Endow SAM with Keen Eyes: Temporal-spatial Prompt Learning for Video Camouflaged Object Detection

Wenjun Hui[1,2], Zhenfeng Zhu[1,2]*, Shuai Zheng[1,2], Yao Zhao[1,2]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology
{wenjunhui, zhfzhu, zs1997, yzhao}@bjtu.edu.cn

Abstract

*The Segment Anything Model (SAM), a prompt-driven foundational model, has demonstrated remarkable performance in natural image segmentation. However, its application in video camouflaged object detection (VCOD) encounters challenges, chiefly stemming from the overlooked temporal-spatial associations and the unreliability of user-provided prompts for camouflaged objects that are difficult to discern with the naked eye. To tackle the above issues, we endow SAM with keen eyes and propose the Temporal-spatial Prompt SAM (TSP-SAM), a novel approach tailored for VCOD via an ingenious prompted learning scheme. Firstly, motion-driven self-prompt learning is employed to capture the camouflaged object, thereby bypassing the need for user-provided prompts. With the detected subtle motion cues across consecutive video frames, the overall movement of the camouflaged object is captured for more precise spatial localization. Subsequently, to eliminate the prompt bias resulting from inter-frame discontinuities, the long-range consistency within the video sequences is taken into account to promote the robustness of the self-prompts. It is also injected into the encoder of SAM to enhance the representational capabilities. Extensive experimental results on two benchmarks demonstrate that the proposed TSP-SAM achieves a significant improvement over the state-of-the-art methods. With the mIoU metric increasing by **7.8%** and **9.6%**, TSP-SAM emerges as a groundbreaking step forward in the field of VCOD.*

1. Introduction

Camouflaged object detection, a pivotal task in computer vision, aims to identify the target object seamlessly blending with its surroundings in images [19, 44]. This task plays an important role in various fields, including surveillance and security [29], wildlife protection [28], and medical image processing [11, 41]. In recent years, single-image camouflaged object detection has made significant progress [16, 19, 45]. However, studies [26, 43] have discerned that observers exhibit heightened perception towards the camouflaged object in motion compared to the static scenarios. This observation has sparked a growing interest in video camouflaged object detection (VCOD) [5].

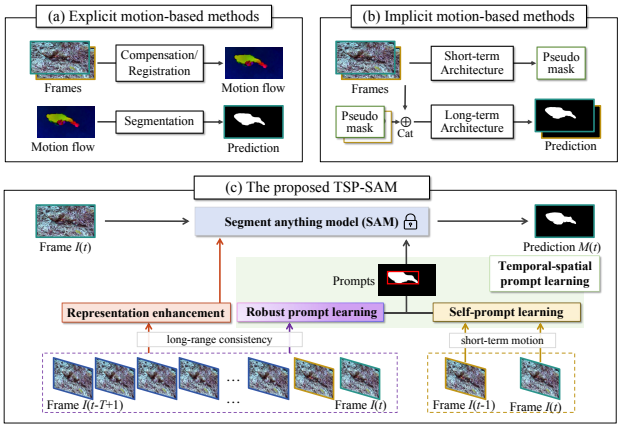

Figure 1. Illustration of previous VCOD methods (a-b) and the proposed TSP-SAM (c). In contrast to previous methods with two-stage architecture, the proposed TSP-SAM is trained in an end-to-end manner, incorporating short-long term temporal-spatial relationships to collaboratively learn the reliable prompts. Meanwhile, the long-range consistency is also used to enhance the representational capability of SAM.

Existing VCOD methods [1, 2, 5, 26] commonly employ a two-step framework to model the inter-frame correlations implicitly or explicitly for camouflage breaking, as shown in Fig. 1(a)(b). Despite yielding promising results, they are noticeably limited in certain aspects:
- **(i)** Primarily, their two-stage architecture exposes them to **cumulative errors**, thus impacting their overall accuracy.
- **(ii)** Moreover, they exhibit **weak generalization ability** due to the limited training data.
Consequently, there is a growing demand for universal models in video camouflaged object detection.

*Corresponding author

With the emergence and development of the visual foundation model, the segment anything model (SAM) [25] has showcased unprecedented performance in the field of natural image segmentation. This accomplishment is attributed to the impressive representations learned on 11 million images in SA-1B [25] and 636 million parameters of ViT-H [8]. SAM not only exhibits heightened versatility in terms of model capacity, but also potentially yields more consistent results across different tasks [30]. Inspired by the remarkable generality of SAM, this work endeavors to make SAM suitable for VCOD.

However, it is worth noting that the performance of SAM is uncertain when applied to VCOD due to the differences between the camouflaged images and natural images. Specifically, SAM, as a prompt-driven foundational model, requires extra prompts when segmenting specific regions. Yet, the seamless blending of the camouflaged object with its surroundings renders it imperceptible to the naked eye, posing a formidable challenge in providing reliable prompts. Hence, there arises an urgent need for adaptive prompt learning to the camouflaged object.

It is well-known that the temporal-spatial interplay in video sequences plays a pivotal role in video-related tasks. Short-term temporal-spatial relationships accentuate the variations between frames, while long-term temporal-spatial correlations model the temporal-spatial context to characterize the dependencies between the target object and the background. Motivated by these considerations, in this work, we develop a flexible end-to-end temporal-spatial prompt SAM, called TSP-SAM, for VCOD. To the best of our knowledge, the proposed TSP-SAM is the first SAM-based camouflaged object detection framework.

