Learning Unsupervised Video Object Segmentation through Visual Attention

Wenguan Wang [*1,2], Hongmei Song [*1], Shuyang Zhao [1],
Jianbing Shen [†1,2], Sanyuan Zhao [1], Steven C. H. Hoi [3,4], Haibin Ling [5]
[1]Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China
[2]Inception Institute of Artificial Intelligence, UAE [3]Singapore Management University, Singapore
[4]Salesforce Research Asia, Singapore [5]Temple University, USA
wenguanwang.ai@gmail.com, shenjianbingcg@gmail.com
https://github.com/wenguanwang/AGS

Abstract

This paper conducts a systematic study on the role of visual attention in the Unsupervised Video Object Segmentation (UVOS) task. By elaborately annotating three popular video segmentation datasets (DAVIS$_{16}$, Youtube-Objects and SegTrack$_{V2}$) with dynamic eye-tracking data in the 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 judgements during dynamic, task-driven viewing. Such novel observations provide an in-depth insight into the underlying rationale behind UVOS. 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 merits: 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 popular benchmarks show that, even without using expensive video object mask annotations, our model achieves compelling performance in comparison with state-of-the-arts.

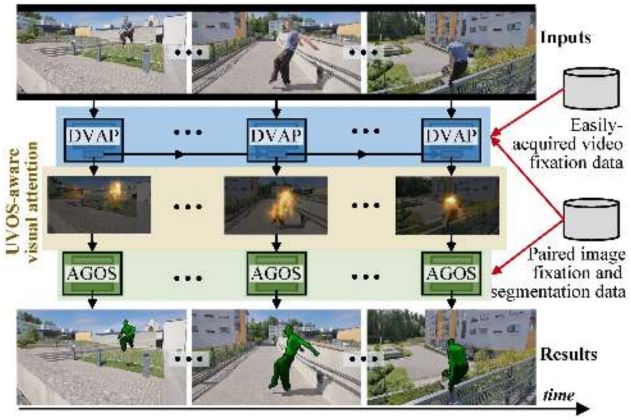
Inputs
DVAP
DVAP
DVAP
Easily-acquired video fixation data
UVOS-aware visual attention
AGOS
AGOS
AGOS
Paired image fixation and segmentation data
Results
time

Figure 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.

1. Introduction

Unsupervised Video Object Segmentation (UVOS), *i.e.*, automatically segmenting primary object regions from the background in videos, has been a long standing research challenge in computer vision [29, 30, 12, 23], and has shown potential benefits for numerous applications, *e.g.*, action recognition [62] and object tracking [50]. 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.

Deep learning has been actively explored for solving UVOS recently. Despite having achieved promising results, current deep learning based UVOS models [64, 45, 33, 67] often rely on expensive pixel-wise video segmentation annotation data [86] to directly map input video frames into corresponding segmentation masks, which are restricted and generally lack of an explicit interpretation about the ra-

[*] Equal contribution.

[†] Corresponding author: *Jianbing Shen*. This work was supported in part by the Beijing Natural Science Foundation under Grant 4182056, the Fok Ying Tung Education Foundation under Grant 141067, and the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission.

tionale behind their choice of the foreground object(s). Similar problems also has been experienced in a closely related research area, video salient object detection (VSOD) [79], which aims to extract a continuous saliency map for each frame that highlights the most visually important area. An biological interpretation for the choice of the salient object regions is essential. The results from video salient object detection are used as a vital cue or pre-processing step for UVOS [64, 77].

In this paper, we emphasize the value of human visual attention in UVOS (and its related task, video salient object detection). According to studies in cognitive psychology [39, 68, 82, 37], 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 the underlying mechanism that drives UVOS*. The foreground in UVOS should be the object(s) that attracts human attention most, as the choice of the object(s) should be consistent with human attention judgements.

To validate this novel hypothesis, we extend three popular video segmentation datasets, DAVIS$_{16}$ [58], Youtube-Objects [60] and SegTrack$_{V2}$ [44], 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 two essential experiments, *i.e.*, quantifying the inter-subject consistency and the correlation between human dynamic attention and explicit object judgement (§4), in which two key observations are found from our quantitative analysis:

- There exist highly consistent attention behaviors among human observers in the UVOS task, though the notion of 'primary object(s)' is sometimes viewed as ill-posed for extremely-diverse dynamic scenes.
- There exists a strong correlation between human fixation and human explicit judgement of primary object(s).

