De-Supervision in Camouflaged Videos

Luca Alessandrini, Antonino Maria Rizzo, Luca Magri, Giacomo Boracchi and Federica Arrigoni
DEIB – Politecnico di Milano, Italy
luca1.alessandrini@polimi.it

Abstract

The aim of Zero-shot Video Camouflaged Object Segmentation (ZVCOS) is to automatically separate the foreground subject from the background, in videos where the subject is camouflaged within the environment, without any user intervention. Camouflaged videos represent the most challenging setting for video object segmentation, due to the minimal appearance-based cues available for the camouflaged subject, which closely resembles its surroundings. Consequently, ZVCOS has received limited research attention, primarily due to the scarcity of annotated datasets, with most existing approaches focusing on the supervised scenario. In this paper we introduce a simple but effective framework, named DeSC-V, that operates in an unsupervised manner. On one side, we exploit prior knowledge on the camouflaged subjects' appearance, roughly estimated from an image segmentation network. On the other side, we enhance such prior knowledge by taking advantage of the temporal information coming from close/distant time frames through the Optical Flow, which enforces global coherence among the estimated masks within the video: this allows us to address the challenge of transferring information from images to a video in a principled way. Experimental results on camouflaged datasets show that DeSC-V is effective, outperforming its closest competitor.

1. Introduction

Video Object Segmentation (VOS) [54] is a fundamental problem that aims to separate the *foreground* from the *background* in a video, providing a consistent segmentation across frames. What should be considered as foreground is also referred to as the *subject* of the video. Most research in VOS focuses on the *One-Shot* scenario [2, 14, 16, 17, 28, 52], where the first frame segmentation is manually provided and used at inference time. Similarly, SAM [18] and SAM2 [31] also require a prompt in input to segment the desired subject. In contrast, in the *Zero-Shot* scenario [5], no user interaction is required, making segmentation significantly more challenging, as no prior subject information is available. This field received considerable attention over recent years, thanks to powerful Deep Learning models achieving compelling performances. The availability of diverse datasets – *e.g.*, DAVIS_{16} [30], FBMS [29], YouTube-VOS [42], SegTrackV2 [20], MOSE [7], MOSEv2 [9], MeViS [6], MeViSv2 [8] – has been fundamental, as most approaches [13, 21, 22, 50, 53] typically rely on multiple benchmarks for training.

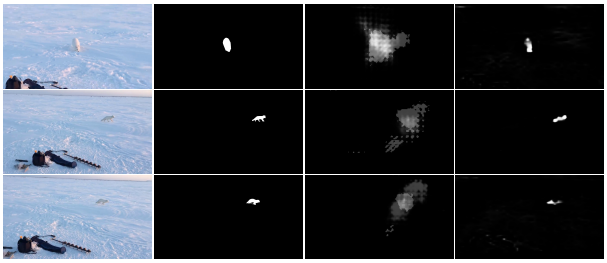

Figure 1. Video Camouflaged Object Segmentation aims to identify camouflaged subjects in a video by estimating a per-frame pixel-wise mask. Sample frames from MoCA-Mask [4] are reported, displaying animals nearly impossible to spot. For each frame the Ground-Truth, results of our closest competitor MG [45], and our results (DeSC-V) are also reported. Reported frames are from sequence arctic_fox_3 (frames 15, 155, 160).

In this work, we address the problem of *Zero-shot Video Camouflaged Object Segmentation* [4], which aims to segment camouflaged subjects in video frames in a Zero-Shot manner. An example of a camouflaged video is the one depicted in Fig. 1, where the subject tends to disguise itself with the surrounding environment. Identifying camouflaged subjects in videos is crucial for several applications [54], such as video editing, autonomous driving, surveillance and medical imaging. Camouflaged Video Object Segmentation inherits all the typical challenges of video segmentation: *i*) occlusions, where the subject is partially or fully obscured by other objects; *ii*) non-rigid deformations of the subject, causing changes in its shape that can hinder consistent segmentation across frames; and *iii*) minimal subject motion or subtle frame-to-frame changes, which make motion cues ineffective to segment the subject from the background. Beyond these, the camouflaged Zero-Shot case represents one of the most challenging test benches for video object seg-

mentation models, as it exacerbates its difficulty by introducing unique challenges. Specifically: *iv*) the scarcity of distinct appearance cues, as camouflaged subjects blend with their environment, and *v*) the insufficiency of extensive labeled datasets, which limits the application of dataset-hungry Zero-Shot methods.

Currently, the only available datasets explicitly tailored for camouflaged object segmentation are MoCA-Mask [4] and the Camouflaged Animal Dataset (CAD) [1], which contain videos of animals that are perhaps the most representative camouflaged subjects. These benchmarks are relatively small in terms of animal variety, and the ground-truth segmentation annotations are sparse, not available for every frame. These limitations undermine the development of novel Deep Learning methods for segmenting camouflaged subjects in videos. All in all, the Zero-Shot *camouflaged* setting is way less explored than the segmentation of general videos. Most methods for Zero-shot Video Camouflaged Object Segmentation consider the supervised setting [4, 15, 27, 41, 51], where video mask annotations are used for training, whereas the unsupervised scenario is highly unexplored and counts only one work [45]. The latter, however, focuses on segmenting *moving* objects, therefore performs poorly in identifying camouflaged animals exhibiting little movement over time. Although the *unsupervised* camouflaged scenario is highly relevant, as it eliminates the need for a large annotated training set of video frames (which are time-consuming and labor-intensive to gather), an effective approach for this setting is still lacking.

1.1. Contribution

In this work, we bridge this gap by introducing a novel approach that operates under the most difficult scenario: *i)* videos will involve camouflage subjects; *ii)* at inference time, no mask is required (zero-shot); *iii)* training is performed without any video mask annotation. Following the terminology from [26], we can view our method as *unsupervised*, as it does not exploit any label in the input (*i.e.*, video) domain during training. However, it is not entirely end-to-end as we use prior knowledge on the appearance of camouflaged subjects, learned from a similar domain, as clarified below. Note that using pre-trained networks [11, 34] as blocks of a modern architecture is a common practice in Deep Learning literature.

