU to 47.5. Using EOF brings mDice to 55.9 and mIoU to 47.9. With all components included, the model performs the best, achieving 64.3 on mDice and 54.6 on mIoU.

Effect of Distance Metric. We compare different distance metrics for k-means clustering in OPG, as shown in Tab. 6. Cosine distance performs better than Euclidean distance, likely due to its effectiveness in grouping camouflaged features by angular relationships rather than direct distances.

Influence of Number of Prototypes k**.** We examine the impact of the number of prototypes k, as shown in Tab. 7. The results show that both fewer or higher numbers

Distance Metric	mDice↑	mIoU↑
Euclidean	61.9	52.7
Cosine	64.3	54.6

Table 6. Impact of using different distance metrics for *k*-means in Object Prototype Generation. Cosine distance shows superiority.

# Prototypes (k)	mDice ↑	mIoU ↑
3	60.2	51.8
5	64.3	54.6
7	60.6	50.8

Table 7. Impact of using different number of prototypes in Object Prototype Generation.

of prototypes will reduce the performance due to *under-representation* or *redundancy*, respectively. It is observed that k=5 is found to be optimal for capturing essential informative details in camouflaged features.

5. Conclusion

In this paper, we introduce the CamSAM2, by equipping SAM2 with the ability to accurately segment and track the camouflaged objects for VCOS. While SAM2 demonstrates strong performance across general segmentation tasks, its performance on VCOS is suboptimal due to a lack of feature optimization and architectural support for considering the challenges of VCOS. To overcome the limitations, we propose to add a learnable decamouflaged token to optimize SAM2's features for VCOS, as well as three key modules: IOF for enhancing memory-conditioned features with implicitly object-aware high-resolution features, EOF for refining features with explicit object details, and OPG for abstracting high-quality features within the object region into informative object prototypes. Our experiments on three popular benchmarks of two camouflaged scenarios demonstrate that CamSAM2 clearly improves VCOS performance over SAM2, especially with point prompts, while fully inheriting SAM2's zero-shot capability. By setting new state-of-the-art performance, CamSAM2 offers a more practical and effective solution for real-world VCOS applications.

References

- Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In CVPR, 2009. 6
- [2] Yevgeny Beiderman, Mina Teicher, Javier Garcia, Vicente Mico, and Zeev Zalevsky. Optical technique for classification, recognition and identification of obscured objects. *Optics communications*, 283(21):4274–4282, 2010. 2
- [3] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, 2016. 2, 5
- [4] Pia Bideau, Erik Learned-Miller, Cordelia Schmid, and Karteek Alahari. The right spin: Learning object motion from rotation-compensated flow fields. *IJCV*, 132(1):40–55, 2024.
- [5] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weaklysupervised camouflaged object detection. arXiv preprint, 2024. 2
- [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, 2024. 3
- [7] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In CVPR, 2022. 2, 5, 6, 8
- [8] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017. 6
- [9] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018. 6
- [10] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In CVPR, 2020. 2, 6, 8
- [11] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, 2020. 6, 8
- [12] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 2021. 2, 6, 8
- [13] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Daguang Xu, and Wenqi Li. A short review and evaluation of sam2's performance in 3d ct image segmentation. *arXiv preprint*, 2024. 3

- [14] Wenjun Hui, Zhenfeng Zhu, Guanghua Gu, Meiqin Liu, and Yao Zhao. Implicit-explicit motion learning for video camouflaged object detection. *IEEE TMM*, 2024. 3
- [15] Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. Endow sam with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection. In CVPR, 2024. 2, 3
- [16] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, 2021. 6, 8
- [17] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. 2, 5
- [18] Ch Kavitha, B Prabhakara Rao, and A Govardhan. An efficient content based image retrieval using color and texture of image sub blocks. *IJEST*, 3(2):1060–1068, 2011. 2
- [19] 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 *ICCV*, 2023. 2, 3
- [20] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil D. B. Bruce, Richard P. Wildes, and Konstantinos G. Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In CVPR, 2022. 2, 3
- [21] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In ACCV, 2020. 5
- [22] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In ICCV, 2023. 3
- [23] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and im*age understanding, 184:45–56, 2019. 2
- [24] Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. arXiv preprint, 2024. 3
- [25] Zelin Lu, Liang Xie, Xing Zhao, Binwei Xu, Haoran Liang, and Ronghua Liang. A weakly-supervised cross-domain query framework for video camouflage object detection. *IEEE TCSVT*, 2024. 2, 3
- [26] Yunqiu Lyu, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In CVPR, 2021. 2
- [27] Mobina Mansoori, Sajjad Shahabodini, Jamshid Abouei, Konstantinos N. Plataniotis, and Arash Mohammadi. Polyp sam 2: Advancing zero shot polyp segmentation in colorectal cancer detection. arXiv preprint, 2024. 3
- [28] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In CVPR, 2014. 6
- [29] Muhammad Nawfal Meeran, Bhanu Pratyush Mantha, et al. Sam-pm: Enhancing video camouflaged object detection using spatio-temporal attention. In *CVPRW*, 2024. 2, 3, 5, 6

