Shifting More Attention to Video Salient Object Detection

Deng-Ping Fan[1] Wenguan Wang[2] Ming-Ming Cheng[1] * Jianbing Shen[2,3]
[1] TKLNDST, CS, Nankai University [2] Inception Institute of Artificial Intelligence [3] Beijing Institute of Technology
http://mmcheng.net/DAVSOD/

Abstract

The last decade has witnessed a growing interest in video salient object detection (VSOD). However, the research community long-term lacked a well-established VSOD dataset representative of real dynamic scenes with high-quality annotations. To address this issue, we elaborately collected a visual-attention-consistent Densely Annotated VSOD (DAVSOD) dataset, which contains 226 videos with 23,938 frames that cover diverse realistic-scenes, objects, instances and motions. With corresponding real human eye-fixation data, we obtain precise ground-truths. This is the first work that explicitly emphasizes the challenge of ***saliency shift****,* i.e.*, the video salient object(s) may dynamically change. To further contribute the community a complete benchmark, we systematically assess 17 representative VSOD algorithms over seven existing VSOD datasets and our DAVSOD with totally ∼84K frames (largest-scale). Utilizing three famous metrics, we then present a comprehensive and insightful performance analysis. Furthermore, we propose a baseline model. It is equipped with a saliency-shift-aware convLSTM, which can efficiently capture video saliency dynamics through learning human attention-shift behavior. Extensive experiments*[1] *open up promising future directions for model development and comparison.*

1. Introduction

Salient object detection (SOD) targets at extracting the most attention-grabbing objects from still images [17] or dynamic videos. This task originates from the cognitive studies of human visual attention behavior, *i.e.*, the astonishing ability of the human visual system (HVS) to quickly orient attention to the most informative parts of visual scenes. Previous studies [6, 45] quantitatively confirmed that there exists a strong correlation between such explicit, object-level saliency judgment (object-saliency) and the implicit visual attention allocation behavior (visual attention mechanism).

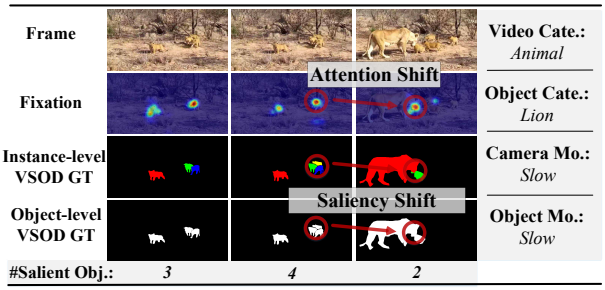

Figure 1: **Annotation examples of our *DAVSOD* dataset**. The rich annotations, including saliency shift, object-/instance-level ground-truths (GT), salient object numbers, scene/object categories, and camera/object motions, provide a solid foundation for VSOD task and benefit a wide range of potential applications.

Video salient object detection (VSOD) is thus significantly essential for understanding the underlying mechanism behind HVS during free-viewing in general and instrumental to a wide range of real-world applications, *e.g.*, video segmentation [74, 83], video captioning [57], video compression [27, 29], autonomous driving [91], robotic interaction [82], weakly supervised attention [95]. Besides its academic value and practical significance, VSOD presents great difficulties due to the challenges carried by video data (diverse motion patterns, occlusions, blur, large object-deformations, *etc*.) and the inherent complexity of human visual attention behavior (*i.e.*, selective attention allocation, attention shift [5, 37, 60]) during dynamic scenes. Thus it invoked dramatically increasing research interest over the past few years [7, 25, 31, 36, 38, 39, 61] (Table 2).

However, in striking contrast with the flourishing development of VSOD modeling, the effort on a standard representative VSOD benchmark still lags behind seriously. Although several datasets [35, 40, 43, 52, 56, 59, 75] are proposed for VSOD, they suffered from the following shortages. First, during dynamic-viewing, the allocation of attentional resources is not only selective but also dynamically varied among different parts of inputs, with the changing of video content. Nevertheless, previous datasets are annotated via *static frames*, without a *dynamic human eye-fixation* guided annotation methodology, and thus do not reveal real human attention behavior during dynamic-

*M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.
[1]Dataset and code are available at: http://dpfan.net/DAVSOD/

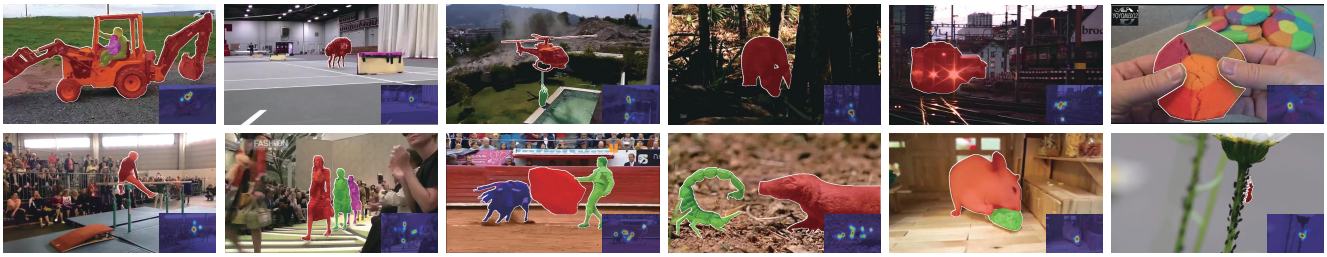

Figure 2: **Sample video sequences from our *DAVSOD* dataset**, with instance-level GT and fixations overlaid.

viewing. Second, they are typically limited in their scalability, coverage, diversity and difficulty. Thus, these limitations of existing datasets inhibit the further development of this branch.

This paper presents two contributions. First, we collect a large-scale ***DAVSOD*** (Densely Annotated Video Salient Object Detection) dataset specifically designed for VSOD.

- It contains 226 video sequences, which were strictly annotated according to real human fixation records (Fig. 2). More importantly, two essential dynamic human attention characteristics, *i.e.*, selective attention and attention shift are both considered. In *DAVSOD*, the salient object(s) may change at different time (Fig. 1), which is more realistic and requires a complete video content understanding. Above efforts result in a visual-attention-consistent VSOD dataset.

- Besides, the videos were carefully selected to cover diverse scene/object categories, motion patterns, and densely annotated with per-frame pixel-accurate ground-truths (GT).

- Another discriminative feature of *DAVSOD* is the availability of both object- and instance-level annotations, benefiting broader potential research directions, such as instance-level VSOD, video salient object subitizing, saliency-aware video captioning, *etc*.

Second, with the established *DAVSOD* dataset and previous 7 VSOD datasets [35,40,43,52,56,59,75], we present a comprehensive evaluation of 17 state-of-the-art models [8, 11,35,41,44,52,53,62,67,68,70,74–76,81,87,92], making it the most complete VSOD benchmark. Additionally, we also propose a baseline model, named **SSAV** (Saliency-Shift Aware VSOD). It learns to predict video saliency by using a saliency-shift-aware convLSTM module, which explicitly models human visual attention-shift behavior in dynamic scenes. The promising results on above benchmark clearly demonstrate its effectiveness.

Our two contributions represent a complete benchmark suite with the necessary tools for a complementary evaluation, bring a more insightful glimpse into the task of VSOD and boost more research efforts towards this direction.