Drawing inspiration from the human experience, a subject which can catch attention typically either moves or is easy to spot due to its appearance. This idea motivates the two main components of our loss. First, to be able to exploit the little appearance information coming from camouflaged subjects, we rely on a Frame Expert [39], which is a network pre-trained in a supervised way to segment camouflaged objects in *images*, and we force the predictions of our network to be close to those (fixed) of the Frame Expert: this constitutes the first part of our training loss. Since the Frame Expert operates on images only, disregarding the motion information leads to coarse predictions. Hence, in the second part of our loss, we use the temporal information within the video to enforce global consistency, resulting in improved mask predictions: we transfer information from other frames (both distant and close) to the current one via Optical Flow [34], thereby using motion to generate appearance-based supervision. In other terms, by exploiting frame-to-frame variations, we compensate for the deficiency of a video segmentation model based on appearance only. Note that temporal consistency was already shown to be effective in other Computer Vision tasks [23, 24, 35], but remains underexplored with camouflaged videos.

Our approach can be viewed as a novel *unsupervised video training framework*: starting from a supervised method that operates on images only, it is possible to combine this image-based knowledge with video-based temporal information, without using Ground-Truth (GT) video annotations, effectively "de-supervising" the training process. For this reason, we call the proposed method DeSC-V (De-Supervision in Camouflaged Videos). Our framework largely improves upon the Frame Expert [39] (when evaluated frame by frame over a video), demonstrating the effectiveness of the proposed unsupervised training procedure and its ability to generalize to unseen sequences. We achieve outstanding performance on the CAD [1] dataset. Furthermore, our experiments demonstrate that DeSC-V significantly outperforms its closest competitor – MG [45] – on the challenging MoCA-Mask benchmark [4], setting the state of the art in unsupervised Zero-shot Video Camouflaged Object Segmentation.

2. Problem Formulation

In this paper, we tackle the problem of Zero-shot Video Camouflaged Object Segmentation. However, before formally defining this task and its settings, we clarify relevant terminology that is often used inconsistently in the literature, which can lead to confusion.

***V**ideo **O**bject **S**egmentation* (VOS) is the general task that provides a per-frame segmentation of the input video by identifying salient objects. We refer the reader to the excellent survey [54] for more references and insights on this topic. A VOS model takes as input a video $\mathbf{v_i}$ (belonging to a video dataset $\mathcal{D}$) consisting of a sequence of T_i RGB frames, namely $\mathbf{v}_i = \left\{F_{t,i} \in \mathbb{R}^{W_i \times H_i \times 3}\right\}_{t=1}^{T_i}$, where W_i and H_i represent width and height of the frame $F_{t,i}$ for video i[1], respectively. VOS methods return as output of each video i a sequence of T_i binary segmentation masks $\mathbf{q}_i = \left\{Q_{t,i} \in \{0,1\}^{W_i \times H_i}\right\}_{t=1}^{T_i}$, often ob-

[1] In the other sections of this paper the subscript i will be omitted for the sake of simplicity.

tained by binarising the corresponding probability maps $\mathbf{m}_i = \left\{\mathcal{M}_{t,i} \in [0,1]^{W_i \times H_i}\right\}_{t=1}^{T_i}$. The typical video segmentation annotation is a frame-wise segmentation annotation $\mathbf{g}_i$ provided for each video $\mathbf{v}_i$. It consists of a sequence of T_i ground-truth binary masks, *i.e.*, $\mathbf{g}_i = \left\{G_{t,i} \in \{0,1\}^{W_i \times H_i}\right\}_{t=1}^{T_i}$, that associates to each frame of the video its corresponding segmentation annotation. We denote the whole annotated dataset as pairs $\langle \mathbf{v}_i, \mathbf{g}_i \rangle \in \mathcal{D}_{\text{Sup}}$.

Inference Approach. Existing methods can be roughly divided into two main categories according to the inference time approach [54], *i.e.*, ***O**ne-shot **V**ideo **O**bject **S**egmentation* (OVOS) and ***Z**ero-shot **V**ideo **O**bject **S**egmentation* (ZVOS). Formally, at inference time, OVOS requires in input the video $\mathbf{v}_i$ togheter with the first frame's manual segmentation $G_{1,i}$, whereas ZVOS processes the video $\mathbf{v}_i$ only. In the past, OVOS and ZVOS have also been improperly called "Supervised" and "Unsupervised" VOS, respectively, even though these names do not describe the *training* supervision type but only the inference approach. Instead, we will use the terms "supervised" and "unsupervised" exclusively to denote the models' training modality.

Training Modality. Orthogonally to the data type and the inference approach, we can define the training supervision type by characterizing the datasets from which training samples are drawn. Following [26]: in *Supervised Learning*, the training set $\mathcal{D}_{\text{Sup}}$ comprises both videos and video segmentation annotations, *i.e.*, $\langle \mathbf{v}_i, \mathbf{g}_i \rangle \in \mathcal{D}_{\text{Sup}}$; in *Unsupervised Learning*, the training set consists of videos without annotations, *i.e.*, $\langle \mathbf{v}_i \rangle \in \mathcal{D}_{\text{Unsup}}$; in *Weakly Supervised Learning*, the training set $\mathcal{D}_{\text{Weakly}}$ is composed of pairs $\langle \mathbf{v}_i, \mathbf{k}_i \rangle$ where $\mathbf{k}_i$ plays a role similar to that of $\mathbf{g}_i$, but instead of being a full segmentation annotation, it is something easier to be acquired for each video frame, *e.g.*, video eye gaze annotations [38].

Our Scenario. In this paper we address ***Z**ero-shot **V**ideo **C**amouflaged **O**bject **S**egmentation* (ZVCOS), that is the ZVOS subcategory where subjects (foreground) blend into the background, which is more challenging and less studied. We address the problem in an *unsupervised* fashion, therefore the input is simply constituted by the videos $\mathbf{v}_i$ during training. Although our approach exploits a model which has been pre-trained on the domain of image segmentation, this has eventually became a common practice in Deep Learning, and our training set contains only unlabeled videos from $\mathcal{D}_{\text{Unsup}}$. Accordingly, we stick with the definition of unsupervised from [26] and consider weakly supervised those methods similar to [38], where an actual effort is needed to gather $\mathbf{k}_i$ in the alternative annotation domain. Other related problems – *not addressed in this paper* – include video object detection [19] (where the task is to roughly identify salient objects via bounding boxes instead of masks) and video semantic segmentation [44] (where the objective is to assign a semantic label to all the pixels).

