A Benchmark Dataset and Saliency-guided Stacked Autoencoder 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 the past decades. However, video-based SOD is much less explored since there lacks 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 (64 minutes). In constructing the dataset, we manually annotate all objects and regions over $7,650$ uniformly sampled keyframes and collect the eye-tracking data of 23 subjects that free-view all videos. From the user data, we find salient objects in 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 the dataset, this paper proposes an unsupervised approach for video-based SOD by using a saliency-guided stacked autoencoder. In the proposed approach, spatiotemporal saliency cues are first extracted at pixel, superpixel and object levels. With these saliency cues, a stacked autoencoder is unsupervisedly trained which can automatically infer a saliency score for each pixel by progressively encoding the high-dimensional saliency cues gathered from the pixel and its spatiotemporal neighbors. Experimental results show that the proposed approach outperforms 19 image-based and 5 video-based models on the proposed dataset. Moreover, the comprehensive benchmarking results show that the proposed dataset is very challenging and have the potential to greatly boost the development of video-based SOD.

***Index Terms*—Salient object detection, Video dataset, Stacked autoencoder, Model benchmarking**

I. Introduction

THE booming of image-based salient object detection (SOD) originates from the presence 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, such large-scale datasets enables the fair comparisons between state-of-the-arts [7], [8]. Actually, the large-scale datasets provide a solid foundation for SOD and consistently guide the direction of this area.

In the past decade, SOD datasets progressively evolve to meet the increasing demands in developing and benchmarking new models. Some researchers argue that images in early datasets like **MSRA-B** [1] and **ASD** [2] are relatively simple and extend image-based datasets in terms of amount [9], [10] or complexity [11]–[14]. Meanwhile, some other researchers extend the concept of SOD to RGBD images [15], image collections [16]–[18] and even videos [19]–[22]. Among these extensions, video-based SOD has invoked great interests since it uniquely re-defines the problem from a spatiotemporal perspective. However, there still lacks large-scale video datasets for comprehensive model comparison, which prevents the fast growth of this branch. For example, the most widely used SegTrack dataset [23] consists of only 6 videos with 21 to 71 frames per video, while the latest dataset for video-based SOD, ViSal [22], contains only 17 videos with 30 to 100 frames per video. In addition, the definition of salient object in video is still not very clear (*e.g.*, manually annotated foreground objects [24], class-specific objects [22] or moving objects [25]). Therefore, it is necessary to construct a large and realistic video dataset with unambiguously defined and annotated salient objects.

To address this issue, this paper proposes **VOS**, a large-scale benchmark dataset for video-based SOD that consists 200 realistic videos (64 minutes, $116,103$ frames in total, see Fig. 1 for representative scenarios in **VOS**). In constructing **VOS**, we first collect two types of user data, including 1) the eye-tracking data of 23 subjects that free-view all the 200 videos and 2) the masks of all objects and regions in $7,650$ uniformly sampled keyframes that are annotated by another 4 subjects. Based on these user data, salient objects in the keyframes of a video are unambiguously annotated as the objects that consistently receive the highest density of fixations throughout the video. After discarding the pure-background keyframes as well as the keyframes in which salient objects are partially occluded by distractors and split into several disjoint regions, we obtain $7,467$ keyframes with binary masks of salient objects.

Given the large-scale dataset, it becomes a feasible solution to directly learn a supervised or unsupervised model. To validate this point, we propose an unsupervised approach that learns a saliency-guided stacked autoencoder for video-based SOD. The proposed approach first extracts multiple spatiotemporal saliency cues at pixel, superpixel and object levels. A stacked autoencoder is then unsupervisedly trained which can automatically infer a saliency score for each pixel by progressively encoding the high-dimensional saliency cues gathered from the pixel and its spatiotemporal neighbors.

J. Li, 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.

J. Li is also with the International Research Institute for Multidisciplinary Science at Beihang University, Beijing, 100191, China.

