Paying Attention to Video Object Pattern Understanding

Wenguan Wang, *Member, IEEE*, Jianbing Shen, *Senior Member, IEEE*, Xiankai Lu, *Member, IEEE*, Steven C. H. Hoi, *Fellow IEEE*, Haibin Ling

Abstract—This paper conducts a systematic study on the role of visual attention in video object pattern understanding. By elaborately annotating three popular video segmentation datasets (DAVIS$_{16}$, Youtube-Objects and SegTrack$_{V_2}$) with dynamic eye-tracking data in the unsupervised video object segmentation (UVOS) setting, for the first time, we quantitatively verified the high consistency of visual attention behavior among human observers, and found strong correlation between human attention and explicit primary object judgments during dynamic, task-driven viewing. Such novel observations provide an in-depth insight of the underlying rationale behind video object pattens. Inspired by these findings, we decouple UVOS into two sub-tasks: UVOS-driven Dynamic Visual Attention Prediction (DVAP) in spatiotemporal domain, and Attention-Guided Object Segmentation (AGOS) in spatial domain. Our UVOS solution enjoys three major advantages: 1) modular training without using expensive video segmentation annotations, instead, using more affordable dynamic fixation data to train the initial video attention module and using existing fixation-segmentation paired static/image data to train the subsequent segmentation module; 2) comprehensive foreground understanding through multi-source learning; and 3) additional interpretability from the biologically-inspired and assessable attention. Experiments on four popular benchmarks show that, even without using expensive video object mask annotations, our model achieves compelling performance compared with state-of-the-arts and enjoys fast processing speed (10 fps on a single GPU). Our collected eye-tracking data and algorithm implementations have been made publicly available at https://github.com/wenguanwang/AGS.

Index Terms—Video Object Pattern Understanding, Unsupervised Video Object Segmentation, Top-Down Visual Attention, Video Salient Object Detection.

1 Introduction

Unsupervised video object segmentation (UVOS), a task for segmenting primary object(s) from the background in videos without any human involvement, has been a long standing research challenge in computer vision [2]–[6]. It has shown potential benefits for numerous applications, *e.g.*, action recognition [7] and object tracking [8].

Due to the lack of user interactions in UVOS, it is very challenging to automatically determine the primary foreground objects from the complex background in real-world scenarios. This calls for an in-depth understanding of foreground object patterns in videos. Although some efforts were made along this direction, the rationale behind their choice of the foreground objects are often inconsistent and intuitive, lacking a theoretical basis and empirical evidence. For example, early video object segmentation datasets, like FBMS$_{59}$ [9] and SegTrack$_{V_2}$ [10], mainly focus on moving object(s), and most of the pioneering efforts in this area started with geometry-based motion analysis. Different from these early *motion-based* datasets, recent benchmarks (*e.g.*, DAVIS$_{16}$ [11] and Youtube-Objects [12]) are *saliency-based*, which are more aware of salient, primary video object segmentation. Similar problems have also been experienced in a closely related research area, video salient object detection (VSOD) [13], which aims to extract a continuous

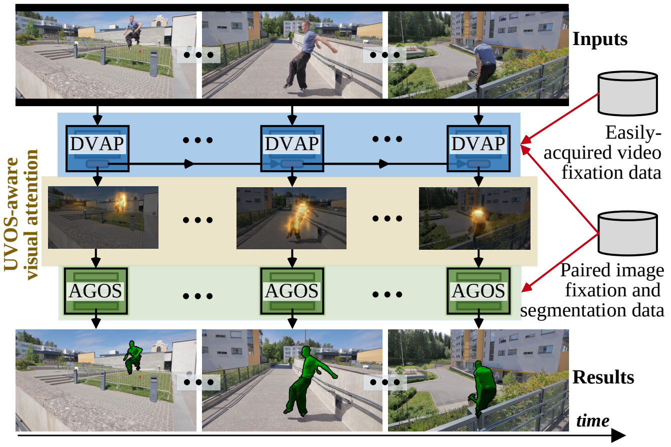

Inputs
DVAP
DVAP
DVAP
Easily-acquired video fixation data
UVOS-aware visual attention
AGOS
AGOS
AGOS
Paired image fixation and segmentation data
Results
time

Fig. 1. **Our UVOS solution** has two key steps: Dynamic Visual Attention Prediction (DVAP, §5.2), cascaded by Attention-Guided Object Segmentation (AGOS, §5.3). The UVOS-aware attention from DVAP acts as an intermediate video object representation, freeing our method from the dependency of expensive video object annotations and bringing better interpretability with human readable and assessable attention maps.

- *W. Wang is with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China, and also with ETH Zurich, Switzerland. (Email: wenguanwang.ai@gmail.com)*
- *J. Shen is with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China, and also with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. (Email:shenjianbingcg@gmail.com)*
- *X. Lu is with Inception Institute of Artificial Intelligence, UAE. (Email: carrierlxk@gmail.com)*
- *S.C.H. Hoi is with the School of Information Systems, Singapore Management University, and Salesforce Research Asia, Singapore. (Email: stevenhoi@gmail.com)*
- *H. Ling is with the Department of Computer Science, Stony Brook University, Strony Brook, NY, USA. (Email: hling@cs.stonybrook.edu)*
- *A preliminary version of this work has appeared in CVPR 2019 [1].*
- *Corresponding author: Jianbing Shen*

saliency map for each frame that highlights the most visually important area. The results from video salient object detection can be used as a critical cue in pre-processing steps for many spatio-temporal vision tasks, such as UVOS [14], [15], video re-timing [16], and thumbnailing [17]. However, for the widely used VSOD datasets (*e.g.*, ViSal [13] and DAVIS$_{16}$ [11]), a biological and testable interpretation for the choice of the salient object regions, though essential, is long-term missing.

In this paper, we emphasize the value of human visual attention in video object pattern modeling. According to studies in cognitive psychology [18]–[21], during visual perception, humans are able to quickly orient attentions to the most important parts of the visual stimuli, allowing them to achieve goals efficiently. We therefore argue that *human visual attention should be an underlying mechanism that drives UVOS and VSOD*. The foreground in UVOS and VSOD should contain the object(s) that attracts human attention most, as the choice of the object(s) should be consistent with human attention judgments. This provides a unified, insightful and verifiable explanation for moving object patterns. Previous studies [22], [23] of visual attention mechanism in cognitive psychology and computer vision have shown empirically that both motion and appearance stimulus are key factors that direct attention allocation in dynamic viewing. Therefore, from the visual attention point of view, previous video segmentation datasets, though collected under different strategies, provide insights into our problem from different angles.

To validate this novel hypothesis, we extend three popular video segmentation datasets, DAVIS$_{16}$ [11], Youtube-Objects [12] and SegTrack$_{V_2}$ [10], with real human fixation annotation in the UVOS setting. The gaze data are collected over a total of 190 video sequences with 25,049 frames from 20 human observers using professional eye-tracking instruments (§3). To the best of our knowledge, this is the first attempt to collect UVOS-aware human attention data. Such comprehensive datasets facilitate us to perform three essential experiments, *i.e.*, (1) quantifying the inter-subject consistency, (2) studying the correlation between human dynamic attention and explicit object judgment, and (3) analysing the stability of human fixation patterns in the presence of challenging video processing factors. Three key observations are found from our empirical analyses:

- There exist highly consistent attention behaviors among human observers in the UVOS task, though the notion of 'primary object(s)' is sometimes ambiguous for extremely-diverse dynamic scenes.
- There exists a strong correlation between human fixation and human explicit judgment of primary object(s).
- Fixation patterns among subjects present significant stability across different video processing challenges and the ability of human beings of capturing moving objects is different in the scenes with different challenges.

These findings offer an insightful glimpse into the rationale behind UVOS from human attention perspective. Thus inspired, we decompose UVOS into two sub-tasks: *dynamic visual attention prediction* (DVAP) and *attention-guided object segmentation* (AGOS). Accordingly, we devise a novel UVOS model with two tightly coupled components for DVAP and AGOS (see Fig. 1). One extra advantage of such task decomposition lies in modular training and data acquisition. Instead of using expensive video segmentation annotation, the relatively easily-acquired dynamic fixation data can be used to train DVAP, and existing large-scale fixation-segmentation paired annotations (*e.g.*, [24], [25]) can be used to train the AGOS module.[1] This is because AGOS learns to map an individual input frame and fixation data to a segmentation mask, thus only needs static image data. Roughly speaking, visual attention acts as a middle-level representation that bridges dynamic foreground characteristic modeling and static attention-aware object segmentation. Such design naturally reflects real-world human behavior [26], *i.e.*, first orienting rough attention to important areas during dynamic viewing, and then focusing on fine-grained, pixel-wise object segmentation.

In our UVOS model, the DVAP module is built upon a CNN-convLSTM architecture, where the convLSTM takes static CNN feature sequence as input and learns to capture the dynamic visual attention, and the AGOS module is based on an FCN architecture. Intuitively, DVAP informs AGOS where the objects are located in each frame, then AGOS performs refined object segmentation. Besides, our model brings several important beneficial characteristics:

- *Fully-differentiable and supervised attention mechanism.* For AGOS, the attention from DVAP is used as a neural attention mechanism, thus the whole model is fully-differentiable and end-to-end trainable. At high level, DVAP can be viewed as an attention network, which provides an explicit spatiotemporal attention mechanism to AGOS and is trained in a supervised manner.
- *Comprehensive foreground understanding through multi-source learning and weight sharing.* Our experiments with dynamic gaze-tracking data confirm a strong correlation between eye movements and primary video objects perception. Training with both fixation and segmentation data allows more comprehensive foreground understanding. Moreover, by sharing several initial convolutional layers between DVAP and AGOS, information can be exchanged efficiently.
- *Learning from large-scale affordable data.* Deep learning models are often hungry for large-scale datasets for training, but segmentation annotation on such datasets can be profitably expensive. Our model leverages more affordable dynamic gaze data and existing large-scale attention-segmentation paired image data to achieve the same goal. Our experiments show that our model yields promising segmentation results without training on the ground-truth video segmentation data.
- *Biologically-inspired and assessable interpretability.* The attention learned from DVAP not only enables our model attend to the important object(s), but also offers an extra dimension to interpret where our model focuses on. Such interpretability is meaningful (biologically-inspired) and assessable (w.r.t. human gaze records).

[1] According to the statistics offered by the DAVIS committee, it took around 30 minutes per-frame to annotation with 5 specialists. In contrast, with eye-tracker equipment, annotating each frame takes only 1∼2 seconds.

1.1 Our Contributions

Our contributions in this paper are summarized as follows:

1) We complement existing famous dynamic computer vision annotated datasets (*i.e.*, DAVIS$_{16}$ [11], Youtube-Objects [12], and SegTrack$_{V_2}$ [10]) with human eye movements collected under the ecological constraints of the UVOS task (§3). To the best of our knowledge, these are the first large human eye tracking datasets specifically collected and publicly shared for UVOS.
2) With our collected top-down visual eye tracking data, for the first time, we systemically conduct studies on the human attention behaviors in the context of UVOS task (§4). Our findings underline the remarkable stability of patterns of visual search among subjects and verify the strong correlation between human visual fixation allocation and video object determination, at least within the class of the datasets we studied. We also explore human gaze patterns with different video processing challenges. Our studies shed light on the rationale behind UVOS (and its related task: VSOD) from a view of top-down human visual attention mechanism.
3) With our findings that suggest a remarkable degree of consistency between dynamic fixation and video object determination patterns of human subjects, we propose a powerful, fully differentiable, and biologically-inspired UVOS model that fully exploits the value of visual attention (§5). Our model produces state-of-the-art results on both motion-based dataset (*i.e.*, FBMS$_{59}$ [9]) and saliency-based benchmarks (*i.e.*, DAVIS$_{16}$ [11], DAVIS$_{17}$ [27] and Youtube-Objects [12]), with the use of acquired video fixation data (§6). Such results empirically verify that the above datasets, despite of different definitions of video objects, capture essential characteristics of moving object patterns.