3. Related Work

While our work belongs to ZVCOS methods, for the sake of completeness in this section we also review existing ZVOS methods in addition to ZVCOS techniques. Special attention is devoted to weakly-supervised and unsupervised approaches – developed either for camouflaged or general subjects – which are the closest to our framework.

Automatically segment subjects - ZVOS. Zero-shot Video Object Segmentation, *i.e.*, segmenting subjects in a video without requiring any hint from the final user at inference time, has always been a challenging task of interest. Methods addressing this problem are typically supervised, *i.e.*, they exploit the hand-curated segmentation ground-truth of the available datasets [7, 20, 29, 30, 42], thereby heavily relying on human annotations. Such benchmarks comprise videos from general categories, without considering the camouflaged challenges. A preliminary work was [12], which considers Optical Flow's motion boundaries to detect moving objects in videos to get segmentation proposals, and integrate a deep learning module to discard bad ones. Later on, the first end-to-end trainable supervised methods appeared, some of which (*e.g.*, [33]) exploiting recurrent structures to deal with videos. In the meanwhile, two-branches networks gained attention [3, 10, 22, 32, 53]: these architectures aim to explore motion with one branch, and appearance with another branch. With the advent of Siamese Networks, several methods [13, 25, 48, 50] employed them to simultaneously process frames at an arbitrary distance. A valid alternative to Siamese Networks for considering distant frames is represented by [37], where Graph Neural Networks are used to model the video as a fully connected graph, in which nodes represent frames.

Removal of Appearance - ZVCOS. Unlike the standard ZVOS setting, ZVCOS focuses on detecting camouflaged foreground objects that blend into the background. This task is particularly challenging due to the limited availability of distinguishing appearance information. For this setting, the two reference datasets are MoCA-Mask [4] and CAD [1], which only provide sparse hand-crafted ground truths. The authors of [4] also propose a *supervised* method named SLT-Net, exploiting Short-Term correlation blocks to leverage short-term relationships and a "sequence-to-sequence" model to promote long-term consistency. Other supervised works focusing on the camouflaged video aspect include [15, 27, 41, 51].

Un-/Weakly-Supervised Methods. Addressing ZVCOS or ZVOS without requiring segmentation annotations in training is less explored than the supervised counterpart. Concerning the camouflaged domain (the focus of this paper) there is only one work in the literature, namely MG [45],

which aims to detect moving objects in an unsupervised way with Optical Flow. Focusing on the layered representation [36] of a video, according to which a video can be decomposed into different levels with simple motions, MG [45] separates the flow image into two distinct levels: one representing the foreground and the other the background. The learning process is supervised by the reconstruction of the input flow, together with a consistency term among close frames. However, MG [45] belongs to the category of *Video Motion Segmentation* [43], whose objective is to segment objects in motion. Differently, our framework aims to complement motion information and (little) appearance to detect also nearly stationary camouflaged objects, failure cases for MG [45]. Still, this work represents our closest competitor (see Sec. 5 for an experimental comparison).

Other methods not requiring video ground-truth masks for training are available in the literature [26, 38, 46, 47], either approaching the problem in a weakly supervised or unsupervised manner, but they consider general (non camouflaged) subjects. Wang *et al*. [38] exploit "eye-gaze" attention maps to perform a weakly supervised training: this represents an improvement compared to supervised approaches, but efforts and equipments are required to track users' gaze. The authors of MuG [26] propose both an unsupervised method (when frame supervisory masks come from a traditional saliency method) and a weakly supervised approach (when using CAM maps, exploiting image classification [49]), where they inspect videos at different temporal granularities. Unlike us, particular annotations (*i.e.*, captions for the chosen CAM module) and complex components exploiting features embeddings are needed. DeSC-V, instead, has a simpler redesigned cross-frame consistency loss. The authors of CIS [46] and DyStaB [47] focus on moving objects and train the proposed unsupervised method in an adversarial way through optical flow inpainting: a first network tries to mask the optical flow, and another one tries to reconstruct it within the masked region. These approaches [46, 47] suffer from limited performance when considering nearly static subjects, similarly to MG [45]. Note that the unsupervised methods discussed in this paragraph [26, 46, 47] focus on general subjects, and do not introduce specific measures for camouflaged videos, hence they are not directly comparable to our approach. Anyway, [46] is included in our experiments as a reference (Sec. 5).

4. Proposed Method

This section is devoted to the proposed approach, named DeSC-V, for performing ZVCOS in an unsupervised way (*i.e.*, without segmentation annotations for training).

Our DeSC-V framework, illustrated in Fig. 2, consists of a to-be-trained model $\mathcal{T}$ and a collection of "experts" that provide supervisory signals, namely: a *Frame Expert* $\mathcal{B}$, a *long term expert* $\mathcal{H}$ (divided in forward $\mathcal{H}_f$ and backward $\mathcal{H}_b$), a *short term expert* $\mathcal{S}$ (divided in forward $\mathcal{S}_f$ and backward $\mathcal{S}_b$), and a warping module $\mathcal{W}$. We chose a Pyramid Vision Transformer V2 [39] both as $\mathcal{T}$ and $\mathcal{B}$, as described in the sequel. $\mathcal{T}$ predicts a segmentation mask for the current frame, and all experts provide supervision to $\mathcal{T}$ at the *frame* level in the form of segmentation supervisory masks to estimate the likelihood of each pixel belonging to the foreground. Temporal consistency from these supervisory masks, predicted at different time steps, is leveraged to improve the accuracy of the probability mask at time t. The warping module is responsible for mask alignment across frames, integrating information from both short-term and long-term frames to refine the current prediction.