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).

In the comprehensive model benchmarking on **VOS**, the proposed approach outperforms 19 image-based and 5 video-based models. Moreover, the benchmarking results also validate that **VOS** is challenging dataset with plenty realistic videos that can greatly boost the development of this area.

Our main contributions are summarized as follows: 1) we propose a large, realistic and challenging dataset for video-based SOD, which will be released to facilitate the development of this area, 2) a saliency-guided stacked autoencoder is learned for video-based SOD, which outperforms 19 image-based and 5 video-based models and demonstrates the feasibility of directly learning a video-based SOD model in the data-driven manner, and 3) we provide a comprehensive benchmark of 24 state-of-the-arts, which reveals several key challenges in video-based SOD and further validates the effectiveness of the proposed dataset.

The rest of this paper is organized as follows: Section 2 reviews existing datasets and models. Section 3 presents the dataset for video-based SOD. In Section 4, we propose a saliency-guided stacked autoencoder for video-based SOD. Section 5 benchmarks the proposed model and 24 state-of-the-arts, and the paper is concluded in Section 6.

II. Related Work

Video-based SOD is tightly correlated with many video object segmentation areas like foreground object detection, primary object discovery, moving object segmentation, etc. In this section, we briefly review the most related works from all these areas. Since there exist many datasets for video object segmentation, we will first briefly review the most popular datasets that are frequently used in foreground video object segmentation. After that, we will provide a brief survey of state-of-the-art video object segmentation models that are tightly correlated with video-based SOD.

A. Datasets

SegTrack and SegTrack V2 are two popular datasets widely used in many researches of video object segmentation. SegTrack [23] contains 6 videos about animal and human with 244 frames in total (21 to 71 frame per video), while only one foreground object is manually annotated per frame. Videos in SegTrack are intentionally collected for benchmarking models with predefined challenges. SegTrack V2 [26] extends SegTrack from two perspectives. First, additional annotations of foreground objects are provided for the six videos in SegTrack. Second, 8 new videos are carefully chosen to cover more challenges. In total, SegTrack V2 contains 14 videos about bird, animal, car and human with $1,066$ densely annotated frames.

Freiburg-Berkeley Motion Segmentation dataset is designed for motion segmentation (*i.e.*, segmenting regions with similar motion). It is first proposed in [25] with 26 videos, and then Ochs *et al.* [27] extends the dataset with another 33 videos. In total, the extended dataset contains 59 videos with 720 annotated sparsely annotated frames. Although the dataset is much larger than SegTrack and SegTrack V2, the scenarios it covers is still far from sufficient [24]. Moreover, moving object is not equivalent to salient object, especially in scenes with complex contents.

DAVIS is one of the latest dataset for video object segmentation, which contains 50 high quality videos about human, animal, vehicle, object and action with $3,455$ densely annotated frames. Each video has Full HD 1080p spatial resolution and lasts about 2 to 4 seconds with a frame rate of 24fps. Each clip in DAVIS contains one foreground object or two spatially connected objects.

ViSal is a pioneer dataset for video-based SOD. It is proposed in [22] and contains 17 videos about human, animal, motorbike, etc. Each video contains 30 to 100 frames and salient objects are manually annotated according to the semantic classes of videos. In other words, this dataset assumes that salient objects is equivalent to the primary objects within videos of predefined tags.

Generally speaking, the datasets introduced above have greatly boosted the researches in video object segmentation but still have several drawbacks in benchmarking salient object detection models.

First, these datasets are still a little small for state-of-the-art learning algorithms like Convolutional Neural Networks. Although thousands of frames in SegTrack V2 and DAVIS are densely annotated, the rich redundancy in consecutive frames of the same videos may increase the over-fitting risk in model training.

Second, videos in these datasets are often selected so as to maximally cover several predefined challenges in video object segmentation. However, such intentionally selected videos may make the dataset less realistic.