These findings offer an insightful glimpse into the rationale behind UVOS from human attention perspective. Inspired by this, 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.*, [87, 47]) 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 needing 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, *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 fine-grained object segmentation. Besides, our model also enjoys several important 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 learning on multi-source data and sharing weights.* 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 data, but a large video segmentation annotation data is very 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).

In summary, we propose a powerful, fully differentiable, and biologically-inspired UVOS model that fully exploits

[1] Taking the DAVIS dataset as an example, it took several minutes per-frame to annotation with 5 specialists, while with eye-tracker equipment, annotating each frame only takes 1$\sim$2 seconds.

Dataset	Pub.	Year	#Videos	#Viewers	Task
CRCNS [31]	TIP	2004	50	15	scene unders.
Hollywood-2 [52]	TPAMI	2012	1,707	19	action recog.
UCF sports [52]	TPAMI	2012	150	19	action recog.
SFU [21]	TIP	2012	12	15	free-view
DHF1K [75]	CVPR	2018	1,000	17	free-view
DAVIS$_{16}$ (**Ours**)	-	2018	50	20	UVOS
Youtube-Objects (**Ours**)		2018	126	20	UVOS
SegTrack$_{V2}$ (**Ours**)		2018	14	20	UVOS

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$_{16}$ [58], Youtube-Objects [60], and SegTrack$_{V2}$ [44] datasets with extra UVOS-aware gaze data.

the value of visual attention. The proposed model produces state-of-the-art results on popular benchmarks. We expect this work, together with our newly collected data, to provide a deeper insight into the underlying mechanism behind UVOS and video salient object detection, and inspire more research along this direction.

2. Related Work

Unsupervised Video Object Segmentation. Early UVOS methods are typically based on handcrafted features and heuristics such as long-term point trajectory [54, 5, 17, 53, 9], motion boundary [56], objectness [43, 51, 89, 18, 59, 83, 40, 45], and saliency [13, 77, 76, 34, 27]. Later, with the renaissance of neural network, many deep learning based models were proposed, which typically use multilayer perceptron based moving objectness detector, adopt two-stream architecture [67, 33], or CNN encoder-decoder structure [66, 11, 45, 46, 64]. These deep UVOS models generally achieve promising performance, due to the strong learning ability of deep neural networks.

Although a handful of UVOS models [13, 77, 56, 81, 27, 64] 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. Additionally, 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.

Video Salient Object Detection. VSOD is a very close topic to UVOS. VSOD [16, 49, 79, 77, 80] aims to give a gray saliency value for each pixel in the videos sequence. The continuous saliency maps are valuable for a wide range of applications, such as cropping, object tracking, and video object segmentation. However, previous VSOD simply use the UVOS datasets for benchmarking, which lacks a biological evidence for such choice. In this work, through demonstrating the consistency between human fixations and explicit object judgement, we given an in-depth glimpse into both UVOS and VSOD, which share a unified basis, *i.e.*, top-down task-driven visual attention mechanism.

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 [32]. According to the underlying mechanism, attention models can be categorized as either *bottom-up* (stimuli-inspired) or *top-down* (task-driven). Early attention models [42, 90, 19, 6, 22, 25, 36, 15, 20, 26, 61, 22] are based on biologically-inspired features (color, edge, optical flow, *etc.*) and cognitive theories about visual attention (attention shift [39], feature integration theory [68], guided search [82], *etc.*). Recently, deep learning based attention models [71, 28, 55, 73, 75] were proposed and generally yield better performance.

However, most previous methods use static, bottom-up models and none of them is specially designed for modeling UVOS-driven, top-down attention in dynamic scenes. Previous dynamic eye-tracking datasets [31, 52, 21, 75] were constructed under free-viewing or other task-driven settings (see Table 1). In this work, numerous eye gaze data on popular video segmentation datasets [58, 60, 44] are carefully collected in the UVOS setting. 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 and visual attention prediction.