This paper builds upon our conference paper [1] and significantly extends it in various aspects. First, we perform in-depth analyses on the changes of the stability of human fixation patterns and the alignment between human visual attention and video object determination, with different video processing challenging factors. Second, we provide more details regarding our eye-tracking data collection, the formulation and implementation of our proposed UVOS algorithm. Third, we report much more extensive experimental results with an additional large-scale dataset, FBMS$_{59}$ [9], for further validation. Last but not least, to further demonstrate the effectiveness of our model, we examine its performance on DAVIS$_{17}$ [27] with an instance-level segmentation setting. We expect this work, together with our newly collected data, to provide deep insight into the underlying mechanism behind UVOS and VSOD, and to inspire more related studies. Our eye-tracking data, algorithm implementations and all the segmentation results are made publicly available to the research community at https://github.com/wenguanwang/AGS.

2 Related Work

2.1 Unsupervised Video Object Segmentation

The problem of automatic binary foreground/background video segmentation has been widely addressed in the vision community for more than two decades [2], [3]. Earlier methods mainly focus on *motion analysis*, *i.e.*, extracting information from sequential images to describe movement. They were typically *geometry-based*, constrained to specific families of background-induced motion patterns [5], [28]. Then, *trajectory-based* methods [9], [29]–[34] were proposed to capture long-term motion information and segment all objects which are moving (in 3D) relative to the scene background. These methods, however, are constrained to the accuracy of optical flow estimation, thus easily suffer difficulties in the presence of highly non-rigid motions. Later, more research efforts were devoted to the task of *video segmentation* without human interaction, or more precisely, UVOS. UVOS addresses segmenting the prominent foreground objects, rather than all moving objects, in unconstrained videos. Please note that in UVOS setting, a training set can still be used for fully supervised learning methods, but there is no any interaction or annotation used during the testing phase. Earlier UVOS methods typically employ saliency cues [13], [15], [35], [36] or objectness information [37]–[42] for better identifying the main objects, and rely on certain heuristic priors (*e.g.*, local motion difference [32], background prior [43]). However, they suffer from significant feature engineering and meet problems in case of violation of their underlying assumptions.

More recently, with the renaissance of artificial neural networks, Fragkiadaki *et al.* [40] proposed to learn a multi-layer perceptron based moving objectness detector (MOD) to assist UVOS. This work represents an early attempt towards applying deep learning techniques to this area. However, since the MOD is built upon a fully connected network (FCN), it cannot model spatial information and suffers from heavy computation burden. Later, many fully convolutional networks based models were proposed, which typically adopt two-stream architecture [44], [45] or CNN encoder-decoder structure [14], [46]–[48]. More recently, the Siamese network [49] and graph neural network [50] were introduced to address the issue of multi-frame information mining. The above efforts mainly concentrated on object-level setting, only few recent work [51] addressed instance-level video object segmentation. These representative deep UVOS models generally achieve promising performance, due to the strong learning ability of deep neural networks.

Although UVOS has been extensively studies in several years, a clear definition of 'primary video objects' is still missing. Additionally, many UVOS models [14], [15], [32], [35], [36], [52], [53] use saliency (or foreground-map, a similar notion), they are either heuristic methods lacking end-to-end trainability or based on object-level saliency cues, instead of an explicit, biologically-inspired visual attention representation. None of them quantifies the consistency between visual attention and explicit primary video object determination. Furthermore, previous deep UVOS models are limited to the availability of large-scale well-annotated video data. By contrast, via leveraging dynamic visual attention as an intermediate video object representation, our approach offers a feasible way to alleviate this problem.

2.2 Video Salient Object Detection

VSOD targets at highlighting the most visually important object regions from dynamic videos, *i.e.*, giving a saliency

TABLE 1
Statistics of dynamic eye-tracking datasets. Previous datasets are either collected for bottom-up attention during free-viewing or related to other tasks. By contrast, we extend existing DAVIS$_{16}$ [11], Youtube-Objects [12], and SegTrack$_{V_2}$ [10] datasets with extra UVOS-aware gaze data.

Dataset	Pub.	Year	#Videos	#Viewers	Task
CRCNS [70]	TIP	2004	50	15	scene unders.
Hollywood-2 [71]	TPAMI	2012	1707	19	action recog.
UCF sports [71]	TPAMI	2012	150	19	action recog.
SFU [72]	TIP	2012	12	15	free-view
DHF1K [73]	CVPR	2018	1000	17	free-view
DAVIS$_{16}$ (**Ours**)	-	2019	50	20	UVOS
Youtube-Objects (**Ours**)		2019	126	20	UVOS
SegTrack$_{V_2}$ (**Ours**)		2019	14	20	UVOS

value for each pixel in the videos sequences. This problem recently attracted a great deal of research interest in computer vision since the continuous saliency maps are valuable for a wide range of object-level applications, such as content-aware media re-targeting [54], [55], object tracking [56], and video object segmentation [15]. Being a relatively new task, VSOD can be traced back to the pioneer works in [57] and [58], which are inspired by the studies in visual attention modeling [59]. *Early methods* [13], [15], [60]–[62] largely relied on hand-crafted, low-level features (*e.g.*, color, optical flow, HOG, *etc.*), theories of visual attention in cognitive area (*e.g.*, feature integration theory [19], guided search [20], *etc.*) and heuristics for the salient objects or background (*e.g.*, color contrast [63], background prior [64]). Later, deep learning-based methods become dominant. Some representative works [14], [65]–[67] are built upon well-designed network architectures for semantic segmentation (*e.g.*, FCNs [68], DeepLab [69], *etc.*).

Though the importance of visual attention patterns is emphasized in this task, previous VSOD models rarely explored eye fixation information. In addition, they are typically benchmarked on existing UVOS datasets [10]–[12], while lacking support from experimental evidence on the agreement between video object annotation and visual attention deployment. In this work, we perform behavioral studies that investigate how humans select video primary objects explicitly and how these judgments relate to eye movements during dynamic task-driven viewing. Our findings give an in-depth glimpse into both UVOS and VSOD based on visual attention patterns. Moreover, with the integration of a neural attention mechanism and human eye fixation data, our model goes one step further towards a more biologically inspired VSOD solution.

2.3 Visual Attention Prediction

Human attention mechanism plays an essential role in visual information perception and processing. In the past decade, the computer vision community has made active research efforts on computationally modeling such selective attention process [74]. According to the underlying mechanism, attention models can be categorized as either *bottom-up* (exogenous) or *top-down* (endogenous). The former one is concerned solely with the stimuli and independent of the state of human mind. In such case, there should be some intrinsic property that predicts which stimuli will win and which will lose the competition for attention [75]. In contrast, the willed attentional effects are under clear voluntary control, *i.e.*, the voluntary allocation of attention to certain features, objects, or regions in space. Early attention models [76]–[86] are based on biologically-inspired features (*e.g.*, color, edge, optical flow, *etc.*) and cognitive studies about visual attention (*e.g.*, attention shift [18], feature integration theory [19], guided search [20], *etc.*). Recently, deep learning based attention models [73], [87]–[90] were proposed and generally yield better performance.

Despite the rapid development in this area, most previous methods are static, bottom-up attention models and none of them is specially designed for modeling UVOS-driven, top-down attention in dynamic scenes. Though some eye-tracking datasets [70]–[73] were established and greatly advanced the development of attention modeling in dynamic scenes, they are constructed under free-viewing [72], [73] or other task-driven settings [70], [71] (see Table 1). Differently, in this work, numerous eye gaze data on popular video segmentation datasets [10]–[12] are carefully collected under the ecological constraints of the UVOS task. Consequently, for the first time, a dynamic, top-down attention model is learned for guiding UVOS. With above efforts, we expect to establish a closer link between UVOS, VSOD and visual attention prediction.

2.4 Trainable Attention in Neural Networks

Recent years have witnessed rapid growth of research towards integrating neural networks with fully-differentiable attention mechanism. The neural attention stimulates the human selective attention mechanism and allows the network focus on the most task-relevant parts of the input. It is built upon weighted average instead of hard selection and thus is deterministic. It has shown wide successes in natural language processing and computer vision tasks, such as machine translation [91], image captioning [92], visual question answering [93], and image classification [94], [95], to list a few. Those neural attentions are learned in an implicit, goal-oriented and end-to-end way.

Our DVAP module can be viewed as a neural attention mechanism, as it is end-to-end trainable and used for soft-weighting the feature of AGOS models. It differs from the others in its UVOS-aware nature, explicitly-training ability (with the availability of ground-truth data), and spatiotemporal application domain.

3 UVOS-Aware Eye-Tracking Data Collection

One objective of our work is to contribute extra eye-fixation annotations to three public video segmentation datasets [10]–[12]. Fig. 2 shows some example frames with our UVOS-aware eye-tracking annotation, along with visual attention distributions over each dataset.

Stimuli: In our eye tracking study, the dynamic stimuli are from DAVIS$_{16}$ [11], Youtube-Objects [12], and SegTrack$_{V_2}$ [10]. DAVIS$_{16}$ is a popular UVOS benchmark containing 50 video sequences with totally 3455 frames. Youtube-Objects is a large dataset with 126 videos covering 10 common object categories, with 20,647 frames in total. SegTrack$_{V_2}$ consists of 14 short videos with totally 947 frames.

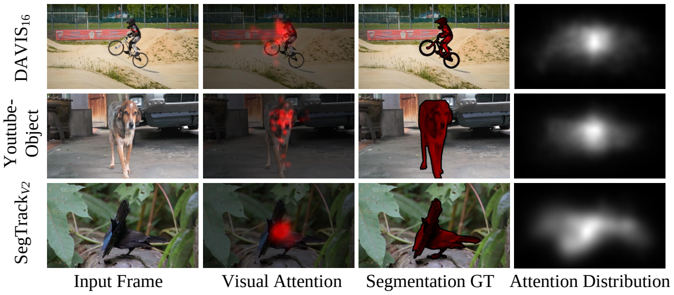


Fig. 2. **Example frames from three datasets** [10]–[12] with our eye-tracking annotation (§3). The last column shows the average attention maps of these datasets. We quantitatively verify (§4) the high consistency between human attention behavior (2nd column) and primary-object determination (3rd column).

TABLE 2
Quantitative results of inter-subject consistency (ISC) and inter-task correlation (ITC), measured by AUC-Juddy. The high ISC-scores over all the three datasets are significantly higher above *chance* (*i.e.*, AUC-Juddy of a random map is 0.5), suggesting the consistency subjects' attention behavior. Moreover, the ITC scores, which are significantly higher than *chance*, demonstrating the strong correlation between top-down dynamic visual attention and human's determination of primary objects. See §4 for details.