Third, the salient/foreground objects are determined only by an annotator, which may incorporate strong subjective bias. For example, only the monkey in a video with both dog and monkey is annotated in SegTrack, while SegTrack V2 has the dog annotated as foreground objects as well. Actually, the manual annotations of salient objects from different subjects often conflict with each other in videos with multiple candidate objects [28], leading to ambiguous annotations in complex scenarios.

To sum up, existing datasets are still somehow insufficient to benchmark video-based SOD models due to the limited video number as well as the ambiguous definition and annotation of salient/foreground/moving objects. To further boost the development of this area, it is necessary to construct a large dataset that covers a wide variety of realistic scenarios and contains unambiguously defined and annotated salient objects.

B. Models

Image-based SOD becomes popular due to the presence of large-scale image datasets [1], [2], while video-based SOD is much less explored due to the lack of large video datasets. For example, Liu *et al.* [29] extended their image-based SOD model [1] to the spatiotemporal domain for salient object sequence detection. In their approach, salient object sequence detection was formulated as energy minimization problem in a conditional random field framework. In [30], 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.* [19] proposed to integrate 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 approaches like [29] 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.* [31] manually annotated the salient objects in 10 videos with about 100 frames per video. They also proposed an approach to detect temporally coherent salient objects using regional dynamic contrast features in the spatiotemporal domain of color, texture and motion. Their approach demonstrated impressive performance in processing videos with only one salient object. In [32], 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 encouraging 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 its surrounding background in a good fraction of the video. Wang *et al.* [33] 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 can be segmented by using an energy minimization framework. In their later work [22], Wang *et al.*proposed to utilize 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.

Despite the performance and benchmarking methodologies, these single video-based approaches have provide us an intuitive definition of salient objects. That is, salient objects in video should be spatiotemporally consistent and visually distinct from background regions. However, in realistic videos the assumptions like color/texture dissimilarity and motion irregularity may not always hold. A more general definition of salient objects in 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 [34] proposed a generative model for multi-class video co-segmentation. A global appearance model was learned to connect the segments from the same class so as to segment the foreground targets shared by different videos. Fu *et al.* [21] proposed to detect multiple foreground objects shared by a set of videos. Category-independent object proposals were first extracted and multi-state selection graph was then adopted to handle multiple foreground objects. Although video co-segmentation brings us a interesting new direction for studying video-based SOD, detecting salient objects in a single video is still the most popular scenario in many realistic applications.

III. A Large-scale Dataset for Video-based SOD

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

A. Video Collection

To build the dataset, we first collect hundreds of long videos from Internet and volunteers. After that, we randomly sample a set of short clips from them, and only the clips that contain objects in most frames are kept for further annotation. Finally, we obtain 200 realistic videos that last 64 minutes in total (*i.e.*, $116,103$ frames at 30fps). The video height falls in $[312, 800]$

and the width falls in $[408, 800]$. Representative scenarios in these videos can be found in Fig. 1. By observing these videos, we group them into two subsets according to the content complexity, including:

VOS-E. This subset contains 97 *easy* videos (27 minutes, $49,206$ frames) that are taken by volunteers with 83 to 962 frames per video. As shown in Fig. 1, a video in this subset usually contains an obvious foreground objects with slow camera motion. This subset serves as a baseline to explore the inherent differences and correlations between image-based and video-based SOD.

VOS-N. This subset contains 103 *normal* videos (37 minutes, $66,897$ frames) collected from Internet and volunteers with 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 background regions, etc. This subset is very challenging and can be used to benchmark models in realistic scenarios.

B. User Data Collection

The direct manual annotation of salient objects often lead to ambiguities in complex scenes and may bring in strong subjective bias. Inspired by the image dataset proposed in [13], we collect two types of user data, including object masks and human fixations, to unambiguously define and annotate salient objects in videos.

