A Benchmark Dataset and Saliency-Guided Stacked Autoencoders for Video-Based Salient Object Detection

Jia Li, *Senior Member, IEEE*, Changqun Xia, and Xiaowu Chen, *Senior Member, IEEE*

***Abstract*— Image-based salient object detection (SOD) has been extensively studied in past decades. However, video-based SOD is much less explored due to the lack of large-scale video datasets within which salient objects are unambiguously defined and annotated. Toward this end, this paper proposes a video-based SOD dataset that consists of 200 videos. In constructing the dataset, we manually annotate all objects and regions over 7650 uniformly sampled keyframes and collect the eye-tracking data of 23 subjects who free-view all videos. From the user data, we find that salient objects in a video can be defined as objects that consistently pop-out throughout the video, and objects with such attributes can be unambiguously annotated by combining manually annotated object/region masks with eye-tracking data of multiple subjects. To the best of our knowledge, it is currently the largest dataset for video-based salient object detection. Based on this dataset, this paper proposes an unsupervised baseline approach for video-based SOD by using saliency-guided stacked autoencoders. In the proposed approach, multiple spatiotemporal saliency cues are first extracted at the pixel, superpixel, and object levels. With these saliency cues, stacked autoencoders are constructed in an unsupervised manner that automatically infers a saliency score for each pixel by progressively encoding the high-dimensional saliency cues gathered from the pixel and its spatiotemporal neighbors. In experiments, the proposed unsupervised approach is compared with 31 state-of-the-art models on the proposed dataset and outperforms 30 of them, including 19 image-based classic (unsupervised or non-deep learning) models, six image-based deep learning models, and five video-based unsupervised models. Moreover, benchmarking results show that the proposed dataset is very challenging and has the potential to boost the development of video-based SOD.**

***Index Terms*— Salient object detection, video dataset, stacked autoencoders, model benchmarking.**

Manuscript received May 4, 2017; revised July 17, 2017 and August 31, 2017; accepted October 6, 2017. Date of publication October 12, 2017; date of current version November 3, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61672072, Grant 61532003, and Grant 61421003, and in part by the Beijing Nova Program and the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Amit K. Roy Chowdhury. *(Corresponding authors: Xiaowu Chen; Jia Li.)*

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, also with the International Research Institute for Multidisciplinary Science, Beihang University, Beijing 100191, China (e-mail: jiali@buaa.edu.cn).

C. Xia and X. Chen are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: xiacq@buaa.edu.cn; chen@buaa.edu.cn).

I. Introduction

THE rapid development of image-based salient object detection (SOD) originates from the availability of large-scale benchmark datasets [1], [2]. With these datasets, it becomes feasible to construct complex models with machine learning algorithms (*e.g.*, random forest regressor [3], bootstrap learning [4], multi-instance learning [5] and deep learning [6]). Moreover, the availability of such datasets enables fair comparisons between state-of-the-art models [7], [8]. Large-scale datasets provide a solid foundation for SOD and consistently guide the development of this area.

Currently, SOD datasets are evolving to meet the increasing demands in developing and benchmarking models. Some researchers argue that images in early datasets such as **ASD** [2] and **MSRA-B** [1] are relatively small and simple. They extend such datasets in terms of size [9], [10] or complexity [11]–[13]. Meanwhile, the concept of SOD has been extended to RGBD images [14], image collections [15]–[17] and videos [18]–[21]. Among these extensions, video-based SOD has attracted great research interest since it re-defines the problem from a spatiotemporal perspective. However, there is still a lack of large-scale video datasets for comprehensive model comparison, which prevent the fast growth of this branch. For example, the widely used **SegTrack** dataset [22] consists of only six videos with 21 to 71 frames per video, while a recent dataset **ViSal** [21] contains only 17 videos with 30 to 100 frames per video. In addition, the definition of a salient object in videos is still not very clear (*e.g.*, manually annotated foreground objects [23], class-specific objects [21] or moving objects [24]). It is necessary to construct a large video dataset with unambiguously defined salient objects.

To address this issue, this paper proposes **VOS**, which is a large-scale dataset with 200 indoor/outdoor videos for video-based SOD (64 minutes and 116, 103 frames; see Fig. 1). In constructing **VOS**, we collect two types of user data: the eye-tracking data of 23 subjects who free-view all 200 videos, and the masks of all objects and regions in 7, 650 uniformly sampled keyframes annotated by another four subjects. With these data, salient objects in a video can be unambiguously annotated as the objects that consistently receive the highest density of fixations throughout the video. By discarding pure-background keyframes (defined as frames containing only

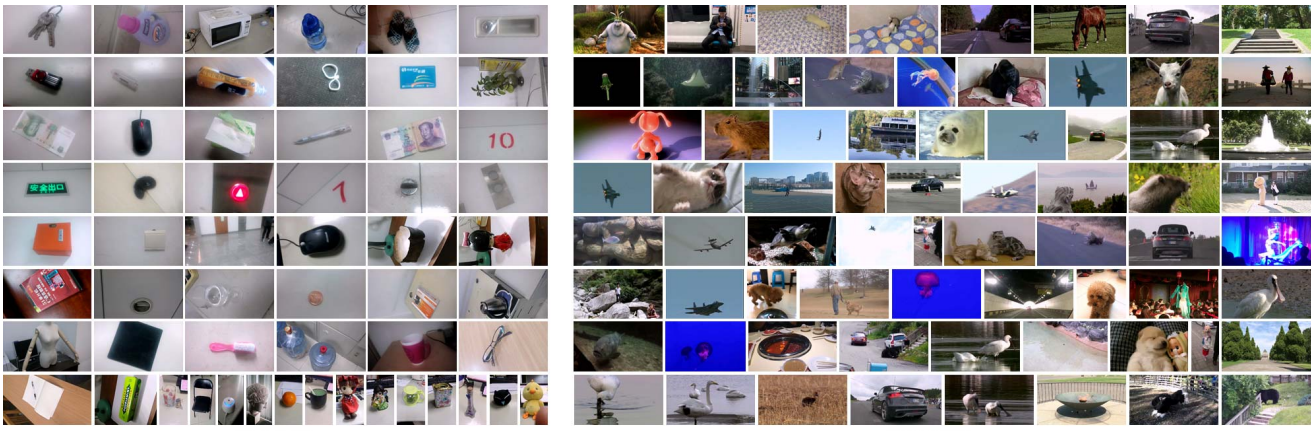


Fig. 1. Representative scenarios in **VOS**. The 200 videos in **VOS** are grouped into two subsets according to the complexity of foreground, background and motion, including **VOS-E** (*easy* subset, 97 videos) and **VOS-N** (*normal* subset, 103 videos).

widely recognized background regions such as blue sky, white wall or green grassland) as well as keyframes in which salient objects are partially occluded or divided into several disjoint parts, we obtain 7,467 keyframes with binary salient object masks.

Based on this dataset, we propose an unsupervised model for video-based SOD by constructing saliency-guided stacked autoencoders. Different from the fixation prediction task, which aims to roughly detect where the human being is looking, and the image-based SOD task, which aims to segment only the most spatially salient objects, the video-based SOD task focuses on detecting and segmenting the objects that consistently pop-out throughout a video from a spatiotemporal perspective. Inspired by this, the proposed approach first extracts multiple spatiotemporal saliency cues at the pixel, superpixel and object levels. Stacked autoencoders are then trained in an unsupervised manner to automatically infer a saliency score for each pixel by progressively encoding the high-dimensional saliency cues gathered from the pixel and its spatiotemporal neighbors. In the comprehensive model benchmarking on **VOS**, the proposed approach is compared with 31 state-of-the-art models and outperforms 30 of them. Moreover, the benchmarking results show that **VOS** is a challenging dataset with the potential to greatly boost the development of this area.

Our main contributions are summarized as follows: 1) we propose a large and challenging dataset for video-based SOD, which can be useful for the development of this area; 2) we propose saliency-guided stacked autoencoders for video-based SOD, which outperform 30 state-of-the-art models; and 3) we provide a comprehensive benchmark of our approach and massive state-of-the-art models, which reveals several key challenges in video-based SOD and further validates the usefulness of the proposed dataset.

The rest of this paper is organized as follows: Section II reviews related works and Section III presents a new dataset. In Section IV, we propose saliency-guided stacked autoencoders for video-based SOD. Section V benchmarks the proposed model and the state-of-the-art models, and the paper is concluded in Section VI.

II. Related Work

Video-based SOD is closely related to image-based SOD, foreground/primary object detection and moving object segmentation. In this section, we will review the most relevant datasets and models from these areas.

A. Datasets

1) SegTrack [22] is a popular dataset for evaluating the segmentation accuracy in video tracking. It contains six videos about animals and humans with 244 frames in total, and the videos are intentionally collected with predefined challenges such as color overlap between target and background appearance, inter-frame motion and change in target shape. Only one foreground object is annotated per frame.

2) SegTrack V2 [27] enhances **SegTrack** by adding additional annotations of foreground objects for the six videos in **SegTrack**. Moreover, eight new videos are carefully chosen to cover additional challenges such as motion blur, appearance change, complex deformation, slow motion, occlusion and multiple adjacent/interacting objects. In total, **SegTrack V2** contains 14 videos about birds, animals, cars and humans with 1,065 densely annotated frames.

3) Freiburg-Berkeley Motion Segmentation (FBMS) [24], [28] is designed for motion segmentation (*i.e.*, segmenting regions with similar motion). It was first proposed in [24] with 26 videos, and then Ochs *et al.* [28] extended the dataset with another 33 videos. In total, this dataset contains 59 videos with 720 sparsely annotated frames with typical challenges such as multiple objects, various motion types, occlusion and changing lighting conditions. Although the dataset is much larger than **SegTrack** and **SegTrack V2**, the scenarios it covers are still far from sufficient [23]. Moreover, moving objects are not equivalent to salient objects, especially in a scene with complex content.