Dataset	Year	#Vi.	#AF.	DL	AS	FP	EF	IL
SegV2 [40]	2013	14	1,065	✓				
FBMS [56]	2014	59	720					
MCL [35]	2015	9	463					
ViSal [75]	2015	17	193					
DAVIS [59]	2016	50	3,455	✓				
UVSD [52]	2017	18	3,262	✓				
VOS [43]	2018	200	7,467			✓		
DAVSOD	2019	**226**	**23,938**	✓	✓	✓	✓	✓

Table 1: **Statistics of previous VSOD datasets and the proposed *DAVSOD* dataset**, showing *DAVSOD* provides much richer annotations. **#Vi.**: number of videos. **#AF.**: number of annotated frames. **DL**: whether provide densely (per-frame) labeling. **AS**: whether consider attention shift. **FP**: whether annotate salient objects according to eye fixation records. **EF**: whether offer the eye fixation records for annotated salient object(s). **IL**: whether provide instance-level annotation.

2. Related Work

VSOD Datasets. Over the past few years, several datasets (Table 1) have been created or introduced into VSOD. Specifically, *SegV2* [40] and *FBMS* [56] are two early adopted datasets. Since they are designed for their specific purposes, they are not very suitable for VSOD task. Another dataset *MCL* [35] only has 9 simple video examples. *ViSal* [75] is the first specially designed VSOD dataset, while only containing 17 video sequences with obvious objects. More recently, Wang *et al*. [76] introduced *DAVIS* [59], a famous video segmentation dataset with 50 challenging scenes, for VSOD. Although above datasets advanced the field of VSOD to various degrees, they are severely limited to small scales (only dozens of videos). In addition, those datasets do not consider real human attention during dynamic scenes instead arbitrarily and manually identify the salient objects by only a few annotators. The annotation is performed over each frame *individually*, failed in accounting temporal characteristics in complex dynamic scenes. A recent larger scale *VOS* [43] dataset partially remedied above limitations. But its diversity and generality are quite limited as it contains many simple indoor, stable-camera scenarios.

Overall, our *DAVSOD* significantly discriminate from above datasets: i) Through in-depth analyzing real human dynamic attention behavior, we observe visual attention-shift phenomenon, and thus, for the first time, emphasize the shift of salient objects in dynamic scenes and provide

No.	Model	Year	Pub.	#Training	Training Set	Basic	Type	OF	SP	S-measure	PCT	Code
1	SIVM [62]	2010	ECCV			CRF, statistic	T			0.481∼0.606	72.4*	M&C++
2	DCSM [36]	2011	TCSVT			SORM distance	T				0.023*	C++
3	RDCM [47]	2013	TCSVT			gabor, region contrast	T	✓			9.8*	N/A
4	SPVM [53]	2014	TCSVT			SP, histogram	T		✓	0.470∼0.724	56.1*	M&C++
5	CDVM [20]	2014	TCSVT			compressed domain	T				1.73*	M
6	TIMP [92]	2014	CVPR			time-mapping	T	✓		0.539∼0.667	69.2*	M&C++
7	STUW [21]	2014	TIP			uncertainty weighting	T	✓			50.7*	M
8	EBSG [55]	2015	CVPR			gestalt principle	T	✓				N/A
9	SAGM [74]	2015	CVPR			geodesic distance	T	✓	✓	0.615∼0.749	45.4*	M&C++
10	ETPM [64]	2015	CVPR			eye tracking prior	T	✓				N/A
11	RWRV [35]	2015	TIP			random walk	T			0.330∼0.595	18.3*	M
12	GFVM [75]	2015	TIP			gradient flow	T	✓	✓	0.613∼0.757	53.7*	M&C++
13	MB+M [87]	2015	ICCV			minimum barrier distance	T			0.552∼0.726	0.02*	M&C++
14	MSTM [70]	2016	CVPR			minimum spanning tree	T			0.540∼ 0.657	0.02*	M&C++
15	SGSP [52]	2017	TCSVT			histogram, graph	T	✓	✓	0.557∼0.706	51.7*	M&C++
16	SFLR [8]	2017	TIP			low-rank coherency	T	✓	✓	0.470∼0.724	119.4*	M&C++
17	STBP [81]	2017	TIP			background priors	T		✓	0.533∼0.752	49.49*	M&C++
18	VSOP [28]	2017	TYCB			object proposals	T	✓	✓			M&C++
19	DSR3 [38]	2017	BMVC	44 (6+8+30) clips	10C+S2+DV	RCL [48]	D					Py&Ca
20	VQCU [3]	2018	TMM			spectral, graph structure	T		✓		0.78*	M
21	CSGM [77]	2018	TCSVT			joint video co-saliency	T	✓	✓		3.86*	M&C++
22	STUM [2]	2018	TIP			local spatiotemporal neighborhood cues	T					N.A.
23	SAVM [78]	2018	TPAMI			geodesic distance	T	✓	✓	0.615∼0.749	45.4*	M&C++
24	bMRF [7]	2018	TMM			MRF	T	✓	✓		2.63*	N/A
25	LESR [93]	2018	TMM			localized estimation, spatiotemporal	T	✓	✓		5.93*	N/A
26	TVPI [61]	2018	TIP			geodesic distance, CRF	T		✓		2.78*	M&C
27	SDVM [4]	2018	TIP			spatiotemporal decomposition	T					N/A
28	SCOM [11]	2018	TIP	∼10K frame pairs	MK	DCL [42]	D	✓	✓	0.555∼0.832	38.8	N/A
29	STCR [39]	2018	TIP	44 (6+8+30) clips	10C+S2+DV	CRF	D		✓			N/A
30	DLVS [76]	2018	TIP	∼18K frame pairs	MK+DO+S2+FS	FCN [54]	D	✓	✓	0.682∼0.881	0.47	Py&Ca
31	SCNN [68]	2018	TCSVT	∼11K frame pairs	MK+S2+FS	VGGNet [66]	D	✓	✓	0.674∼0.794	38.5	N/A
32	FGRN [41]	2018	CVPR	∼10K frame pairs	S2+FS+DV	LSTM	D	✓		0.693∼0.861	0.09	Py&Ca
33	SCOV [33]	2018	ECCV			BOW [22], proposal, FCIS [46]	T	✓	✓		3.44	N/A
34	MBNM [44]	2018	ECCV	∼13K frame pairs	Voc12 + Coco [49] + DV	motion based, DeepLab [9]	D	✓		0.637∼0.898	2.63	N/A
35	PDBM [67]	2018	ECCV	∼18K frame pairs	MK+DO+DV	DC [85]	D			0.698∼0.907	0.05	Py&Ca
36	UVOS [31]	2018	ECCV			standard edge detector	D	✓	✓			N/A
37	**SSAV (Ours)**	2019	CVPR	∼13K frame pairs	DAVSOD val + DO +DV	SSLSTM, PDC [67]	D			0.724∼0.941	0.05	Py&Ca

Table 2: **Summarizing of 36 previous representative VSOD methods and the proposed SSAV model. Training Set:** 10C = *10-Clips* [24]. S2 = *SegV2* [40]. DV = *DAVIS* [59]. DO = *DUT-OMRON* [84]. MK = *MSRA10K* [12]. MB = *MSRA-B* [51]. FS = *FBMS* [56]. Voc12= PASCAL VOC2012 [16]. **Basic:** CRF = Conditional Random Field. SP = superpixel. SORM = self-ordinal resemblance measure. MRF = Markov Random Field. **Type:** T = Traditional. D = Deep learning. **OF:** Whether use optical flow. **SP:** Whether use superpixel over-segmentation. **S-measure [18]:** The range of scores over the 8 datasets in Table 4. **PCT:** Per-frames Computation Time (second). Since [3, 7, 11, 33, 44, 47, 68, 93] did not release implementations, corresponding PCTs are borrowed from their papers or provided by authors. **Code:** M = Matlab. Py = Python. Ca= Caffe. N/A = Not Available in the literature. "*" indicates CPU time.

the unique annotations of visual-attention-consistent property. ii) Its diversity, large-scale dense annotation, as well as comprehensive object-/instance-level salient object annotations, rich attribute annotations (*e.g.*, object numbers, motion patterns, scene/object categories), altogether make a solid and unique foundation for VSOD.