Our main contributions can be summarized as follows:

- Instead of user-provided prompts, we propose motion-driven self-prompt learning to capture the camouflaged object. It perceives the overall motion of the camouflaged object by establishing inter-frame associations in the frequency domain, facilitating its spatial identification.
- To eliminate the prompt bias stemming from inter-frame discontinuities, robust prompt learning based on long-range consistency is proposed. By modeling the long-range temporal-spatial dependencies within video sequences, the robustness of self-prompts is well promoted.
- In order to enhance the representational capability of SAM, we propose temporal-spatial injection that incorporates the long-range temporal-spatial consistency into the encoder of SAM. This enhancement contributes to more precise detection within the SAM framework.
- Extensive experiments on MoCA-Mask and CAD2016 datasets demonstrate that the proposed TSP-SAM achieves substantial performance improvements over the state-of-the-art method, with mIoU metric increasing by **7.8%** and **9.6%**.

2. Related work

2.1. Camouflaged object detection

Camouflaged object detection (COD) aims to identify objects blending into their surroundings in images. Existing CNN-based methods address this task with various strategies. Fan *et al.* [10, 12] introduced a two-stage process, involving a search stage for localization followed by a segmentation stage for refinement. In multi-task learning, Yang *et al.* [43] and Ji *et al.* [21] incorporated an auxiliary task to derive shared context representations for enhancing the COD task. Other strategies, such as frequency cues, have also been explored. These works [6, 45] combined the static frequency spectrum with the spatial representations to discern the subtle differences between the camouflaged object and its surroundings. In contrast to the static frequency perception, our TSP-SAM perceives the frequency energy variants to depict the implicit motion between frames, facilitating the spatial identification of the camouflaged object.

2.2. Video camouflaged object detection

Compared to the static appearance of images, the motion between video frames is deemed a significant clue for breaking camouflage. Existing video camouflaged object detection methods [1, 2, 5, 26] implicitly or explicitly modeled the temporal-spatial correlations through a two-stage architecture to detect camouflaged objects. Explicit motion-based methods [1, 2, 26] segmented the optical flow field after compensation or registration into object and background. In contrast, implicit feature correspondences between frames are learned by the neural network for an effective alignment [5]. Nonetheless, their two-stage architecture leads to cumulative errors. Different from the two-stage architecture, our TSP-SAM is trained in an end-to-end manner, incorporating the short-long term temporal-spatial relationships to collaboratively learn reliable prompts for SAM.

2.3. Segment anything model

The segment anything model (SAM) [25], a prompt-driven large-scale foundation model, has exhibited unprecedented performance in natural image segmentation. The core components of SAM comprise an image encoder, a flexible prompt encoder and a lightweight mask decoder. Despite exhibiting excellent zero-shot segmentation capability, SAM performs unstable in application-specific tasks [4]. It fails to segment medical images [18], camouflaged objects [37] and concealed scenes [22]. To apply SAM to medical images, MedSAM [30] and SAM-Adapter [4, 40] incorporated specific domain knowledge into SAM. Yet, up to now, there has been no work on adapting SAM to be tailored for camouflaged object detection. To fill this gap, TSP-SAM explores the temporal-spatial relationships, presenting the first SAM-based camouflaged object detection framework.

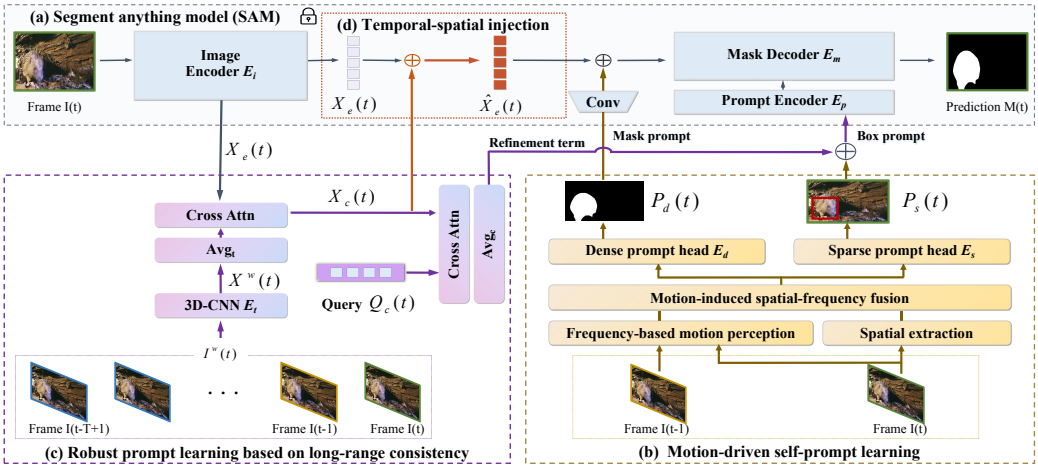

Figure 2. The framework of the proposed TSP-SAM. In motion-driven self-prompt learning (b), the implicit motion between consecutive frames is perceived to facilitate the spatial identification of the camouflaged object, thereby learning the dense self-prompt $P_d(t)$ and sparse self-prompt $P_s(t)$. Subsequently, in robust prompt learning (c), the long-range consistency $X_c(t)$ is modeled to refine the sparse self-prompt $P_s(t)$. Moreover, in temporal-spatial injection (d), the long-range consistency $X_c(t)$ is injected into the image embedding $X_e(t)$ of SAM for representation enhancement. During the training phase, SAM (a) is frozen, including the image encoder E_i, prompt encoder E_p and the mask decoder E_m.

3. Overview of the proposed TSP-SAM

3.1. Notations

In the subsequent sections, we establish the key notations for clarity and consistency in our discourse. Given a sequence of camouflaged images, denote by $I(t) \in \mathbb{R}^{H\times W\times 3}$ the t-th frame and $I^w(t) = [I(t-T+1), \cdots, I(t)] \in \mathbb{R}^{T\times H\times W\times 3}$ the corresponding windowed sequence containing T frames. For a tensor $A \in \mathbb{R}^{H\times W\times C}$, $\mathrm{Mat}(A) \in \mathbb{R}^{HW\times C}$ denotes its matrixization by spatial-wise concatenation. For a matrix $B \in \mathbb{R}^{HW\times C}$, $\mathrm{Ten}(B) \in \mathbb{R}^{H\times W\times C}$ represents its tensorization by spatial-wise division.