4) DAVIS [23] is a video dataset about humans, animals, vehicles, objects and actions (50 videos, with 3,455 densely annotated frames). Each video has Full HD 1080p resolution and lasts approximately 2 to 4 seconds. Typically, a video contains one foreground object or two spatially connected

TABLE I

COMPARISON BETWEEN **VOS** (SUBSETS: **VOS-E** AND **VOS-N**) AND 12 REPRESENTATIVE IMAGE/VIDEO OBJECT SEGMENTATION DATASETS. THE COLUMNS #AVG. OBJ. AND OBJ. AREA (%) ARE THE AVERAGE NUMBER AND AREA OF FOREGROUND OBJECTS PER IMAGE OR FRAME, RESPECTIVELY

	Dataset	#Vid.	Resolution (in pixels)			#Orig. Frames		#Labeled Frames		#Avg. Obj.	Obj. Area (%)
			Width	Height	Max Res.	Total	Avg.	Total	Avg.		
Image-based	**ASD** [2]	-	[188, 400]	[165, 400]	400 × 400	1000	-	1000	-	1.16±0.87	19.9 ± 9.52
	ECSSD [11]	-	[222, 400]	[139, 400]	400 × 400	1000	-	1000	-	1.16±0.56	28.5 ± 20.5
	DUT-O [10]	-	[167, 401]	[89, 401]	401 × 401	5168	-	5168	-	1.20±0.69	14.9 ± 12.2
	PASCAL-S [13]	-	[191, 500]	[151, 500]	500 × 500	850	-	850	-	1.52±1.11	24.7 ± 16.0
	MSRA10K [9]	-	[179, 400]	[165, 400]	400 × 400	10000	-	10000	-	1.05±0.46	22.2 ± 10.1
	HKU-IS [25]	-	[100, 401]	[100, 401]	400 × 400	4447	-	4447	-	1.60±0.82	19.1 ± 10.9
	XPIE [26]	-	[128, 300]	[128, 300]	300 × 300	10000	-	10000	-	1.15±0.46	19.4 ± 14.4
Video-based	**SegTrack** [22]	6	[259, 414]	[212, 352]	414 × 352	244	41±18	244	41±18	1.00±0.00	3.46 ± 2.84
	SegTrack V2 [27]	14	[259, 640]	[212, 360]	640 × 360	1065	76±82	1065	76±82	1.38±1.01	7.38 ± 7.89
	FBMS [28]	59	[350, 960]	[253, 540]	960 × 540	13860	235±193	720	12±8	1.78±1.54	14.4 ± 13.7
	DAVIS [23]	50	[1600, 1920]	[900, 1080]	1920 × 1080	3455	69±19	3455	69±19	5.39±22.87	8.10 ± 6.44
	ViSal [21]	17	[320, 512]	[240, 288]	512 × 288	963	57±20	193	11±4	1.16±0.40	10.5 ± 6.51
	VOS-E	97	[408, 800]	[448, 800]	800 × 640	49206	507±130	3236	33±9	1.02±0.18	18.4 ± 12.8
	VOS-N	103	[448, 800]	[312, 800]	800 × 800	66897	649±510	4231	41±33	1.25±0.54	8.92 ± 10.8
	VOS	200	[408, 800]	[312, 800]	800 × 800	116103	581±383	7467	37±25	1.15±0.44	13.0 ± 12.6

* Objects are counted as disconnected foreground regions. In **DAVIS**, a semantic object may be divided into hundreds of disconnected parts (*e.g.*, a bus occluded by a tree), leading to an extremely high mean and standard deviation in the number of foreground "objects" per frame.

objects (an object may be split into many small regions due to occlusion), covering challenges such as occlusions, motion blur and appearance changes.

5) ViSal [21] is a pioneering video dataset that intends to provide a deeper exploration of video-based SOD. It contains 17 videos about humans, animals, motorbikes, etc. Each video contains 30 to 100 frames, in which salient objects are annotated according to the semantic classes of videos. In other words, this dataset assumes that salient objects are equivalent to the primary objects within the videos annotated by semantic tags. Major challenges in these videos include complex color distributions, highly cluttered backgrounds, various object motion patterns, rapid topology changes and camera motion.

To facilitate the comparison between these datasets and our **VOS** dataset, we show in Table I some dataset statistics. Moreover, we list the details of seven image-based SOD datasets to provide an intuitive comparison between image- and video-based SOD. Generally, the previous datasets reviewed above have greatly boosted the research in video object segmentation, but have several drawbacks.

First, these datasets are slightly smaller for modern learning algorithms such as Convolutional Neural Networks (CNN). As shown in Table I, the numbers of annotated frames in most previous video datasets are much smaller than those in the image-based SOD datasets and **VOS**. Although thousands of frames in **SegTrack V2** and **DAVIS** are densely annotated, the rich redundancy in consecutive frames may increase the over-fitting risk in model training.

Second, videos in some datasets are selected to maximally cover predefined challenges in video object segmentation (*e.g.*, **SegTrack** and **SegTrack V2**). However, such intentionally selected videos may make these datasets not very "realistic" (*i.e.*, different from the videos in real-world scenarios). Moreover, such datasets may favor models that are particularly designed to "over-fit" the limited scenarios. In contrast, our **VOS** dataset is much larger, so the over-fitting risk can be largely alleviated.

Third, foreground objects in previous datasets are often manually annotated by only one or several annotators, which may result in the incorporation of strong subjective bias into these datasets. For example, in a video with both a dog and monkey, only the monkey is annotated in **SegTrack**, while **SegTrack V2** has the dog annotated as well. Moreover, manual annotations from different subjects often conflict with one another [29] and cause ambiguity. To alleviate such ambiguity, previous works such as [30]–[32] have tried to locate salient targets by averaging rectangles that were manually annotated by 23 subjects [30] or collecting human fixations via an eye-tracking apparatus [31], [32]. However, these datasets cannot be directly used in video-based SOD due to a lack of pixel-wise annotations of salient objects. Pixel-wise annotation is the most time-consuming step in constructing video-based SOD datasets.

To sum up, existing datasets are insufficient for benchmarking video-based SOD models due to the limited numbers of videos as well as the ambiguous definition and annotation processes of salient/foreground/moving objects. To further boost the development of this area, it is necessary to construct a large-scale dataset that covers a wide variety of real-world scenarios and contains salient objects that are unambiguously defined and annotated.

B. Models

Hundreds of image-based SOD models [3], [33]–[37] have been proposed in the past decade. With the development of deep learning methodology and the availability of large-scale datasets [10], [25], [38], many deep models [6], [39]–[41] have been proposed for image-based SOD. For example, Han *et al.* [42] proposed multi-stream stacked denoising autoencoders that can detect salient regions by measuring the reconstruction residuals that reflect the distinction between background and salient regions. He *et al.* [43] adopted CNNs to characterize superpixels with hierarchical features to detect salient objects on multiple scales, and such superpixel-based saliency computation was used by [25] and [44] as well.

Considering that the task of fixation prediction is closely related to SOD, a unified deep network was proposed in [45] for simultaneous fixation prediction and image-based SOD.

The state-of-the-art deep SOD models often adopt recurrent frameworks that can achieve impressive performance. For example, Liu and Han [46] adopted hierarchical recurrent CNNs to progressively refine the details of salient objects. In [47], a coarse saliency map was first generated by using the convolution-deconvolution networks. Then, it was refined by iteratively enhancing the results in various sub-regions. Wang *et al.* [48] iteratively delivered the intermediate predictions back to the recurrent CNNs to refine saliency maps. In this way, salient objects can be gradually enhanced, while distractors can be suppressed.

Compared with image-based SOD, video-based SOD is less explored due to the lack of large video datasets. For example, Liu *et al.* [49] extended their image-based SOD model [1] to the spatiotemporal domain for salient object sequence detection. In [50], visual attention (*i.e.*, the estimated fixation density) was used as prior knowledge to guide the segmentation of salient regions in video. Rahtu *et al.* [18] proposed the integration of local contrast features in illumination, color and motion channels with a statistical framework. A conditional random field was then adopted to recover salient objects from images and video frames. Due to the lack of large-scale benchmarking datasets, most of these early approaches only provide qualitative comparisons, and only a few works such as [49] have provided quantitative comparisons on a small dataset within which salient objects are roughly annotated with rectangles.

To conduct quantitative comparisons in single-video-based SOD, Bin *et al.* [51] manually annotated the salient objects in 10 videos with approximately 100 frames per video. They also proposed an approach to detect temporally coherent salient objects using regional dynamic contrast features in the spatiotemporal domains of color, texture and motion. Their approach demonstrated impressive performance in processing videos with only one salient object. In [52], Papazoglou and Ferrari proposed an approach for the fast segmentation of foreground objects from background regions. They first estimated an initial foreground map with respect to the motion information, which was then refined by building the foreground/background appearance models and enhancing the spatiotemporal smoothness of foreground objects over the whole video. The main assumption required by their approach was that foreground objects should move differently from the surrounding background in a large fraction of the video. Wang *et al.* [53] proposed an unsupervised approach for video-based SOD. In their approach, frame-wise saliency maps were first generated and refined with respect to the geodesic distances between regions in the current frame and subsequent frames. After that, global appearance models and dynamic location models were constructed so that the spatially and temporally coherent salient objects could be segmented by using an energy minimization framework. In their later work [21], Wang *et al.* proposed the utilization of the inter-frame and intra-frame information in a gradient flow field. By extracting the local and global saliency measures, an energy function was then adopted to enhance the spatiotemporal consistency of the output saliency maps.

In addition to the impressive performance, these single-video-based approaches have provided us with an intuitive definition of salient objects: salient objects in a video should be spatiotemporally consistent and visually distinct from background regions. However, in real-world scenarios, assumptions such as color/texture dissimilarity and motion irregularity may not always hold. A more general definition of salient objects in a video is required to guide the annotation and detection processes.

Beyond single-video-based approaches, some approaches extend the idea of image co-segmentation to the video domain. For example, Chiu and Fritz [54] proposed a generative model for multi-class video co-segmentation. A global appearance model was learned to connect the segments from the same class to segment the foreground targets shared by different videos. Fu *et al.* [20] proposed to detect multiple foreground objects shared by a set of videos. Category-independent object proposals were first extracted, and a multi-state selection graph was then adopted to handle multiple foreground objects. Although video co-segmentation provides an interesting new direction for studying video-based SOD, detecting salient objects in a single video is still the most common requirement in many real-world applications, such as video compression [55], summarization [56], retrieval [57] and editing [58], as well as action recognition [59].

III. A Large-Scale Dataset for Video-Based SOD

A good dataset should cover many real-world scenarios and the annotation process should contain little subjective bias. In this section, we will introduce the details of the dataset construction process and discuss how salient objects can be unambiguously defined and annotated in videos.