Trainable Attention in Neural Networks. Recent years have witnessed 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 has shown wide successes in natural language processing and computer vision tasks, such as machine translation [2], image captioning [85], visual question answering [88], human object interaction [14], and image classification [72], to list a few. Those neural attentions are learned in an implicit, goal-driven and end-to-end way.

Our DVAP module can also 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 [58, 60, 44]. Fig. 2 shows some example frames with our UVOS-aware eye-tracking annotation, along with

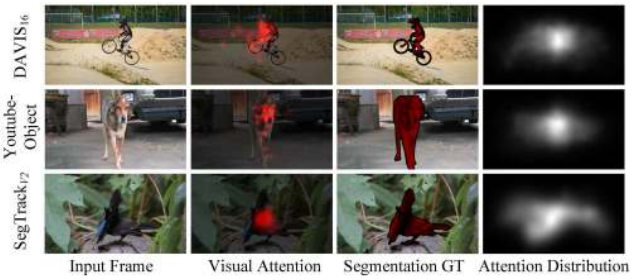

Figure 2. **Example frames from three datasets** ([58, 60, 44]) with our eye-tracking annotation (§3). The last column shows the average attention maps of these datasets. We quantitatively verify (§4) the high consistency between human attention behavior (2nd column) and primary-object determination (3rd column).

visual attention distributions over each dataset.

Stimuli: The dynamic stimuli are from DAVIS$_{16}$ [58], Youtube-Objects [60], and SegTrack$_{V2}$ [44]. DAVIS$_{16}$ is a popular UVOS benchmark containing 50 video sequences with totally 3,455 frames. Youtube-Objects is a large dataset with 126 videos covering 10 common object categories, with 20,647 frames in total. SegTrack$_{V2}$ consists of 14 short videos with totally 947 frames.

Apparatus: Observer eye movements were recorded using a 250 Hz SMI RED250 eye tracker (SensoMotoric Instruments). The dynamic stimuli were displayed on a 19” computer monitor at a resolution of 1440×900 and in their original speeds. A headrest was used to maintain a viewing distance of about 68 cm, as advised by the product manual.

Participants: Twenty participants (12 males and 8 females, aging between 21 and 30), 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 experimenters first ran the standard SMI calibration routine with recommended settings for the best results. 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 [21]. 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.

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 [47], 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-Judd [7], 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 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$_{V2}$) 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 [70, 1, 78], there do exist some ‘universally-agreed’ visually important clues that attract human attentions stably and consistently.

Aspect	Metric	DAVIS$_{16}$ [58]	Youtube-Object [60]	SegTrack$_{V2}$ [44]
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

Table 2. **Quantitative results of inter-subject consistency (ISC) and inter-task correlation (ITC)**, measured by AUC-Juddy.

Correlation between visual attention and video object determination: It is essential to study whether human visual attention and video primary object judgement agree with each other, which has never been explored before. Here we apply the experimental protocol suggested by [4] 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 set. 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.

5. Proposed UVOS Method

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 [64, 46, 33, 67] learn a DNN as a mapping function $\mathcal{F}_{\text{UVOS}}: \mathbb{R}^{W\times H\times 3\times T} \mapsto \{0,1\}^{W\times H\times T}$ that directly maps the input into the segmentation masks:

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

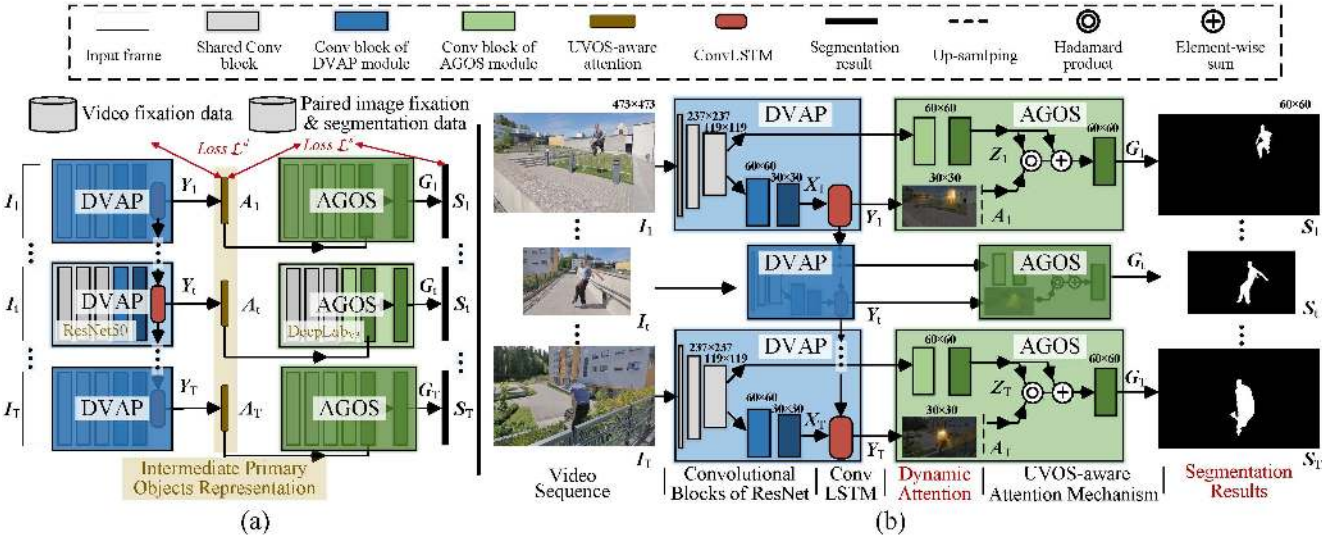

Figure 3. **Illustration of the proposed UVOS model.** (a) Simplified schematization of our model that solves UVOS in a two-step manner, without the need of training with expensive precise video object masks. (b) Detailed network architecture, where the DVAP (§5.2) and AGOS (§5.3) modules share the weights of two bottom conv 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.

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 to tackle UVOS. Specifically, a DVAP module $\mathcal{F}_{\text{DVAP}}$ is first designed to predict dynamic UVOS-aware visual attentions $\{\mathbf{A}_t \in [0,1]^{W'\times H'\times 1}\}_{t=1}^{T}$:

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

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 $\mathbf{S}_t$:

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

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.

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 [24]. 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 [63], 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:

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

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. 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:

$$\mathbf{a} = f_{\mathrm{A}}(\mathbf{i}), \quad \mathbf{z} = f_{\mathrm{Z}}(\mathbf{i}), \quad \mathbf{g} = \mathbf{a} \odot \mathbf{z}, \tag{5}$$

where $\odot$ is element-wise multiplication, and f_{Z} indicates a feature extraction network. Some neural attention models 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:

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

where $\mathcal{F}_{\mathrm{Z}}$ 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 d-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 [24, 72], the feature enhancement step in Eq. 6 is enhanced 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 [24] and modified with the ASPP module proposed in DeepLab$_{V3}$ [10]. 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 $\mathcal{F}_{\mathrm{ASPP}}$. The attention map $\mathbf{A}_t$ is also ×2 upsampled by bilinear interpolation. Finally, our AGOS module in Eq. 6 is implemented as:

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

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 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.

5.4. Implementation Details

Training loss: For DAVP, given an input frame $\mathbf{I} \in R^{473\times473\times3}$, it predicts an attention map $\mathbf{A} \in [0,1]^{30\times30}$. 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 [28], the loss function $\mathcal{L}_{\mathrm{DVAP}}$ for DAVP is designed as:

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

where the $\mathcal{L}_{\mathrm{CE}}$ indicates the classic *cross entropy* loss, and $\mathcal{L}_{\mathrm{CC}}, \mathcal{L}_{\mathrm{NSS}}, \mathcal{L}_{\mathrm{SIM}}$ 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 [28]. We use $\mathcal{L}_{\mathrm{CE}}$ 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 $\mathcal{L}_{\mathrm{AGOS}}$ of the AGOS module is formulated as:

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

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)):

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

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

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

The image training batch contains several attention-segmentation paired image masks, which is 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 $\mathcal{L}_{\text{AGOS}}$:

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

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. 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.