The main idea underpinning our solution is to exploit prior knowledge from a similar task: the *Frame Expert* $\mathcal{B}$ is a pre-trained object segmentation network for camouflaged subjects *in images*. We observed that the tentative supervisory mask returned by $\mathcal{B}$ are very coarse and inaccurate since $\mathcal{B}$ has been pre-trained on an image dataset. To overcome this drawback, we exploit temporal information by incorporating additional experts, namely a *Short-Term Expert* and a *Long-Term Expert*, to promote short-term regularization and long-term temporal consistency, respectively. In particular, since subjects in videos often exhibit continuous movements, we exploit the Optical Flow [34] to leverage supervision through the warping module, aligning subjects' positions. Supervisory masks predicted by all the experts are compared with the current frame prediction of our to-be-trained network $\mathcal{T}$. Note that such temporal compensation eliminates the need for *video* segmentation annotations. In the following paragraphs, we overview the primary purpose and details of all the experts and the warping module.

Segmentation Loss. Before illustrating each module, we describe how to compare the prediction of our network $\mathcal{T}$ against each supervisory mask coming from the different experts. This comparison will be made through the SLT-Net's [4] short-term module loss:

$$\mathcal{L} = \mathcal{L}^w_{\text{ce}} + \mathcal{L}^w_{\text{iou}} \tag{1}$$

which sums a weighted Cross-Entropy loss $\mathcal{L}^w_{\text{ce}}$ and a weighted Intersection-over-Union loss $\mathcal{L}^w_{\text{iou}}$, each of which takes in input two segmentation masks and evaluate their similarity. As explained in [40], the idea of $\mathcal{L}$ is to weigh each pixel differently, based on the surrounding pixels' properties, since gathering the correct labelling around boundaries or elongated areas is more complicated. This loss will be used by all modules of our architecture.

Frame Expert. It is an Image Segmentation network, denoted as $\mathcal{B}$ in Fig. 2, used as a black-box expert, that takes in input the current frame F_t to provide the supervisory mask $\mathcal{M}^{\mathcal{B}}_t$ for the current frame. Its purpose is to provide a meaningful initialization for the network prediction, and to act as a fixed reference that prevents the network $\mathcal{T}$ from drifting

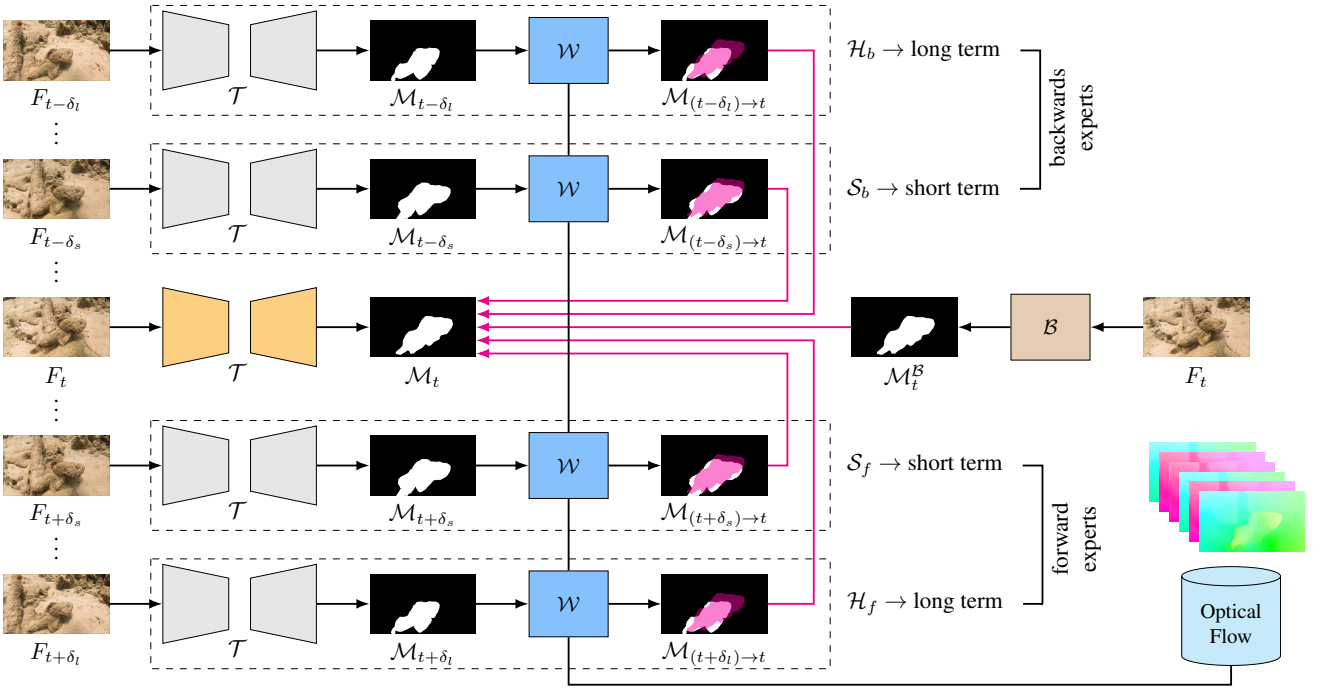

Figure 2. Our training framework, DeSC-V. At each iteration, the training model $\mathcal{T}$ (in yellow) predicts the mask $\mathcal{M}_t$ for the frame at time t. Four frozen copies of the model at that training step (in grey) predict the masks for frames $t-\delta_l$, $t-\delta_s$, $t+\delta_s$, $t+\delta_l$, which are then warped at time t by a warping module $\mathcal{W}$ using pre-computed Optical Flow, constituting the Short-Term expert $\mathcal{S}$ and the Long-Term expert $\mathcal{H}$ (splitted in the forward (f) and backward (b) counterparts). The warped masks $\mathcal{M}_{h\rightarrow t}$ from time h to t – in magenta – are used to supervise the prediction for frame t together with the output supervisory mask $\mathcal{M}_t^{\mathcal{B}}$ of the black-box architecture $\mathcal{B}$ (Frame Expert), as indicated by the supervisory signals (in magenta). Our framework exploits both motion information and appearance in an unsupervised way. *Best viewed in color.*