- [30] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal endoscopy, 93(4):960–967, 2021. 5
- [31] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003. 5
- [32] Feiyu Pan, Hao Fang, Runmin Cong, Wei Zhang, and Xi-ankai Lu. Video object segmentation via sam 2: The 4th solution for Isvos challenge vos track. arXiv preprint, 2024.
- [33] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE TPAMI*, 2024. 2, 6, 8
- [34] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. *arXiv preprint*, 2024. 2
- [35] 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. NeurIPS, 2019. 5
- [36] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In CVPR, 2012. 6
- [37] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundaryaware salient object detection. In CVPR, 2019. 6, 8
- [38] Osher Rafaeli, Tal Svoray, Roni Blushtein-Livnon, and Ariel Nahlieli. Prompt-based segmentation at multiple resolutions and lighting conditions using segment anything model 2. *arXiv preprint*, 2024. 3
- [39] 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*, 2024. 2, 3, 4, 6, 8
- [40] Arunabha Mohan Roy, Jayabrata Bhaduri, Teerath Kumar, and Kislay Raj. A computer vision-based object localization model for endangered wildlife detection. *Ecological Eco*nomics, Forthcoming, 2022. 1
- [41] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023. 4
- [42] Yiqing Shen, Hao Ding, Xinyuan Shao, and Mathias Unberath. Performance and non-adversarial robustness of the segment anything model 2 in surgical video segmentation. *arXiv preprint*, 2024. 3
- [43] Tomasz Stanczyk and Francois Bremond. Masks and boxes: Combining the best of both worlds for multi-object tracking. arXiv preprint, 2024. 3

- [44] PJ Stephenson. Technological advances in biodiversity monitoring: Applicability, opportunities and challenges. *Current Opinion in Environmental Sustainability*, 45:36–41, 2020.
- [45] Wei Sun, Chengao Liu, Linyan Zhang, Yu Li, Pengxu Wei, Chang Liu, Jialing Zou, Jianbin Jiao, and Qixiang Ye. Dqnet: Cross-model detail querying for camouflaged object detection. arXiv preprint, 2022. 2
- [46] George Tang, William Zhao, Logan Ford, David Benhaim, and Paul Zhang. Segment any mesh: Zero-shot mesh part segmentation via lifting segment anything 2 to 3d. *arXiv* preprint, 2024. 3
- [47] Lv Tang and Bo Li. Evaluating sam2's role in camouflaged object detection: From sam to sam2. *arXiv preprint*, 2024.
- [48] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In CVPR, 2019. 6. 8
- [49] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. *NeurIPS*, 2022. 3
- [50] Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is sam (and flow). In ACCV, 2024. 3
- [51] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. arXiv preprint, 2024. 3
- [52] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *ICCV*, 2019. 6, 8
- [53] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 6, 8
- [54] 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 TPAMI*, 2024. 2
- [55] Andrew Seohwan Yu, Mohsen Hariri, Xuecen Zhang, Mingrui Yang, Vipin Chaudhary, and Xiaojuan Li. Novel adaptation of video segmentation to 3d mri: efficient zero-shot knee segmentation with sam2. *arXiv preprint*, 2024. 3
- [56] Jieming Yu, An Wang, Wenzhen Dong, Mengya Xu, Mobarakol Islam, Jie Wang, Long Bai, and Hongliang Ren. Sam 2 in robotic surgery: An empirical evaluation for robustness and generalization in surgical video segmentation. *arXiv* preprint, 2024. 3
- [57] Zifan Yu, Erfan Bank Tavakoli, Meida Chen, Suya You, Raghuveer Rao, Sanjeev Agarwal, and Fengbo Ren. Tokenmotion: Motion-guided vision transformer for video camouflaged object detection via learnable token selection. In ICASSP, 2024. 3
- [58] Jin Zhang, Ruiheng Zhang, Yanjiao Shi, Zhe Cao, Nian Liu, and Fahad Shahbaz Khan. Learning camouflaged object detection from noisy pseudo label. arXiv preprint, 2024. 2

- [59] Jianwei Zhao, Xin Li, Fan Yang, Qiang Zhai, Ao Luo, Zicheng Jiao, and Hong Cheng. Focusdiffuser: Perceiving local disparities for camouflaged object detection. *arXiv* preprint, 2024. 2
- [60] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, 2019. 6, 8
- [61] Yuli Zhou, Guolei Sun, Yawei Li, Luca Benini, and Ender Konukoglu. When sam2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation. *arXiv preprint*, 2024. 2, 3