Object masks. Four subjects (2 males and 2 females, aged between 24 and 34) manually annotate the accurate boundaries of all objects/regions in video frames. Since it consumes too much time to process all frames, we uniformly sample only one keyframe out of every 15 frames and manually annotate only the $7,650$ keyframes. In the annotation, an object will maintain the same label throughout a video, and the holes in objects are filled to speed up the annotation. Since moving objects may merge or split several times in a short period and it is difficult to consistently assign different labels to them (*e.g.*, the kissing bear and the fighting cat in the third row of Fig. 2), we assign the same label to objects if they become indistinguishable in certain frames (*e.g.*, the cats and bears in Fig. 2) or difficult to be re-identified (*e.g.*, the jelly fishes 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.

Human fixations. Twenty-three subjects (16 males and 7 females, aged between 21 and 29) participate in the eye-tracking experiments. Note that none of them participates in the annotation of object mask. Each subject is asked to free-view all the 200 videos displayed on a 22-inch color monitor with a resolution of 1680×1050. A chin rest is adopted to reduce head movements and enforce a viewing distance of 75cm. During the free-viewing process, an eye-tracking apparatus with a sample rate of 500Hz is used to record eye movements. Finally, we keep only the *fixations* and denote the set of fixations received by a video $\mathcal{V}$ as $\mathbb{F}_{\mathcal{V}}$, in which a fixation 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 that f starts (a fixation lasts about two milliseconds, see Fig. 3 for the recorded fixations).

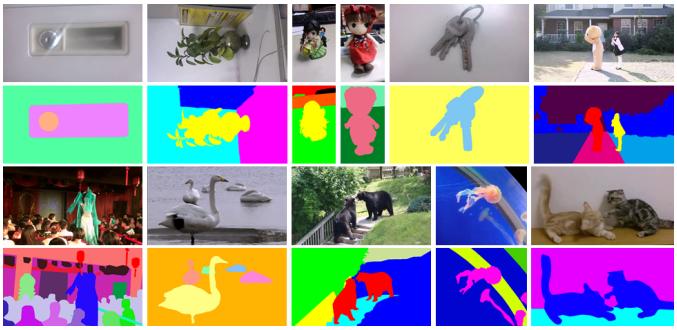

Fig. 2: Masks of objects and regions annotated by 4 subjects. Holes are filled up to speed up the annotation process (*e.g.*, the key in the first row), and multiple objects will be assigned the same labels throughout the video if they cannot be easily separated in certain frames (*e.g.*, the kissing bears and the fighting cats) or difficult to be re-identified (*e.g.*, the jelly fishes which frequently appear and disappear near screen borders).

Fig. 3: Human fixations (red dot) of 23 subjects on consecutive video keyframes. However, these fixations are still insufficient to directly annotate salient objects frame by frame. (a) Insufficient fixations to distinguish multiple salient objects and distractors; (b) Fixations fall outside small moving objects; (c) Fixations distracted by visually surprising regions; (d) Salient objects occluded by background regions, leading to background-only frames.

C. Definition and Annotation of Salient Objects in Video

In early datasets with simple images, salient objects can be manually annotated without much ambiguity. However, in a complex video there may exist several candidate objects, and different subjects may have different biases in determining which ones are the most salient. As a result, such subjective biases prevent the direct manual annotation of salient objects in complex videos.

To alleviate the subjective bias, the fixations of multiple subjects can be used to find the most salient objects. For example, Li *et al.* [13] collect fixations from 8 subjects that free-view the same image for 2 seconds. After that, salient objects are defined as the objects that receive 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.*, 30ms) 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 surprise and then return to the salient objects after a short period. In this case, the surprising background regions will be mistakenly 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 are purely background. If salient objects are defined by fixations received only by these frames, background regions in these frames will be mistakenly annotated as salient (*e.g.*, the girl is occluded by background regions in Fig. 3(d)).