6. Experiments

Training data: During training, we use the video sequences and corresponding fixation data from the training split of DAVIS$_{16}$ [58] and the whole SegTrack$_{V2}$ [44] dataset, leading to totally 54 video sequences with 6,526 frames. Additionally, two image salient object segmentation datasets, DUT-O [87] and PASCAL-S [47], 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. Therefore, 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 §6.2, 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 obtaining the corresponding primary object predictions. Following the common protocol [66, 8, 84, 57] in video segmentation, the fully-connected CRF [41] 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.

6.1. Performance of DVAP module

Test datasets: We evaluate our DVAP module on the test set of DAVIS$_{16}$ [58] and the full Youtube-Objects [60], with the gaze-tracking ground-truth 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 [3] for details).

Quantitative and qualitative results: We compare our DVAP module with 12 state-of-the-art visual attention models, including 5 deep models [75, 35, 73, 55, 28] and 7 traditional models [15, 20, 26, 61, 22, 32]. Quantitative results over the test set of DAVIS$_{16}$ [58] and Youtube-Objects [60] are summarized in Tables 3 and 4, respectively. As seen, our DVAP generally outperforms other competitors, as none of them is specifically designed for UVOS-aware attention prediction. Our DVAP can guide our UVOS model to accurately attend to visually attractive regions in videos.

Dataset	Methods	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
DAVIS$_{16}$	ACL [75]	**0.901**	**0.453**	0.617	**0.559**	**2.252**
	OMCNN [35]	0.889	**0.408**	0.621	**0.518**	**2.101**
	DVA [73]	0.885	0.382	**0.647**	0.494	1.906
	DeepNet [55]	0.880	0.318	**0.644**	0.470	1.866
	ShallowNet [55]	0.874	0.293	0.622	0.471	1.871
	SALICON [28]	0.818	0.276	0.628	0.352	1.432
	STUW [15]	**0.892**	0.363	0.636	0.508	2.019
	PQFT [20]	0.685	0.202	0.584	0.191	0.821
	Seo *et al.* [61]	0.724	0.234	0.582	0.222	0.923
	Hou *et al.* [26]	0.782	0.263	0.581	0.273	1.119
	GBVS [22]	0.882	0.294	0.617	0.442	1.683
	ITTI [32]	0.820	0.249	0.621	0.354	1.332
	Ours	**0.909**	**0.504**	**0.667**	**0.620**	**2.507**

Table 3. **Quantitative comparison of visual attention models on the test set of DAVIS$_{16}$ [58]** (§6.1). The three best scores are indicated in **red**, **blue** and **green**, respectively (same for other tables).

Dataset	Methods	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
Youtube-Objects	ACL [75]	**0.912**	**0.405**	0.711	**0.531**	**2.627**
	OMCNN [35]	0.889	0.326	0.698	0.461	**2.307**
	DVA [73]	**0.905**	**0.372**	**0.741**	**0.526**	2.294
	DeepNet [55]	0.894	0.268	**0.737**	0.448	2.182
	ShallowNet [55]	0.890	0.252	0.704	0.436	2.069
	SALICON [28]	0.840	0.265	0.692	0.380	1.956
	STUW [15]	0.869	0.264	0.666	0.388	1.876
	PQFT [20]	0.730	0.170	0.646	0.210	1.061
	Hou *et al.* [26]	0.786	0.221	0.639	0.243	1.223
	Seo *et al.* [61]	0.763	0.210	0.605	0.224	1.118
	GBVS [22]	0.881	0.244	0.706	0.395	1.919
	ITTI [32]	0.837	0.214	0.709	0.339	1.638
	Ours	**0.914**	**0.419**	**0.747**	**0.543**	**2.700**

Table 4. **Quantitative comparison of different visual attention models on Youtube-Objects [60]** (§6.1).

6.2. Performance of full UVOS model

Test datasets: The test sets of DAVIS$_{16}$ [58] and the full Youtube-Objects [60] are used for assessing the performance of our full UVOS model.

Evaluation metrics: For the UVOS task, we use three standard metrics suggested by [58], *i.e.*, region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and time stability $\mathcal{T}$.