VSOD Models. *Early VSOD models* [8, 26, 28, 35, 52, 53, 62, 63, 74, 75] are built upon hand-crafted features (color, motion, *etc.*), and largely rely on classic heuristics in image salient object detection area (*e.g.*, center-surround contrast [12], background prior [79]) and cognitive theories of visual attention (*e.g.*, feature integration theory [69], guided search [80]). They also explored the way of integrating spatial and temporal saliency features through different computational mechanisms, such as gradient flow field [75], geodesic distance [74], restarted random walk [35], and spectral graph structure [3]. Traditional VSOD models are bound to significant feature engineering and limited expression ability of hand-features. See Table 2 for more details.

More recently, *deep learning based VSOD models* [31, 38, 39, 41, 67, 68, 76] have gained more attention inspired by the success of applying deep neural networks on image saliency detection [13–15, 32, 50, 71, 72, 86, 88–90, 94]. More specifically, the work of Wang *et al.* [76] represents an early attempt that trains a fully convolutional neural network for VSOD. Another concurrent work [38] uses a 3D filter to incorporate both spatial and temporal information in a spatiotemporal CRF framework. Later, spatiotemporal deep feature [39], RNN [41], pyramid dilated convLSTM [67] are proposed for better capturing spatial and temporal saliency characteristics. These deep VSOD models generally achieved better performance due to the strong learning ability of neural network. However, these models ignored the saliency shift phenomenon which is quite important for understanding the human visual attention mechanism. In contrast, our SSAV model utilizes the saliency shift cue explicitly, yielding a competitive VSOD model.

In this work, we systematically benchmark 17 state-of-the-art VSOD models on seven previous datasetsand the proposed *DAVSOD* dataset, which represents the largest

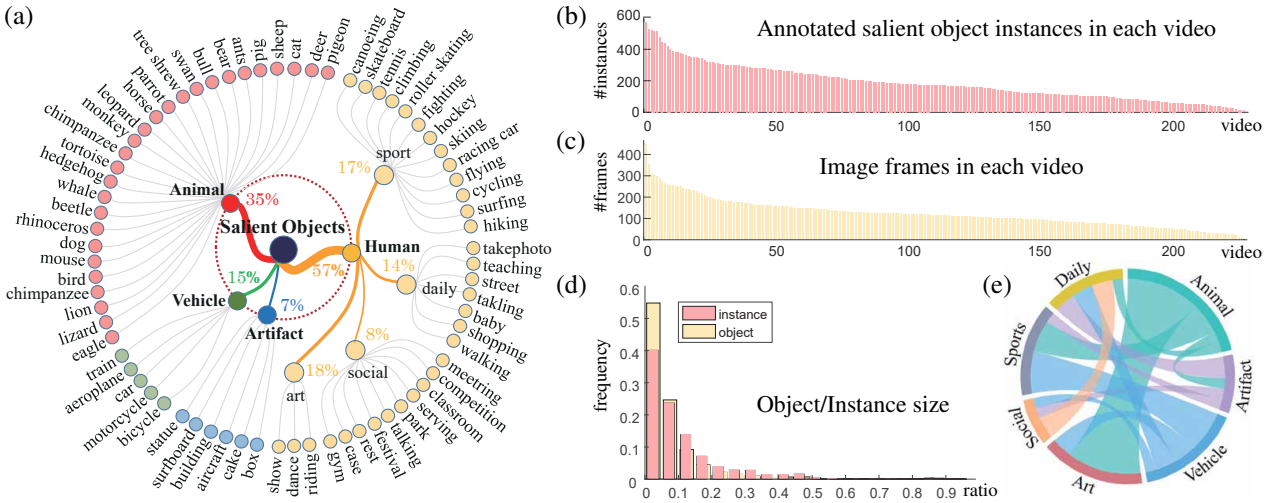

Figure 3: **Statistics of the proposed *DAVSOD* dataset**. (a) Scene/object categories. (b, c) Distribution of annotated instances and image frames, respectively. (d) Ratio distribution of the objects/instances. (e) Mutual dependencies among scene categories in (a).

performance evaluation in VSOD area so far. With our extensively quantitative results, we present deep insights into VSOD and point out some promising research directions.

3. Proposed Dataset

Some example frames can be found in Fig. 1 and Fig. 2. See our website for details. We will show details of *DAVSOD* from the following 4 key aspects.

3.1. Stimuli Collection

The stimuli of *DAVSOD* come from *DHF1K* [73], which is the current largest-scale dynamic eye-tracking dataset. There are several advantages of using *DHF1K* create our dataset. *DHF1K*[2] is collected from *Youtube* and covers diverse realistic-scenes, different object appearances and motion patterns, various object categories, and large span of major challenges in dynamic scenarios, providing us a solid basis to build a large-scale and representative benchmark. More essentially, the companied visual fixation record allows us to produce reasonable and biologically-inspired object-level saliency annotations. We manually trim the videos into shot clips (Fig. 3(c)) and remove dark-screen transitions. In this way, we finally reach a large-scale dataset, containing 226 video sequences with totally 23, 938 frames and 798 seconds duration.

3.2. Data Annotation

Saliency Shift Annotation. Human attention behavior is more complex during realistic, dynamic scenes [37,60], *i.e.*, selective attention allocation and overt attention shift (due to abrupt onsets, new dynamic events, *etc.*) may both happening. With the eye-tracking record of *DHF1K*, we also observe stimulus-driven attention-shifts [23] are ubiquitous, as shown in Fig. 1. However, none of the previous work in the VSOD area explicitly emphasizes such essential visual attention behavior. In *DAVSOD*, we annotate the salient objects according to real human fixations, and the temporal location at which attention shift occurs, for the first time, emphasizing the challenge of *saliency shift*[3] in this field.

Scene/Object Category Labeling. Consistent with [73], each video is manually labeled with a category (*i.e.*, *Animal*, *Vehicle*, *Artifact*, *Human Activity*). *Human Activity* has four sub-classes: *Sports*, *Daily*-, *Social*-, and *Art-Activity*. For object class, following *MSCOCO* [49], only "thing" categories instead of "stuff" are included. Then we built a list of about 70 most frequently present scenes/objects. In Fig. 3(a)&(e), we show the scene/object categories and their mutual dependencies, respectively. Five annotators were asked to annotate the object labels.

Instance-/Object-Level Salient Object Annotation. Twenty human annotators, who were pre-trained with ten video examples, are instructed to select up to five objects per-frame according to the corresponding fixation records and carefully annotate them (by tracing boundaries instead of rough polygons). They are also asked to differentiate instances and annotate them individually, resulting in totally 23,938 object-level ground-truth masks and 39,498 instance-level salient object annotations.

3.3. Dataset Features and Statistics

To offer deeper insights into the proposed *DAVSOD*, we discuss its several important characteristics.