3.2. Framework

Fig. 2 illustrates the framework of the proposed TSP-SAM. The core components consist mainly of the following three modules:

(a) Motion-driven self-prompt learning. The motion-driven self-prompt learning is to use the implicit inter-frame motion $\hat{X}_{(t-1)\to t}$ in the frequency domain to facilitate the spatial identification of the camouflaged object, thereby learning the self-prompts for SAM.

(b) Robust prompt learning based on long-range consistency. To eliminate the prompt bias stemming from underlying inter-frame discontinuities, the long-range temporal-spatial consistency $X_c(t)$ in video sequences is modeled to promote the robustness of the self-prompts.

(c) Temporal-spatial injection for representation enhancement. To enhance the representational capabilities of SAM, the long-range temporal-spatial consistency $X_c(t)$ is injected into the image embedding $X_e(t)$ of SAM, contributing to more precise detection.

4. Methodology

In this section, we detail the proposed TSP-SAM method for video camouflaged object detection.

4.1. Motion-driven self-prompt learning

It is well-known that SAM requires user-provided prompts when segmenting specific regions. Nevertheless, the difficulty in distinguishing the camouflaged object with the naked eye makes the requirement of providing visual prompts a significant challenge. Consequently, a compelling need arises to devise a prompt learning network, capable of adaptively identifying the camouflaged object.

4.1.1 Frequency-based motion perception

Inspired by the two-stream hypothesis in neuroscience [14], which highlights the importance of motion in visual perception, we consider the motion as a crucial clue for breaking camouflage. Building on this premise, several studies [7, 23] indicate the motion between frames can be expressed exhaustively in the frequency domain through the variants in frequency energy. Hence, frequency-based motion perception is designed to facilitate the spatial identification of the camouflaged object.

Specifically, as shown in Fig. 3, given consecutive video frames $I(t-1) \in \mathbb{R}^{H\times W\times 3}$ and $I(t) \in \mathbb{R}^{H\times W\times 3}$, we first map them into the frequency domain. Following [45], the image features are divided into a set of $s \times s$ patches and each patch is processed by Discrete Cosine Transform

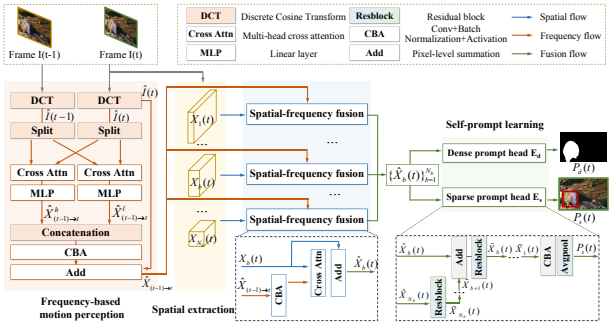

Figure 3. The framework of the motion-driven self-prompt learning. The spatial representations $\{X_b(t)\}_{b=1}^{N_b}$ are extracted by PVT [39], where N_b is the number of attention blocks.

(DCT) into frequency spectrum, where s is the size of patches. To group all components of the same frequency into one channel, we flatten and reshape the spectrum of these patches, forming the frequency features $\hat{I}(t-1) \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 3s^2}$ and $\hat{I}(t) \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 3s^2}$.

In order to perceive the overall motion of the camouflaged object, the inter-frame associations are established on high frequency and low frequency bands respectively. Specifically, the frequency spectrum $\{\hat{I}(t-1), \hat{I}(t)\}$ is divided into two groups in the channel dimension, namely low frequency bands $\{\hat{I}^l(t-1), \hat{I}^l(t)\}$ and high frequency bands $\{\hat{I}^h(t-1), \hat{I}^h(t)\}$. Hence, we have:

$$\hat{X}^l_{(t-1)\to t} = \mathrm{MLP}(\mathrm{ATTN}(\mathrm{Mat}(\hat{I}^l(t-1)), \mathrm{Mat}(\hat{I}^l(t)))) \quad (1)$$

$$\hat{X}^h_{(t-1)\to t} = \mathrm{MLP}(\mathrm{ATTN}(\mathrm{Mat}(\hat{I}^h(t-1)), \mathrm{Mat}(\hat{I}^h(t)))) \quad (2)$$

where $\hat{X}^l_{(t-1)\to t} \in \mathbb{R}^{\frac{HW}{s^2}\times\frac{3s^2}{2}}$ and $\mathrm{MLP}(\cdot)$ is the linear layer. $\mathrm{ATTN}(\cdot,\cdot)$ is multi-head cross attention [38] to extract the inter-frame correspondences, in which the $(t-1)$-th frame is taken as query and the t-th frame serves as key and value. Note that Eq.1 tends to highlight the global deformations of the camouflaged object by characterizing the feature correspondences on the low-frequency bands. Meanwhile, Eq.2 implicitly reveals the local edge motion of the camouflaged object on high-frequency bands.

Subsequently, the complete implicit motion in the frequency domain is obtained by:

$$\hat{X}_{(t-1)\to t} = \hat{I}(t) + \mathrm{CBA}(\mathrm{Cat}(\mathrm{Ten}(\hat{X}^l_{(t-1)\to t}), \mathrm{Ten}(\hat{X}^h_{(t-1)\to t}))) \quad (3)$$

where $\mathrm{CBA}(\cdot)$ is the convolution operation followed by batch normalization and activation and $\mathrm{Cat}(\cdot)$ is the channel-wise concatenation. Importantly, $\hat{X}_{(t-1)\to t} \in \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times 3s^2}$ implicitly depicts the overall motion of the camouflaged object by perceiving the variants in the frequency energy, enabling to facilitate the discovery of the camouflaged object.