A. Video Collection

We first collect hundreds of long videos from the Internet (*e.g.*, recommended videos from video-sharing websites such as YouTube) and volunteers. Note that no instruction is given on what types of videos are required since we aim to collect more "realistic" daily videos. After that, we randomly sample short clips from long videos and keep only the clips that contain objects in most frames. Finally, we obtain 200 indoor/outdoor videos that last 64 minutes in total (116, 103 frames at 30fps). These videos are grouped into two subsets according to the content complexity (determined through voting by the authors):

1) VOS-E: This subset contains 97 *easy* videos (27 minutes, 49, 206 frames, 83 to 962 frames per video). As shown in Fig. 1, a video in this subset usually contains obvious foreground objects with many slow camera motion types. This subset serves as a baseline to explore the inherent relationship between image- and video-based SOD.

2) VOS-N: This subset contains 103 *normal* videos (37 minutes, 66, 897 frames, 710 to 2, 249 frames per video). As shown in Fig. 1, videos in this subset contain complex or highly dynamic foreground objects, dynamic or cluttered

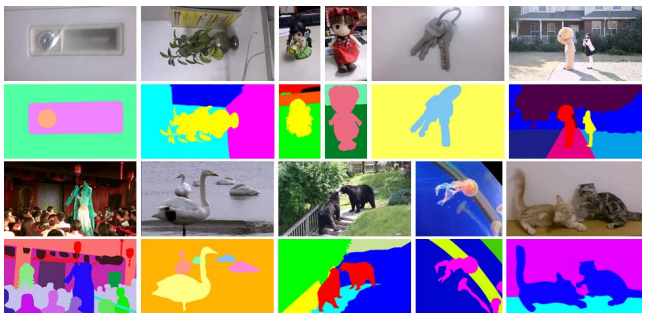


Fig. 2. Masks of objects and regions annotated by four subjects. Holes are filled up to speed up the annotation process (*e.g.*, the key ring in the first row), and multiple objects will be assigned the same labels throughout the video if they cannot be easily separated in certain frames (*e.g.*, the fighting bears and cats) or difficult to be re-identified (*e.g.*, the jelly fish which frequently appears and disappears near screen borders).

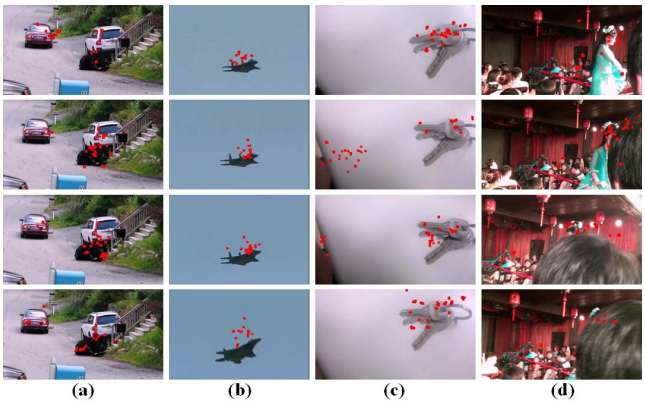


Fig. 3. Human fixations (red dots) of 23 subjects on consecutive keyframes. These fixations are insufficient to directly annotate salient objects frame by frame. (a) Insufficient fixations to separate salient objects and distractors; (b) Fixations fall outside small objects; (c) Fixations distracted by visual surprises; (d) Salient objects occluded by background regions.

background regions, etc. This subset is very challenging and can be used to benchmark models in realistic scenarios.

B. User Data Collection

The manual annotation of salient objects often generates ambiguity and subjective bias in complex scenes. Inspired by the solution used in [13], we collect two types of user data, namely, object masks and human fixations, to alleviate the ambiguity in defining and annotating salient objects.

1) Object Masks: Four subjects (2 males and 2 females, aged between 24 and 34) manually annotate the boundaries of all objects and regions in video frames. Since it consumes too much time to annotate all frames, we uniformly sample only one keyframe out of every 15 frames and manually annotate the $7,650$ keyframes. In the annotation, an object maintains the same label throughout a video, and the holes in objects are filled to speed up the annotation (*e.g.*, the background region inside the key ring in Fig. 2). Since moving objects may merge or split several times in a short period, it is difficult to consistently assign different labels to them (*e.g.*, the fighting bears and cats in Fig. 2), we assign the same label to objects if they become indistinguishable in certain frames (*e.g.*, the bears and cats in Fig. 2) or difficult to re-identify (*e.g.*, the jelly fish in Fig. 2 frequently appear and disappear near screen borders). Finally, regions smaller than 16 pixels are ignored and we obtain the accurate boundaries of $53,478$ objects and regions.

2) Human Fixations: Twenty-three subjects (16 males and 7 females, aged between 21 and 29) participate in the eye-tracking experiments. None of them participate in annotating the object/region masks. Each subject is asked to free-view all 200 videos on a 22-inch color monitor with a resolution of 1680×1050. A chin rest is utilized to reduce head movements and enforce a viewing distance of 75 cm. Considering that the non-stop watching of 200 videos (64 minutes) would be very tiring, we randomly divide videos into subgroups and adopt an interlaced schedule for different subjects who free-view the same subgroup of videos. In this manner, each subject has sufficient time to rest after watching a small collection of videos, making the eye-tracking data more reliable. During the free-viewing process, an eye-tracking apparatus with a sample rate of 500 Hz (SMI RED 500) is used to record various types of eye movements. Finally, we keep only the *fixations* and denote the set of eye positions on a video $\mathcal{V}$ as $\mathbb{F}_{\mathcal{V}}$, in which a sampled eye position f is represented by a triplet (x_f, y_f, t_f). Note that x_f and y_f are the coordinates of f and t_f is the time stamp at the start of f (an eye position sampled by the 500 HZ eye-tracker lasts approximately two milliseconds; see Fig. 3 for some examples).

C. Definition and Annotation of Salient Objects in Video

In a simple scene, salient objects can be manually annotated without much ambiguity. However, in a complex scene, there may exist several candidate objects, and different subjects may have different biases in determining which objects are salient. To alleviate the subjective bias, the fixations of multiple subjects can be used to find salient objects. For example, Li *et al.* [13] collected fixations from eight subjects who free-viewed the same image for 2 seconds. Then, salient objects were defined as the objects that received the highest number of fixations. This solution provides a less ambiguous definition of salient objects in images, but may fail on videos due to four reasons:

1) Insufficient Viewing Time: The viewing time of a frame (*e.g.*, 33 ms) is much shorter than that of an image. As a result, the fixations received by a frame are often insufficient to fully distinguish the most salient objects, especially when there exist multiple candidates in the same video frame (*e.g.*, the cars and bears in Fig. 3 (a)).

2) Inaccurate Fixations: Human fixations may fall outside moving objects and small objects (*e.g.*, the fast moving aircraft in Fig. 3 (b)).

3) Rapid Attention Shift: Human attention can be suddenly distracted by visual surprises for a short period. In this case, some distractor stimuli, which are not the subject of the video, will be recognized as salient if only the fixations in this short period are considered in defining salient objects (*e.g.*, the black region in Fig. 3 (c)).

4) Background-Only Frames: Some frames contain no obvious salient object. If salient objects are defined according to the fixations received in only these frames, background regions in these frames will be mistakenly annotated as salient (*e.g.*, the girl is occluded in Fig. 3 (d)).

For these reasons, it is difficult to directly define and annotate salient objects separately on each frame. Inspired by the idea of co-saliency [20], [54], we propose to define salient objects at the scale of whole videos. That is, salient objects in videos are defined as *the objects that consistently receive the highest fixation densities throughout a video*. The highest *density* of fixations is used in defining salient objects in videos, rather than the highest *number* of fixations. In this manner, we can avoid mistakenly assigning high saliency values to large background regions when salient objects are small (*e.g.*, the aircraft in Fig. 3 (b)).

D. Generation of Salient Object Masks

Based on the proposed definition, we can generate masks of salient objects for each video. We first compute the fixation density at each object in the manually annotated keyframes. Considering that the fixations received by each keyframe are very sparse, we consider the fixations that are recorded in a short period after the keyframe is displayed. Let $\mathcal{I}_t \in \mathcal{V}$ be a frame presented at time t and $\mathcal{O} \in \mathcal{I}_t$ be an annotated object. We measure the fixation density at $\mathcal{O}$, denoted as $S_0(\mathcal{O})$, as

$$S_0(\mathcal{O}) = \frac{1}{\|\mathcal{O}\|} \sum_{f \in \mathbb{F}_\mathcal{V}} \delta(t_f > t) \cdot \left(\sum_{p \in \mathcal{O}} \mathrm{Dist}(f, p) \cdot \exp\left(-\frac{(t_f - t)^2}{2\sigma_t^2} \right) \right), \quad (1)$$

where p is a pixel at (x_p, y_p) and $\|\mathcal{O}\|$ is the number of pixels in $\mathcal{O}$. The indicator function $\delta(t_f > t)$ equals 1 if $t_f > t$ holds and 0 otherwise. $\mathrm{Dist}(f, p)$ measures the spatial distance between the fixation f and the pixel p, which can be computed as

$$\mathrm{Dist}(f, p) = \exp\left(-\frac{(x_f - x_p)^2 + (y_f - y_p)^2}{2\sigma_s^2} \right). \quad (2)$$

From Eq. (1) and Eq. (2), we can see that the influence of the fixation f on the fixation density at the object $\mathcal{O}$ gradually decreases when the spatial or temporal distances between f and pixels in $\mathcal{O}$ increase. This influence is controlled by σ_s and σ_t which are empirically set to 3% of the video width (or video height if it is larger than the width) and 0.1 s, respectively.

Based on the fixation density $S_0(\mathcal{O})$, we can compute its saliency score $S(\mathcal{O})$ from the global perspective:

$$S(\mathcal{O}) = \frac{\sum_{\mathcal{I}_t \in \mathcal{V}} \delta(\mathcal{O} \in \mathcal{I}_t) \cdot S_0(\mathcal{O})}{\sum_{\mathcal{I}_t \in \mathcal{V}} \delta(\mathcal{O} \in \mathcal{I}_t)}. \quad (3)$$

In Eq. (3), the saliency of an object is defined as its average fixation density throughout the video. After that, we select the objects with saliency scores above an empirical threshold of 50 (or the object with the highest saliency score if it is smaller than 50). Note that such a threshold is empirically selected based on the subjectively assessed object completeness as well as the consistency between segmented salient objects and all recorded fixations. That is, we first overlay the recorded fixations onto all video frames to infer which objects attract the majority of subjects in the eye-tracking experiments. After that, we generate salient objects by enumerating a set of predefined thresholds (*e.g.*, 25, 50, 75, 100, 128, 150 and 200) or adaptive thresholds (*e.g.*, max, median, mean and twice the mean). We find that the fixed threshold of 50 provides the best subjective impression in segmenting the most attractive objects according to the fixations of 23 subjects. At this threshold, we obtain a set of salient objects for each video, represented by a sequence of binary masks at keyframes. In particular, a keyframe that contains only background or a salient object that splits into several disconnected parts due to the occlusion of background distractors will be discarded. Finally, we obtain 7,467 binary masks of keyframes (3,236 for the 97 videos in **VOS-E** and 4,231 for the 103 videos in **VOS-N**). Representative masks of salient objects can be found in Fig. 4.