Sufficient Salient Object Diversity. The salient objects in *DAVSOD* span a large set of classes (Fig. 3 (a)) such as animals (*e.g.*, lion, bird), vehicles (*e.g.*, car, bicycle), artifacts (*e.g.*, box, building), and humans in various activities (*e.g.*, dancer, rider), enabling a comprehensive understanding of object-level saliency in dynamic scenes.

[2]Download: https://github.com/wenguanwang/DHF1K

[3] **Notion of saliency shift.** The saliency shift is not just represented as a binary signal, w.r.t., whether it happens in a certain frame. Since we focus on an object-level task, we change the saliency values of different objects according to the shift of human attention.

DAVSOD	Camera Mo.		Object Mo.			# Object Instances			
	slow	fast	stable	slow	fast	1	2	3	≥ 4
# videos	102	124	117	72	37	134	125	46	33

Table 3: **Statistics regarding camera/object motions and salient object instance numbers** in *DAVSOD* dataset.

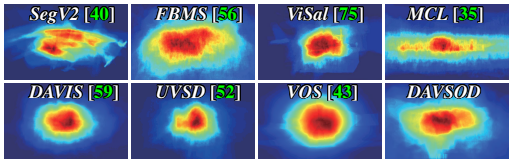

Figure 4: **Center bias** of *DAVSOD* and existing VSOD datasets.

Amount of Salient Object Instances. Existing datasets fall in short of limited numbers of salient object instances (Table 1). However, previous studies [34] showed human could accurately enumerate up to five objects at a glance without counting. In Table 3, *DAVSOD* is therefore designed to contain more salient objects ($\leq$ 5 salient object instances per-frame, avg.: 1.65). The distribution of annotated instances in each video can be found in Fig. 3(b).

Size of Salient Objects. The size of object-level salient object is defined as the proportion of foreground object pixels to the image. In Fig. 3(d), the ratio distribution in *DAVSOD* are $0.29\% \sim 91.3\%$ (avg.: 11.5%), yielding a broader range.

Varied Camera Motion Patterns. *DAVSOD* contains diverse camera motions (summarized in Table 3). Algorithms trained on such data could potentially handle realistic dynamic scenes better and thus are more practical.

Diverse Object Motion Patterns. *DAVSOD* inherits the advantage of *DHF1K* that covers diverse (Table 3) realistic dynamic scenes (*e.g.*, object motion from stable to fast). It is crucial to avoid over-fitting and benchmark algorithms objectively and precisely.

Center Bias. To depict the degree of center bias, we compute the average saliency map over all frames for each dataset. The center bias of *DAVSOD* and existing datasets [35, 40, 43, 52, 56, 59, 75] are presented in Fig. 4.

3.4. Dataset Splits

Existing datasets do not maintain a preserved test set, easily leading to model over-fitting. Thus, our videos are split into separate training, validation and test sets in the ratio of 4:2:4. Following random selection, we arrive at a unique split containing 90 training and 46 validation videos with released annotations, and 90 test videos with preserved annotations for benchmarking. The test set is further divided into 35 easy, 30 normal, and 25 difficult subsets according to the degree of difficulty of the VSOD task.

4. Proposed Approach

4.1. Saliency-Shift-Aware VSOD Model

Overview of Model. The proposed **SSAV** model has two essential components: *pyramid dilated convolution* (PDC) [67], and *saliency-shift-aware convLSTM* (SSLSTM). The former is for robust static saliency representation learning. The latter one extends traditional convLSTM [65] with saliency-shift-aware attention (SSAA) mechanism. It takes the static feature sequence from PDC module as input and produces corresponding VSOD results with considering temporal dynamics and saliency-shift simultaneously.

Pyramid Dilated Convolution (PDC) Module. Recent advance [10, 67] in semantic segmentation and VSOD showed that stacking a set of parallel dilated convolution layer with sampling rates can bring better performance, due to the exploit of multi-scale information and the preservation of spatial details. We use the PDC module [67] as our static feature extractor. Formally, let $\mathbf{Q} \in \mathbb{R}^{W \times H \times C}$ denote a 3D feature tensor of an input frame $\mathbf{I} \in \mathbb{R}^{w \times h \times 3}$. A dilated conv layer $\mathcal{D}_d$ with the dilated rate $d > 1$ can be applied to $\mathbf{Q}$ to obtain an output feature $\mathbf{P} \in \mathbb{R}^{W \times H \times C'}$, which maintains original spatial resolution while considering a larger receptive field (with sampling step d). The PDC is achieved by arranging a set of K dilated conv layers $\{\mathcal{D}_{d_k}\}_{k=1}^{K}$ with different dilated rates $\{d_k\}_{k=1}^{K}$ in parallel:

$$\mathbf{X} = [\mathbf{Q},\ \mathbf{P}_1, \ldots,\ \mathbf{P}_k, \ldots,\ \mathbf{P}_K], \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{W \times H \times (C + KC')}$ and $\mathbf{P}_k = \mathcal{D}_{d_k}(\mathbf{Q})$. $[.,.]$ indicates the concatenation operation. The PDC-enhanced feature $\mathbf{X}$ is a more robust representation (by leveraging multi-scale information) and preserves original information $\mathbf{Q}$ (through residual connection).

Saliency-Shift-Aware convLSTM (SSLSTM). We propose a saliency-shift-aware convLSTM, which equips convLSTM [65] with a saliency-shift-aware attention mechanism. It is a powerful recurrent model that not only captures temporal dynamics but also discriminates salient objects from the background as well as encodes attention-shift information. More specifically, through the PDC module, we obtain the static representations $\{\mathbf{X}_t\}_{t=1}^{T}$ of an input video with T frames. At time step t, given $\mathbf{X}_t$, the saliency-shift-aware convLSTM outputs the corresponding salient object mask $\mathbf{S}_t \in [0,1]^{W \times H}$:

$$\begin{aligned}
\text{Hidden state:} \quad & \mathbf{H}_t = \text{convLSTM}(\mathbf{X}_t, \mathbf{H}_{t-1}), \\
\text{Saliency-shift-aware attention:} \quad & \mathbf{A}_t = \mathcal{F}^A(\{\mathbf{X}_1, \cdots, \mathbf{X}_t\}), \\
\text{Attention-enhanced feature:} \quad & \mathbf{G}_{m,t} = \mathbf{A}_t \odot \mathbf{H}_{m,t}, \\
\text{Salient object prediction:} \quad & \mathbf{S}_t = \sigma(\mathbf{w}^S \otimes \mathbf{G}_t),
\end{aligned} \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^{W \times H \times M}$ indicates the 3D-tensor hidden state. The attention map $\mathbf{A} \in [0,1]^{W \times H}$ is computed from a saliency-shift-aware attention network $\mathcal{F}^A$, which takes previous frames into account. $\mathbf{G_t} \in \mathbb{R}^{W \times H \times M}$ indicates the attention-enhanced feature in time t. $\mathbf{G_{m,t}} \in \mathbb{R}^{W \times H}$ indicates the 2D feature slice of $\mathbf{G_t}$ in the m-th channel ($m \in [1, M]$). $\odot$ is element-wise multiplication. $\mathbf{w}^S \in \mathbb{R}^{1 \times 1 \times M}$, a 1×1 conv kernel, is adopted as a salient object readout