From these reasons, it is difficult to directly define and annotate salient objects separately on each frame. Inspired by the idea of co-saliency, 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 video other than the highest *number* of fixations. In this manner, we can avoid mistakenly assigning large background regions high saliency values when salient objects are very small (*e.g.*, the aircraft in Fig. 3(b)).

D. Generation of Salient Object Mask

Based on the proposed definition, we can thus generate masks of salient objects for each video. We first compute the fixation density at each object in manually annotated keyframes. Considering that the fixations received by each keyframe are very sparse, we take the fixations recorded in a short period after the keyframe is displayed into consideration. 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

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

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 to 1 if $t_f > t$ and 0 otherwise. $\text{Dist}(f, p)$ measures the spatial distance between the fixation f and the pixel p, which can be computed as

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

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

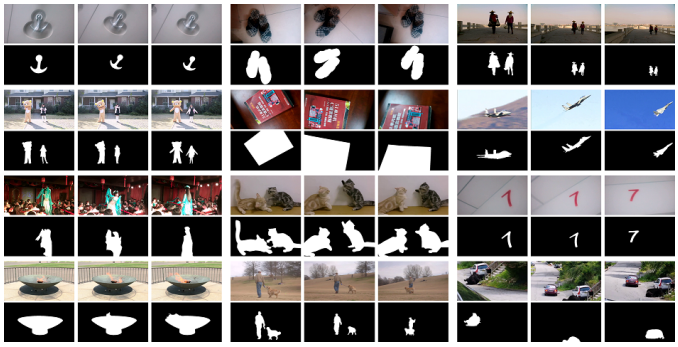

Fig. 4: Representative keyframes and masks of salient objects.

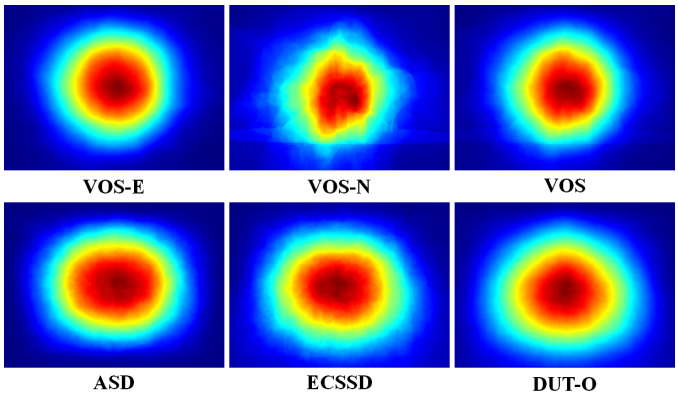

Fig. 5: The average annotation maps of 6 datasets.

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

$$S(\mathcal{O}) = \frac{\sum_{\mathcal{I}_t \in \mathcal{V}} \sum_{\mathcal{O} \in \mathcal{I}_t} S_0(\mathcal{O})}{\sum_{\mathcal{I}_t \in \mathcal{V}} \sum_{\mathcal{O} \in \mathcal{I}_t} 1}. \quad (3)$$

In (3), the saliency of an object is defined as its average fixation density throughout a 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). Finally, we generate a set of salient objects for each video, represented by a sequence of binary masks at keyframes. In particular, a keyframe which contains only background or a salient object that splits into several 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.

E. Dataset Statistics

To reveal the main characteristics of **VOS**, we shown 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]). Note that the AAMs of **VOS-E**, **VOS-N** and **VOS** are generated by averaging the AAM of each video that