Quantitative and qualitative results: The quantitative comparison results over above two datasets are reported in Tables 5 and 6, respectively. We can observe that the proposed model outperforms other competitors over most metrics across all the datasets. This is significant and distinguishes our model from previous deep UVOS models [40, 46, 67, 33, 66, 11] since our model is trained without precise segmentation mask ground-truths. Some qualitative results are shown in Fig. 4, validating our model yields high-quality results with interpretable dynamic attentions.

Dataset	Metric	**Ours**	PDB [64]	ARP [40]	LVO [67]	FSEG [33]	LMP [66]	SFL [11]	FST [56]	CUT [38]	NLC [13]	MSG [53]	KEY [43]	CVOS [65]	TRC [17]
DAVIS$_{16}$	Mean ↑	**79.7**	**77.2**	**76.2**	75.9	70.7	70.0	67.4	55.8	55.2	55.1	53.3	49.8	48.2	47.3
	$\mathcal{J}$ Recall ↑	**91.1**	**90.1**	**91.1**	89.1	83.5	85.0	81.4	64.9	57.5	55.8	61.6	59.1	54.0	49.3
	Decay ↓	**0.0**	0.9	7.0	**0.0**	1.5	1.3	6.2	**0.0**	2.2	12.6	2.4	14.1	10.5	8.3
	Mean ↑	**77.4**	**74.5**	70.6	**72.1**	65.3	65.9	66.7	51.1	55.2	52.3	50.8	42.7	44.7	44.1
	$\mathcal{F}$ Recall ↑	**85.8**	**84.4**	**83.5**	83.4	73.8	79.2	77.1	51.6	61.0	51.9	60.0	37.5	52.6	43.6
	Decay ↓	**0.0**	**-0.2**	7.9	**1.3**	1.8	2.5	5.1	2.9	3.4	11.4	5.1	10.6	11.7	12.9
	$\mathcal{T}$ Mean ↓	44.5	29.1	39.3	**26.5**	32.8	57.2	28.2	36.6	27.7	42.5	30.2	**26.9**	**25.0**	39.1

Table 5. **Quantitative UVOS results on the test sequences of DAVIS$_{16}$** [58]. The results selected from the public leaderboard (`https://davischallenge.org/davis2016/soa_compare.html`) maintained by the DAVIS challenge. See §6.2 for details.

Dataset	Category	**Ours**	PDB [64]	ARP [40]	LVO [67]	SFL [11]	FSEG [33]	FST [56]	COSEG [69]	LTV [54]
Youtube -Object	Airplane	**87.7**	78.0	73.6	**86.2**	65.6	**81.7**	70.9	69.3	13.7
	Bird	**76.7**	**80.0**	56.1	**81.0**	65.4	63.8	70.6	76.0	12.2
	Boat	**72.2**	58.9	57.8	**68.5**	59.9	**72.3**	42.5	53.5	10.8
	Car	**78.6**	**76.5**	33.9	69.3	64.0	**74.9**	65.2	70.4	23.7
	Cat	**69.2**	63.0	30.5	58.8	58.9	**68.4**	52.1	**66.8**	18.6
	Cow	**64.6**	64.1	41.8	**68.5**	51.2	**68.0**	44.5	49.0	16.3
	Dog	**73.3**	**70.1**	36.8	61.7	54.1	**69.4**	65.3	47.5	18.2
	Horse	**64.4**	**67.6**	44.3	53.9	**64.8**	60.4	53.5	55.7	11.5
	Motorbike	**62.1**	58.4	48.9	**60.8**	52.6	**62.7**	44.2	39.5	10.6
	Train	48.2	35.3	39.2	**66.3**	34.0	**62.2**	29.6	**53.4**	19.6
	$\mathcal{J}$ Mean ↑	**69.7**	65.5	46.2	**67.5**	57.1	**68.4**	53.8	58.1	15.5

Table 6. **Quantitative UVOS results on Youtube-Objects** [60]. Performance over each category and the average score are reported.