Aspect	Metric	DAVIS$_{16}$ [11]	Youtube-Object [12]	SegTrack$_{V_2}$ [10]
ISC	AUC-J *(chance=0.5)*	0.899±0.029	0.876±0.056	0.883±0.036
ITC	AUC-J *(chance=0.5)*	0.704±0.078	0.733±0.105	0.747±0.071

Apparatus: Observer eye movements were recorded using a table-mounted, video-based SMI RED250 eye tracker (SensoMotoric Instruments), with observers' gaze paths at 250Hz as they viewed the series of videos. The dynamic stimuli were displayed on a 19″ computer monitor at a resolution of 1440×900px and in their original speeds. A headrest was used to stabilize the observers' head and maintain a viewing distance of about 68 cm, as advised by the product manual.

Participants: Twenty participants (12 males and 8 females, age range 21-30 years), who passed the eye tracker calibration with less than 10% fixation dropping rate, were qualified for our experiment. All had normal/corrected-to-normal vision and never seen the stimuli before.

Recording protocol: The subjects were informed that they would watch a series of unrelated silent video clips. The stimuli were equally partitioned into 10 non-overlapping sessions and their original frame rates and aspect ratios are maintained during gaze-data capturing. The experimenters first ran the standard SMI calibration routine with recommended settings for the best results. The calibration procedure was repeated until an acceptable calibration was obtained as determined by means of validation procedure offered by the product. This procedure expected participants to look at four small circles near the middle of the screen. The calibration was considered to be acceptable if a fixation was shown for each circle and no fixation appeared in an obvious outlier position. During viewing, the stimulus videos were displayed in random order and *the participants were instructed to identify the primary object occurring in each stimulus*. Since we aim to explore human attention behavior in UVOS setting, each stimulus was repeatedly displayed three times to help the participants better capture the video content. Such data capturing design is inspired by the protocol in [72]. To avoid eye fatigue, 5-second black screen was intercalated between each. Additionally, the stimuli were split into 5 sessions. After undergoing a session of videos, the participant can take a rest. Finally, a total of 12 318 862 fixations were recorded from 20 subjects on 190 videos. To obtain a continuous *fixation map* from the eye tracking data, for each frame, we convolve all the fixation locations with a small Gaussian filter. As suggested by [96], [97], the size of the Gaussian is chosen as one degree of visual angle (~ 30 image pixels in our case). The fixation map is normalized to range between zero and one.

4 In-depth Data Analysis

Inter-subject consistency: We first conduct experiments to analyze eye movement consistency within subjects. To quantify such inter-subject consistency (ISC), following the protocols in [25], data from half of the subjects are randomly selected as the test subset, leaving the rest as the new ground-truth subset. After that, AUC-Juddy [98], a classic visual attention evaluation metric, is employed to the test subset to measure ISC. The experimental results are shown in Table 2. It is interesting to find that there exists high consistency of top-down attention behaviors among human subjects, across all the three datasets. The correlation scores (0.899 on DAVIS$_{16}$, 0.876 on Youtube-Object, 0.883 on SegTrack$_{V_2}$) are significantly above chance (0.5). The chance level is the accuracy of a random map with value of each pixel drawn uniformly random between 0 and 1. This novel observation further suggests that, even though 'unsupervised video object(s)' is often considered as ill-defined [99]–[101], there do exist some 'universally-agreed' visually important clues that attract human attentions stably and consistently.

Correlation between visual attention and video object determination: It is essential to study whether human visual attention and video primary object judgment agree with each other, which has never been explored before. Here we apply the experimental protocol suggested by [102] to calculate the inter-task correlation (ITC). More specifically, we use the segmentation mask to explain the fixation map. During the computation of AUC-Juddy metric, human fixations are considered as the positive set and some points sampled from other non-fixation positions as the negative one. The segmentation mask is then used as a binary classifier to separate positive samples from negative samples. The results are reported in Table 2, showing that visual attention does not fall on the background significantly higher than its corresponding chance level. Taking Youtube-Objects as an example, the correlation score 0.733 (std = 0.105) is significantly above chance using t-test ($p < 0.05$). This observation reveals the strong correlation between human dynamic visual attention and video object determination. It is also in alignment with our expectation that foreground object(s) should be resided in visually important regions that attract human attention most.

TABLE 3
Attribute-based aggregate results of inter-subject consistency (ISC) and inter-task correlation (ITC) on DAVIS$_{16}$ with AUC-Juddy. Significant performance drops (more than 1% decrease compared with the average score) are denoted by underline. See §4 for details.

Dataset	Metric	Type	Attribute															Avg.
			AC	BC	CS	DB	DEF	EA	FM	HO	IO	LR	MB	OCC	OV	SC	SV	
DAVIS$_{16}$	AUC-J	ISC	0.899	0.893	0.909	0.910	0.902	0.903	0.898	0.901	0.898	0.906	0.899	0.903	0.903	0.896	0.902	0.899
		ITC	0.702	0.737	0.709	0.711	0.699	0.702	0.678	0.689	0.682	0.661	0.675	0.690	0.686	0.681	0.688	0.704

Performance change of ISC and ITC with video processing challenges: In DAVIS$_{16}$, the videos are categorized according to their various attributes, *i.e.*, appearance change (AC), background clutter (BC), camera shake (CS), dynamic background (DB), deformation (DEF), edge ambiguity (EA), fast motion (FM), heterogeneous objects (HO), interesting objects (IO), low resolution (LR), motion blur (MB), occlusion (OCC), out-of-view (OV), shape complexity (SC), and scale-variation (SV). The attributes are major challenges typically faced in video processing, and allow us to further study the influence of these challenging factors on ISC and ITC. In Table 3 we report the AUC-Juddy scores of ISC and ITC on different subsets of DAVIS$_{16}$ characterized by a particular attribute and underline the significant performance drops with 1% decrease with respective to the average scores.

From the attribute-based analysis, we first observe that the willed attentional effects among human subjects are still highly correlated, and the scores of ISC did not change significantly over different challenges. This implies that our recorded attention is in a top-down form, *i.e.*, driven by human endogenous expectations (or values), being independent of the extrinsic property of stimuli. This also suggests the importance of expectations in the operation of top-down attentional biasing. Based on the pre-defined task instruction, the human beings share similar values that determine consistent top-down biases [75], ignoring the challenging factors in external circumstances. We then find that some challenging factors (*i.e.*, FM, MB) decrease the ITC score significantly. This demonstrates that, even for human beings, it is hard to handle some challenges. In addition, the significant drop of ITC score on the FM subset further supports the view in cognitive science that human dynamic attention, as a tracking mechanism, estimates the position of targets with some uncertainty [103]. As the speed at the target move increases, the positional uncertainty increases, leading to inferior ITC score. Another interesting phenomena is that the observers perform very well with the scenes with BC. This seems to imply that human visual attention handles BC well or may be just because the scenes with BC in DAVIS$_{16}$ happen to be relatively easy. Further, human beings have little difficulty distinguishing foreground objects from the background when facing some challenges such as AC, CS, and EA.

5 Proposed UVOS Method

According to our analyses in §4, human top-down visual attention and video object determination are highly correlated. We therefore decompose UVOS into two cascaded sub-tasks (§5.1): predicting UVOS-aware visual attention in spatiotemporal domain and then performing attention-guided object segmentation for each individual frame. Accordingly, our model is devised as a unity of two tightly coupled sub-networks (see Fig. 3 (a)): *dynamic visual attention prediction* (DVAP) module (§5.2), and *attention-guided object segmentation* (AGOS) module (§5.3). Specifically, DVAP captures spatiotemporal foreground characteristics via dynamic UVOS-aware attention. It provides an explicit attention mechanism to allow AGOS focus on visually important areas and produce precise object segments. DVAP benefits from our collected UVOS-aware gaze-tracking data, while AGOS can be trained with attention-segmentation paired groundtruth from existing image (*instead of video*) segmentation datasets [24], [25]. The two modules share weights of several initial convolution layers, leading to a unified, end-to-end trainable UVOS model without using expensive pixel-wise video segmentation annotation (§5.4).

5.1 Problem Formulation

Denote an input video with T frames as $\{\mathbf{I}_t\in\mathbb{R}^{W\times H\times 3}\}_{t=1}^T$, then the goal of UVOS is to generate the corresponding sequences of binary video object segmentation-masks $\{\mathbf{S}_t\in\{0,1\}^{W\times H}\}_{t=1}^T$. Many recently proposed UVOS methods [14], [44], [45], [48] learn a DNN as a mapping function $\mathcal{F}_{\text{UVOS}}\!:\mathbb{R}^{W\times H\times 3\times T}\!\rightarrow\!\{0,1\}^{W\times H\times T}$ that directly maps the input into the segmentation masks:

$$\{\mathbf{S}_t\}_{t=1}^T = \mathcal{F}_{\text{UVOS}}(\{\mathbf{I}_t\}_{t=1}^T). \tag{1}$$

To learn such *direct input-output mapping* $\mathcal{F}_{\text{UVOS}}$, numerous pixel-wise video segmentation annotations are needed, which are however very expensive to obtain.

In this work, we instead propose an *input-attention-output mapping* strategy for UVOS. Specifically, a DVAP module $\mathcal{F}_{\text{DVAP}}$ is first designed to predict dynamic UVOS-aware visual attentions $\{\mathbf{A}_t\in[0,1]^{W'\times H'\times}\}_{t=1}^T$:

$$\{\mathbf{A}_t\}_{t=1}^T = \mathcal{F}_{\text{DVAP}}(\{\mathbf{I}_t\}_{t=1}^T). \tag{2}$$

An AGOS module $\mathcal{F}_{\text{AGOS}}$, which takes a single frame image $\mathbf{I}_t$ and corresponding attention map $\mathbf{A}_t$ as input, is then used to generate final segmentation result $\mathbf{S}_t$:

$$\mathbf{S}_t = \mathcal{F}_{\text{AGOS}}(\mathbf{I}_t, \mathbf{A}_t), \qquad t\in\{1,2,\ldots,T\}. \tag{3}$$

As shown in Fig. 3 (a), $\{\mathbf{A}_t\}_{t=1}^T$ encode both static object infomation and temporal dynamics, enabling AGOS to focus on fine-grained segmentation in spatial domain, *i.e.*, applying AGOS for each frame individually. Essentially, the visual attention, as a biologically-inspired visual cue and intermediate object representation, links DVAP and AGOS together, and offers an explicit interpretation by telling where our model is looking at. Next we elaborate our DVAP and AGOS modules in details.