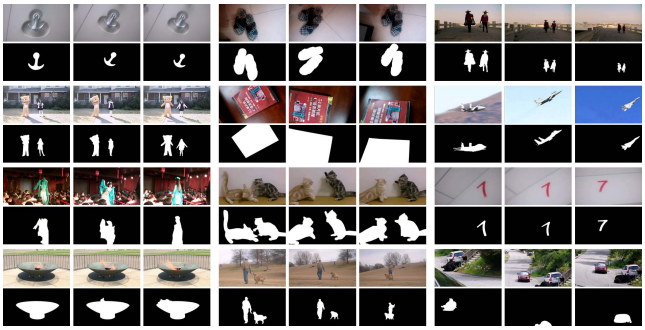


Fig. 4. Representative keyframes and masks of salient objects.

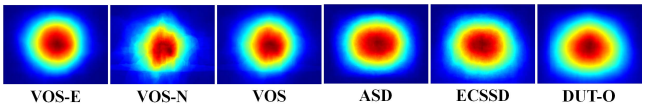


Fig. 5. The average annotation maps of 6 datasets.

E. Dataset Statistics

To reveal the main characteristics of **VOS**, we show in Fig. 5 the average annotation maps (AAMs) of **VOS-E**, **VOS-N**, **VOS** and three image datasets (*i.e.*, **ASD** [2], **ECSSD** [11] and **DUT-O** [10]). Similar to [8], the AAM of an image-based SOD dataset is computed by 1) resizing all ground-truth masks from the dataset to the same resolution, 2) summing the resized masks pixel by pixel, and 3) normalizing the resulting map to a maximum value of 1.0. For a video-based SOD dataset (*e.g.*, **VOS-E**, **VOS-N** and **VOS**), an AAM is first computed over each video, while the AAMs from all videos are fused following the same three steps to obtain the final AAM. In this manner, we can provide a better view of the distributions of salient objects in different videos (otherwise, the AAMs will be heavily influenced by long videos).

From Fig. 5, we can see that the distributions of salient objects in **VOS** and its two subsets are both center-biased,

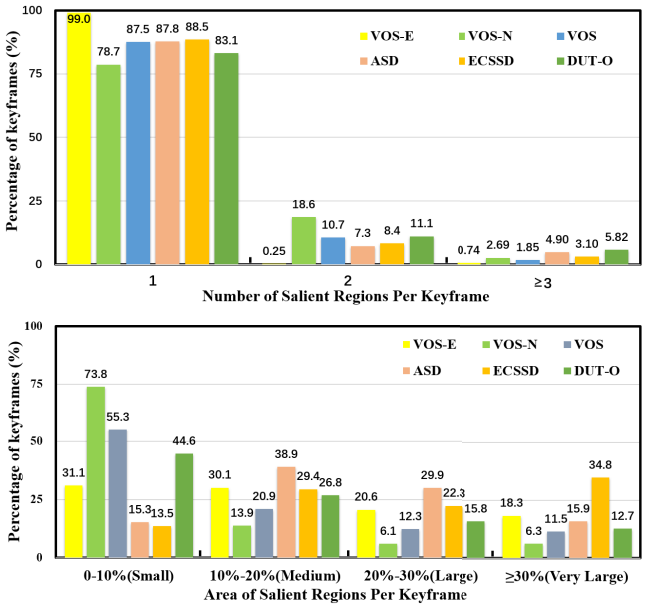


Fig. 6. Histograms of the number and area of salient objects.

and the degree of center bias is a little stronger than in **ASD**, **ECSSD** and **DUT-O**. This is mainly because photographers often have a strong tendency to place salient targets near the center of the view when taking videos. Moreover, in Eq. (2), a Gaussian blob is constructed at each fixation location (x_p, y_p). The saliency of a pixel is then computed by summing various blobs, which will add some center bias to the salient objects since center pixels are more likely to have more neighboring fixations than border pixels. This problem is also faced in many image-based SOD research works that make use of an eye-tracking apparatus. The existence of center bias may imply that image-based and video-based SODs are inherently related, and it is possible to transfer some useful saliency cues from the spatial domain to the spatiotemporal domain (*e.g.*, the background prior [3], [37] obtained from the boundary pixels).

Moreover, Figure 6 shows the histograms of the number and area of salient objects. We see that the number and area of salient objects in **VOS** are similar to those in **DUT-O**. This implies that **VOS**, like the **DUT-O** dataset, is very challenging, as it reflects many realistic scenarios. In particular, most keyframes from **VOS-E** contain only one salient object, similar to the famous dataset **SegTrack**. However, the sizes of salient objects in **VOS-E** are distributed almost uniformly among the Small (31.1%), Medium (30.1%), Large (20.6%) and Very Large (18.3%) categories, making **VOS-E** more challenging than **SegTrack**. Considering that the 97 videos in **VOS-E**, like those in **SegTrack**, contain varying numbers of distractors and cover many camera motion types, we believe **VOS-E** can serve as a good baseline dataset to benchmark video-based SOD models.

IV. A Baseline Model for Video-Based SOD with Saliency-Guided Stacked Autoencoders

A. The Framework

To construct a baseline model for **VOS**, we propose an unsupervised approach that learns saliency-guided stacked autoencoders. The framework of the proposed approach is shown in Fig. 7. We first transform each frame from **VOS** into several color spaces and extract object proposals as well as the motion information (*e.g.*, optical flow). After that, we extract three spatiotemporal saliency cues from each frame at the pixel, superpixel and object levels, and such cues reveal the presence of salient objects from different perspectives. Considering that salient objects are often spatially smooth and temporally consistent in consecutive frames, we characterize each pixel with a high-dimensional feature vector, which consists of the saliency cues collected from the pixel, its spatial neighbors and the corresponding pixel in the subsequent frame.

With the guidance of saliency cues in the high dimensional feature vector at each pixel, stacked autoencoders can be learned in an unsupervised manner, which contain only one hidden node in the last encoding layer (see Fig. 7). Since the saliency cues within a pixel and its spatiotemporal neighbors can be well reconstructed from the output of this layer, we can safely assume that the degree of saliency at each pixel is strongly related to the output score. By computing the output scores and the linear correlation coefficient with the input saliency cues, we can derive an initial saliency map for each frame that is spatially smooth and temporally consistent. Finally, several simple post-processing operations are applied to further enhance salient objects and suppress distractors.

B. Extracting Multi-Scale Saliency Cues

To extract saliency cues, we first resize a frame $\mathcal{I}_t$ to the maximum side length of 400 pixels and convert it to the Lab and HSV color spaces. After that, we estimate the optical flow [60] between $\mathcal{I}_t$ and $\mathcal{I}_{t+1}$ and compute the inter-frame flicker as the absolute in-place difference of intensity between $\mathcal{I}_t$ and $\mathcal{I}_{t-1}$. For the sake of simplification, we use a space XYT, which is formed by combining the optical flow and the flicker to represent the variations along the horizontal, vertical and temporal directions. Finally, each frame is represented by 12 feature channels from the RGB, Lab, HSV and XYT spaces. Based on these channels, we extract three types of saliency cues:

1) Pixel-Based Saliency: To efficiently extract the pixel-based saliency, we utilize the unsupervised algorithm proposed in [37] that computes the minimum barrier distance from a pixel to the image boundary (one pixel width). In the computation, we discard the Hue channel since the difference between hue values cannot always reflect the color contrast. Moreover, we also discard the RGB channels and the Value channel in HVS, which are redundant to the other channels. For the remaining four spatial and three temporal channels, the minimum barrier distances from all pixels to the image boundary are separately computed over each channel. Such distances are then summed across channels to initialize a pixel-based saliency map $\mathbf{S}_t^{pix}$. Moreover, we extract a backgroundness map as in [37] and multiply it with $\mathbf{S}_t^{pix}$ to further enhance salient regions and suppress probable background regions. Finally, we conduct a morphological smoothing step over the pixel-based saliency map to smooth $\mathbf{S}_t^{pix}$ while preserving the details of significant boundaries. As shown in Fig. 8 (c),

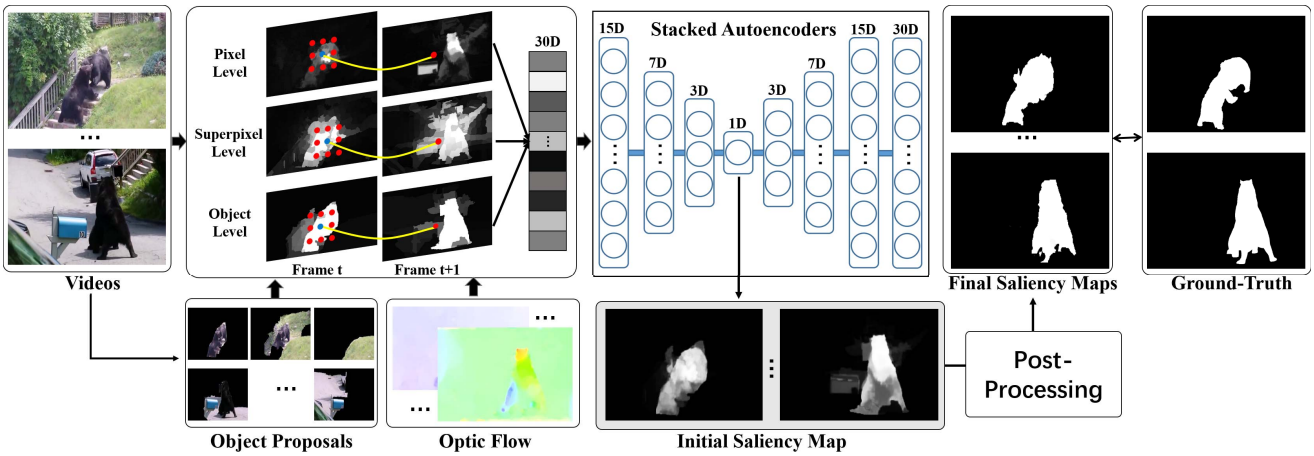


Fig. 7. The framework of the proposed saliency-guided stacked autoencoders.

the pixel-based saliency can be efficiently computed but is sensitive to noise.