in its prediction during training. In fact, without the Frame Expert, a trivial solution that satisfies the temporal consistency would be the mask $1^{W\times H}$ or $0^{W\times H}$ for all the frames (see Tab. 2). The Frame-Expert loss can be formalized as:

$$\mathcal{L}_{\mathcal{B}} = \mathcal{L}\left(\mathcal{M}_t, \mathcal{M}_t^{\mathcal{B}}\right) \tag{2}$$

where $\mathcal{M}_t$ is our model's prediction for the frame at time t, $\mathcal{M}_t^{\mathcal{B}}$ is the supervisory mask generated by $\mathcal{B}$ for frame t, and $\mathcal{L}$ comes from Eq. (1). In our experiments, we set $\mathcal{B}$ as the initial to-be-trained network $\mathcal{T}$, *i.e.*, a Pyramid Vision Transformer V2 [39] pre-trained on COD10K [11]. It basically accomplishes the Camouflaged Object Detection [11] task in the image domain, and it is useful to recall the network's initial knowledge during the learning process.

Warping Block. Represented as $\mathcal{W}$ in Fig. 2, it leverages Optical Flow from frame F_t to $F_{t+\delta}$ (denoted as $\text{Flow}_{t\rightarrow t+\delta}$) to *warp* predictions from the time instant $t+\delta$, toward the current frame at time t. The purpose is to align the predicted segmentation $\mathcal{M}_{t+\delta}$ with the subject's position in frame t. To align positions, the warping block adjusts each pixel in $\mathcal{M}_{t+\delta}$ according to the movement observed between frames t and $t+\delta$, using Optical Flow to capture this dense motion between pixels in different frames. This operation can be formalized as follows:

$$\mathcal{M}_{(t+\delta)\rightarrow t} = \mathcal{W}\left(\mathcal{M}_{t+\delta}, \text{Flow}_{t\rightarrow t+\delta}\right) \tag{3}$$

where $\mathcal{M}_{(t+\delta)\rightarrow t}$ is the prediction of the network $\mathcal{T}$ for frame $F_{t+\delta}$ aligned with the position of the subject at time t, $\mathcal{M}_{t+\delta}$ is the prediction of $\mathcal{T}$ for frame $F_{t+\delta}$, and $\text{Flow}_{t\rightarrow t+\delta}$ is the Optical Flow among frames F_t and $F_{t+\delta}$. Notice that since the Optical Flow is obtained from similarities among regions, its performances are independent of the content of the considered area, regardless of the fact that it may contain part of the camouflaged subject. We use RAFT [34] as Optical Flow computator, since we obeserved overall good performances in the analyzed videos.

Short and Long-Term Experts. They are introduced for two different reasons. The Short-Term expert makes the prediction smoother in time throughout the video, whereas the Long-Term one enriches the information available at time t exploiting more complex dynamics, as the subject may exhibit more intricate motion patterns over longer intervals. They both exploit the warping block $\mathcal{W}$ to compare the masks for two frames to the one of the current time t, with the only difference that the Short-Term one works

with frames "*close*" to the current time t at distance $\pm\delta_s$, whereas the Long-Term one considers "*distant*" frames, at distance $\pm\delta_l$. They both compare mask predictions from the past, namely *backwards* ($\mathcal{S}_b$ for the Short-Term and $\mathcal{H}_b$ for the Long-Term), and from the future, namely *forward* ($\mathcal{S}_f$ for the Short-Term, and $\mathcal{H}_b$ for the Long-Term), with the current frame. We used $\delta_s = 1$ and $\delta_l = 5$ for short-term and long-term time shifts. The Short-Term expert comprises the 2nd and 4th rows of Fig. 2, whereas the Long-Term one is represented in the 1st and 5th rows of Fig. 2. They can be formalized with a single equation as follows:

$$\mathcal{L}_{\text{temporal}} = f_w \underbrace{\mathcal{L}\left(\mathcal{M}_t, \mathcal{M}_{(t+\delta)\to t}\right)}_{\substack{\mathcal{S}_f \text{ if } \delta = \delta_s \\ \mathcal{H}_f \text{ if } \delta = \delta_l}} + b_w \underbrace{\mathcal{L}\left(\mathcal{M}_t, \mathcal{M}_{(t-\delta)\to t}\right)}_{\substack{\mathcal{S}_b \text{ if } \delta = \delta_s \\ \mathcal{H}_b \text{ if } \delta = \delta_l}} \tag{4}$$

where $\mathcal{L}$ is the segmentation loss defined in Eq. (1), $\mathcal{M}_{(t\pm\delta)\to t}$ is the prediction warped to time t and comes from Eq. (3) where $\delta_l > \delta_s$, $\mathcal{S}$ indicates the Short-Term expert, $\mathcal{H}$ indicates the Long-Term expert, and b_w and f_w are the backwards weight and the forwards weight, respectively. Such weights are computed as the Intersection over Union (IoU) between the current warped mask and the (fixed) frame expert mask. They are essential to allow the network to recall its initial knowledge, as shown in Tab. 2.

Global Loss. It becomes:

$$\mathcal{L}_{\text{Global}} = \alpha\mathcal{L}_{\mathcal{B}} + \beta\mathcal{L}_s + \gamma\mathcal{L}_l \tag{5}$$

where $\mathcal{L}_{\mathcal{B}}$, $\mathcal{L}_s$, and $\mathcal{L}_l$ are defined respectively in Eq. (2) and Eq. (4) with $\delta = \delta_s$, Eq. (4) with $\delta = \delta_l$, and $\alpha, \beta, \gamma > 0$ are weights. We used $\alpha = 0.4$, and $\beta = \gamma = 0.15$ in our experiments. Eq. (5) strikes a balance between the accuracy of our predictions (relative to the tentative masks provided by the Frame Expert) and the temporal smoothness (based on Optical Flow warping at short/long-term levels).