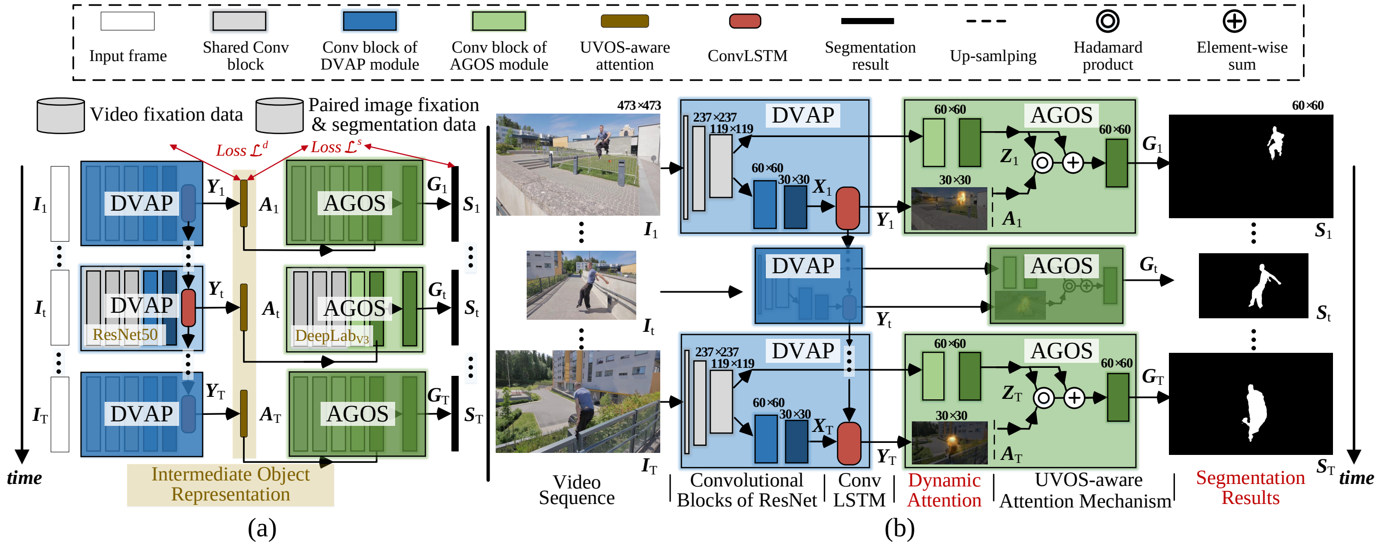


Fig. 3. **Illustration of the proposed UVOS model.** (a) Simplified schematization of our model that solves UVOS in a two-step manner, without the need of training with expensive precise video object masks. (b) Detailed network architecture, where the DVAP (§5.2) and AGOS (§5.3) modules share the weights of three bottom convolution blocks. The UVOS-aware attention acts as an intermediate object representation that connects the two modules densely. Best viewed in color. Zoom in for details.

5.2 DVAP Module

The DVAP module is built on a CNN-convLSTM architecture (see Fig. 3 (b)), where the CNN layers are borrowed from the first five convolutional blocks of ResNet101 [104]. To preserve more spatial details, we reduce the stride of the last block to 1. Given the input video sequence $\{\mathbf{I}_t\}_{t=1}^T$ with typical 473×473 spatial resolution, the spatial feature sequence $\{\mathbf{X}_t\in\mathbb{R}^{30\times30\times2048}\}_{t=1}^T$ from the top-layer of the CNN network is fed into a convLSTM for learning the dynamic visual attention. ConvLSTM [105], proposed as a convolutional counterpart of conventional fully connected LSTM, introduces convolution operation into input-to-state and state-to-state transitions. ConvLSTM is favored here as it preserves spatial details as well as modeling temporal dynamics simultaneously. Our DVAP module $\mathcal{F}_{\text{DVAP}}$ can be formulated as follows:

$$\begin{aligned}\mathbf{X}_t &= \text{CNN}(\mathbf{I}_t),\\ \mathbf{Y}_t &= \text{convLSTM}(\mathbf{X}_t,\mathbf{Y}_{t-1}),\\ \mathbf{A}_t &= \mathcal{R}(\mathbf{Y}_t),\end{aligned}\tag{4}$$

where $\mathbf{Y}_t$ indicates the 3D-tensor hidden state (with 32 channels) of convLSTM at time step t. $\mathcal{R}$ is a readout function that produces the attention map from the hidden state, implemented as a 1×1 convolution layer with the *sigmoid* activation function.

In the next section, we employ DVAP as an attention mechanism to guide AGOS to concentrate more on the visually important regions. An extra advantage of such design lies in disentangling spatial and temporal characteristics of foreground objects, as DVAP captures temporal information by learning from dynamic-gaze data, and thus allows AGOS to focus on pixel-wise segmentation only in spatial domain (benefiting from existing large-scale image datasets with paired fixation and object segmentation annotation).

5.3 AGOS Module

The attention obtained from DVAP suggests the location of the primary object(s), offering informative cue to AGOS for pixel-wise segmentation, as achieved by a neural attention architecture. Before going deep into our model, we first give a general formulation of neural attention mechanisms.

General neural attention mechanism: A neural attention mechanism equips a network with the ability to focus on a subset of input feature. It computes a soft-mask to enhance the feature by multiplication operation. Formally, let $\mathbf{i}\in\mathbb{R}^d$ be an input vector, $\mathbf{z}\in\mathbb{R}^k$ a feature vector, $\mathbf{a}\in[0,1]^k$ an attention vector, $\mathbf{g}\in\mathbb{R}^k$ an attention-enhanced feature and f_{A} an attention network. The neural attention is implemented as:

$$\begin{aligned}\textit{Attention}:\quad \mathbf{a} &= f_{\text{A}}(\mathbf{i}),\\ \textit{Feature enhancement}:\quad \mathbf{z} &= f_{\text{Z}}(\mathbf{i}),\\ \mathbf{g} &= \mathbf{a}\odot\mathbf{z},\end{aligned}\tag{5}$$

where $\odot$ is element-wise multiplication, and f_{Z} indicates a feature extraction network. Some neural attention models [92] equip attention function f_{A} with *soft-max* to constraint the values of attention between 0 and 1. Since the above attention framework is fully differentiable, it is end-to-end trainable. However, due to the lack of 'ground-truth' of the attention, it is trained in an *implicit* way.

Explicit, spatiotemporal, and UVOS-aware attention mechanism: We integrate DVAP into AGOS as an attention mechanism. Let $\mathbf{Z}_t,\mathbf{G}_t$ denote respectively a segmentation feature and an attention glimpse with the same dimensions, our UVOS-aware attention is formulated as:

$$\begin{aligned}\textit{Spatiotemporal attention}:\quad \{\mathbf{A}_t\}_{t=1}^T &= \mathcal{F}_{\text{DVAP}}(\{\mathbf{I}_t\}_{t=1}^T),\\ \textit{Spatial feature enhancement}:\quad \mathbf{Z}_t &= \mathcal{F}_{\text{Z}}(\mathbf{I}_t),\\ \mathbf{G}_t^c &= \mathbf{A}_t\odot\mathbf{Z}_t^c,\end{aligned}\tag{6}$$

where $\mathcal{F}_{\text{Z}}$ extracts segmentation features from the input frame $\mathbf{I}_t$ (will be detailed latter). $\mathbf{G}^c$ and $\mathbf{Z}^c$ indicate the feature slices of $\mathbf{G}$ and $\mathbf{Z}$ in the c-th channel, respectively. As seen, our UVOS-aware attention encodes spatial foreground information as well as temporal characteristics, enabling the AGOS module perform object segmentation over each frame individually. For the position with an attention value close to 0, the corresponding feature response will be suppressed greatly. This may lose some meaningful information. Inspired by [95], [104], the feature enhancement step in Eq. 6 is improved with a residual form (see Fig. 3 (b)):

$$\mathbf{G}_t^c = (1 + \mathbf{A}_t) \odot \mathbf{Z}_t^c. \tag{7}$$

This strategy retains the original information (even with a very small attention value), while enhances object-relevant features efficiently. Besides, due to the availability of the ground-truth gaze data, our UVOS-aware attention mechanism is trained in an *explicit* manner (detailed in §5.4).

The AGOS module is also built upon convolutional blocks of ResNet101 [104] and modified with the ASPP module proposed in DeepLab$_{V3}$ [69]. With an input frame image $\mathbf{I}_t \in \mathbb{R}^{473\times473\times3}$, a segmentation feature $\mathbf{Z}_t \in \mathbb{R}^{60\times60\times1536}$ can be extracted from the ASPP module $\mathcal{F}_{\text{ASPP}}$. The attention map $\mathbf{A}_t$ is also ×2 upsampled by bilinear interpolation. Finally, our AGOS module in Eq. 6 is implemented as:

$$\begin{array}{lrl} \textit{Spatiotemporal attention:} & \{\mathbf{A}_t\}_{t=1}^T &= \mathcal{F}_{\text{DVAP}}(\{\mathbf{I}_t\}_{t=1}^T), \\ \textit{Spatial feature enhancement:} & \mathbf{Z}_t &= \mathcal{F}_{\text{ASPP}}(\mathbf{I}_t), \\ & \mathbf{G}_t^c &= (1 + \mathbf{A}_t) \odot \mathbf{Z}_t^c. \end{array} \tag{8}$$

Knowledge sharing between DVAP and AGOS: DVAP and AGOS modules share similar underlying network architectures (*conv1-conv5* of ResNet101), while capturing object information from different perspectives. We develop a technique to encourage knowledge sharing between the two networks, rather than learning each of them separately. In particular, we allow the two modules to share the weights of the first three convolutional blocks (*conv1*, *conv2*, and *conv3*), and then learn other higher-level layers separately. This is because the bottom-layers typically capture low-level information (edge, corner, *etc*.), while the top-layers tend to learn high-level, task-specific knowledge. Moreover, such weight-sharing strategy improves our computational efficiency and decreases parameter storage. See §6.6 for more detailed experiments regarding our knowledge sharing strategy.

5.4 Implementation Details

Training loss: For DAVP, given an input frame $\mathbf{I} \in R^{473\times473\times3}$, it predicts an attention map $\mathbf{A} \in [0,1]^{30\times30}$. Here the temporal subscript is omitted for simplicity. Denote by $\mathbf{P} \in [0,1]^{30\times30}$ and $\mathbf{F} \in \{0,1\}^{30\times30}$ the ground-truth continuous attention map and the binary fixation map, respectively. $\mathbf{F}$ is a discrete map, recording whether a pixel receives human-eye fixation position, and $\mathbf{P}$ is obtained by blurring $\mathbf{F}$ with a small Gaussian filter. Inspired by [87], the loss function $\mathcal{L}_{\text{DVAP}}$ for DAVP is designed as:

$$\begin{aligned} \mathcal{L}_{\text{DVAP}}(\mathbf{A},\mathbf{P},\mathbf{F}) =& \mathcal{L}_{\text{CE}}(\mathbf{A},\mathbf{P}) + \alpha_1\mathcal{L}_{\text{NSS}}(\mathbf{A},\mathbf{F}) + \\ & \alpha_2\mathcal{L}_{\text{SIM}}(\mathbf{A},\mathbf{F}) + \alpha_3\mathcal{L}_{\text{CC}}(\mathbf{A},\mathbf{P}), \end{aligned} \tag{9}$$

where the $\mathcal{L}_{\text{CE}}$ indicates the classic *cross entropy* loss, and $\mathcal{L}_{\text{CC}}, \mathcal{L}_{\text{NSS}}, \mathcal{L}_{\text{SIM}}$ are derived respectively from three widely-used visual attention evaluation metrics named *Normalized Scanpath Saliency (NSS)*, *Similarity Metric (SIM)* and *Linear Correlation Coefficient (CC)*. Such combination leads to improved performance due to comprehensive consideration of different quantification factors as in [87]. We use $\mathcal{L}_{\text{CE}}$ as the primary loss, and set $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$.