2) Superpixel-Based Saliency: In image-based SOD, superpixels are often used as the basic units for feature extraction and saliency computation since they contain much more structural information than pixels. In this study, we adopt the approach proposed in [61] to extract the superpixel-based saliency in an unsupervised manner. This approach first divides a frame $\mathcal{I}_t$ into superpixels, based on which the sparse and low-rank properties are utilized to decompose the feature matrix of superpixels to obtain their saliency scores. In this process, prior knowledge on location (*i.e.*, center bias), color and background is used to refine the superpixel-based saliency. Finally, the saliency value of a superpixel is mapped back to all pixels it contains to generate a saliency map $\mathbf{S}_t^{sup}$. As shown in Fig. 8 (d), the superpixel-based saliency can be used to detect a large salient object as a whole (*e.g.*, the tissue in the third row of Fig. 8 (d)).

3) Object-Based Saliency: Inspired by the construction process of **VOS**, we adopt the Multiscale Combinatorial Grouping algorithm [62] to generate a set of object proposals for the frame $\mathcal{I}_t$ and estimate an objectness score for each proposal. After that, we adopt the unsupervised fixation prediction model proposed in [63] to generate three fixation density maps in the Lab, HSV and XYT spaces. Let $\mathbb{O}$ be the set of objects with the highest objectness scores and $\mathbf{F}_{lab}$, $\mathbf{F}_{hsv}$ and $\mathbf{F}_{xyt}$ be the three fixation density maps. The object-based saliency at a pixel p can be computed as

$$\mathbf{S}_t^{obj}(p) = \sum_{\mathcal{O}\in\mathbb{O}} \delta(p \in \mathcal{O}) \cdot \mathbf{F}_{lab}(\mathcal{O}) \cdot \mathbf{F}_{hsv}(\mathcal{O}) \cdot \mathbf{F}_{xyt}(\mathcal{O}), \quad (4)$$

where $\delta(p \in \mathcal{O})$ is an indicator function that equals 1 if $p \in \mathcal{O}$ and 0 otherwise. Let $\mathbb{O}$ be the set of objects used for computing the object-based saliency maps. We set $\|\mathbb{O}\| = 50$ in the experiments. $\mathbf{F}_{lab}(\mathcal{O})$ (or $\mathbf{F}_{hsv}(\mathcal{O}), \mathbf{F}_{xyt}(\mathcal{O})$) indicates the ratio of fixations received by $\mathcal{O}$ over the fixation density map $\mathbf{F}_{lab}$, which is computed as:

$$\mathbf{F}_{lab}(\mathcal{O}) = \frac{\sum_{p\in\mathcal{O}} \mathbf{F}_{lab}(p)}{\sum_{p\in\mathcal{I}_t} \mathbf{F}_{lab}(p)}. \quad (5)$$

As shown in Fig. 8 (e), the object-based saliency cues can be used to extract whole large salient objects, but the extracted regions often contain the background regions near the object.

C. Learning Stacked Autoencoders

Given the saliency cues, we have to estimate a non-negative saliency score for each pixel, which, statistically, has a positive correlation with the saliency cues. Moreover, as stated in many previous works [21], [52], [53], the estimated saliency scores should have the following attributes:

1) Spatial Smoothness: Similar pixels that are spatially adjacent to each other should have similar saliency scores.

2) Temporal Consistency: Corresponding pixels in adjacent frames should have similar saliency scores so that salient objects can be consistently detected throughout a video.

To develop a model with such attributes, we train stacked autoencoders that take saliency cues at a pixel and its spatiotemporal neighbors as the input, so that the spatial smoothness and temporal consistency of the predicted saliency scores can be guaranteed. Considering the computational efficiency, for each pixel, we adopt its eight spatial neighbors and only one temporal neighbor in the subsequent frame defined by the optical flow. Each pixel is then represented by a feature vector with $3 \times 10 = 30$ saliency cues.

With the guidance of the high-dimensional saliency cues, we collect the feature vectors from $N = 500,000$ randomly selected pixels in **VOS**, which are denoted as $\{\mathbf{x}_n^1\}_{n=1}^N$. With these data, we train stacked autoencoders with T encoding layers and the same number of decoding layers with logistic sigmoid transfer functions. In the training process, no ground-truth data are used. The tth encoding layer f_t, $t \in \{1, \ldots, T\}$, and its corresponding decoding layer $\hat{f}_t$ are trained by minimizing

$$\min_{f_t, \hat{f}_t} \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n^t - \hat{f}_t\left(f_t(\mathbf{x}_n^t)\right)\|_2^2 + \lambda_w \Omega_w + \lambda_s \Omega_s, \quad (6)$$

where Ω_w is an ℓ-2 regularization term that can be used to penalize the ℓ-2 norm of the weights in the encoding and decoding layers (we empirically set $\lambda_w = 0.001$ in this study).

Ω_s is a sparsity regularizer, which is defined as the Kullback-Leibler divergence between the average output of each neuron in f_t and a predefined score ρ (we empirically set $\rho = 0.05$ and $\lambda_s = 1.0$).

In minimizing Eq. (6), the first encoding layer takes the sampled feature vectors of saliency cues as the input data, while other encoding layers take the output of previous encoding layers as the input. That is, in training the tth encoding/decoding layer, we have

$$\mathbf{x}_n^t = \mathcal{N}\left(f_{t-1}(\mathbf{x}_n^{t-1})\right), \ \forall\, t \in \{2, \ldots, T\}, \tag{7}$$

where $\mathcal{N}(\cdot)$ denotes the normalization operation that forces each dimension of the input data that enters an encoding layer to fall in the same dynamic range of $[-1, 1]$. In this study, we use $T = 4$ encoding layers with $15, 7, 3$ and 1 neurons in each layer, and each layer is trained for 100 epochs. Note that the Tth layer contains only one neuron, and by using its output score the input saliency cues within a pixel and its spatiotemporal neighbors can be well reconstructed by the decoding layers. As a result, we can safely assume that such output scores $\{\mathbf{x}_n^{T+1}\}_{n=1}^N$ are closely related to the input saliency cues $\{\mathbf{x}_n^1\}_{n=1}^N$, and the degree of correlation c can be measured by averaging the linear correlation coefficients between $\{\mathbf{x}_n^{T+1}\}_{n=1}^N$ and every dimension of $\{\mathbf{x}_n^1\}_{n=1}^N$. As a result, the saliency score of a pixel p, given its feature vector $\mathbf{v}_p$, which contains the saliency cues from p and its spatiotemporal neighbors, can be computed as

$$\mathbf{S}(p) = \text{sign}(c) \cdot f_T(\mathcal{N}(\cdots f_1(\mathcal{N}(\mathbf{v}_p)))). \tag{8}$$

After computing the saliency score for each pixel with Eq. (8), we can initialize a saliency map for each frame in **VOS** with the saliency values normalized to $[0, 255]$. As shown in Fig. 8 (f), such a saliency map already performs impressively in highlighting salient objects and suppressing distractors. To further enhance salient objects and suppress distractors, we perform three post-processing operations:

1) Apply temporal smoothing between adjacent frames to reduce the inter-frame flicker. We adopt a Gaussian filter with a width of 3 and $\sigma = 0.75$.
2) Enhance the foreground/background contrast by using the sigmoid function proposed in [37].
3) Binarize the saliency map with the average value of the whole saliency map and suppress the connected components that are extremely small.

As shown in Fig. 8 (g), these post-processing operations can generate compact and precise salient objects. Note that operations such as center-biased re-weighting and spatial smoothing are not adopted here because the autoencoders, which have been learned in an unsupervised manner over a large-scale dataset, already have the capability to accurately detect various types of salient objects, regardless of their positions and sizes.

V. Experiments

In this section, we compare the proposed Saliency-guided Stacked Autoencoders (**SSA**) with the state-of-the-art models on **VOS**. The main objectives are two-fold: 1) validate the effectiveness of the dataset **VOS** and the baseline model **SSA**, and 2) provide a comprehensive benchmark to reveal the key challenges in video-based SOD. This section will first introduce the experimental settings and then discuss the results.

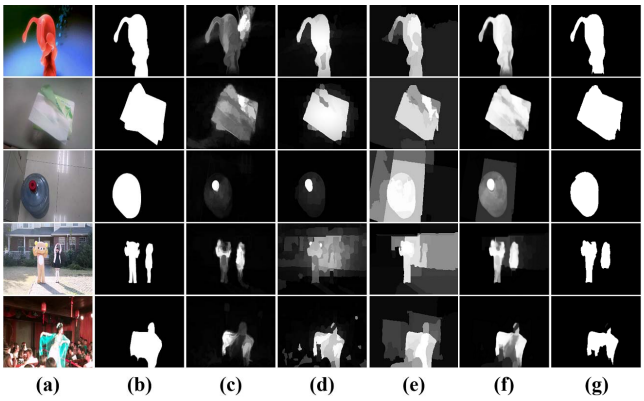


Fig. 8. Saliency cues and the estimated saliency maps. (a) Frames, (b) ground-truth, (c) pixel-based saliency, (d) superpixel-based saliency, (e) object-based saliency, (f) initial saliency maps obtained by the saliency-guided stacked autoencoders, (g) final saliency maps obtained after post-processing.