4.1.2 Implicit motion-induced spatial-frequency fusion

To translate the perceived motion into the spatial domain, the implicit motion-induced spatial-frequency representation fusion is established. This process involves the integration between the implicit motion $\hat{X}_{(t-1)\to t}$ in the frequency domain and the spatial representations $X_b(t) \in \mathbb{R}^{\frac{H}{2^{b+1}}\times\frac{W}{2^{b+1}}\times C}, b \in \{1,\cdots,N_b\}$ extracted by PVT [39], where N_b is the number of attention blocks. It can be expressed as follows:

$$\hat{X}_b(t) = X_b(t) + \mathrm{Ten}(\mathrm{ATTN}(\mathrm{Mat}(\mathrm{CBA}(\hat{X}_{(t-1)\to t})), \mathrm{Mat}(X_b(t)))) \quad (4)$$

where $\hat{X}_b(t) \in \mathbb{R}^{\frac{H}{2^{b+1}}\times\frac{W}{2^{b+1}}\times C}$. Note that the spatial perception of the camouflaged object could be strengthened by establishing cross attention between the implicit motion as query and the spatial representation as key and value.

4.1.3 Self-prompt learning

Based on the co-representations $\{\hat{X}_b(t)\}_{b=1}^{N_b}$, two prompt heads are adopted to generate the dense self-prompt $P_d(t)$ and sparse self-prompt $P_s(t)$, respectively. Specifically, we adopt the decoder of SLTNet [5] as dense prompt head E_d as follows:

$$P_d(t) = E_d(\{\hat{X}_b(t)\}_{b=1}^{N_b}) \quad (5)$$

where $P_d(t)$ is the dense self-prompt. Meanwhile, as illustrated in Fig. 3, to obtain the sparse prompt, we adopt the residual block of Resnet [17] to process the co-representations in a bottom-up manner:

$$\tilde{X}_b(t) = \begin{cases} \mathrm{Resblock}(\hat{X}_b(t)) & b = N_b \\ \mathrm{Resblock}(\hat{X}_b(t) + \tilde{X}_{b+1}(t)) & b \in \{1,\cdots,N_b-1\} \end{cases} \quad (6)$$

Finally, the sparse self-prompt is represented as follows:

$$P_s(t) = \mathrm{Avgpool}(\mathrm{CBA}(\tilde{X}_1(t))) \quad (7)$$

where $\mathrm{Avgpool}(\cdot)$ is global average pooling operation.

4.2. Robust prompt learning based on long-range consistency

The self-prompts of the camouflaged object are well learned by using the perceived inter-frame motion to facilitate its spatial identification. However, the susceptibility of inter-frame motion to underlying temporal-spatial discontinuities, such as occlusions, shaking, scene changes, etc., brings a potential lack of reliability in the learned self-prompts. It is well known that the long-range dependencies within the video sequences enable to characterize the temporal-spatial context to cope with the short-term temporal-spatial discontinuities. Hence, it is necessary to explore the long-range temporal-spatial consistency within the video sequences to improve the robustness of the self-prompts, which likes endowing SAM with keen eyes.

4.2.1 Long-range consistency squeezing

As illustrated in Fig. 2(c), given the windowed sequence $I^w(t) \in \mathbb{R}^{T\times H\times W\times 3}$ corresponding to the current frame $I(t)$, we adopt a 3D convolution neural network E_t to extract the long-range temporal-spatial relationships:

$$X^w(t) = E_t(I^w(t)) \tag{8}$$

where $X^w(t) \in \mathbb{R}^{T\times \frac{H}{4}\times \frac{W}{4}\times C}$ fully reveals the temporal-spatial dependencies within the video sequences. In our work, we adopt 3D Resnet18 [15] as E_t.

Subsequently, to further extract the long-range temporal-spatial consistency of the camouflaged object, we take the camouflage representation $X_e(t)$ from the encoder of SAM as query and the long-range temporal-spatial dependencies $X^w(t)$ averaged along the temporal dimension as key and value to establish the cross attention:

$$X_c(t) = \mathrm{ATTN}(\mathrm{Mat}(X_e(t)), \mathrm{Mat}(\mathrm{Avg_t}(X^w(t)))) \tag{9}$$

where $\mathrm{Avg_t}(\cdot)$ is an average operation to squeeze the temporal dimension. $X_c(t) \in \mathbb{R}^{\frac{HW}{16}\times C}$ characterizes the long-range consistency specific to the camouflaged object.

4.2.2 Robust prompt learning

To enhance the robustness of the self-prompts to inter-frame temporal-spatial discontinuities, the long-range temporal-spatial consistency $X_c(t)$ should be induced into the prompt learning. Specifically, in order to map the long-range consistency into the coordinate space, a learnable query $Q_c(t) \in \mathbb{R}^{4\times C}$ is defined to establish cross attention with $X_c(t)$. It is formulated as:

$$P_r(t) = \mathrm{Avg_c}(\mathrm{ATTN}(Q_c(t), X_c(t))) \tag{10}$$

where $\mathrm{Avg_c}(\cdot)$ is an average operation along the channel dimension. $P_r(t) \in \mathbf{R}^4$ is a refinement term derived from long-range consistency to correct the sparse self-prompt:

$$P_s^r(t) = P_s(t) + P_r(t) \tag{11}$$

Through incorporating long-range consistency into the prompt learning, the visual prompts could be more robust to the temporal-spatial discontinuities, which means endowing SAM with a pair of keen eyes.

4.3. Temporal-spatial injection for representation enhancement

It is known that the remarkable performance of SAM relies not only on well learned robust prompts but also on the representational capability of the image encoder. However, due to the fact that the image embedding of SAM lacks the temporal-spatial information, the final predictions of consecutive frames may be inconsistent. Hence, it is necessary to inject the long-range temporal-spatial consistency into SAM. As shown in Fig. 2(d), it is represented as follows:

$$\hat{X}_e(t) = X_e(t) + \mathrm{Ten}(X_c(t)) \tag{12}$$

where $X_c(t)$ is the long-range consistency from Eq. 9. In this way, based on the enhanced image representations and the robust prompts called keen eyes, SAM achieves more precise camouflaged object detection in videos.