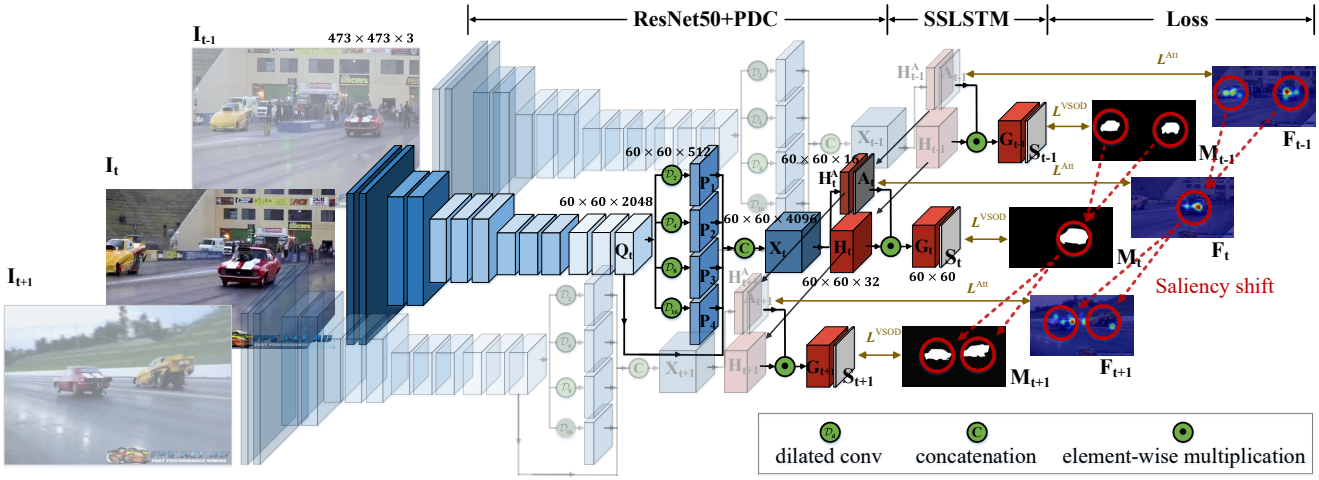

Figure 5: **Overall architecture of the proposed SSAV model**. SSAV consists of two components: pyramid dilated convolution (PDC) module and saliency-shift-aware convLSTM (SSLSTM) module. The former is for efficient static saliency learning, and the latter captures temporal dynamics and saliency-shift simultaneously. See § 4 for details.

function, $\otimes$ indicates conv operation and σ is the *sigmoid* activation function.

The key component of above module is the saliency-shift-aware attention network $\mathcal{F}^A$. Clearly, it acts as a neural attention mechanism since it is utilized to weight the output feature $\mathbf{H}$ of the convLSTM. Besides, it is desired to be effective enough to model the human attention-shift behavior. Considering such task is also different, a small convLSTM is introduced to build $\mathcal{F}^A$, generating a *convLSTM in convLSTM* structure:

$$\begin{aligned} \text{Saliency-shift-aware attention: } & \mathbf{A}_t = \mathcal{F}^A(\{\mathbf{X}_1, \cdots, \mathbf{X}_t\}), \\ \text{Attention feature extraction: } & \mathbf{H}_t^A = \text{convLSTM}^A(\mathbf{X}_t, \mathbf{H}_{t-1}^A), \\ \text{Attention mapping: } & \mathbf{A}_t = \sigma(\mathbf{w}^A \otimes \mathbf{H}_t^A), \end{aligned} \tag{3}$$

note that the first equation is formulated by the last two equations. Where $\mathbf{w}^A \in \mathbb{R}^{1\times1\times M}$ indicates a 1×1 conv kernel that maps the attention feature $\mathbf{H}^A$ as a significance matrix and *sigmoid* σ maps the significance value into $[0, 1]$. Then the attention $\mathbf{A}_t$ is employed to enhance the salient object segmentation feature $\mathbf{H}$ in Eq. 2. Due to the apply of convLSTM^A, our attention module gains strong learning ability, which provides a solid foundation for learning attention-shift in both explicit and implicit manners. Let $\{\mathbf{I}_t \in \mathbb{R}^{w\times h\times 3}\}_{t=1}^T$ denote a training video with T frames, $\{\mathbf{F}_t \in [0,1]^{W\times H}\}_{t=1}^T$ human eye-tracking annotation sequence and $\{\mathbf{M}_t \in \{0,1\}^{W\times H}\}_{t=1}^T$ video salient object ground-truth, we adopt a loss defined over the output $\{\mathbf{A}_t \in \{0,1\}^{W\times H}\}_{t=1}^T$ of the attention model and the final video salient object estimation $\{\mathbf{S}_t \in \{0,1\}^{W\times H}\}_{t=1}^T$:

$$\mathcal{L} = \sum_{t=1}^{T} \Big(\ell(\mathbf{I}_t) \cdot \mathcal{L}^{\text{Att}}(\mathbf{A}_t, \mathbf{F}_t) + \mathcal{L}^{\text{VSOD}}(\mathbf{S}_t, \mathbf{M}_t) \Big), \tag{4}$$

where $\mathcal{L}^{\text{Att}}$ and $\mathcal{L}^{\text{VSOD}}$ are both *cross entropy* loss. $\ell(\cdot) \in \{0,1\}$ indicates whether the attention annotation is available (since most current VSOD datasets lack eye-fixation record, see Table 1). When the corresponding attention annotation is missing, the error cannot be propagated back. More importantly, when $\ell(\cdot) = 0$, the saliency-shift-aware attention model $\mathcal{F}^A$ in Eq. 3 is trained *implicitly*, which can be viewed as a typical neural attention mechanism. When the ground-truth attention is available ($\ell(\cdot) = 1$), $\mathcal{F}^A$ is trained in an *explicit* way. With the convLSTM structure, $\mathcal{F}^A$ is powerful enough to accurately shift the attention of our VSOD model to the important objects (see Fig. 6).

4.2. Implementation Details

The base CNN network of PDC model is borrowed from the conv blocks from ResNet-50 [30] and the conv strides of the last two blocks are changed to 1. All the input frame images are resized into 473×473 spatial resolution, and $\mathbf{Q} \in \mathbb{R}^{60\times60\times2048}$. Following [67], we set $K = 4$, $C = 512$ and $d_k = 2^k$ ($k \in \{1, \cdots, 4\}$). For the convLSTM in Eq. 2, we use a $3 \times 3 \times 32$ conv kernel. The convLSTM^A in Eq. 3 utilizes a $3 \times 3 \times 16$ conv kernel. For training protocol, we follow the same settings in [67] (exclude MSRA-10k [12] dataset). In addition, we further exploit the validation set of *DAVSOD* to train the saliency-shift-aware attention module explicitly.

5. Benchmark Evaluation Results

5.1. Experimental Settings

Evaluation Metrics. To quantitatively assess the model performance, we adopt 2 popular evaluation metrics: Mean Absolute Error (**MAE**) $\mathcal{M}$ [58], **F-measure** $\mathcal{F}$ [1], and the recent released structure measure **S-measure** $\mathcal{S}$ [18].

Benchmark Models. We benchmark 17 models in total (11 traditional methods, 6 deep learning based models). These models were selected based on the two criteria: i) having released implementations, and ii) being representative.