For AGOS, given $\mathbf{I}$, it produces the final segmentation prediction[2] $\mathbf{S} \in [0,1]^{60\times60}$. Let $\mathbf{M} \in \{0,1\}^{60\times60}$ denote the ground-truth binary segmentation mask, the loss function $\mathcal{L}_{\text{AGOS}}$ of the AGOS module is formulated as:

$$\mathcal{L}_{\text{AGOS}}(\mathbf{S},\mathbf{M}) = \mathcal{L}_{\text{CE}}(\mathbf{S},\mathbf{M}). \tag{10}$$

Training protocol: We leverage both video gaze data and attention-segmentation paired image data to train our whole UVOS model. The training process is iteratively performed on a video training batch and an image train batch. Specifically, in the video training batch, we use dynamic gaze data to train the DVAP module only. Given the training video sequence $\{\mathbf{I}_t\}_{t=1}^T$, let $\{\mathbf{A}_t, \mathbf{P}_t, \mathbf{F}_t\}_{t=1}^T$ denote the corresponding attention predictions, ground-truth continuous attention maps and discrete fixation maps, we train our model by minimizing the following loss (see Fig. 3 (a)):

$$\mathcal{L}^d = \sum\nolimits_{t=1}^{T} \mathcal{L}_{\text{DVAP}}(\mathbf{A}_t^d, \mathbf{P}_t^d, \mathbf{F}_t^d), \tag{11}$$

where the superscript 'd' represents *dynamic* video data. Note that we do not consider $\mathcal{L}_{\text{AGOS}}$ loss to save the expensive pixel-wise segmentation ground-truth.

The image training batch contains several attention-segmentation paired image masks, which are used to train both DVAP and AGOS modules simultaneously. Let $\{\mathbf{I},\mathbf{S},\mathbf{F},\mathbf{M}\}$ denote a training sample in the image training batch, which includes a static image and corresponding ground-truth (*i.e.*, continuous attention map, binary fixation map, and segmentation mask). The overall loss function combines both $\mathcal{L}_{\text{DVAP}}$ and $\mathcal{L}_{\text{AGOS}}$:

$$\mathcal{L}^s = \mathcal{L}_{\text{DVAP}}(\mathbf{A}^s, \mathbf{P}^s, \mathbf{F}^s) + \mathcal{L}_{\text{AGOS}}(\mathbf{S}^s, \mathbf{M}^s), \tag{12}$$

where a superscript 's' is used to emphasize the *static* nature. By using static data, the total time span of convLSTM in DVAP is set to 1. This way, the convLSTM can be viewed as a convolution layer, and the weights of the dynamic state transition matrix are skipped. The rationale here is intuitive: although the DVAP module is a dynamic attention model, it should also work well for static scenes. Such a design also improves the generalization ability of the DVAP module.

Our model is implemented in a modified version of Caffe, and trained using the SGD optimizer. The learning rate is set to 0.0001 and is decreased by a factor of 10 every 2 epochs. The network is trained for 10 epochs. Each video training batch uses 2 videos, each with 3 consecutive frames. Both the videos and the start frames are randomly selected. Each image training batch contains 6 randomly sampled images. Data augmentation (*e.g.*, flipping, cropping) is also performed. The whole training procedure takes around ∼30 hours on a Titan-X GPU.

6 Experiments

In this section, we first elaborate our training and testing protocols in §6.1. Then we investigate the performance of

[2] We slightly reuse $\mathbf{S}$ for representing the segmentation prediction.

TABLE 4
Statistics of the training and testing datasets, where 'Fix.', 'O.-Seg.' and 'I.-Seg., indicate 'eye fixation annotation', 'object-level segmentation annotation', and 'instance-level segmentation annotation', respectively. During training, we only employ dynamic fixation video data and fixation-segmentation paired static image data, saving our method from the labor-expensive video segmentation annotation data.

Datasets	Video Datasets													Image Datasets			
	DAVIS$_{16}$ [11]				DAVIS$_{17}$ [27]			FBMS$_{59}$ [9]		Youtube-Object [12]		Seg-Track$_{V2}$ [10]		DUT-O [24]		PASCAL-S [25]	
	Train		Test		Train	Val	Test-dev	Train	Test								
#Videos	30		20		60	30	30	29	30	126		14		-		-	
#Images	2,079		1,376		4,209	1,999	2,294	6,554	7,306	20,647		4,447		5,168		850	
Annotation	Fix.	O.-Seg.	Fix.	O.-Seg.	I.-Seg.	I.-Seg.	I.-Seg.	O.-Seg.	O.-Seg.	Fix.	O.-Seg.	Fix.	O.-Seg.	Fix.	O.-Seg.	Fix.	O.-Seg.
Train Phrase	✓											✓		✓	✓	✓	✓
Test Phrase			✓	✓		✓	✓		✓	✓	✓						

our DVAP module for the dynamic fixation prediction task (§6.2). We further evaluate the performance of our whole model for the object-level UVOS task (§6.3). After that, in §6.4, we conduct experiments on the instance-level UVOS setting. Latter, in §6.5, we test our model on the VSOD setting. To gain a deeper insight into our model, we conduct detailed ablative studies in §6.6. Finally, run time analysis is presented in §6.7. For fairness, all the experiments are performed on a workstation with an Intel Core i7-6700 @3.4GHz CPU and a GTX 1080Ti GPU.

6.1 Experimental Setting

Training data: During training, we use the video sequences and corresponding fixation data from the training split of DAVIS$_{16}$ [11] and the whole SegTrack$_{V_2}$ [10], leading to totally 54 videos with 6,526 frames. Additionally, two image salient object datasets, DUT-O [24] and PASCAL-S [25], offer both static gaze data and segmentation annotations, and are thus also used in our training phase, resulting in totally 6,018 static training examples. More statistics about our training data can be found in Table 4. As seen, our model is trained without labor-intensive pixel-wise video segmentation masks, by leveraging easily-acquired dynamic gaze data and static attention-segmentation annotation pairs. In §6.3, we quantitatively demonstrate that, even without training on video segmentation annotations, the suggested model is still able to achieve state-of-the-art performance.

Testing phase: Given a test video, all the frames are uniformly resized to 473 × 473 and fed into our model for the corresponding primary object predictions. Following the common protocol [14], [46], [106] in video segmentation, the fully-connected CRF [107] is employed to obtain the final binary segmentation results. For each frame, the forward propagation of our network takes about 0.1s, while the CRF-based post-processing takes about 0.5s.

6.2 Performance of DVAP Module

Test datasets: We evaluate our DVAP module on the test set of DAVIS$_{16}$ [11] and the full Youtube-Objects [12], with the gaze-tracking ground-truth (see details in Table 4), and there is no overlap between the training and test data.

Evaluation metrics: Five standard metrics: AUC-Judd (AUC-J), shuffled AUC (s-AUC), NSS, SIM, and CC, are used for comprehensive study (see [109] for details).

Quantitative and qualitative results: We compare our DVAP module with 12 state-of-the-art visual attention models, including 5 deep models [73], [87], [88], [90], [108] and 6 traditional ones [74], [80], [83]–[86]. The results are obtained through running their publicly released codes with default parameters. Quantitative results are summarized in Tables 5 and 6. As seen, our DVAP generally outperforms other competitors, as none of them is specifically designed for UVOS-aware attention prediction. Our promising quantitative results, as well as qualitative results in the middle row in Fig. 4, verify that DVAP can guide our UVOS model to accurately attend to visually attractive regions in videos.

TABLE 5
Quantitative comparison of visual attention models on the test set of DAVIS$_{16}$ [11] (§6.2). (The three best scores are indicated in **red**, **blue** and **green**, respectively. * indicates deep learning based models. These notes are the same for Tables 6,7,8,9,10 and 12.)

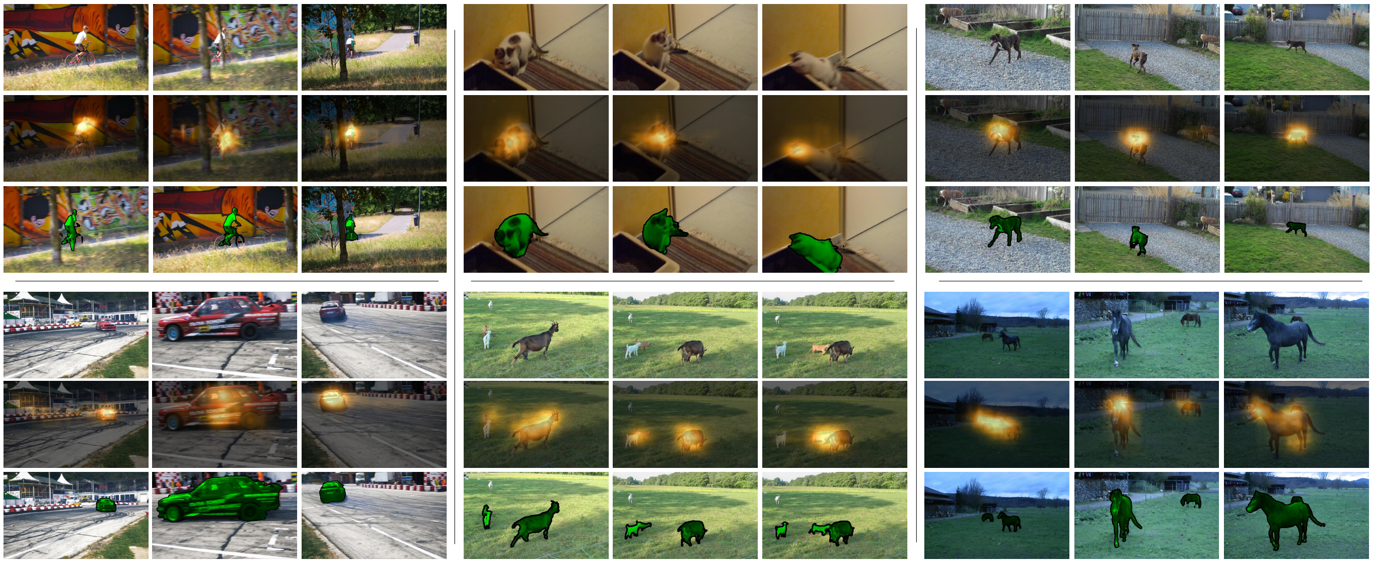
Dataset	Methods	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
DAVIS$_{16}$	*ACL [73]	**0.901**	**0.453**	0.617	**0.559**	**2.252**
	*OMCNN [108]	0.889	**0.408**	0.621	**0.518**	**2.101**
	*DVA [90]	0.885	0.382	**0.647**	0.494	1.906
	*DeepNet [88]	0.880	0.318	**0.644**	0.470	1.866
	*ShallowNet [88]	0.874	0.293	0.622	0.471	1.871
	*SALICON [87]	0.818	0.276	0.628	0.352	1.432
	STUW [83]	**0.892**	0.363	0.636	0.508	2.019
	PQFT [84]	0.685	0.202	0.584	0.191	0.821
	Seo *et al.* [86]	0.724	0.234	0.582	0.222	0.923
	Hou *et al.* [85]	0.782	0.263	0.581	0.273	1.119
	GBVS [80]	0.882	0.294	0.617	0.442	1.683
	ITTI [74]	0.820	0.249	0.621	0.354	1.332
	***Ours**	**0.909**	**0.504**	**0.667**	**0.620**	**2.507**

TABLE 6
Quantitative comparison of different visual attention models on Youtube-Objects [12] (§6.2).