4.4. Model optimization

We train the proposed TSP-SAM in an end-to-end manner by minimizing the joint loss below:

$$\mathcal{L}(t) = \mathcal{L}_{SAM}(t) + \alpha\mathcal{L}_p^s(t) + \beta\mathcal{L}_p^d(t) \tag{13}$$

where $\mathcal{L}_{SAM}$, $\mathcal{L}_p^s$ and $\mathcal{L}_p^d$ supervise the final prediction, the sparse prompt and the dense prompt, respectively.

Following [3], the supervision on the sparse prompt $\mathcal{L}_p^s(t)$ is as follows:

$$\mathcal{L}_p^s(t) = \mathrm{BCE}(P_s^r(t), G(t)) + \mathrm{MSE}(P_s^r(t), G(t)) \tag{14}$$

where BCE is binary cross-entropy loss to supervise the coordinates and MSE is mean square error to supervise the size of the box prompt or the scale of the point prompt. For the dense prompt loss $\mathcal{L}_p^d$ and the final prediction loss $\mathcal{L}_{SAM}$, we adopt the hybrid loss [13] for supervision:

$$\mathcal{L}_{hybrid} = \mathcal{L}_{ce}^w + \mathcal{L}_{iou}^w \tag{15}$$

where $\mathcal{L}_{ce}^w$ is the weighted cross-entropy loss and $\mathcal{L}_{iou}^w$ is the weighted intersection-over-union loss.

5. Experiment results and analysis

5.1. Experiment settings

Datasets. Following previous methods [5], we have conducted a comprehensive evaluation on two widely used datasets: MoCA-Mask [5] and CAD2016 [1]. The MoCA-Mask dataset [5] is a video camouflaged object detection dataset, covering camouflaged animals moving in natural scenes. It contains 87 video sequences (22939 frames), of which 71 video sequences are used for training and 16 sequences for testing. It is well annotated with bounding boxes and pixel-level segmentation masks on every fifth frames. Meanwhile, Camouflaged Animal Dataset [1] (CAD2016) is a small video camouflaged object detection dataset extracted from YouTube videos. It consists of nine short video clips in total and is well annotated with pixel-level semantic masks on every fifth frames.

Metrics. During the testing phase, we assess the prediction masks using six widely-used metrics: S-measure (S_α) [9], Weighted F-measure (F_β^w) [31], Enhanced-alignment measure (E_ϕ) [13], Mean absolute error (MAE, $\mathcal{M}$) [33],

Method	Year	Input	MoCA-Mask						CAD2016					
			$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	mDice↑	mIoU↑	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	mDice↑	mIoU↑
SINet[10]	2020-CVPR	Image	0.574	0.185	0.655	0.030	0.221	0.156	0.601	0.204	0.589	0.089	0.289	0.209
SINet-v2[12]	2021-TPAMI	Image	0.571	0.175	0.608	0.035	0.211	0.153	0.544	0.181	0.546	0.049	0.170	0.110
ZoomNet[32]	2022-CVPR	Image	0.582	0.211	0.536	0.033	0.224	0.167	0.587	0.225	0.594	0.063	0.246	0.166
BGNet [36]	2022-IJCAI	Image	0.590	0.203	0.647	0.023	0.225	0.167	0.607	0.203	0.666	0.089	0.345	0.256
FEDERNet[16]	2023-CVPR	Image	0.555	0.158	0.542	0.049	0.192	0.132	0.607	0.246	0.725	0.061	0.361	0.257
FSPNet[19]	2023-CVPR	Image	0.594	0.182	0.608	0.044	0.238	0.167	0.539	0.220	0.553	0.145	0.309	0.212
PUENet[44]	2023-TIP	Image	0.594	0.204	0.619	0.037	0.302	0.212	0.673	0.427	0.803	0.034	0.499	0.389
RCRNet[42]	2019-ICCV	Video	0.597	0.174	0.583	0.025	0.194	0.137	-	-	-	-	-	-
PNS-Net[20]	2021-MICCAI	Video	0.576	0.134	0.562	0.038	0.189	0.133	0.678	0.369	0.720	0.043	0.409	0.308
MG[43]	2021-ICCV	Video	0.547	0.165	0.537	0.095	0.197	0.141	0.484	0.314	0.558	0.370	0.351	0.260
SLT-Net[5]	2022-CVPR	Video	0.656	0.357	0.785	0.021	0.387	0.310	0.679	0.420	0.805	0.033	0.445	0.342
TSP-SAM(M+P)	Ours	Video	0.673	0.400	0.766	0.012	0.421	0.345	0.681	0.500	0.853	0.031	0.496	0.393
TSP-SAM(M+B)	Ours	Video	0.689	0.444	0.808	0.008	0.458	0.388	0.704	0.524	0.912	0.028	0.543	0.438

Table 1. Quantitative comparisons over two benchmark datasets. The top three results are highlighted in red, green, and blue. M+P: Combination of mask and point prompts. M+B: Combination of mask and box prompts.