Benchmark Protocols. To provide a comprehensive

	Metric	2010-2015							2016-2017				2018						
		SIVM [62]	TIMP [92]	SPVM [53]	RWRV [35]	MB+M [87]	SAGM [74]	GFVM [75]	MSTM [70]	STBP [81]	SGSP [52]	SFLR [8]	SCOM [11]†	SCNN [68]†	DLVS [76]†	FGRN [41]†	MBNM [44]†	PDBM [67]†	**SSAV**†
ViSal	max $\mathcal{F}\uparrow$	.522	.479	.700	.440	.692	.688	.683	.673	.622	.677	.779	.831	.831	.852	.848	.883	.888	***.939***
	$\mathcal{S}\uparrow$	.606	.612	.724	.595	.726	.749	.757	.749	.629	.706	.814	.762	.847	.881	.861	.898	.907	***.943***
	$\mathcal{M}\downarrow$	.197	.170	.133	.188	.129	.105	.107	.095	.163	.165	.062	.122	.071	.048	.045	.020	.032	***.020***
FBMS-T	max $\mathcal{F}\uparrow$	.426	.456	.330	.336	.487	.564	.571	.500	.595	.630	.660	.797	.762	.759	.767	.816	.821	***.865***
	$\mathcal{S}\uparrow$	.545	.576	.515	.521	.609	.659	.651	.613	.627	.661	.699	.794	.794	.794	.809	.857	.851	***.879***
	$\mathcal{M}\downarrow$	.236	.192	.209	.242	.206	.161	.160	.177	.152	.172	.117	.079	.095	.091	.088	.047	.064	***.040***
DAVIS-T	max $\mathcal{F}\uparrow$	.450	.488	.390	.345	.470	.515	.569	.429	.544	.655	.727	.783	.714	.708	.783	.861	.855	***.861***
	$\mathcal{S}\uparrow$	.557	.593	.592	.556	.597	.676	.687	.583	.677	.692	.790	.832	.783	.794	.838	.887	.882	***.893***
	$\mathcal{M}\downarrow$	.212	.172	.146	.199	.177	.103	.103	.165	.096	.138	.056	.048	.064	.061	.043	.031	.028	***.028***
SegV2	max $\mathcal{F}\uparrow$	.581	.573	.618	.438	.554	.634	.592	.526	.640	.673	.745	.764	**	**	**	.716	*.800*	***.801***
	$\mathcal{S}\uparrow$	.605	.644	.668	.583	.618	.719	.699	.643	.735	.681	.804	.815	**	**	**	.809	***.864***	.851
	$\mathcal{M}\downarrow$	.251	.116	.108	.162	.146	.081	.091	.114	.061	.124	.037	.030	**	**	**	.026	.024	***.023***
UVSD	max $\mathcal{F}\uparrow$	.293	.338	.404	.281	.339	.414	.426	.336	.403	.544	.562	.420	.550	.564	.630	.550	***.863***	.801
	$\mathcal{S}\uparrow$	.481	.537	.581	.536	.563	.629	.628	.551	.614	.601	.713	.555	.712	.721	.745	.698	***.901***	.861
	$\mathcal{M}\downarrow$	.260	.178	.146	.180	.169	.111	.106	.145	.105	.165	.059	.206	.075	.060	.042	.079	***.018***	.025
MCL	max $\mathcal{F}\uparrow$	.420	.598	.595	.446	.261	.422	.406	.313	.607	.645	.669	.422	.628	.551	.625	.698	***.798***	.774
	$\mathcal{S}\uparrow$	.548	.642	.665	.577	.539	.615	.613	.540	.700	.679	.734	.569	.730	.682	.709	.755	***.856***	.819
	$\mathcal{M}\downarrow$	.185	.113	.105	.167	.178	.136	.132	.171	.078	.100	.054	.204	.054	.060	.044	.119	***.021***	.027
VOS-T	max $\mathcal{F}\uparrow$	.439	.401	.351	.422	.562	.482	.506	.567	.526	.426	.546	.690	.609	.675	.669	.670	.742	***.742***
	$\mathcal{S}\uparrow$	.558	.575	.511	.552	.661	.619	.615	.657	.576	.557	.624	.712	.704	.760	.715	.742	.818	***.819***
	$\mathcal{M}\downarrow$	.217	.215	.223	.211	.158	.172	.162	.144	.163	.236	.145	.162	.109	.099	.097	.099	.078	***.073***
DAVSOD-T	max $\mathcal{F}\uparrow$	.298	.395	.358	.283	.342	.370	.334	.344	.410	.426	.478	.464	.532	.521	.573	.520	.572	***.603***
	$\mathcal{S}\uparrow$	.486	.563	.538	.504	.538	.565	.553	.532	.568	.577	.624	.599	.674	.657	.693	.637	.698	***.724***
	$\mathcal{M}\downarrow$	.288	.195	.202	.245	.228	.184	.167	.211	.160	.207	.132	.220	.128	.129	.098	.159	.116	***.092***

Table 4: **Benchmarking results of 17 state-of-the-art VSOD models on 7 datasets**: *SegV2* [40], *FBMS* [56], *ViSal* [75], *MCL* [35], *DAVIS* [59], *UVSD* [52], *VOS* [43] and the proposed *DAVSOD* (35 easy test set). Note that TIMP was only tested on 9 short sequences of *VOS* because it cannot handle long videos. "**" indicates the model has been trained on this dataset. "-T" indicates the results on the test set of this dataset. "†" indicates deep learning model. Darker color indicates better performance. The best scores are marked in ***bold***.

benchmark, we evaluate 17 representative methods on existing 7 datasets and the proposed *DAVSOD* dataset. The test sets of *FBMS* [56] (30 clips), *DAVIS* [59] (20 clips), *DAVSOD* (35 easy clips) datasets, and the whole *ViSal* [75] (17 clips), *MCL* [35] (9 clips), *SegV2* [40] (13 clips), *UVSD* [52] (18 clips) datasets are used for testing. For *VOS* [43] dataset, we randomly select 40 sequences as test set. There are total 182 videos with 848,340 (47,130×18) frames.

5.2. Performance Comparison and Data Analysis

In this section, we provide some interesting findings which would benefit the further research.

Performance of Traditional Models. Based on the different metrics in Table 4, we conclude that: *"SFLR [8], SGSP [52], and STBP [81] are the top 3 non-deep learning models for VSOD."* Both SFLR and SGSP explicitly consider the optical flow strategy to extract the motion features. However, the computational cost is usually expensive (see Table 2). One noteworthy finding is that all these models utilize the superpixel technology to integrate spatiotemporal features on region level.

Performance of Deep Models. The top 3 models in this benchmark (*i.e.*, SSAV, PDBM [67], MBNM [44]) are all based on deep learning technique, which demonstrates the strong learning power of neural networks. For *ViSal* dataset (the first specifically-designed dataset for VSOD), their average performance (*e.g.*, max E-measure [19], max F-measure, or S-measure) is even higher than 0.9.

Traditional *vs* Deep VSOD Models. In Table 4, almost all of the deep models outperform traditional algorithms, as more powerful saliency representations can be extracted from networks. Another interesting finding is the classic leading method (SFLR [8]) performs better than some deep models (*e.g.*, SCOM [11]) on *MCL*, *UVSD*, *ViSal*, and *DAVSOD* datasets. It indicates that investigating more effective deep learning architectures with the exploit of human prior knowledge for VSOD is a promising direction.

Dataset Analysis. We mark the scores with gray color in Table 4. Darker colors mean better performance for specific metrics (*e.g.*, max $\mathcal{F}, \mathcal{S}$, and $\mathcal{M}$). We find that *ViSal* and *UVSD* datasets are relatively easy, since the top 2 models: SSAV and PDBM [67] gained very high performance (*e.g.*, $\mathcal{S} > 0.9$). However, for more challenging datasets like *DAVSOD*, the performance of VOSD models decrease dramatically ($\mathcal{S} < 0.73$). It reveals that both the overall and individual performance of VOSD models leave abundant room for future research.