Dataset	Methods	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
Youtube-Objects	*ACL [73]	**0.912**	**0.405**	0.711	**0.531**	**2.627**
	*OMCNN [108]	0.889	0.326	0.698	0.461	**2.307**
	*DVA [90]	**0.905**	**0.372**	**0.741**	**0.526**	2.294
	*DeepNet [88]	0.894	0.268	**0.737**	0.448	2.182
	*ShallowNet [88]	0.890	0.252	0.704	0.436	2.069
	*SALICON [87]	0.840	0.265	0.692	0.380	1.956
	STUW [83]	0.869	0.264	0.666	0.388	1.876
	PQFT [84]	0.730	0.170	0.646	0.210	1.061
	Hou *et al.* [85]	0.786	0.221	0.639	0.243	1.223
	Seo *et al.* [86]	0.763	0.210	0.605	0.224	1.118
	GBVS [80]	0.881	0.244	0.706	0.395	1.919
	ITTI [74]	0.837	0.214	0.709	0.339	1.638
	***Ours**	**0.914**	**0.419**	**0.747**	**0.543**	**2.700**

6.3 Performance of Our Full UVOS Model

Test datasets: Following the common protocol in [14], [44], [67], the test sets of DAVIS$_{16}$ [11] and FBMS$_{59}$ [9], and the whole Youtube-Objects [12] are used for assessing the performance of our full UVOS model (see Table 4). Again, there is no overlap between the training and test data.



DAVIS$_{16}$ [11] Youtube-Objects [12] FBMS$_{59}$ [9]

Fig. 4. **Visual results of predicted attention and primary video object mask on three datasets (each with two example videos).** For each video, the dynamic attention results from our DVAP module are shown in the second row, which are biologically-inspired and used to guide our AGOS module for fine-grained UVOS (see the last row).

TABLE 7
Quantitative UVOS results on the test sequences of DAVIS$_{16}$ [11]. The results are selected from the public leaderboard (https://davischallenge.org/davis2016/soa_compare.html) maintained by the DAVIS challenge (until Aug. 2019). See §6.3 for details.

Dataset	Metric	*Ours	*MOT [110]	*LSMO [106]	*PDB [14]	*ARP [42]	*LVO [44]	*FSEG [45]	*LMP [46]	*SFL [47]
DAVIS$_{16}$	Mean ↑	**79.7**	**77.2**	**78.2**	**77.2**	76.2	75.9	70.7	70.0	67.4
	$\mathcal{J}$ Recall ↑	**91.1**	87.8	**89.1**	**90.1**	**91.1**	**89.1**	83.5	85.0	81.4
	Decay ↓	1.9	5.0	4.1	**0.9**	7.0	**0.0**	1.5	**1.3**	6.2
	Mean ↑	**77.4**	**77.4**	**75.9**	**74.5**	70.6	72.1	65.3	65.9	66.7
	$\mathcal{F}$ Recall ↑	**85.8**	**84.4**	**84.7**	**84.4**	83.5	83.4	73.8	79.2	77.1
	Decay ↓	**0.0**	3.3	3.5	**-0.2**	7.9	**1.3**	1.8	2.5	5.1
	$\mathcal{T}$ Mean ↓	26.7	27.9	**21.2**	29.1	39.3	**26.5**	32.8	57.2	28.2

Dataset	Metric	*Ours	FST [32]	CUT [34]	NLC [35]	MSG [30]	KEY [37]	CVOS [111]	TRC [31]	SAG [15]
DAVIS$_{16}$	Mean ↑	**79.7**	55.8	55.2	55.1	53.3	49.8	48.2	47.3	42.6
	$\mathcal{J}$ Recall ↑	**91.1**	64.9	57.5	55.8	61.6	59.1	54.0	49.3	38.6
	Decay ↓	1.9	**0.0**	2.2	12.6	2.4	14.1	10.5	8.3	8.4
	Mean ↑	**77.4**	51.1	55.2	52.3	50.8	42.7	44.7	44.1	38.3
	$\mathcal{F}$ Recall ↑	**85.8**	51.6	61.0	51.9	60.0	37.5	52.6	43.6	26.4
	Decay ↓	**0.0**	2.9	3.4	11.4	5.1	10.6	11.7	12.9	7.2
	$\mathcal{T}$ Mean ↓	26.7	36.6	27.7	42.5	30.2	26.9	**25.0**	39.1	60.0

Evaluation metrics: For the UVOS task, we use three standard metrics suggested by [11], *i.e.*, region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and time stability $\mathcal{T}$. In addition, in FBMS$_{59}$, we report F-measure [9], $\bar{\mathcal{F}} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$.

Quantitative and qualitative results: For DAVIS$_{16}$, we compare our model with 16 state-of-the-art UVOS methods [14], [15], [30]–[32], [34], [35], [37], [42], [44]–[47], [106], [110], [111], which are selected from the public leaderboard maintained by the DAVIS challenge. For Youtube-Objects and FBMS$_{59}$, the competitors include [9], [14], [32], [42], [44], [45], [47], [106], [110], [112] and [14], [32], [35], [36], [39], [42], [45], [47], [48], [106], [110], [113], respectively. The results are obtained from their literatures or by running their implementations with default parameter settings. The quantitative comparison results over above three datasets are reported in Tables 7, 8 and 9, respectively. We can observe that our model outperforms other contenders over most metrics across all the datasets. This is significant and distinguishes our model from previous deep UVOS models since our model is trained without precise segmentation mask ground-truths.

It is worth noticing that some methods, such as MOT [110], LSMO [106] and ARP [42], though showing promising results on DAVIS$_{16}$, suffer from performance drops on Youtube-Objects and FBMS$_{59}$. This observation suggests that current promising methods may suffer from an overfitting issue on DAVIS$_{16}$. In contrast, our method consistently achieves state-of-the-art over all the three datasets, which demonstrating our strong generalization capability.

Some qualitative results on the three datasets are shown in the last row in Fig. 4, validating our model yields high-quality UVOS results with interpretable dynamic attentions.

Attribute-based study: Table 10 lists the per-attribute based evaluation on the test set of DAVIS$_{16}$ [11]. Among all the 15 video attribute categories, our method gets best scores on 10, the second best on 5, and the best average performance. It attains significantly increased robustness against background clutter (BC), camera shake (CS), fast motion (FM), heterogeneous objects (HO), low resolution (LR), motion blur (MB), and out-of-view (OV). Overall, our method handles well various challenges present in videos.

6.4 Performance on Instance-Level UVOS

Test datasets: To completely examine the performance of our model, we modify and test our model on the instance-level UVOS object setting (also referred as "multi-object unsupervised video segmentation" [27]). Different from previous general UVOS setting, which focus on object-level foreground/background binary separation, instance-level UVOS further concerns the extraction of each individual

TABLE 8
Quantitative UVOS results on Youtube-Objects [12] with the region similarity $\mathcal{J}$. Performance over each category and the average score are reported. See §6.3 for details.

Dataset	Metric	Category	*Ours	*MOT [110]	*LSMO [106]	*PDB [14]	*ARP [42]	*LVO [44]	*SFL [47]	*FSEG [45]	FST [32]	CSEG [112]	LTV [9]
Youtube-Object	$\mathcal{J}$ Mean ↑	Airplane	**87.7**	77.2	60.5	78.0	73.6	**86.2**	65.6	**81.7**	70.9	69.3	13.7
		Bird	**76.7**	42.2	59.3	**80.0**	56.1	**81.0**	65.4	63.8	70.6	76.0	12.2
		Boat	**72.2**	49.3	62.1	58.9	57.8	**68.5**	59.9	**72.3**	42.5	53.5	10.8
		Car	**78.6**	68.6	72.3	**76.5**	33.9	69.3	64.0	**74.9**	65.2	70.4	23.7
		Cat	**69.2**	46.3	66.3	63.0	30.5	58.8	58.9	**68.4**	52.1	**66.8**	18.6
		Cow	64.6	64.2	**67.9**	64.1	41.8	**68.5**	51.2	**68.0**	44.5	49.0	16.3
		Dog	**73.3**	66.1	**70.0**	**70.1**	36.8	61.7	54.1	69.4	65.3	47.5	18.2
		Horse	64.4	**64.8**	**65.4**	**67.6**	44.3	53.9	**64.8**	60.4	53.5	55.7	11.5
		Moto.	**62.1**	44.6	55.5	58.4	48.9	**60.8**	52.6	**62.7**	44.2	39.5	10.6
		Train	48.2	42.3	38.0	35.3	39.2	**66.3**	34.0	**62.2**	29.6	**53.4**	19.6
		Avg.	**69.7**	58.1	64.3	65.5	46.2	**67.5**	57.1	**68.4**	53.8	58.1	15.5

TABLE 9
Quantitative UVOS results on the test sequences of FBMS$_{59}$ [9] with the region similarity $\mathcal{J}$ and F-measure $\bar{\mathcal{F}}$. See §6.3 for details.

Dataset	Metric	*Ours	*MOT [110]	*LSMO [106]	*PDB [14]	*ARP [42]	*LVO [44]	*OBN [48]	*IET [113]	*SFL [47]	*FSEG [45]	ACO [36]	FST [32]	STO [39]	NLC [35]
FBMS$_{59}$	$\mathcal{J}$ Mean ↑	**76.0**	67.1	**72.4**	72.3	59.8	64.7	**73.9**	71.9	56.0	68.4	54.2	55.5	47.3	44.5
	$\bar{\mathcal{F}}$ Mean ↑	**87.4**	79.0	**83.5**	**84.5**	71.3	77.4	83.2	82.8	63.4	-	-	69.2	-	-

TABLE 10
Attribute-based aggregate UVOS performance on the test set of DAVIS$_{16}$ [11] with the region similarity $\mathcal{J}$. See §6.3 for details.