TABLE II
Models for Benchmarking (Symbols: [I] for Image-Based, [V] for Video-Based; [C] for Classic Unsupervised or Non-Deep Learning, [D] for Deep Learning, [U] for Unsupervised)

Model	Pub. & Year &	Type	Model	Pub. & Year &	Type
SIV [18]	ECCV 2010	[V+U]	**CB** [64]	BMVC 2011	[I+C]
RC [33]	CVPR 2011	[I+C]	**ULR** [34]	CVPR 2012	[I+C]
LMLC [65]	TIP 2013	[I+C]	**DRFI** [3]	CVPR 2013	[I+C]
GMR [10]	CVPR 2013	[I+C]	**HS** [11]	CVPR 2013	[I+C]
PCA [66]	CVPR 2013	[I+C]	**CHM** [67]	ICCV 2013	[I+C]
DSR [68]	ICCV 2013	[I+C]	**MC** [35]	ICCV 2013	[I+C]
FST [52]	ICCV 2013	[V+U]	**HDCT** [69]	CVPR 2014	[I+C]
RBD [36]	CVPR 2014	[I+C]	**NLC** [70]	BMVC 2014	[V+U]
BL [4]	CVPR 2015	[I+C]	**BSCA** [71]	CVPR 2015	[I+C]
LEGS [39]	CVPR 2015	[I+D]	**MCDL** [6]	CVPR 2015	[I+D]
MDF [25]	CVPR 2015	[I+D]	**SAG** [53]	CVPR 2015	[V+U]
GP [72]	ICCV 2015	[I+C]	**MB** [37]	ICCV 2015	[I+C]
MB+ [37]	ICCV 2015	[I+C]	**GF** [21]	TIP 2015	[V+U]
ELD [44]	CVPR 2016	[I+D]	**DCL** [40]	CVPR 2016	[I+D]
RFCN [48]	ECCV 2016	[I+D]	**DHSNet** [46]	CVPR 2016	[I+D]
SMD [61]	PAMI 2017	[I+C]	**SSA**	Our approach	[V+U]

A. Settings

As shown in Table II, 32 state-of-the-art models, including the proposed baseline model **SSA**, are tested on the **VOS** dataset (19 image-based classic unsupervised or non-deep learning models, seven image-based deep learning models, and six video-based unsupervised models). Similar to many image-based SOD works, we also adopt Recall, Precision, F_β and Mean Absolute Error (MAE) as the evaluation metrics. Let G be the ground-truth binary mask of a keyframe and S be the saliency map predicted by a model. The MAE score can be computed as the average absolute difference between all pixels in S and G to directly reflect the visual difference [8], [44]. Moreover, the Recall and Precision scores can be computed by

TABLE III

PERFORMANCE BENCHMARKING OF OUR APPROACH AND 31 STATE-OF-THE-ART MODELS ON **VOS** AND ITS TWO SUBSETS **VOS-E** AND **VOS-N**. TOP THREE SCORES IN EACH COLUMN ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY. SYMBOLS OF MODEL CATEGORIES: [I+C] FOR IMAGE-BASED CLASSIC UNSUPERVISED OR NON-DEEP LEARNING, [I+D] FOR IMAGE-BASED DEEP LEARNING, [V+U] FOR VIDEO-BASED UNSUPERVISED

	Models	VOS-E				VOS-N				VOS			
		MAP	MAR	F_β	MAE	MAP	MAR	F_β	MAE	MAP	MAR	F_β	MAE
[I+C]	**CB** [64]	0.755	0.791	0.763	0.145	0.463	0.563	0.483	0.229	0.605	0.674	0.619	0.188
	RC [33]	0.738	0.677	0.723	0.171	0.465	0.561	0.484	0.221	0.597	0.617	0.602	0.197
	ULR [34]	0.693	0.737	0.703	0.158	0.390	0.675	0.432	0.168	0.537	0.705	0.568	0.163
	LMLC [65]	0.687	0.736	0.697	0.154	0.408	0.501	0.426	0.262	0.543	0.615	0.558	0.210
	GMR [10]	0.813	0.697	0.783	0.140	0.500	0.611	0.522	0.195	0.652	0.653	0.652	0.168
	HS [11]	0.755	0.615	0.717	0.141	0.497	0.521	0.502	0.262	0.622	0.567	0.608	0.203
	CHM [67]	0.756	0.765	0.758	0.124	0.409	0.611	0.443	0.186	0.578	0.685	0.599	0.156
	DRFI [3]	0.762	0.837	0.778	0.114	0.442	0.733	0.486	0.150	0.597	0.783	0.632	0.132
	PCA [66]	0.750	0.725	0.744	0.143	0.420	0.696	0.462	0.142	0.580	0.710	0.606	0.143
	DSR [68]	0.765	0.748	0.761	0.112	0.450	0.679	0.488	0.140	0.603	0.713	0.625	0.127
	MC [35]	0.819	0.737	0.799	0.140	0.499	0.665	0.530	0.192	0.655	0.700	0.664	0.167
	HDCT [69]	0.711	0.791	0.728	0.128	0.420	0.677	0.460	0.142	0.561	0.733	0.593	0.136
	RBD [36]	0.799	0.782	0.795	0.091	0.516	0.709	0.550	0.145	0.653	0.745	0.672	0.119
	GP [72]	0.743	0.788	0.753	0.141	0.405	0.704	0.449	0.227	0.569	0.745	0.602	0.185
	MB [37]	0.814	0.735	0.794	0.107	0.480	0.696	0.517	0.151	0.642	0.715	0.657	0.129
	MB+ [37]	0.803	0.792	0.801	0.096	0.492	0.754	0.535	0.162	0.643	0.772	0.669	0.130
	BL [4]	0.765	0.777	0.768	0.165	0.477	0.658	0.509	0.220	0.617	0.716	0.637	0.194
	BSCA [71]	0.766	0.758	0.764	0.133	0.457	0.663	0.493	0.195	0.607	0.709	0.628	0.165
	SMD [61]	0.811	0.789	0.806	0.096	0.528	0.688	0.558	0.148	0.665	0.737	0.681	0.123
[I+D]	**LEGS** [39]	0.820	0.685	0.784	0.193	0.556	0.593	0.564	0.215	0.684	0.638	0.673	0.204
	MCDL [6]	0.831	0.787	0.821	0.081	0.570	0.645	0.586	0.085	0.697	0.714	0.701	0.083
	MDF [25]	0.740	0.848	0.762	0.100	0.527	0.742	0.565	0.098	0.630	0.793	0.661	0.099
	ELD [44]	0.790	0.884	0.810	0.060	0.569	0.838	0.615	0.081	0.676	0.861	0.712	0.071
	DCL [40]	0.864	0.735	0.830	0.084	0.583	0.809	0.624	0.079	0.719	0.773	0.731	0.081
	RFCN [48]	0.834	0.820	0.831	0.075	0.614	0.783	0.646	0.080	0.721	0.801	0.738	0.078
	DHSNet [46]	0.863	0.905	0.872	0.049	0.649	0.851	0.686	0.055	0.753	0.877	0.778	0.052
[V+U]	**SIV** [18]	0.693	0.543	0.651	0.204	0.451	0.523	0.466	0.201	0.568	0.533	0.560	0.203
	FST [52]	0.781	0.903	0.806	0.076	0.619	0.691	0.634	0.117	0.697	0.794	0.718	0.097
	NLC* [70]	0.439	0.421	0.435	0.204	0.561	0.610	0.572	0.123	0.502	0.518	0.505	0.162
	SAG [53]	0.709	0.814	0.731	0.129	0.354	0.742	0.402	0.150	0.526	0.777	0.568	0.140
	GF [21]	0.712	0.798	0.730	0.153	0.346	0.738	0.394	0.331	0.523	0.767	0.565	0.244
	SSA	0.875	0.776	0.850	0.062	0.660	0.682	0.665	0.103	0.764	0.728	0.755	0.083

* The executable of NLC only output valid results on 187 videos (91 from **VOS-E** and 96 from **VOS-N**).

converting S into a binary mask M and comparing it with G:

$$\text{Recall} = \frac{\#(\text{Non-zeros in } M \cap G)}{\#(\text{Non-zeros in } G)},$$
$$\text{Precision} = \frac{\#(\text{Non-zeros in } M \cap G)}{\#(\text{Non-zeros in } M)}. \quad (9)$$

Intuitively, the overall performance of a model on **VOS** can be assessed by directly computing the average Recall and Precision over all keyframes. However, this solution will over-emphasize the performance on long videos and ignore the performance on short videos (*e.g.*, a video with 100 keyframes will overwhelm a video with only 10 keyframes). To avoid that, we first compute the average Recall, Precision and MAE separately over each video. Then, the mean values of the average Recall, Precision and MAE are computed over all videos. In this manner, the Mean Average Recall (MAR), Mean Average Precision (MAP) and MAE can well reflect the performance of a model by equivalently considering its performances over all videos. Correspondingly, F_β is computed by fusing MAR and MAP to quantify the overall performance:

$$F_\beta = \frac{(1+\beta^2) \cdot \text{MAP} \cdot \text{MAR}}{\beta^2 \cdot \text{MAP} + \text{MAR}}. \quad (10)$$

Here, we set $\beta^2 = 0.3$, which is the value used in most existing image-based models [2], [8] for performance evaluation.

Another problem in assessing models with MAP, MAR and F_β is determining how to turn a gray-scale saliency map S into a binary mask M. Similar to image-based SOD, we adopt the adaptive threshold proposed in [2], which is computed as twice the average value of S, to generate a binary mask from each saliency map. Note that we set this threshold to the maximal saliency value if it exceeds the maximal value. In this manner, unique MAR, MAP and F_β scores can be generated to measure the overall performance of a model.

B. Model Benchmarking

The performances of the baseline model **SSA** and the other state-of-the-art models on **VOS-E**, **VOS-N** and **VOS** are illustrated in Table III. Some representative results of the best models from the three model categories are shown in Fig. 9, including **SMD** (image-based classic unsupervised or non-deep learning), **DHSNet** (image-based deep learning) and **SSA/FST** (video-based unsupervised). Based on Table III and Fig. 9, we conduct three comparisons:

*1) Comparisons Between **SSA** and the Other Models:* From Table III, we can see that **SSA** outperforms

Fig. 9. Representative results of the best models from the three model categories. The four models are **SMD** (image-based classic unsupervised or non-deep learning), **DHSNet** (image-based deep learning) and **SSA/FST** (video-based unsupervised).

30 state-of-the-art models in terms of F_β, including six image-based deep learning models (except **DHSNet**) and five video-based models. Note that no ground-truth data in any form has been used in **SSA**, while the other deep models often make use of VGGNet [73] pre-trained on a massive number of images with semantic tags, and fine-tune their SOD models on thousands of images with manually annotated salient objects (*e.g.*, **DHSNet** starts with VGGNet and then takes 9500 images from two datasets for model fine-tuning). Even in such a challenging setting, the unsupervised shallow model **SSA**, which only utilizes four layers of stacked autoencoders, still outperforms all deep models in terms of MAP, and outperforms the other six deep learning models (**LEGS**, **MCDL**, **MDF**, **ELD**, **DCL** and **RFCN**) in terms of F_β score. This result validates the effectiveness of the saliency-guided autoencoding scheme in video-based SOD.