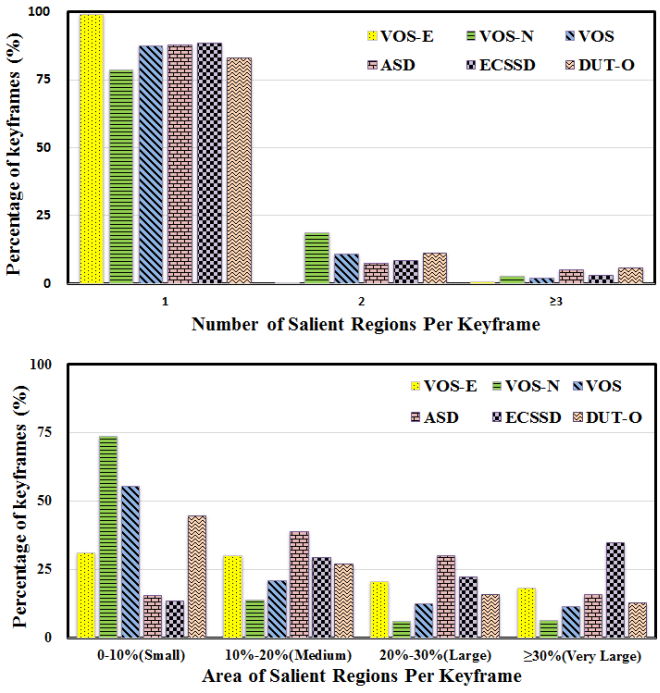

Fig. 6: The number and area of salient objects of 6 datasets.

is produced by summing up the binary masks of all keyframes. In this manner, we can provide a better view of the distribution 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, while the degree of center-bias is a little stronger than that in **ASD**, **ECSSD** and **DUT-O**. This is caused by the fact that photographers often have strong tendency to place salient targets near the center of the view in taking videos. This implies that image-based and video-based SOD are inherently correlated, and it is possible to directly transfer some useful saliency cues from the spatial domain to the spatiotemporal domain (*e.g.*, the background prior [3], [35] obtained from the boundaries pixels).

Moreover, we show in Fig. 6 the number and area of salient objects. As shown in Fig. 6, the number and area of salient objects in **VOS** is similar to those in **DUT-O**. This implies that **VOS** is a challenging dataset that well reflects many realistic scenarios. In particular, almost all keyframes from **VOS-E** contain only one salient object, while the sizes of such salient objects distribute almost uniformly in the categories of Small (31.1%), Medium (30.1%), Large (20.6%) and Very Large (18.3%). In addition, all videos in **VOS-E** have nearly static salient objects and camera motion. This finding indicates that **VOS-E** serves as a good baseline to benchmark video-based models.

IV. Learning a Saliency-guided Stacked Autoencoder for Video-based SOD

A. The Framework

For video-based SOD, we propose an unsupervised approach that learns a saliency-guided stacked autoencoder. The framework of the proposed approach is shown in Fig. 7. We first turn each frame from **VOS** into several color spaces and extract object proposals as well as the motion information (*e.g.*, optical flows). After that, we extract three spatiotemporal saliency cues from each frame at pixel, superpixel and object levels, while such saliency 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 eight 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, a stacked autoencoder can be unsupervisedly learned which contains 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 correlated with 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 which is spatially smooth and temporally consistent. Finally, several simple post-processing operations are applied to further pop-out salient objects and suppress distractors.

B. Multi-scale Saliency Cues Extraction

To extract saliency cues, we first resize a frame $\mathcal{I}_t$ to the maximum side length of 300 pixels and convert it to the Lab and HSV color spaces. After that, we estimate the optical flow [36] 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 formed by combining the optical flow and the flicker to indicate the variations along 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, including:

1) Pixel-based saliency. To efficiently extract the pixel-based saliency, we refer to the algorithm proposed in [35] that computes the minimum barrier distance from a pixel to image boundary (one pixel width). In the computation, we discard the Hue channel since the substraction between hue values can not always reflect the color contrast. Moreover, we also discard the RGB channels and the Value channel in HVS, which are somehow redundant to the other channels. For the rest 4 spatial and 3 temporal channels, the minimum barrier distances from all pixels to image boundary are separately computed over each channel. Such distances are then summed up across channels to initialize a pixel-based saliency map