Dataset	Metric	Method	Attribute															Avg.
			AC	BC	CS	DB	DEF	EA	FM	HO	IO	LR	MB	OCC	OV	SC	SV	
DAVIS$_{16}$	$\mathcal{J}$ Mean ↑	SAG [15]	36.6	37.9	34.8	28.6	41.4	34.7	32.9	38.0	38.5	31.4	33.6	38.4	35.5	38.5	30.1	42.6
		TRC [31]	39.3	48.9	59.3	44.6	44.6	48.5	43.3	40.7	47.2	44.6	36.3	40.7	36.4	36.6	42.2	47.3
		CVOS [111]	43.0	45.7	45.1	28.5	44.7	40.9	31.6	42.2	54.7	35.1	38.8	37.2	32.9	39.8	39.3	48.2
		KEY [37]	44.8	53.2	44.0	44.1	49.5	49.6	42.0	45.1	49.0	40.3	50.9	47.8	45.1	47.9	44.0	49.8
		MSG [30]	56.9	60.8	53.6	47.1	48.4	51.4	44.1	48.8	56.1	53.7	39.8	43.0	46.4	42.7	51.4	53.3
		NLC [35]	60.8	34.4	50.0	48.3	58.3	45.6	56.5	55.2	55.0	57.2	53.6	68.5	52.4	52.9	47.6	55.1
		CUT [34]	58.0	52.1	67.1	35.4	55.9	49.3	52.3	51.2	59.6	54.1	51.0	40.8	58.0	46.8	48.1	55.2
		FST [32]	56.4	56.0	56.2	46.9	52.1	55.3	56.2	52.2	51.0	57.0	50.1	50.3	58.7	47.9	50.3	55.8
		*SFL [47]	59.9	76.3	73.3	27.0	66.6	67.5	61.6	61.2	66.5	66.8	65.6	67.9	65.4	63.8	63.8	67.4
		*LMP [46]	71.5	72.8	71.8	58.3	70.7	67.4	65.5	66.7	67.7	67.2	63.4	66.6	60.9	62.3	65.9	70.0
		*FSEG [45]	70.0	76.7	76.7	50.0	69.5	69.0	69.9	65.2	66.9	71.8	65.4	64.3	72.3	61.5	65.5	70.7
		*LVO [44]	74.0	**78.2**	**80.7**	55.2	75.2	73.7	70.5	71.9	75.1	75.0	71.1	73.6	71.5	70.5	72.9	75.9
		*ARP [42]	**78.1**	70.7	76.0	**71.7**	**76.6**	71.1	75.3	**75.2**	**76.7**	74.3	72.9	74.3	**79.6**	71.1	**74.7**	76.2
		*PDB [14]	77.0	76.9	78.4	**62.4**	76.3	**75.9**	**76.4**	73.9	74.6	**77.7**	**74.0**	**77.9**	**77.6**	72.2	74.4	**77.2**
		*LSMO [106]	**77.3**	**77.1**	**81.8**	55.3	**78.7**	75.4	**75.7**	**74.9**	**77.0**	**78.7**	**74.6**	76.8	77.3	**73.3**	**75.1**	**78.2**
		*MOT [110]	76.5	76.1	80.5	60.6	76.2	**78.0**	**76.4**	73.9	74.0	**81.4**	73.5	**80.4**	74.3	**74.3**	**75.1**	**77.2**
		***Ours**	**79.9**	**80.7**	**82.4**	**66.5**	**77.7**	**77.6**	**79.1**	**76.2**	**77.4**	**81.4**	**76.2**	**78.3**	**81.2**	**73.6**	**77.6**	**79.7**

TABLE 11
Quantitative instance-level UVOS results on DAVIS$_{17}$ (DAVIS$_{19}$ challenge [27]) `val` and `test-dev` sets. The best scores are marked in **red**. See §6.4 for details.

Dataset	Methods	$\mathcal{J}$&$\mathcal{F}$ ↑	$\mathcal{J}$			$\mathcal{F}$		
			Mean ↑	Recall ↑	Decay ↓	Mean ↑	Recall ↑	Decay ↓
DAVIS$_{17}$ `val`	*RVOS [51]	41.2	36.8	40.2	**0.5**	45.7	46.4	**1.7**
	*PDB [14]	55.1	53.2	58.9	4.9	57.0	60.2	6.8
	***Ours**	**57.5**	**55.5**	**61.6**	7.0	**59.5**	**62.8**	9.0
DAVIS$_{17}$ `test-dev`	*RVOS [51]	22.5	17.7	16.2	**1.6**	27.3	24.8	**1.8**
	*PDB [14]	40.4	37.7	42.6	4.0	43.0	44.6	3.7
	***Ours**	**45.6**	**42.1**	**48.5**	2.6	**49.0**	**51.5**	2.6

foreground objects from the background. We conduct experiments on the DAVIS$_{17}$ dataset, which has three subsets: `train`, `val`, and `test-dev`, containing 60, 30, and 30 video sequences, respectively (see Table 4). We test our method on the `val` and `test-dev` sets. As the original annotations of DAVIS$_{17}$ dataset are biased towards the semi-supervised scenario, DAVIS$_{19}$ challenge [27] re-annotates DAVIS$_{17}$ for facilitating the unsupervised setting. Therefore, we adopt these new annotations as the groundtruth.

Modification of our model: To adapt our model to the instance-level (multi-object) setting, we made the following modifications. First, for each testing video frame, we apply mask-RCNN [121] to generate a set of category agnostic object proposals. Then we run our model over the whole video sequence and generate a binary mask for the primary object(s) in each frame. By combining the object bounding-box proposals and binary object-level segmentation masks, we produce pixel-wise, instance-level video segmentation results for each frame. Finally, we use [122] to link the object instances across different frames. Please note that we do not use any data in the `train` set to fine-tune our model.

Evaluation metrics: Following the standard evaluation setting, we reports the performance in terms of region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and the mixed metric $\mathcal{J}$&$\mathcal{F}$. The evaluation scores on the `test-dev` set are obtained from the evaluation server of DAVIS$_{19}$ challenge, as the annotations of `test-dev` set are preserved.

Quantitative results: Since instance-level UVOS is a new topic, currently there is only one method, RVOS [51], that reports the performance on DAVIS$_{17}$ `val` and `test-dev` sets. For completeness, we involve a recent top-performing

TABLE 12
Quantitative VSOD results on the test sequences of DAVIS$_{16}$ [11] and FBMS [9], and whole Youtube-Objects [12] with max F-measure and MAE. Our method consistently improves over previous salient object detection results. See §6.5 for details.

Dataset	Metric	Video SOD									Image SOD						
		***Ours**	*PDB [14]	*FGRNE [67]	*FCNS [65]	SGSP [62]	GAFL [13]	SAG [15]	STUW [60]	SP [61]	*ASNet [114]	*Amulet [115]	*SRM [116]	*UCF [117]	*DSS [118]	*NLDF [119]	*DCL [120]
DAVIS$_{16}$	$F^{\max}$ ↑	**0.870**	**0.849**	**0.786**	0.729	0.677	0.578	0.479	0.692	0.601	0.750	0.699	0.779	0.716	0.717	0.723	0.631
	MAE ↓	**0.026**	**0.030**	0.043	0.053	0.128	0.091	0.105	0.098	0.130	0.054	0.082	**0.039**	0.107	0.062	0.056	0.070
Youtube-Object	$F^{\max}$ ↑	**0.805**	**0.788**	-	0.711	-	0.555	0.536	-	-	**0.772**	0.698	0.752	0.686	0.721	0.743	0.683
	MAE ↓	**0.053**	**0.056**	-	0.087	-	0.164	0.156	-	-	0.064	0.092	**0.063**	0.135	0.083	0.071	0.086
FBMS$_{59}$	$F^{\max}$ ↑	**0.844**	**0.815**	0.779	0.735	0.571	0.551	0.581	0.528	0.538	**0.786**	0.725	0.776	0.679	0.764	0.736	0.726
	MAE ↓	**0.049**	**0.069**	**0.083**	0.100	0.171	0.163	0.155	0.143	0.161	0.087	0.110	0.071	0.147	**0.083**	0.086	0.089

TABLE 13
Ablation study on the test sequences of DAVIS$_{16}$ [11] and FBMS [9], and whole Youtube-Objects [12] with the region similarity $\mathcal{J}$. See §6.6 for details.

Aspect	Variant	Dataset		
		DAVIS$_{16}$	*Youtube-Object*	*FBMS$_{59}$*
Full model	DVAP+AGOS+CRF	**79.7**	**69.7**	**76.0**
Component	DVAP+CRF	26.0	24.0	25.6
	DVAP+FCN+CRF	39.2	37.9	40.7
Variant	*w/o* weight sharing	79.5	**69.7**	75.8
Post-process	DVAP+AGOS	78.4	69.5	75.6

TABLE 14
Runtime comparison (seconds/frame) on the test sequences of DAVIS$_{16}$ [11]. Note that our method is faster that other competitors. OA, OF, and OP indicate Online Adaption, Optical Flow, and Object Proposal, respectively. See §6.7 for details.

Method	*Ours	*Ours+CRF	*MOT [110]	*LSMO [106]	*PDB [14]	ARP [42]
Pre-process	-	-	OA, OF	OF	-	OF, OP
Time(s)	**0.1**	**0.1**+0.5	1.0	>2.5	0.7	124.7
Method	***Ours**	***FSEG [45]**	***LMP [46]**	***SFL [47]**	**FST [32]**	**SAG [15]**
Pre-process	-	OF	OF	OF	OF	OF
Time(s)	**0.1**	7.2	18.3	7.9	51.4	54.0

object-level UVOS method, PDB [14], which is modified for instance-level UVOS in a similar way. We compare the performance of our methods with RVOS and PDB in Table 11. The results clearly demonstrate that our model outperforms both RVOS and PDB by a large margin. For instance, in terms of $\mathcal{J}\&\mathcal{F}$, mean $\mathcal{J}$, and mean $\mathcal{F}$, our method dramatically surpasses RVOS by 23.1%, 24.4% and 21.7%, respectively, on the `test-dev` set.

6.5 Performance on the Setting of VSOD

Test datasets: We further test the performance of our model in the setting of VSOD, *i.e.*, generating continuous saliency maps that highlight the salient video objects. We reproduce the results of our model by omitting the CRF binarization. Following [13]–[15], [65], [67], the test sets of DAVIS$_{16}$ [11] and FBMS$_{59}$ [9], and the whole Youtube-Objects [12] are used for evaluation.

Evaluation metrics: Standard F-measure and MAE metrics are used for quantitative evaluation [114].

Quantitative results: We compare our model with six well known VSOD models [13]–[15], [60]–[62], [65], [67]. For completeness, nine state-of-the-art image salient object detection models [114]–[120] are also included in our experiment. The results of these methods are obtained by running their publicly available codes with default settings or directly obtained from the authors. As shown in Table 12, our model outperforms all the previous VSOD models [13]–[15], [60]–[62], [65], [113] with interpretable attention maps. This verifies the strong correlation between VSOD and UVOS from a view of top-down attention mechanism.

6.6 Ablation Study

We investigate the effectiveness of essential components of our model (§5) by presenting an ablation study in Table 13. First, to validate the contribution of our AGOS module (§5.3), we provide two variants: *DVAP+CRF* and *DVAP+FCN+CRF*, where the first baseline interprets the attention generated by the DVAP module (§5.2) as prediction maps of the primary video objects, while the latter one replaces the AGOS module by a small FCN that has 3 convolutional layers: Conv($3\times3, 128$) → ReLU → Conv($3\times3, 64$) → ReLU → Conv($1\times1, 1$) → *sigmoid*. As seen, directly removing the AGOS module brings significant performance drops (*e.g.*, 79.7→26.0 on DAVIS$_{16}$ *test*), which clearly shows the AGOS module is indeed useful. We also observe a non-trivial performance degradation when employing *DVAP+FCN+CRF* (*e.g.*, 40.5 mIoU point decrease on DAVIS$_{16}$ *test*), suggesting the effectiveness of the network design of the AGOS module.

Then, we study the effect of sharing the first three convolutional blocks (*conv1*, *conv2*, and *conv3*) between DVAP and AGOS modules (see §5.3). To do this, we train DVAP and AGOS modules separately and observe stable (or even slightly worse) performance of the resulting baseline *w/o weight sharing*. This shows the weight-sharing strategy inspires more representative feature learning. Such design is more favored when considering the extra advantages of reducing the computation time and model size. This experiment also evidences the close correlation between human visual attention and explicit object judgement from another point of view.