Runtime Analysis. Table 2 reports the computation time of previous VSOD methods and the proposed SSAV approach

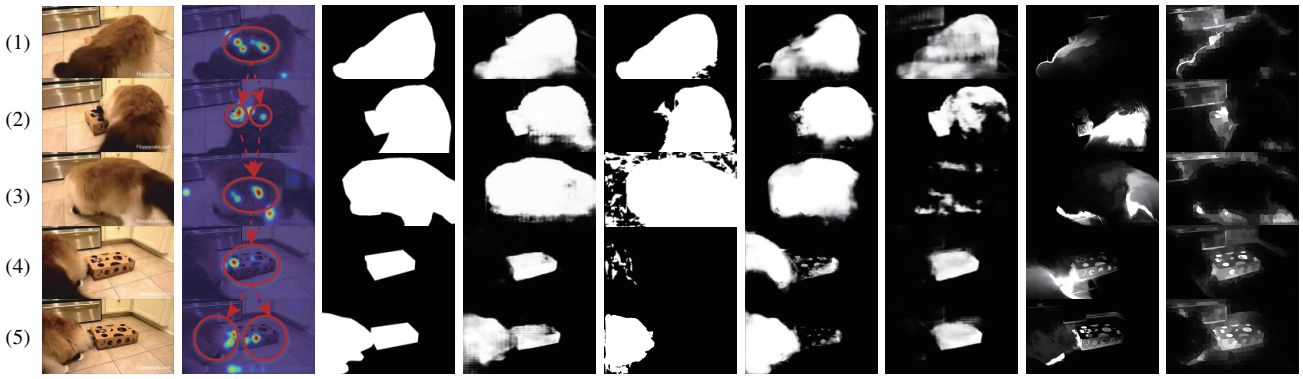

Figure 6: **Visual comparisons with top 3 deep (MBNM [44], FGRN [41], PDBM [67]) models and 2 traditional classical (SFLR [8], SAGM [74]) models on the proposed *DAVSOD* dataset**. Our SSAV model captures the saliency shift phenomenon successfully.

(in **PCT** column). For the models with released codes, the timings are tested on the same platform: Intel Xeon(R) E5-2676v3 @2.4GHz×24 and GTX TITAN X. The rest of the timings are borrowed from their papers. Note that the proposed model does not apply any pre-/post-processing (*e.g.*, CRF), thus the processing speed only takes about 0.05s.

5.3. Ablation Study

Implicit *vs* Explicit Saliency-Shift-Aware Attention Mechanism. To study the influence of different training strategies of the proposed SSAA module, we derive 2 baselines: *explicit* and *implicit*, refer to the proposed SSAV model trained explicitly or implicitly. We obtain the *implicit* baseline by only using VSOD annotations (exclude *DAVSOD*). We observe that SSAV with explicit attention is better than the one with implicit attention, according to the statistics in Table 5. It demonstrates that utilizing fixation data can help our model to better capture saliency shift and thus further boost final VSOD performance.

Effectiveness of Saliency-Shift-Aware convLSTM. To study the effectiveness of SSLSTM (§ 4), we provide another baseline: *w/o SSLSTM*, which excludes SSLSTM module from the proposed SSAV model. From Table 5, we observe a performance decrease (*e.g.*, $\mathcal{S}: 0.724 \rightarrow 0.667$), which confirms that the proposed SSLSTM module is effective to learn both selective attention allocation and attention shift cues from the challenging data.

Comparison with State-of-the-Arts. In Table 4, we compare the proposed SSAV model with current 17 state-of-the-art VSOD algorithms. The proposed baseline method performs better against other competitors over most existing datasets. More specifically, our model obtains significant performance improvements on *ViSal* and *FBMS* datasets. It also obtains comparable performance on *VOS*, *SegV2* and *DAVIS* datasets.

5.4. Analysis for the saliency shift challenge

For the proposed challenging *DAVSOD* dataset, the SSAV model also gains the best performance. We attribute the promising performance to the introduce of SSLSTM, which efficiently captures saliency allocations in dynamic scenes and guides our model to accurately attend to those visually important regions. Fig. 6 shows that the proposed SSAV approach obtains more visually favorable results than other top competitors. Our SSAV model captures the saliency shift successfully (from frame-1 to frame-5: $cat \rightarrow [cat, box] \rightarrow cat \rightarrow box \rightarrow [cat, box]$). However, the other top-performance VSOD models either do not highlight the whole salient objects (*e.g.*, SFLR, SAGM) or only capture the moving cat (*e.g.*, MBNM). We envision our SSAV model would open up promising future directions for model development.

Type	Baseline	$\mathcal{S} \uparrow$	max $\mathcal{F} \uparrow$	$\mathcal{M} \downarrow$
SSAA	*explicit*	**0.724**	**0.603**	**0.092**
	implicit	0.684	0.593	0.103
SSLSTM	*w/o SSLSTM*	0.667	0.541	0.132

Table 5: **Ablation studies of the SSAV on *DAVSOD* dataset.**

6. Conclusion

We have presented a comprehensive study on VSOD by creating a new visual-attention-consistent *DAVSOD* dataset, building up the largest-scale benchmark, and proposing a SSAV baseline model. Compared with other competing traditional or deep learning models, the proposed SSAV model achieves superior performance and produces more visually favorable results. Extensive experiments verified that even considering top performing models, VSOD remain seems far from being solved. The above contributions and in-depth analyses would benefit the develop of this area and be helpful to stimulate broader potential research, *e.g.*, saliency-aware video captioning, video salient object subitizing and instance-level VSOD.

Acknowledgements. This research was supported by NSFC (61620106008, 61572264), the national youth talent support program, the Fundamental Research Funds for the Central Universities (Nankai University, NO. 63191501) and Tianjin Natural Science Foundation (17JCJQJC43700, 18ZXZNGX00110).

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009. 6

[2] Tariq Alshawi, Zhiling Long, and Ghassan AlRegib. Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection. *IEEE TIP*, pages 2818–2827, 2018. 3

[3] Çağlar Aytekin, Horst Possegger, Thomas Mauthner, Serkan Kiranyaz, Horst Bischof, and Moncef Gabbouj. Spatiotemporal saliency estimation by spectral foreground detection. *IEEE TMM*, 20(1):82–95, 2018. 3

[4] Saumik Bhattacharya, K Subramanian Venkatesh, and Sumana Gupta. Visual saliency detection using spatiotemporal decomposition. *IEEE TIP*, 27(4):1665–1675, 2018. 3

[5] Marc D. Binder, Nobutaka Hirokawa, and Uwe Windhorst, editors. *Gaze Shift*, pages 1676–1676. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 1

[6] 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. 1

[7] Chenglizhao Chen, Shuai Li, Hong Qin, Zhenkuan Pan, and Guowei Yang. Bi-level feature learning for video saliency detection. *IEEE TMM*, 2018. 1, 3

[8] Chenglizhao Chen, Shuai Li, Yongguang Wang, Hong Qin, and Aimin Hao. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE TIP*, 26(7):3156–3170, 2017. 2, 3, 7, 8

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 5

[11] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. Scom: Spatiotemporal constrained optimization for salient object detection. *IEEE TIP*, 27(7):3345–3357, 2018. 2, 3, 7