Method	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	mDice↑	mIoU↑
SINet[10]	0.636	0.346	0.775	0.041	0.381	0.283
SINet-v2[12]	0.653	0.382	0.762	0.039	0.413	0.318
ZoomNet[32]	0.633	0.349	0.601	0.033	0.349	0.273
FEDERNet[16]	0.573	0.272	0.609	0.051	0.260	0.199
FSPNet[19]	0.681	0.401	0.716	0.044	0.446	0.332
PUENet[44]	0.691	0.485	0.795	0.034	0.514	0.396
RCRNet[42]	0.627	0.287	0.666	0.048	0.309	0.229
PNS-Net[20]	0.655	0.325	0.673	0.048	0.384	0.290
MG[43]	0.594	0.336	0.691	0.059	0.368	0.268
SLT-Net[5]	0.696	0.481	0.845	0.030	0.493	0.401
TSP-SAM(M+P)	0.705	0.565	0.836	0.027	0.531	0.422
TSP-SAM(M+B)	**0.751**	**0.628**	**0.851**	**0.021**	**0.603**	**0.496**

Table 2. Generalization comparisons on CAD2016 dataset. The best results are highlighted in **bold**. M+P: Combination of mask and point prompts. M+B: Combination of mask and box prompts.

meanDice (mDice), meanIoU (mIoU). Overall, a better VCOD method has larger S_α, F_β^w, E_ϕ, mDice and mIoU scores, but a smaller MAE score.

Implementation details. In our experiments, we freeze SAM and train our TSP-SAM in an end-to-end manner with a maximum of 20 epochs. Specifically, for a fair comparison, both the training and testing images are resize to 352×352. We leverage Adam [24] as our optimizer and set the weight decay to 1×10^{-5}. On the MoCA-Mask dataset, the learning rate is initialized to 1×10^{-5} for the spatial extractor of motion-driven self-prompt learning and 5×10^{-5} for other trainable parameters. After 10 epochs, the learning rate is reduced by 90%. On the CAD2016 dataset, we adopt a ratio of 9:1 to divide all classes into training and testing sets and report the mean value over the testing set. We set the initial learning rate to 1×10^{-4} for the spatial extractor of the motion-driven self-prompt learning and 5×10^{-4} for other trainable parameters. The other settings are the same as those on MoCA-Mask dataset.

Method	Tuning parameters (M)
FSPNet	274.24
SLT-Net	82.41
TSP-SAM(ours)	89.78

Table 3. The number of tuning parameters.

Method	Time (s) ↓	Speed (fps) ↑
FSPNet	568.70	1.31
SLT-Net	253.40	2.94
Overall (ours)	294.99	2.53
Prompt learning	73.28	10.17
SAM	221.71	3.36

Table 4. Comparisons on inference time and speed.

5.2. Performance comparison

1) Baselines: To evaluate the performance of the proposed TSP-SAM, we compare it with a range of state-of-the-art methods. These methods fall into two categories:

- **Single-image camouflaged object detection.** These methods analyze the static appearance of images to discern the subtle differences between camouflaged objects and their surroundings. The compared methods in this type include SINet [10], SINet-v2 [12], ZoomNet [32], BGNet [36], FEDERNet [16], FSPNet[19], and PUENet [44].
- **Video object segmentation.** This set of methods explores the temporal-spatial relationships between frames for the purpose of segmenting objects with specific attributes. The compared methods in this type encompass RCRNet [42], PNS-Net [20], MG[43], and SLT-Net[5].

2) Quantitative Results: We evaluate TSP-SAM on MoCA-Mask and CAD2016 datasets. From Table 1, we can conclude that, (i) compared to the combination of mask and point prompts, the combination of mask and box prompts is more reliable for SAM. This is attributed to the limitations of the point prompt in conveying the boundary information. In contrast, bounding box excels in indicating the boundaries of the camouflaged object, even in the presence of bias. (ii) the strong contrast between the performance of TSP-SAM and single-image camouflaged object detection methods points to the importance of temporal-spatial relationships in breaking camouflage. (iii) the proposed TSP-SAM

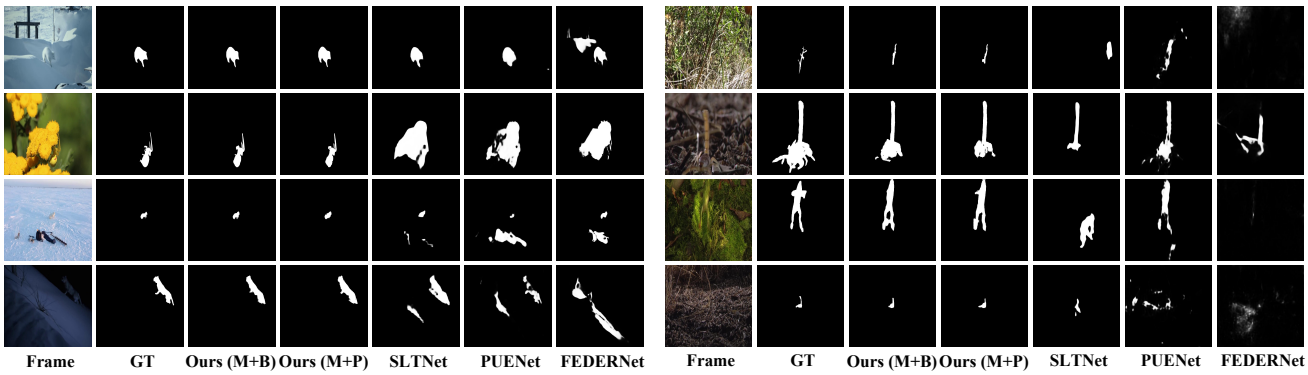

Figure 4. Visualization of our proposed TSP-SAM and baseline methods on MoCA-Mask and CAD2016 dataset. From left to right: frames (1st column), ground truth (2nd column), prediction of our TSP-SAM with mask and box prompts (3rd column), prediction of our TSP-SAM with mask and point prompts (4th column), and predictions of compared methods (5th-7th columns).