In addition, on **VOS** and its two subsets, **SSA** always has the best Precision (MAP $= 0.764$ on **VOS**), while its MAR scores are even lower than those of some unsupervised image-based models such as **MB+** and **RBD**. This may be because such models adopt bottom-up frameworks that tend to detect almost all regions that are different from the predefined context (*i.e.*, image boundaries in **MB+** and **RBD**), leading to high Recall rates. However, the suppression of distractors is given less

emphasis in such frameworks, making their Precision much lower than that obtained with **SSA**. In the SOD task, it is widely recognized that high Precision is much more difficult to obtain than high Recall [29], [38], and a frequently used trade-off is to gain a remarkable increase in Precision at the cost of slightly decreasing Recall. That is why the computation of F_β in this work and in almost all the image-based models places more emphasis on Precision than Recall. Although a higher Recall usually leads to a better subjective impression in qualitative comparisons, the overall performance, especially in terms of F_β score, may be not very satisfactory due to the emphasis of Precision in computing F_β. This result also poses a challenge for the proposed **VOS** dataset: how can the Recall rate be further improved while maintaining the high Precision?

2) Comparisons Between (Non-Deep) Image-Based and Video-Based Models: Beyond analyzing the best models, another issue that is worth discussing is the performance of image-based and video-based models, especially the non-deep models. Interestingly, video-based models such as **GF** and **SAG** may sometimes perform even worse than image-based models (*e.g.*, **SMD**, **RBD** and **MB**+). This may be due to two reasons. First, the impact of incorporating temporal information into the visual saliency computation is not always positive. In some videos, the salient objects, as defined by many video-based models, have specific motion patterns that are remarkably different from those of the distractors (*e.g.*, the dancing bear & girl in the second row of Fig. 4). However, this may not always be the case when processing the "realistic" videos from **VOS**. For example, in some videos with global camera motion and static salient objects/distractors (*e.g.*, the shoes and book in the second column of Fig. 4), the temporal information acts as a kind of noise and often leads to unsatisfactory results. Second, the parameters of most video-based models are manually fine-tuned on small datasets and may become "over-fitting" to specific scenarios. Given a new scenario contained in **VOS**, these parameters may lead to unsatisfactory results, either by emphasizing the wrong feature channels or by propagating the wrong results from some frames to the entire video in an energy-based optimization framework.

3) Comparisons Between Image-Based Deep And Non-Deep Models: From Table III, we also find that image-based deep models often perform remarkably better than image-based models with classic unsupervised or non-deep learning frameworks. This may be because deep models can become very complex to make use of massive training data. Taking the seven deep models compared in Table III as examples, we can create a ranked list in decreasing order of F_β on **VOS**. The ranked list, as well as the corresponding training data, is given as follows: 1) **DHSNet**: 9500 from **MSRA10K** and **DUT-O**, 2) **RFCN**: 10000 from **MSRA10K**, 3) **DCL**: 2500 from **MSRA-B**, 4) **ELD**: 9000 from **MSRA10K**, 5) **MCDL**: 8000 from **MSRA10K**, 6) **LEGS**: 3340 from **MSRA-B** and **PASCAL-S**, and 7) **MDF**: 2500 from **MSRA-B**. Note that the scenarios in **DUT-O** and **PASCAL-S** are much more challenging than those from **MSRA-B** and **MSRA10K** (many images of **MSRA-B** are also contained in **MSRA10K**). From this ranked list, we can conclude that, except for an outlier (**DCL**), the more training data and training sources, the better the performance of a deep model. This finding is quite interesting and may help explain the success of some top-ranked deep models such as **DHSNet** and **RFCN**. Moreover, the top-ranked models often adopt a recurrent mechanism in detecting salient objects. Such mechanisms can help iteratively discover salient objects and suppress probable distractors. For video-based SOD, the success of such deep models shows a feasible way to develop better spatiotemporal models by using image-based training data as well as the recurrent architecture. Furthermore, it is necessary to develop an unsupervised baseline model that utilizes no training data in any form to provide fair comparisons for the other unsupervised and supervised models. Therefore, we propose **SSA**, an unsupervised model with the potential of being widely used as the baseline model on **VOS**.

C. Performance Analysis of *SSA*

Beyond model benchmarking, we also conducted several experiments to analyze the performance of **SSA**, including scalability and speed tests, influences of various components and the temporal window size, and the failure cases.

1) Scalability Test: One concern about **SSA** may be its scalability to other datasets. To examine this, we reuse the stacked autoencoders generated on **VOS** on a new dataset **ViSal** [21]. On **ViSal**, the performances of **SSA** and the other nine models (*i.e.*, the top three models on **VOS** from each model category) are reported in Table IV. We find that the overall performance of **SSA**, although not fine-tuned on **ViSal**, still ranks second place on this dataset (it is only outperformed by the deep model **DHSNet**). In particular, its MAE score is ranked higher than on **VOS**, which may be because **VOS** is a large dataset that covers a variety of scenarios (*e.g.*, **VOS-N** contains many outdoor scenarios involving animals and airplanes that are also present in **ViSal**). Moreover, the unsupervised architecture often has better performance in scalability tests and can be generalized to new scenarios. This can be further demonstrated by the model **FST**, which ranks third place in terms of F_β on **ViSal** (which is higher than its rank on **VOS**). To sum up, **VOS** contains a large number of real-world scenarios, which may help reduce the over-fitting risk. Moreover, the unsupervised framework of **SSA** makes it a scalable model that can be generalized to other scenarios without a remarkable drop in performance.

2) Influence of Various Components: **SSA** involves three types of saliency, and we aim to explore which types contribute most to the performance of **SSA**. Toward this end, we conduct an experiment to examine the performance of **SSA** on **VOS** when some types of saliency are ignored. For fair comparisons, we adopt the same architecture of stacked autoencoders, but set some saliency cues to zero when training and testing **SSA**. As shown in Table V, the pixel-based saliency has the best Precision, while the object-based saliency has the best Recall. Integrating all three types of saliency leads to the best overall performance. An interesting phenomenon is that in the superpixel-only setting, **SSA** outperforms **SMD** in both Recall and Precision, while **SMD** is exactly the model used in

TABLE IV

PERFORMANCE SCORES OF OUR APPROACH AND THE OTHER 9 MODELS ON **ViSal**. THE TOP 3 SCORES IN EACH COLUMN ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY

	Models	MAP	MAR	F_β	MAE
[I+C]	**MB+** [37]	0.551	0.887	0.604	0.145
	RBD [36]	0.529	0.787	0.572	0.129
	SMD [61]	0.583	0.886	0.633	0.133
[I+D]	**DCL** [40]	0.718	0.859	0.747	0.261
	RFCN [48]	0.781	0.897	0.805	0.050
	DHSNet [46]	0.816	0.955	0.845	0.027
[V+U]	**GF** [21]	0.556	0.850	0.604	0.108
	SAG [53]	0.538	0.858	0.589	0.104
	FST [52]	0.803	0.815	0.806	0.052
	SSA	0.787	0.884	0.808	0.046

TABLE V

PERFORMANCE OF **SSA** ON **VOS** WHEN DIFFERENT TYPES OF SALIENCY CUES ARE USED

Saliency Cues	MAP	MAR	F_β	MAE
Pixel-only	0.705	0.649	0.691	0.105
Superpixel-only	0.691	0.744	0.703	0.103
Object-only	0.647	**0.811**	0.678	0.155
Pixel + Superpixel	0.677	0.801	0.702	0.100
Pixel + Object	**0.788**	0.564	0.722	0.100
Superpixel + Object	0.720	0.774	0.732	0.091
All	0.764	0.728	**0.755**	**0.083**

computing the superpixel-based saliency. This may be mainly because temporal cues from adjacent frames are incorporated into the auto-encoding processes, which provides an opportunity to refine the results of **SMD** from a temporal perspective. Due to the existence of the temporal dimension in defining and annotating salient video objects, video-based SOD datasets contain something that cannot be obtained from image-based SOD datasets. For example, in the "fighting bears" scenario illustrated in the first two rows of the right column of Fig. 9, the mailbox and cars are considered to be non-salient from the perspective of the entire video, even though in some specific frames they do capture more human fixations than the fighting bears. In other words, the **VOS** dataset provides a new way to explore the influence of spatiotemporal cues (*e.g.*, optical flow and features propagated from adjacent frames) in defining, annotating and detecting salient objects, while in most image-based SOD datasets only spatial cues are involved. We believe the spatiotemporal definition of salient objects in **VOS** may help methods in future works distinguish salient and non-salient objects in the same way that human beings do.

3) Influence of Temporal Window Size: In **SSA**, only one subsequent frame is considered when processing a frame. To evaluate this, we conduct an experiment that gradually incorporates zero or more subsequent frames and examine the F_β variation of **SSA** on **VOS**. In this experiment, we consider the next W frames, where $W = 0, 1, 2, 4, 8$ or 15. As shown in Fig. 10, by considering only the subsequent frame, the F_β score increase from 0.735 (W=0) to 0.755 (W=1). This result implies that the temporal cues can facilitate the detection of salient objects in a frame, even though consecutive frames are highly related. With the incorporation of additional frames

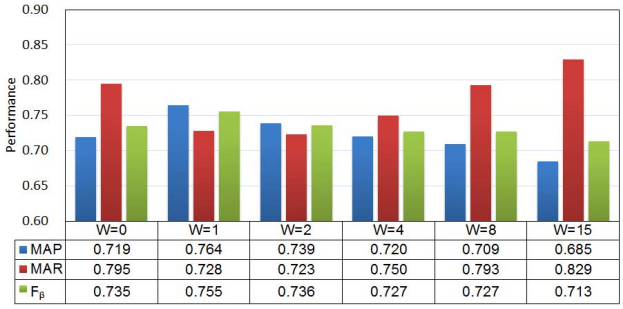


Fig. 10. Performance of **SSA** on **VOS** when temporal windows with different sizes are taken into consideration.

($W = 2, 4, 8$ or 15), the F_β score gradually decreases. This may be because the temporal correspondence between consecutive frames is the most reliable, and the reliability gradually decreases when the temporal gap between two frames increases. Although using a longer temporal window can bring us more cues in detecting salient objects and lead to higher Recall; such long-term temporal correspondences, which are not very reliable, may decrease the Precision. As a result, the F_β score, which places greater emphasis on the Precision, decreases with a longer temporal window. Such an experiment, together with the scalability test, can empirically prove that the over-fitting risk of **SSA** is not very high, even though only one subsequent frame is used as the temporal context of the current frame.