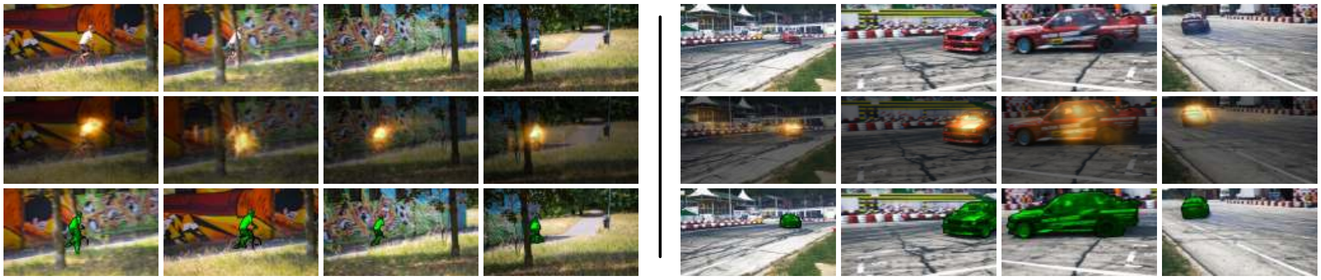

Figure 4. **Visual results on two example videos.** The dynamic attention results from our DVAP module are shown in the second row, which are biologically-inspired and used to guide our AGOS module for fine-grained UVOS (see the last row).

Dataset	**Metric**	**Ours**	PDB [64]	FGRNE [45]	FCNS [80]	SGSP [48]	GAFL [79]	SAGE [77]	STUW [16]	SP [49]
DAVIS$_{16}$	$F^{\max}$ ↑	**0.870**	**0.849**	**0.786**	0.729	0.677	0.578	0.479	0.692	0.601
	MAE ↓	**0.026**	**0.030**	**0.043**	0.053	0.128	0.091	0.105	0.098	0.130

Table 7. **Quantitative VSOD results on the test sequences of DAVIS$_{16}$** [58] with MAE and max F-measure (see §6.3).

6.3. Performance on the VSOD task

Test datasets: The test sets of DAVIS$_{16}$ [58] is used for testing our model in the VSOD setting.

Evaluation metrics: Standard F-measure and MAE metrics are used for quantitative evaluation [74].

Quantitative results: As shown in Table 7, our model (without CRF binaryzation) outperforms previous VSOD models [64, 45, 80, 48, 79, 77, 16, 49] with human readable attention maps. This verifies the strong correlation between VSOD and UVOS from a view of top-down attention mechanism.

7. Conclusion

This work 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 novel 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 the proposed model outperforms other state-of-the-art UVOS methods. The suggested model also gains best performance in the VSOD setting. Therefore, we closely connect the top-down, segmentation-aware visual attention mechanism, UVOS and VSOD tasks, and offer a new glimpse into the rationale behind them.

References

[1] S Avinash Ramakanth and R Venkatesh Babu. SeamSeg: Video object segmentation using patch seams. In *CVPR*, 2014. 4

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[3] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2013. 7

[4] Ali Borji, Dicky N Sihite, and Laurent Itti. What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, 91:62–77, 2013. 4

[5] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 3

[6] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *NIPS*, 2006. 3

[7] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. http://saliency.mit.edu/. 4

[8] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 7

[9] Lin Chen, Jianbing Shen, Wenguan Wang, and Bingbing Ni. Video object segmentation via dense trajectories. *IEEE TMM*, 17(12):2225–2234, 2015. 3

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[11] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, 2017. 3, 7, 8

[12] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002. 1

[13] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014. 3, 8

[14] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018. 3

[15] Yuming Fang, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin. A video saliency detection model in compressed domain. *IEEE TCSVT*, 24(1):27–38, 2014. 3, 7

[16] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE TIP*, 23(9):3910–3921, 2014. 3, 8

[17] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 3, 8

[18] Huazhu Fu, Dong Xu, Bao Zhang, and Stephen Lin. Object-based multiple foreground video co-segmentation. In *CVPR*, 2014. 3

[19] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *NIPS*, 2005. 3

[20] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1):185–198, 2010. 3, 7

[21] Hadi Hadizadeh, Mario J Enriquez, and Ivan V Bajic. Eye-tracking database for a set of standard video sequences. *IEEE TIP*, 21(2):898–903, 2012. 3, 4

[22] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, 2007. 3, 7

[23] Eric Hayman and Jan-Olof Eklundh. Statistical background subtraction for a mobile observer. In *ICCV*, 2003. 1

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[25] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 3