Finally, we examine the performance gain brought by the CRF post-processing. We find that CRF provides 0.2∼1.3 mIoU boost over the three datasets. Overall, all components introduced in our approach lead to the state-of-the-art results on DAVIS$_{16}$ [11], FBMS$_{59}$ [9], and Youtube-Objects [12].

6.7 Runtime Comparison

It is clear that run time efficiency has great impact on the usability of UVOS algorithms. We conduct running-time comparisons on the test set of DAVIS$_{16}$ [11] with 480p resolution. We include several representative UVOS models [14], [15], [32], [42], [45]–[47], [106], [110] for providing a comprehensive comparison of execution time costs of existing approaches. The time cost comparison results summarized in Table 14 show that our model achieves a fast processing speed of 10 fps (without using CRF), which is faster than the other methods. This advantage of time efficiency is mainly because **(i)** our DVAP and AGOS modules share multi-layer weights; and **(ii)** our model does not need any other time-consuming pre-processing step, such as online adaption [110], optical flow computation [15], [32], [42], [45]–[47], [106], [110] and object proposal generation [42].

7 Conclusion

In this work we systematically studied the role of visual attention in UVOS and its related task, VSOD. We extended three popular video object segmentation datasets with real human eye-tacking records. Through in-depth analysis, for the first time, we quantitatively validated that human visual attention mechanism plays an essential role in UVOS and VSOD tasks. With this novel insight, we proposed a visual attention-driven UVOS model, where the DVAP module, mimicking human attention behavior in the dynamic UVOS setting, is used as a supervised neural attention to guide the subsequent AGOS module for fine-grained video object segmentation. With the visual attention as an intermediate representation, our model is able to produce promising results without training on expensive pixel-wise video segmentation ground-truths, and it gains better post-hoc, biologically-consistent interpretability. Experimental results demonstrated that the proposed model outperforms other state-of-the-art UVOS competitors on popular benchmarks. The suggested model also gains the best performance in the VSOD setting. Therefore, it closely connects the top-down, segmentation-aware visual attention mechanism, UVOS and VSOD tasks, and offers a new glimpse into the rationale behind them.

Acknowledgment

The authors would like to thank the reviewers and editors for their time and valuable comments on this work. Dr. Wang was supported partially by CCF-Tencent Open Fund, Tencent AI Lab Rhino-Bird Focused Research Program, and Zhijiang Lab's International Talent Fund for Young Professionals. Prof. Ling was supported partially by the Yahoo Faculty Research and Engagement Program Award and the Amazon AWS Machine Learning Research Award.

References

[1] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. H. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3064–3074.

[2] D. Sinclair, "Motion segmentation and local structure," in *International Conference on Computer Vision*, 1993, pp. 366–373.

[3] M. Irani, B. Rousso, and S. Peleg, "Computing occluding and transparent motions," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 5–16, 1994.

[4] M. Irani and P. Anandan, "A unified approach to moving object detection in 2d and 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 577–589, 1998.

[5] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1151–1163, 2002.

[6] E. Hayman and J.-O. Eklundh, "Statistical background subtraction for a mobile observer," in *IEEE International Conference on Computer Vision*, 2003, pp. 67–74.

[7] D. Shao, Y. Xiong, Y. Zhao, Q. Huang, Y. Qiao, and D. Lin, "Find and focus: Retrieve and localize video events with natural language queries," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 200–216.

[8] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 353–369.

[9] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187–1200, 2014.

[10] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *IEEE International Conference on Computer Vision*, 2013, pp. 2192–2199.

[11] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 724–732.

[12] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3282–3289.

[13] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.

[14] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convLSTM for video salient object detection," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 715–731.

[15] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

[16] F. Zhou, S. Bing Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3358–3365.

[17] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 8, pp. 2014–2027, 2017.

[18] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*, 1987, pp. 115–141.

[19] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[20] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 3, p. 419, 1989.

[21] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014.

[22] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, 2011.

[23] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[24] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[25] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[26] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 742–756, 2015.

[27] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool, "The 2019 davis challenge on vos: Unsupervised multi-object segmentation," *arXiv preprint arXiv:1905.00737*, 2019.

[28] S. Brutzer, B. Höferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1937–1944.

[29] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 282–295.

[30] P. Ochs and T. Brox, "Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions," in *IEEE International Conference on Computer Vision*, 2011, pp. 1583–1590.

[31] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1846–1853.

[32] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *IEEE International Conference on Computer Vision*, 2013, pp. 1777–1784.

[33] L. Chen, J. Shen, W. Wang, and B. Ni, "Video object segmentation via dense trajectories," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2225–2234, 2015.

[34] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *IEEE International Conference on Computer Vision*, 2015, pp. 3271–3279.

[35] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proceedings of the British Machine Vision Conference*, 2014.

[36] W.-D. Jang, C. Lee, and C.-S. Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 696–704.

[37] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *IEEE International Conference on Computer Vision*, 2011, pp. 1995–2002.

[38] T. Ma and L. J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 670–677.

[39] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 628–635.

[40] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4083–4090.

[41] F. Xiao and Y. Jae Lee, "Track and segment: An iterative unsupervised approach for video object proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 933–942.

[42] Y. J. Koh and C.-S. Kim, "Primary object segmentation in videos based on region augmentation and reduction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7417–7425.

[43] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[44] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *IEEE International Conference on Computer Vision*, 2017, pp. 4481–4490.

[45] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2126.

[46] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3386–3394.

[47] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *IEEE International Conference on Computer Vision*, 2017, pp. 686–695.

[48] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. Jay Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 207–223.

[49] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3623–3632.

[50] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *International Conference on Computer Vision*, 2019.

[51] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i Nieto, "Rvos: End-to-end recurrent network for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5277–5286.

[52] W. Wang, J. Shen, H. Sun, and L. Shao, "Video co-saliency guided co-segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1727–1736, 2018.

[53] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 786–802.

[54] D.-Y. Chen and Y.-S. Luo, "Preserving motion-tolerant contextual visual saliency for video resizing," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1616–1627, 2013.

[55] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2018.

[56] V. Mahadevan and N. Vasconcelos, "On the connections between saliency and tracking," in *Advances in Neural Information Processing Systems*, 2012, pp. 1664–1672.

[57] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*, 2010, pp. 366–379.

[58] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, 2011.

[59] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *arXiv preprint arXiv:1904.09146*, 2019.

[60] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting." *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[61] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, 2014.

[62] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2527–2542, 2017.

[63] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[64] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 29–42.

[65] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2018.

[66] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8554–8564.

[67] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3243–3252.

[68] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[69] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[70] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[71] S. Mathe and C. Sminchisescu, "Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, 2015.

[72] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic, "Eye-tracking database for a set of standard video sequences," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 898–903, 2012.

[73] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[74] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[75] R. S. Frackowiak, *Human brain function*. Elsevier, 2004.

[76] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[77] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32–32, 2008.

[78] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in Neural Information Processing Systems*, 2005, pp. 481–488.

[79] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2006, pp. 155–162.

[80] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[81] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[82] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[83] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27–38, 2014.

[84] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression." *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[85] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2009, pp. 681–688.

[86] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 15–15, 2009.

[87] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[88] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.

[89] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2019.

[90] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[91] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.

[92] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015.

[93] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[94] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4271–4280.

[95] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6450–6458.

[96] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

[97] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, pp. 18–18, 2008.

[98] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.

[99] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 268–281.

[100] S. Avinash Ramakanth and R. Venkatesh Babu, "SeamSeg: Video object segmentation using patch seams," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 376–383.

[101] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 985–998, 2019.

[102] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? A study of human explicit saliency judgment," *Vision Research*, vol. 91, pp. 62–77, 2013.

[103] G. A. Alvarez and S. L. Franconeri, "How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism," *Journal of vision*, vol. 7, no. 13, pp. 14–14, 2007.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[105] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[106] P. Tokmakov, C. Schmid, and K. Alahari, "Learning to segment moving objects," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 282–301, 2019.

[107] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, 2011, pp. 109–117.

[108] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 602–617.

[109] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[110] M. Siam, C. Jiang, S. Lu, L. Petrich, M. Gamal, M. Elhoseiny, and M. Jagersand, "Video segmentation using teacher-student adaptation in a human robot interaction (hri) setting," in *IEEE International Conference on Robotics and Automation*, 2019.

[111] B. Taylor, V. Karasev, and S. Soatto, "Causal video object segmentation from persistence of occlusions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4268–4276.

[112] Y.-H. Tsai, G. Zhong, and M.-H. Yang, "Semantic co-segmentation in videos," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 760–775.

[113] S. Li, B. Seybold, A. Vorobyov, A. Fathi, Q. Huang, and C.-C. Jay Kuo, "Instance embedding transfer to unsupervised video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6526–6535.

[114] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[115] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object

detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[116] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4019–4028.

[117] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 212–221.

[118] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5300–5309.

[119] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6609–6617.

[120] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[121] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[122] J. Luiten, P. Voigtlaender, and B. Leibe, "Premvos: Proposal-generation, refinement and merging for video object segmentation," in *Asian Conference on Computer Vision*, 2018, pp. 565–580.

Wenguan Wang received his Ph.D. degree from Beijing Institute of Technology in 2018. He is currently a postdoc researcher at ETH Zurich, Switzerland. From 2016 to 2018, he was a joint Ph.D. candidate in University of California, Los Angeles, directed by Prof. Song-Chun Zhu. From 2018 to 2019, he was a senior scientist at Inception Institute of Artificial Intelligence, UAE. His current research interests include computer vision, image processing and deep learning.

Jianbing Shen (M'11-SM'12) is currently acting as the Lead Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He is also a Full Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers such as *IEEE TPAMI*, *CVPR*, and *ICCV*. He has obtained many honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. His research interests include computer vision and deep learning. He is an Associate Editor of *IEEE TNNLS*, *IEEE TIP* and other journals.

Xiankai Lu received the Ph.D. degree from Shanghai Jiao Tong University in 2018. Currently he is a research associate with Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interest includes object tracking, video object segmentation and deep learning.

Steven C. H. Hoi is an Associate Professor of the School of Information Sytems, Singapore Management Unviersity, Singapore. Prior to joining SMU, he was an Associate Professor with Nanyang Technological University, Singapore. He received his Bachelor degree from Tsinghua University, P.R. China, in 2002, and his Ph.D degree in computer science and engineering from The Chinese University of Hong Kong, in 2006. His research interests are machine learning and data mining and their applications to multimedia information retrieval (image and video retrieval), social media and web mining, and computational finance, *etc*., and he has published over 150 refereed papers in top conferences and journals in these related areas. Currently he is the Editor-in-Chief of *Neurocomputing*, Associate Editor of *IEEE TPAMI*.

Haibin Ling received the B.S. and M.S. degrees from Peking University in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. In 2007, he joined Siemens Corporate Research as a research scientist. From 2008 to 2019, he worked as a faculty member of the Department of Computer Sciences at Temple University. In fall 2019, he joined Stony Brook University as a SUNY Empire Innovation Professor in the Department of Computer Science. His research interests include computer vision, augmented reality, medical image analysis, and human computer interaction. He received Best Student Paper Award at ACM UIST in 2003, and NSF CAREER Award in 2014. He serves as Associate Editors for several journals including IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Pattern Recognition (PR), and Computer Vision and Image Understanding (CVIU). He has served as Area Chairs for CVPR 2014, 2016, 2019 and 2020.