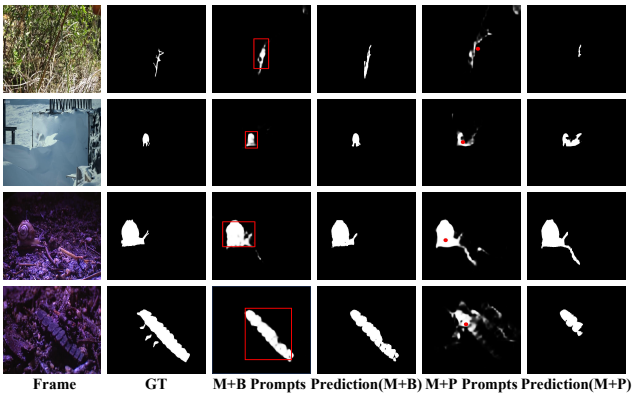

Figure 5. Visualization of several failure cases with mask and point prompts. From left to right: frames (1st column), ground truth (2nd column), mask and box prompts (3rd column), prediction with mask and box prompts (4th column), mask and point prompts (5th column), and prediction with mask and point prompts (6th column).

surpasses all video object segmentation methods. Notably, TSP-SAM exhibits a 1.3% - 8.7% improvement over SLT-Net [5] on MoCA-Mask dataset, meaning that the temporal-spatial exploration in TSP-SAM is more effective.

To assess the generalization performance, we train on the MoCA-Mask dataset and report the test performance on the CAD2016 dataset. It can be observed from Table 2 that the proposed TSP-SAM outperforms all baselines. Compared with SLTNet [5], TSP-SAM surpasses it by 0.9%-14.7%, indicating that TSP-SAM is capable of generalizing to the unseen data more effectively. Meanwhile, we also test the performance of TSP-SAM on general video object segmentation datasets. On DAVIS2016 dataset [34], our model achieves comparable performance to the SOTA SLT-Net. While on SegTrack-v2 dataset [27] with multiple objects, our method obtains 62.84% for mIoU, surpassing SLT-Net by 9.9%. Surprisingly, our model is even comparable to the recent SMTC [35] on SegTrack-v2.

Besides, we report the number of tuning parameters, the inference time and speed. From Table 3, compared with SLT-Net, our model achieves 9.6% improvements with a slight parameter increase. From Table 4, our model is comparable to the latest video-oriented SLT-Net. Decomposing the overall inference time, SAM consumes 75% of it.

3) Qualitative Results: To provide more intuitive evaluations of the proposed TSP-SAM, we conduct two qualitative discussions. Firstly, as shown in Fig. 4, we present the segmentation results of several examples to intuitively compare the proposed TSP-SAM with the baselines. It can be seen that: (i) some confusing surrounding regions interfere with the baseline methods, while our TSP-SAM is more robust in identifying the camouflaged object. This demonstrates the necessity of short-long term temporal-spatial relationships. (ii) compared with the baselines, our TSP-SAM exhibits superior capabilities in boundary segmentation in terms of discriminability and details.

Secondly, to further analyze the influence of the type of sparse prompts on overall performance, we visualize the several failure cases with mask and point prompts. From Fig. 5, it can be found that: (i) since the target object is not always a regular shape, the learned point prompt is not on the target object although it is near the center. (ii) under the condition with unclear boundaries, the point prompt is more likely to segment the regions of background. (iii) the point prompt may induce the model to segment the local areas instead of the complete target object. In contrast, the box prompt can avoid the above defects and achieve more precise segmentation results. Meanwhile, the complete localization under the learned mask and box prompts verifies the effectiveness of the temporal-spatial prompt learning.

5.3. Ablation study

To gain a comprehensive understanding of the efficacy of core components and the impact of hyper parameters, we delve into a detailed analysis.

Method	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	mDice↑	mIoU↑
SAM+SSP	0.637	0.330	0.768	0.367	0.311
SAM+MSP	0.657	0.353	**0.823**	0.407	0.340
SAM+MSP+LCP	0.665	0.409	0.757	0.429	0.366
SAM+MSP+LCP+TSI	**0.689**	**0.444**	0.808	**0.458**	**0.388**

Table 5. Ablation studies of the core components on MoCA-Mask dataset. SSP: Spatial-based self-prompt learning. MSP: Motion-driven self-prompt learning. LCP: Long-range consistency-based robust prompt learning. TSI: Temporal-spatial injection for representation enhancement.

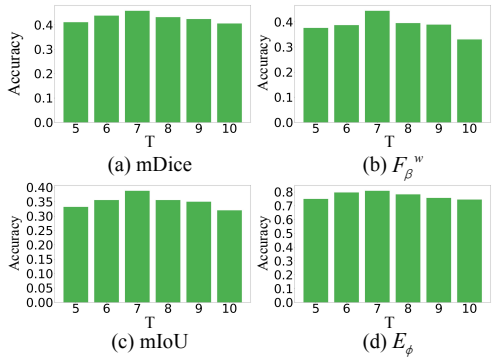

Figure 6. Sensitivity analysis of the size T of the windowed sequence on MoCA-Mask dataset.

Ablation analysis: Table 5 present the segmentation results by progressively applying each part in our model. Compared to the spatial-based self-prompt learning (SSP) comprising a spatial extractor and two prompt heads, motion-driven self-prompt learning (MSP) improves the mDice metric from 36.7% to 40.7% to a large degree. This affirms the effectiveness of the inter-frame motion in camouflage breaking. Subsequently, the long-range consistency-based robust prompt learning (LCP) brings 2.2% gains for mDice, indicating the effectiveness of long-range consistency in improving the robustness of self-prompts. Meanwhile, the temporal-spatial injection (TSI) further boosts the mDice metric by 2.9%, once again verifying the significance of long-range consistency.

To further analyze the effectiveness of frequency-based motion perception in MSP, we replace it with spatial-based motion perception, frequency-based static appearance perception, and spatial-based static appearance perception to conduct the ablation studies. As can be seen from Table 6, the frequency-based motion perception yields the best performance. While verifying the significant role of inter-frame motion in breaking camouflage, it also demonstrates that inter-frame motion can be expressed more exhaustively in the frequency domain than in the spatial domain.