Implementation Details. The chosen network $\mathcal{T}$ of our architecture constitutes a specific block of SLT-Net [4], a consolidated baseline for supervised ZVCOS. Specifically, we employed the part of [4] responsible for the single-frame analysis, *i.e.*, a model for image segmentation. This module, taking a single frame as input and returning its mask, is built upon the Pyramid Vision Transformer v2 (PVTv2) [39], a consolidated backbone often adopted in different state-of-the-art architectures. The encoder, together with the PVTv2 [39], also comprises Texture Enhancing Modules [11], each including four parallel residual branches. Its features are fed into Guided Reversal Attention blocks [11], where the Neighbor Connection Decoder provides reversal guidance about the subject location. Note that, in our experiments, we never used weights from SLT-Net [4], as they have undergone supervised training on MoCA-Mask [4]. To provide initial knowledge to DeSC-V and to have meaningful supervisory masks from the beginning, our $\mathcal{T}$ has been pre-trained on the training set of COD10K [11], a dataset designed for Camouflaged Object Segmentation. Note that it is a dataset of images, not videos. After pre-training, we obtained our starting network $\mathcal{T}$.

Memory Optimizations. The design choices and the implementation of our framework have been performed to make DeSC-V lightweight: pre-computing both the optical flow and the predictions of the (fixed) frame expert, in addition to the design choice of a shared architecture among the temporal experts and the to-be-trained network, allows us to perform the training even on a laptop with 16GB of RAM, an i7-8750-H, and an Nvidia GTX 1050 mobile.

5. Experiments

In this section we report our experiments run on the MoCA-Mask [4] dataset and on the Camouflaged Animal Dataset (CAD) [1]. Our code is available online[2].

5.1. MoCA-Mask Dataset

MoCA-Mask[3] [4] is a reference benchmark for Camouflaged Video Object Segmentation, which contains 87 sequences with 22 939 annotated frames in total. The segmentation masks are provided as hand-curated masks only every 5th frame of the original sequence, with the rest interpolated from these. The dataset contains 71 different sequences for the training set, and 16 for the test set. The number of hand-curated segmentation annotations varies from 10 to 154. We provide quantitative results only on frames with hand-curated masks, as evaluations on interpolated ones would produce a bias according to the interpolation method. The same holds also for qualitative results provided in Fig. 1 and Fig. 3. Our final model (trained in an unsupervised way on MoCA-Mask [4] for 150 epochs) is denoted by DeSC-V.

Competing methods. As reviewed in Sec. 3, the closest competitor to our DeSC-V is MG [45], which addresses Zero-shot Video Camouflaged Object Segmentation (ZVCOS) in an unsupervised fashion. Other unsupervised ZVCOS methods are not available. However, in order to enrich the evaluation, we also consider CIS [46], an unsupervised approach developed for general subjects (ZVOS) without considering the camouflaged scenario. Both MG [45] and CIS [46] focus on moving objects[4]. To perform a fair comparison, we re-trained CIS [46] on MoCA-Mask [4] using the authors' code, and we considered the predictions of MG [45] publicly provided by [4]. In order to give further insights, we also report the performances obtained with our initial network $\mathcal{T}$ (*i.e.*, a Pyramid Vision Transformer V2 [39]) *prior* to training with our DeSC-V framework. We

[2] https://github.com/alessandriniluca/DeSC-V

[3] https://xueliancheng.github.io/SLT-Net-project/

[4] We do not consider other unsupervised ZVOS methods developed for general (non camouflaged) subjects, like MuG [26] and DyStaB [47], since the code is either not available or not reproducible.

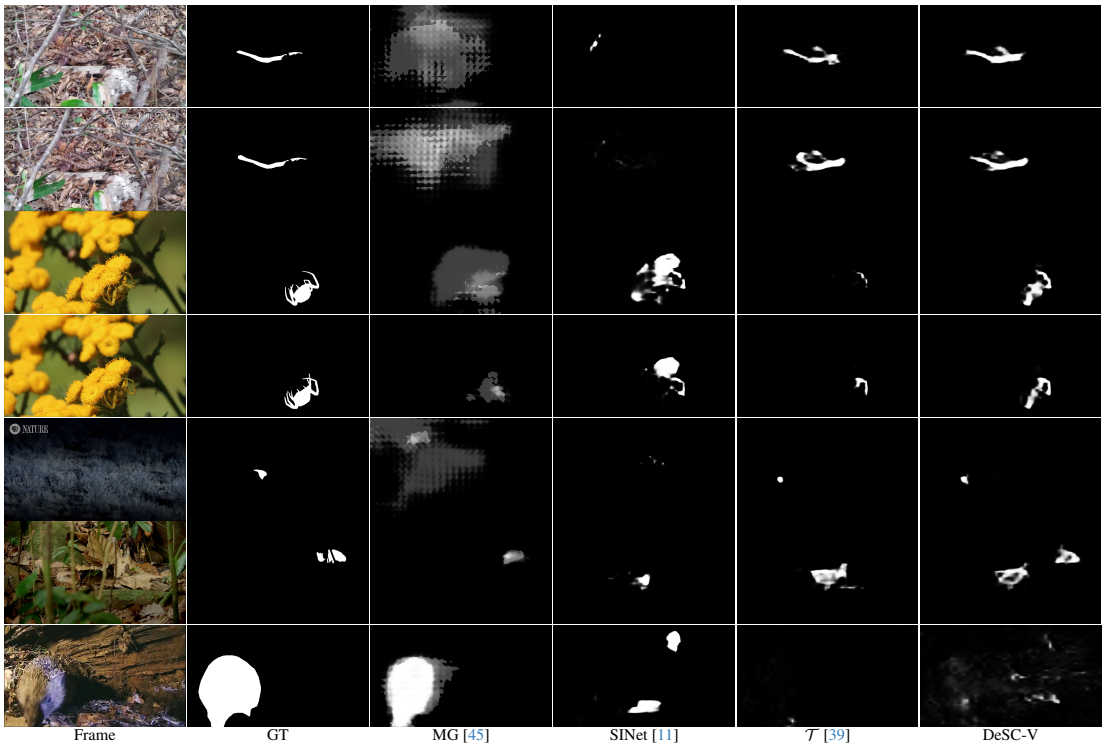

Figure 3. Qualitative results from sample scenes of MoCA-Mask [4]. For each frame, we report the Ground-Truth (GT), results from the competing methods, from the starting network $\mathcal{T}$ and our DeSC-V. Results for MG [45] have been taken from the ones provided by [4]. Sequences, from top to bottom: copperhead_snake (frames 105, 185), flower_crab_spider_2 (frames 45, 75), black_cat_1 (frame 135), rusty_spotted_cat_0 (frame 50), hedgehog_3 (frame 110). Best viewed in colour – zoom in for details –.