[12] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 3, 6

[13] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE TCSVT*, PP(99):1–19, 2018. 3

[14] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE TIP*, 27(2):568–579, 2018. 3

[15] Runmin Cong, Jianjun Lei, Huazhu Fu, Weisi Lin, Qingming Huang, Xiaochun Cao, and Chunping Hou. An iterative co-saliency framework for rgbd images. *IEEE TYCB*, 49(1):233–246, 2019. 3

[16] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 3

[17] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*. Springer, 2018. 1

[18] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *IEEE ICCV*, pages 4548–4557, 2017. 3, 6

[19] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, pages 698–704, 2018. 7

[20] 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

[21] 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

[22] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, pages 524–531, 2005. 3

[23] Steven L Franconeri, Daniel J Simons, and Justin A Junge. Searching for stimulus-driven shifts of attention. *Psychonomic Bulletin & Review*, 11(5):876–881, 2004. 4

[24] Ken Fukuchi, Kouji Miyazato, Akisato Kimura, Shigeru Takagi, and Junji Yamato. Saliency-based video segmentation with graph cuts and sequentially updated priors. In *ICME*, pages 638–641, 2009. 3

[25] Siavash Gorji and James J Clark. Going From Image to Video Saliency: Augmenting Image Salience With Dynamic Attentional Push. In *IEEE CVPR*, pages 7501–7511, 2018. 1

[26] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE CVPR*, pages 1–8, 2008. 3

[27] 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. 1

[28] Fang Guo, Wenguan Wang, Jianbing Shen, Ling Shao, Jian Yang, Dacheng Tao, and Yuan Yan Tang. Video saliency detection using object proposals. *IEEE Transactions on Cybernetics*, 2017. 3

[29] Hadi Hadizadeh and Ivan V Bajic. Saliency-aware video compression. *IEEE TIP*, 23(1):19–33, 2014. 1

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 6

[31] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*. Springer, 2018. 1, 3

[32] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *IEEE CVPR*, pages 7142–7150, 2018. 3

[33] Yeong Jun Koh, Young-Yoon Lee, and Chang-Su Kim. Sequential clique optimization for video object segmentation. In *ECCV*. Springer, 2018. 3

[34] Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949. 5

[35] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE TIP*, 24(8):2552–2564, 2015. 1, 2, 3, 5, 7

[36] Wonjun Kim, Chanho Jung, and Changick Kim. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE TCSVT*, 21(4):446–456, 2011. 1, 3

[37] Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. 1987. 1, 4

[38] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, 2017. 1, 3

[39] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE TIP*, 27(10):5002–5015, 2018. 1, 3

[40] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE ICCV*, pages 2192–2199, 2013. 1, 2, 3, 5, 7

[41] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *IEEE CVPR*, pages 3243–3252, 2018. 2, 3, 7, 8

[42] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *IEEE CVPR*, pages 478–487, 2016. 3

[43] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE TIP*, 27(1):349–364, 2018. 1, 2, 5, 7

[44] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C.-C. Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*. Springer, 2018. 2, 3, 7, 8

[45] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *IEEE CVPR*, pages 280–287, 2014. 1

[46] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE CVPR*, pages 2359–2367, 2017. 3

[47] Yong Li, Bin Sheng, Lizhuang Ma, Wen Wu, and Zhifeng Xie. Temporally coherent video saliency using regional dynamic contrast. *IEEE TCSVT*, 23(12):2067–2076, 2013. 3

[48] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *IEEE CVPR*, pages 3367–3375, 2015. 3

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 3, 4

[50] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *IEEE CVPR*, pages 3089–3098, 2018. 3

[51] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to Detect A Salient Object. In *IEEE CVPR*, pages 1–8, 2007. 3

[52] 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. 1, 2, 3, 5, 7

[53] Zhi Liu, Xiang Zhang, Shuhua Luo, and Olivier Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE TCSVT*, 24(9):1522–1540, 2014. 2, 3, 7

[54] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, pages 3431–3440, 2015. 3

[55] Thomas Mauthner, Horst Possegger, Georg Waltner, and Horst Bischof. Encoding based saliency detection for videos and images. In *IEEE CVPR*, pages 2494–2502, 2015. 3

[56] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2014. 1, 2, 3, 5, 7

[57] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, pages 6504–6512, 2017. 1

[58] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 6

[59] 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 *IEEE CVPR*, pages 724–732, 2016. 1, 2, 3, 5, 7

[60] Matthew S Peterson, Arthur F Kramer, and David E Irwin. Covert shifts of attention precede involuntary eye movements. *Perception & Psychophysics*, 66(3):398–405, 2004. 1, 4

[61] Wenliang Qiu, Xinbo Gao, and Bing Han. Eye fixation assisted video saliency detection via total variation-based pairwise interaction. *IEEE TIP*, pages 4724–4739, 2018. 1, 3

[62] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting salient objects from images and videos. In *ECCV*, pages 366–379. Springer, 2010. 2, 3, 7

[63] 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

[64] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *CVPR*, pages 3241–3250, 2015. 3

[65] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LST-

M network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 5

[66] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[67] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Sheng, and Kin-Man Lam. Pyramid dilated deeper convLSTM for video salient object detection. In *ECCV*. Springer, 2018. 2, 3, 5, 6, 7, 8

[68] Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen, Yang Hua, and Xia Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE TCSVT*, 2018. 2, 3, 7

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

[70] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *IEEE CVPR*, pages 2334–2342, 2016. 2, 3, 7

[71] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *IEEE CVPR*, pages 3127–3135, 2018. 3

[72] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *IEEE CVPR*, pages 1711–1720, 2018. 3

[73] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *IEEE CVPR*, pages 4894–4903, 2018. 4

[74] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *IEEE CVPR*, pages 3395–3402, 2015. 1, 2, 3, 7, 8

[75] 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. 1, 2, 3, 5, 7

[76] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2018. 2, 3, 7

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

[78] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, pages 20–33, 2018. 3

[79] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42. Springer, 2012. 3

[80] 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. 3

[81] Tao Xi, Wei Zhao, Han Wang, and Weisi Lin. Salient object detection with spatiotemporal background priors for video. *IEEE TIP*, 26(7):3425–3436, 2017. 2, 3, 7

[82] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *IEEE CVPR*, pages 373–381, 2016. 1

[83] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018. 1

[84] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE CVPR*, pages 3166–3173, 2013. 3

[85] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

[86] Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, and Ali Borji. Learning to promote saliency detectors. In *IEEE CVPR*, pages 1644–1653, 2018. 3

[87] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *IEEE ICCV*, pages 1404–1412, 2015. 2, 3, 7

[88] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *IEEE CVPR*, pages 9029–9038, 2018. 3

[89] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *IEEE CVPR*, pages 1741–1750, 2018. 3

[90] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *IEEE CVPR*, pages 714–722, 2018. 3

[91] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *IEEE CVPR*, pages 669–677, 2016. 1

[92] Feng Zhou, Sing Bing Kang, and Michael F Cohen. Time-mapping using space-time saliency. In *IEEE CVPR*, pages 3358–3365, 2014. 2, 3, 7

[93] Xiaofei Zhou, Zhi Liu, Chen Gong, and Wei Liu. Improving video saliency detection via localized estimation and spatiotemporal refinement. *IEEE TMM*, pages 2993–3007, 2018. 3

[94] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 3

[95] Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, and Ahmed Elgammal. Learning where to look: Semantic-guided multi-attention localization for zero-shot learning. *arXiv preprint arXiv:1903.00502*, 2019. 1