4) Speed Test: The **SSA** model consists of many feature extraction steps, and their speed analysis will help determine how to further enhance the efficiency. Toward this end, we calculate the time costs of various key steps of **SSA** in processing the first video in **VOS**, and compare them with those of the other five video-based SOD models. Note that the video has an original resolution of 800×448, and we down-sample it to 400×224 for the fair comparison of various models in the speed test. All models are tested on a CPU platform (single core, 3.4 GHz) with 128GB memory. As shown in Table VI, the speed of **SSA** is comparable to those of many previous algorithms, such as **SIV** and **NLC**. By investigating the time cost of each component of **SSA**, we find that approximately 58.8% of the computational resource is consumed in extracting the object proposal, and approximately 22.1% is spent on generating the optical flow. As a result, a probable way to speed up **SSA** is to replace these two components with faster models for object proposal generation and optical flow computation. In addition, a parallel processing mechanism can be explored as well, especially in extracting and encoding frame-wise saliency cues.

5) Failure Cases: Although **SSA** outperforms many state-of-the-art models, we can see that its F_β score is still far from perfect, which is mainly due to the low Recall rate. On **VOS-E**, which contains only simple videos with nearly static salient objects and distractors as well as slow camera motion, **SSA** only reaches an F_β score of 0.850 (still far from perfect), while the performance score drops sharply to 0.665 on **VOS-N**. This implies that the videos from the real-world scenarios are much more challenging than the videos taken in

TABLE VI

SPEED TEST OF **SSA**, ALL ITS COMPONENTS AND THE OTHER 5 VIDEO-BASED SOD MODELS. ALL TESTS ARE TESTED ON THE FIRST VIDEO OF **VOS** WITH 617 FRAMES, WHICH IS DOWN-SAMPLED TO THE RESOLUTION OF 400×224 FOR FAIR COMPARISONS OF ALL MODELS

Models or Key Steps	Average Time (s/frame)
SIV [18]	10.5
FST [52]	5.80
NLC [70]	19.0
SAG [53]	5.37
GF [21]	4.67
Optical Flow	1.84
Object proposal	4.89
Pixel-based Saliency	0.06
Superpixel-based Saliency	0.83
Object-based Saliency	0.38
Auto-encoding & Post-Proc.	0.31
SSA	8.31

the laboratory environment. This is also the main barrier to the usage of existing SOD models in other applications.

To validate this point, we illustrate in Fig. 11 two representative scenarios in which **SSA** fails, which provide two key challenges in video-based SOD. First, salient objects in a keyframe should be defined and detected by considering the entire video, other than the keyframe itself. For example, in some early frames of Fig. 11, it is difficult to determine whether the pen or the notebook is the most salient object. Although the pen is correctly detected in some later frames, it is difficult to transfer such correct results to frames that are separated by a large temporal gap. This indicates that the local spatiotemporal correspondences between pixels used by **SSA** are still insufficient to handle more challenging scenarios, and a salient object should be detected by computing the saliency from the global perspective as well.

Nevertheless, the failure cases in Fig. 11 not only suggest what should be considered in developing new video-based models but also validate the effectiveness of the **VOS** dataset. The indoor/outdoor scenarios from **VOS** are mainly taken by non-professional photographers, which are quite different from those in existing image datasets. For example, the moving crab in Fig. 11 consistently receives the highest density of fixations and becomes the most salient object in the video, even though it is very small. The existence of such scenarios in **VOS** increases the difficulties in transferring the knowledge obtained from existing image datasets (*e.g.*, the deep model **DHSNet**, learned from 9500 images) to the spatiotemporal domain, making video-based SOD on **VOS** an extremely challenging task. With such challenging cases, it is believed that **VOS** can facilitate the development of new models by benchmarking their performances in processing real-world videos.

D. Discussion

From all the results presented above, we draw three major conclusions: First, video-based SOD is much more challenging than image-based SOD. Even the state-of-the-art image-based models perform far from perfectly without fully utilizing the temporal information from both local and global perspectives. Second, there exists an inherent relationship between image-based and video-based SOD, and the **VOS-E** subset serves as a good baseline to help extend existing image-based models to the spatiotemporal domain. Third, real-world scenarios are still very challenging for existing models. In user-generated videos, salient objects may be very small, fast moving, with poor lighting conditions and cluttered dynamic background, etc. By handling such challenging scenarios in **VOS-N**, a model can improve its capability to process real-world scenarios. In particular, fixation prediction models often have impressive performances in detecting the most salient locations even in very complex real-world scenarios [74], [75]. Therefore, developing a better fixation prediction model may be very helpful in handling the **VOS-N** dataset, in which salient objects are annotated with respect to human fixations.

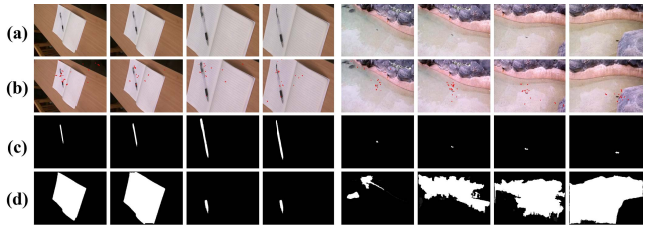


Fig. 11. Failure cases. (a) Frames, (b) the fixations received in 30 ms after a keyframe is displayed, (c) binary masks of salient objects and (d) the estimated saliency maps of **SSA**.

VI. CONCLUSION

Salient object detection is a hot topic in the area of computer vision. In the past five years, hundreds of innovative models have been proposed for detecting salient objects in images, which have gradually evolved from bottom-up models to deep models due to the availability of large-scale image datasets. However, the problem of video-based SOD has not been sufficiently explored due to the lack of large-scale video datasets. The most challenging step in constructing such a dataset is providing a reasonable and unambiguous definition of salient objects from the spatiotemporal perspective.

In this paper, we propose **VOS**, which is a large-scale dataset with 200 videos. Different from existing datasets, salient objects in **VOS** are defined by combining human fixations and manually annotated objects throughout a video. As a result, the definition and annotation of salient objects in videos become less ambiguous. Moreover, we propose saliency-guided stacked autoencoders for video-based SOD, which are compared with massive state-of-the-art models on **VOS** to demonstrate the challenges of video-based SOD as well as its differences from and relationship with image-based SOD. We find that **VOS** is very challenging, as it contains a large number of realistic videos, and its subset **VOS-E** serves as a good baseline for extending existing image-based models to the spatiotemporal domain. Moreover, its subset **VOS-N** covers many real-world scenarios that can facilitate the development of better algorithms. This dataset can be very helpful in video-based SOD, and the unsupervised

saliency-guided stacked autoencoders can be used as a good baseline model for benchmarking new video-based models.

REFERENCES

[1] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[3] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.

[4] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1884–1892.

[5] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2016.

[6] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1265–1274.

[7] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 414–429.

[8] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[9] *MSRA10K and THUR15K*. Accessed: Jul. 2016. [Online]. Available: http://mmcheng.net/gsal/

[10] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[11] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[12] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 742–756, Feb. 2014.

[13] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[14] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[15] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.

[16] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[17] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 594–602.

[18] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.

[19] W.-T. Li, H.-S. Chang, K.-C. Lien, H.-T. Chang, and Y.-C. F. Wang, "Exploring visual and motion saliency for automatic video object extraction," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2600–2610, Jul. 2013.

[20] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-based multiple foreground video co-segmentation via multi-state selection graph," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3415–3424, Nov. 2015.

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

[22] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–11.

[23] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 724–732.

[24] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 282–295.

[25] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5455–5463.

[26] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4142–4150.

[27] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2192–2199.

[28] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, Jun. 2014.

[29] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[30] J. Li, Y. Tian, T. Huang, and W. Gao, "A dataset and evaluation methodology for visual saliency in video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun./Jul. 2009, pp. 442–445.

[31] R. Carmi and L. Itti, "The role of memory in guiding attention during natural vision," *J. Vis.*, vol. 6, no. 9, p. 4, 2006.

[32] T. Vigier, J. Rousseau, M. P. Da Silva, and P. Le Callet, "A new HD and UHD video eye tracking dataset," in *Proc. Int. Conf. Multimedia Syst.*, 2016, pp. 48:1–48:6.

[33] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 569–582.

[34] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 853–860.

[35] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1665–1672.

[36] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.

[37] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1404–1412.

[38] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[39] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3183–3192.

[40] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 478–487.

[41] Q. Hou, M.-M. Cheng, X.-W. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3203–3212.

[42] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[43] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 330–344, 2015.

[44] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 660–668.

[45] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5781–5790.

[46] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 678–686.

[47] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3668–3677.

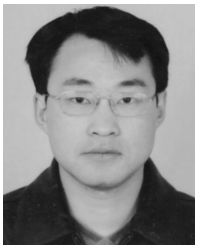
[48] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[49] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: Learning to detect a salient object sequence," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[50] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun./Jul. 2009, pp. 638–641.
[51] Y. Li, B. Sheng, L. Ma, W. Wu, and Z. Xie, "Temporally coherent video saliency using regional dynamic contrast," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2067–2076, Dec. 2013.
[52] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1777–1784.
[53] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3395–3402.
[54] W.-C. Chiu and M. Fritz, "Multi-class video co-segmentation with a generative multi-video model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 321–328.
[55] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.
[56] W.-H. Cheng, C.-W. Wang, and J.-L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.
[57] M. Gitte, H. Bawaskar, S. Sethi, and A. Shinde, "Content based video retrieval system," *Int. J. Res. Eng. Technol.*, vol. 3, no. 6, pp. 1–6, 2014.
[58] H. Jiang, G. Zhang, H. Wang, and H. Bao, "Spatio-Temporal video segmentation of static scenes and its applications," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 3–15, Jan. 2015.
[59] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Oct. 2005, pp. 1395–1402.
[60] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.
[61] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.
[62] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2016.
[63] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.
[64] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Mach. Vis. Conf.*, 2011, vol. 6. no. 7, pp. 1–12.
[65] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.
[66] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1139–1146.
[67] X. Li, Y. Li, C. Shen, A. Dick, and A. van den Hengel, "Contextual hypergraph modeling for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3328–3335.
[68] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2976–2983.
[69] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 883–890.
[70] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proc. Brit. Mach. Vis. Conf.*, 2014, vol. 2. no. 7, pp. 1–12.
[71] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 110–119.
[72] P. Jiang, N. Vasconcelos, and J. Peng, "Generic promotion of diffusion-based salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 217–225.
[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[74] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via SDAEs," *IEEE Trans. cybern.*, vol. 46, no. 2, pp. 487–498, Feb. 2016.
[75] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

Jia Li (M'12–SM'15) received the B.E. degree from Tsinghua University in 2005, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently an Associate Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image/video processing.

Changqun Xia is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image processing.



Xiaowu Chen (M'09–SM'15) received the Ph.D. degree in computer science from Beihang University in 2001. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include virtual reality, augmented reality, computer graphics, and computer vision.