Method	Appearance	Motion	Training	IoU
CIS [46]	✗	✓	Video	0.044
$\mathcal{T}$ [39]	✓	✗	Image	<u>0.173</u>
SINet [11]	✓	✗	Image	0.164
MG [45]	✗	✓	Video	0.130
DeSC-V	✓	✓	Video	**0.200**

Table 1. Results of the analyzed methods on MoCA-Mask [4] in terms of IoU (↑), best in bold, second-best underlined. The focus of each method (motion/appearance) during training is also reported: only their combination (our DeSC-V) reaches the highest IoU measure.

also consider SINet [11], which is a popular network meant to work with images of camouflaged subjects. Both SINet [11] and $\mathcal{T}$ underwent supervised training on COD10K [11] for fair comparison, which consists of images (rather than videos) depicting camouflaged animals; at inference, they are applied frame by frame to the testing videos. Properties of the competitors are summarized in Tab. 1.

Results. To assess the video segmentation performance of the considered methods we adopt the Intersection over Union (IoU) between the obtained mask and the ground-truth one, averaged over frames and sequences, as customary in the literature. The results are given in Tab. 1: the fact that IoU values are relatively low for all methods, it underlines the challenge of segmenting camouflaged animals and is in line with the literature. Nothwistanding this, Tab. 1 clearly shows that our approach outperforms all the analysed methods. This highlights the effectiveness of the proposed framework for addressing Zero-shot Video Camouflaged Object Segmentation without any video ground-truth in training. In particular, DeSC-V is significantly better than MG [45] (our closest competitor) and CIS [46], confirming the inability of methods relying solely on Optical Flow to segment subjects when there is poor motion information, as already observed in Sec. 3. The poor performance achieved by CIS might be due to the fact that it was not specifically developed for camouflaged animals: this highlights the need for specific choices to manage the challenging camouflaged scenario, present instead in our approach. Note that DeSC-V outperforms the initial network $\mathcal{T}$, showing the effectiveness of combining multiple experts. Our method is also superior to SINet [11], thereby further emphasizing the importance of exploiting video information. We also provide qualitative results in Fig. 3, which confirms that DeSC-V outperforms both the starting network $\mathcal{T}$

method	IoU
$\mathcal{T}$ [39]	0.173
$\mathcal{T}$ + short & long term	0.020
$\mathcal{T}$ + short & long term + frame expert	0.164
$\mathcal{T}$ + short & long term + frame expert + IoU weight	**0.200**

Table 2. Ablation study of our framework on MoCA-Mask [4] in terms of IoU (↑). The first row corresponds to the starting network $\mathcal{T}$ (i.e., a Pyramid Vision Transformer V2 [39]). The last row corresponds to the proposed network DeSC-V. The IoU weights correspond to f_w and b_w in Eq. (4). As visible, the contribution of all the components is fundamental.

and its competitors in many cases (CIS [46] has not been reported in the figure, as it often fails in identifying the camouflaged subject). The last two rows show instead failure cases for DeSC-V, where MG [45] succeeds, as the latter benefits from a perfectly static camera and a subject that exhibits strong movement. Further visualizations are given in the supplementary material.

Ablation Analysis. An ablation study of DeSC-V on MoCA-Mask [4] is given in Tab. 2, which demonstrates the effectiveness of all our components, that are the short/long-term experts, the frame expert, and the usage of IoU weights f_w, b_w in Eq. (4). The latter, in particular, have the effect of recalling the network's initial knowledge. Note also that, solely relying on temporal experts disregarding the frame one leads to performance degradation (second row in Tab. 2), due to the network's drift towards trivial solutions minimizing only the mask warping difference.

5.2. Camouflaged Animal Dataset (CAD)

The Camouflaged Animal Dataset (CAD)[5] [1] contains nine sequences with a total of 839 frames. Similarly to MoCA-Mask [4], only a subset of the frames have been manually annotated – we used only such frames to compute the quantitative results. Multiple instances in the annotations have been considered as a single object. Due to the limited amount of sequences, we use this dataset solely for testing purposes, as common practice. As done in Sec. 5.1, we compared our method to CIS [46], the starting network $\mathcal{T}$ [39], and SINet [11]. Here, we do not include MG [45] in our analysis as their weights trained on MoCA-Mask [4] have not been released publicly. Results in terms of IoU are reported in Tab. 3, further validating the results already obtained on MoCA-Mask [4]: our DeSC-Voutperforms other approaches, thanks to the fact that it combines appearance and motion. We observe that here the improvement of our method over the starting network $\mathcal{T}$ is less pronounced compared to the MoCA-Mask dataset, as the latter reports more

[5] https://sites.google.com/view/piabideau/research/structure-from-motion

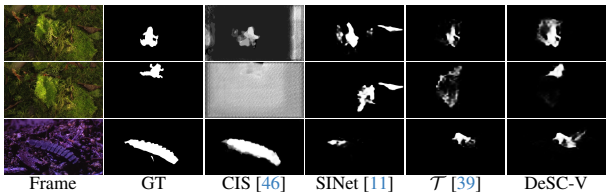

Figure 4. Qualitative results from sample scenes of CAD [1]. For each frame, we report the Ground-Truth (GT), results from the competing methods, from the starting network $\mathcal{T}$ and our DeSC-V. Sequences, from top to bottom: frog (frames 1, 9), glowwormbeatle (frame 1). Best viewed in colour – zoom in for details –.

method	IoU
CIS [46]	0.200
SINet [11]	0.255
$\mathcal{T}$ [39]	<u>0.324</u>
DeSC-V	**0.337**

Table 3. Results of the analyzed methods on the Camouflaged Animal Dataset (CAD) [1] in terms of IoU (↑), best in bold, second-best underlined. DeSC-V reaches the highest IoU.

challenging scenes. Some qualitative examples are available in Fig. 4, showing segmentations of good quality in the first two rows. The last row, instead, reports a failure case of our method, in which CIS [46], instead, benefits from a perfectly distinct background-subject movement. As already observed, DeSC-V is inferior to CIS in situations where the subject exhibits strong movement over time, whereas our approach is typically not influenced by the presence of nearly-stationary animals.