Hyper parameter analysis: To investigate the effect of the size of the windowed sequence on segmentation performance, we conduct the ablation experiments by varying T from $\{5, 6, 7, 8, 9, 10\}$. As illustrated in Fig. 6, the accuracy initially rises with increasing the size of the windowed sequence, reaches the best at $T = 7$ before declining. Moreover, we assess the sensitivity of α and β that are used to balance the sparse and dense prompt loss in Fig. 7. The optimal segmentation performance is observed at $(\alpha, \beta) = (1, 0.6)$.

Method	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	mDice↑	mIoU↑
Spatial appearance	0.637	0.330	0.768	0.367	0.311
Frequency appearance	0.647	0.344	0.793	0.396	0.328
Spatial motion	0.651	0.335	0.730	0.396	0.324
Frequency motion	**0.657**	**0.353**	**0.823**	**0.407**	**0.340**

Table 6. Ablation studies of frequency-based motion perception on MoCA-Mask dataset. Spatial appearance: Spatial-based appearance perception. Frequency appearance: Frequency-based appearance perception. Spatial motion: Spatial-based motion perception. Frequency motion: Frequency-based motion perception.

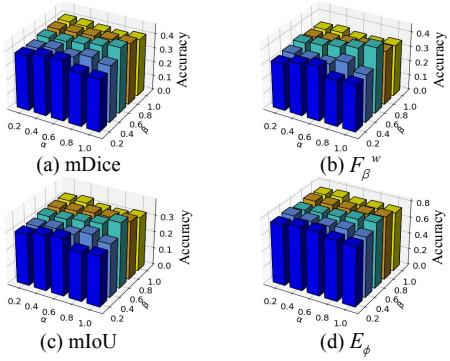

Figure 7. Sensitivity analysis of the balance factor (α, β) in the loss function on MoCA-Mask dataset.

6. Conclusion

In this paper, we propose a flexible end-to-end temporal-spatial prompt SAM (TSP-SAM) for video camouflaged object detection. Instead of user-provided prompts, motion-driven self-prompt learning is proposed to capture the overall motion of the camouflaged object, facilitating its spatial identification. Subsequently, a robust prompt learning based on long-range consistency is proposed to promote the robustness of the self-prompts, which likes endowing SAM with keen eyes. Moreover, the long-range consistency is injected into the encoder of SAM for representation enhancement. Notably, TSP-SAM provides a new perspective, that is temporal-spatial associations, for visual prompt learning.

7. Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No.2021ZD0140407, Beijing Natural Science Foundation under Grant No.7222313, and National High Level Hospital Clinical Research Funding under Grant No.2022-PUMCH-C-041.

References

[1] Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, pages 433–449, 2016. 1, 2, 5

[2] Pia Bideau, Rakesh R Menon, and Erik Learned-Miller. Moa-net: self-supervised motion segmentation. In *ECCV Workshops*, pages 0–0, 2018. 1, 2

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[4] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, 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, and more. *arXiv preprint arXiv:2304.09148*, 2023. 2

[5] 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*, pages 13864–13873, 2022. 1, 2, 4, 5, 6, 7

[6] Runmin Cong, Mengyao Sun, Sanyi Zhang, Xiaofei Zhou, Wei Zhang, and Yao Zhao. Frequency perception network for camouflaged object detection. In *ACM MM*, pages 1179–1189, 2023. 2

[7] Xinyi Cui, Qingshan Liu, and Dimitris Metaxas. Temporal spectral residual: fast motion saliency detection. In *ACM MM*, pages 617–620, 2009. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. 2

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017. 5

[10] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020. 2, 6

[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 *International conference on medical image computing and computer-assisted intervention*, pages 263–273, 2020. 1

[12] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10): 6024–6042, 2021. 2, 6

[13] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6), 2021. 5

[14] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992. 3

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 5

[16] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pages 22046–22055, 2023. 1, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[18] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023. 2

[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 *CVPR*, pages 5557–5566, 2023. 1, 6

[20] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 142–152, 2021. 6

[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. 2

[22] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Ming-Ming Cheng, Bowen Zhou, and Luc Van Gool. Sam struggles in concealed scenes–empirical study on" segment anything". *arXiv preprint arXiv:2304.06022*, 2023. 2

[23] Yang Jia, Jie Yuan, Jinjun Wang, Jun Fang, Qixing Zhang, and Yongming Zhang. A saliency-based method for early smoke detection in video sequences. *Fire technology*, 52: 1271–1292, 2016. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[26] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020. 1, 2

[27] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013. 7

[28] Thomas Lidbetter. Search and rescue in the face of uncertain threats. *European Journal of Operational Research*, 285(3): 1153–1160, 2020. 1

[29] Ting Liu, Yao Zhao, Yunchao Wei, Yufeng Zhao, and Shikui Wei. Concealed object detection for activate millimeter wave image. *IEEE Transactions on Industrial Electronics*, 66(12): 9909–9917, 2019. 1

[30] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[31] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014. 5

[32] 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 *CVPR*, pages 2160–2170, 2022. 6

[33] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 5

[34] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 7

[35] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *ICCV*, pages 16675–16687, 2023. 7

[36] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. *arXiv preprint arXiv:2207.00794*, 2022. 6

[37] Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*, 2023. 2

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4

[39] 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 *ICCV*, pages 568–578, 2021. 4

[40] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2

[41] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE TIP*, 30:3113–3126, 2021. 1

[42] 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*, pages 7284–7293, 2019. 6

[43] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, pages 7177–7188, 2021. 1, 2, 6

[44] Yi Zhang, Jing Zhang, Wassim Hamidouche, and Olivier Deforges. Predictive uncertainty estimation for camouflaged object detection. *IEEE TIP*, 2023. 1, 6

[45] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *CVPR*, pages 4504–4513, 2022. 1, 2, 3