[26] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 2009. 3, 7

[27] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*, 2018. 3

[28] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015. 3, 6, 7

[29] Michal Irani and P Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE TPAMI*, 20(6):577–589, 1998. 1

[30] Michal Irani, Benny Rousso, and Shmuel Peleg. Computing occluding and transparent motions. *IJCV*, 12(1):5–16, 1994. 1

[31] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE TIP*, 13(10):1304–1318, 2004. 3

[32] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 3, 7

[33] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusion-seg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In *CVPR*, 2017. 1, 3, 4, 7, 8

[34] Won-Dong Jang, Chulwoo Lee, and Chang-Su Kim. Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In *CVPR*, 2016. 3

[35] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *ECCV*, 2018. 7

[36] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 3

[37] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, 2014. 2

[38] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *ICCV*, 2015. 8

[39] Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. 1987. 2, 3

[40] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, 2017. 3, 7, 8

[41] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 7

[42] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE TPAMI*, 28(5):802–817, 2006. 3

[43] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *ICCV*, 2011. 3, 8

[44] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 2, 3, 4, 7

[45] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C.-C. Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *CVPR*, 2018. 1, 3, 8

[46] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C.-C. Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, 2018. 3, 4, 7

[47] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 2, 4, 7

[48] Zhi Liu, Junhao Li, Linwei Ye, Guangling Sun, and Liquan Shen. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. *IEEE TCSVT*, 27(12):2527–2542, 2017. 8

[49] Zhi Liu, Xiang Zhang, Shuhua Luo, and Olivier Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE TCSVT*, 24(9):1522–1540, 2014. 3, 8

[50] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *ECCV*, 2018. 1

[51] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012. 3

[52] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE TPAMI*, 37(7):1408–1424, 2015. 3

[53] Peter Ochs and Thomas Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011. 3, 8

[54] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2014. 3, 8

[55] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016. 3, 7

[56] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013. 3, 8

[57] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 7

[58] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 3, 4, 7, 8

[59] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, 2015. 3

[60] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 2, 3, 4, 7, 8

[61] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, 2009. 3, 7

[62] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, 2018. 1

[63] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 5

[64] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convLSTM for video salient object detection. In *ECCV*, 2018. 1, 2, 3, 4, 8

[65] Brian Taylor, Vasiliy Karasev, and Stefano Soatto. Causal video object segmentation from persistence of occlusions. In *CVPR*, 2015. 8

[66] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017. 3, 7, 8

[67] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. 1, 3, 4, 7, 8

[68] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. 2, 3

[69] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *ECCV*, 2016. 8

[70] Amelio Vazquez-Reina, Shai Avidan, Hanspeter Pfister, and Eric Miller. Multiple hypothesis video segmentation from superpixel flows. In *ECCV*, 2010. 4

[71] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014. 3

[72] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 3, 6

[73] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE TIP*, 27(5):2368–2378, 2018. 3, 7

[74] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. Inferring salient objects from human fixations. *IEEE TPAMI*, 2019. 8

[75] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *CVPR*, 2018. 3, 7

[76] Wenguan Wang, Jianbing Shen, Xuelong Li, and Fatih Porikli. Robust video object cosegmentation. *IEEE TIP*, 24(10):3137–3148, 2015. 3

[77] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015. 2, 3, 8

[78] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *IEEE TPAMI*, 2018. 4

[79] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE TIP*, 24(11):4185–4196, 2015. 2, 3, 8

[80] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2018. 3, 8

[81] Wenguan Wang, Jianbing Shen, Hanqiu Sun, and Ling Shao. Video co-saliency guided co-segmentation. *IEEE TCSVT*, 28(8):1727–1736, 2018. 3

[82] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419, 1989. 2, 3

[83] Fanyi Xiao and Yong Jae Lee. Track and segment: An iterative unsupervised approach for video object proposals. In *CVPR*, 2016. 3

[84] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In *CVPR*, 2018. 7

[85] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 3

[86] Wenqiang Xu, Yonglu Li, and Cewu Lu. Srda: Generating instance segmentation annotation via scanning, reasoning and domain adaptation. In *ECCV*, 2018. 1

[87] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2, 7

[88] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 3

[89] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013. 3

[90] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32, 2008. 3