6. Conclusion

In this paper, we tackled the challenging and unexplored task of unsupervised Zero-Shot Video Camouflaged Object Segmentation, aiming to segment camouflaged subjects in a video without any mask annotations at training/inference time. We addressed this problem leveraging a novel framework that combines appearance and motion information: to capture appearance cues, we exploit prior knowledge from a pre-trained image segmentation network; for motion, we employ optical flow to warp predicted masks from various time frames, aligning them with the current frame. Experimental results showed that our method achieves state-of-the-art performances on the MoCa-Mask and CAD datasets. In the future, we plan to extend our framework to other segmentation tasks, thanks to its modularity. We are also interested in analyzing more diverse datasets, such as the recent MeViSv2 [8] and MOSEv2 [9], which include a mixture of general subjects and camouflaged animals.

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.

[2] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *CVPR*, pages 3151–3161, 2024.

[3] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, pages 686–695, 2017.

[4] 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.

[5] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Dual prototype attention for unsupervised video object segmentation. In *CVPR*, pages 19238–19247, 2024.

[6] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–2703, 2023.

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H.S. Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *ICCV*, pages 20224–20234, 2023.

[8] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE TPAMI*, 47(12):11400–11416, 2025.

[9] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip H. S. Torr, and Song Bai. Mosev2: A more challenging dataset for video object segmentation in complex scenes, 2025.

[10] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, pages 3664–3673, 2017.

[11] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10): 6024–6042, 2022.

[12] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving objects in videos. In *CVPR*, pages 4083–4090, 2015.

[13] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. *AAAI*, 34(07):10869–10876, 2020.

[14] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *CVPR*, pages 4144–4154, 2021.

[15] Wenjun Hui, Zhenfeng Zhu, Guanghua Gu, Meiqin Liu, and Yao Zhao. Implicit-explicit motion learning for video camouflaged object detection. *IEEE TMM*, 26:7188–7196, 2024.

[16] Varun Jampani, Raghudeep Gadde, and Peter V. Gehler. Video propagation networks. In *CVPR*, pages 451–461, 2017.

[17] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *IJCV*, 127(9):1175–1197, 2019.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[19] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. *Asian Conference on Computer Vision*, 2020.

[20] 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.

[21] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, pages 3243–3252, 2018.

[22] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019.

[23] Siyuan Li, Yue Luo, Ye Zhu, Xun Zhao, Yu Li, and Ying Shan. Enforcing temporal consistency in video depth estimation. In *ICCV Workshops*, pages 1145–1154, 2021.

[24] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *ECCV*, pages 352–368, 2020.

[25] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, pages 3623–3632, 2019.

[26] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J. Crandall, and Steven C. H. Hoi. Learning video object segmentation from unlabeled videos. In *CVPR*, pages 8960–8970, 2020.

[27] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *CVPR*, pages 17169–17180, 2024.

[28] Muhammad Nawfal Meeran, Gokul Adethya T, and Bhanu Pratyush Mantha. Sam-pm: Enhancing video camouflaged object detection using spatio-temporal attention. In *CVPR Workshops*, pages 1857–1866, 2024.

[29] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2014.

[30] 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*, 2016.

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman

Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv*, 2408.00714, 2024.

[32] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *CVPR*, pages 15455–15464, 2021.

[33] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018.

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020.

[35] Serin Varghese, Yasin Bayzidi, Andreas Bar, Nikhil Kapoor, Sounak Lahiri, Jan David Schneider, Nico M. Schmidt, Peter Schlicht, Fabian Huger, and Tim Fingscheidt. Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *CVPR Workshops*, pages 336–337, 2020.

[36] John Y. A. Wang and Edward H. Adelson. Representing moving images with layers. *IEEE TIP*, 3(5):625–638, 1994.

[37] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, pages 9236–9245, 2019.

[38] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven C. H. Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, pages 3064–3074, 2019.

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

[40] Jun Wei, Shuhui Wang, and Qingming Huang. F³net: Fusion, feedback and focus for salient object detection. *AAAI*, 34 (07):12321–12328, 2020.

[41] Haozhe Xing, Shuyong Gao, Yan Wang, Xujun Wei, Hao Tang, and Wenqiang Zhang. Go closer to see better: Camouflaged object detection via object area amplification and figure-ground conversion. *IEEE TCSVT*, 33(10):5444–5457, 2023.

[42] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018.

[43] Xun Xu, Loong Fah Cheong, and Zhuwen Li. Motion segmentation by exploiting complementary geometric models. In *CVPR*, pages 2859–2867, 2018.

[44] Xiaolong Xu, Lei Zhang, Jiayi Li, Lituan Wang, Yifan Guan, Yu Yan, Leyi Zhang, and Hao Song. Dual-temporal exemplar representation network for video semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10775–10785, 2025.

[45] 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.

[46] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, pages 879–888, 2019.

[47] Yanchao Yang, Brian Lai, and Stefano Soatto. Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In *CVPR*, pages 2826–2836, 2021.

[48] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip H. S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, pages 931–940, 2019.

[49] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *CVPR*, pages 6074–6083, 2019.

[50] Lu Zhang, Jianming Zhang, Zhe Lin, Radomír Měch, Huchuan Lu, and You He. Unsupervised video object segmentation with joint hotspot tracking. In *ECCV*, pages 490–506, 2020.

[51] Peng Zhang, Hong Yu, Haiqing Li, Xin Zhang, Sixue Wei, Wan Tu, Zongyi Yang, Junfeng Wu, and Yuanshan Lin. Msgnet: multi-source guidance network for fish segmentation in underwater videos. *Frontiers in Marine Science*, 10: 1256594, 2023.

[52] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *CVPR*, pages 6949–6958, 2020.

[53] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. *AAAI*, 34(07):13066–13073, 2020.

[54] Tianfei Zhou, Fatih Porikli, David J. Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE TPAMI*, 45(6):7099–7122, 2023.