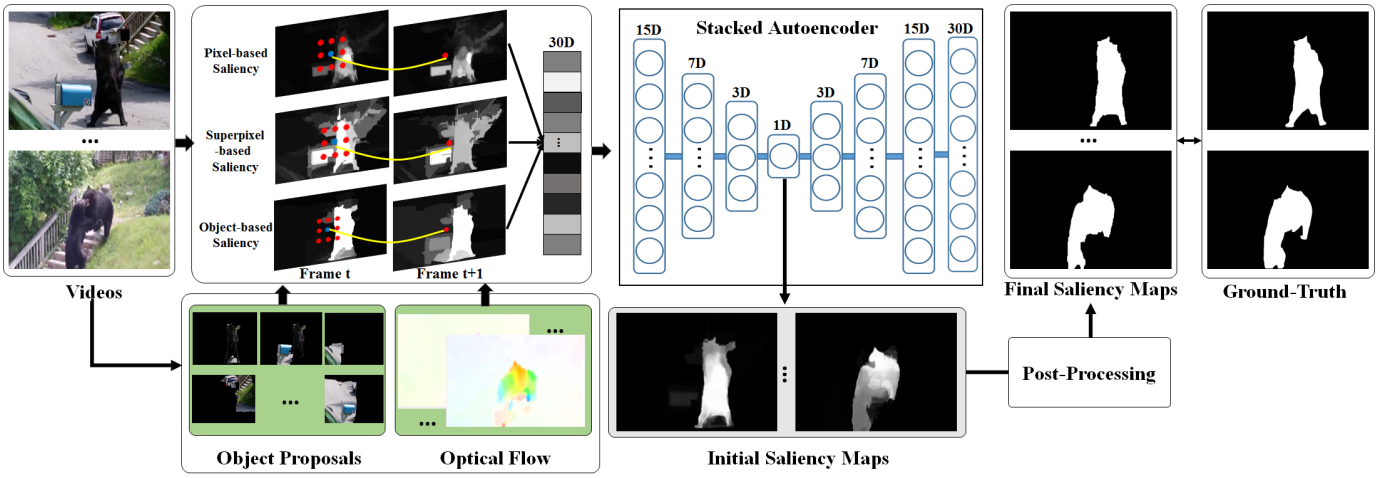

Fig. 7: The framework of the proposed saliency-guided stacked autoencoder.

$\mathbf{S}_t^{pix}$. Moreover, we also extract a backgroundness map as in [35] 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), the pixel-based saliency can be efficiently computed but sensitive to noise.

2) Superpixel-based saliency. In image-based SOD, superpixels are often used as the basic unit for feature extraction and saliency computation since they contain much more structural information than pixels. In this study, we adopt the supervised approach proposed in [37] to extract superpixel-based saliency by using the regression model pre-trained on images. A frame $\mathcal{I}_t$ is first divided into superpixels, from which the regional contrast, regional property, and regional backgroundness descriptors are extracted from the RGB, Lab and HSV color spaces. These descriptors are then combined to infer the saliency value of a superpixel through a random forest regressor pre-trained on images. Different from [37], we only compute the superpixel-based saliency at a single scale to speed up the computation process. 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 is more noisy but can detect a large salient object as a whole (*e.g.*, the tissue and the bucket in the third and fourth rows of Fig. 8(d)).

3) Object-based saliency. Inspired by the construction process of **VOS**, we adopt the Multiscale Combinatorial Grouping algorithm [38] 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 fixation prediction model proposed in [39] to generate three fixation density maps in the Lab, HSV and XYT spaces, respectively. Let $\mathbb{O}$ be the top-ranked objects with the highest objectness scores and $\mathbf{F}_{lab}, \mathbf{F}_{hsv}, \mathbf{F}_{xyt}$ be the three fixation density maps, the object-based saliency at a pixel p can be computed as:

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

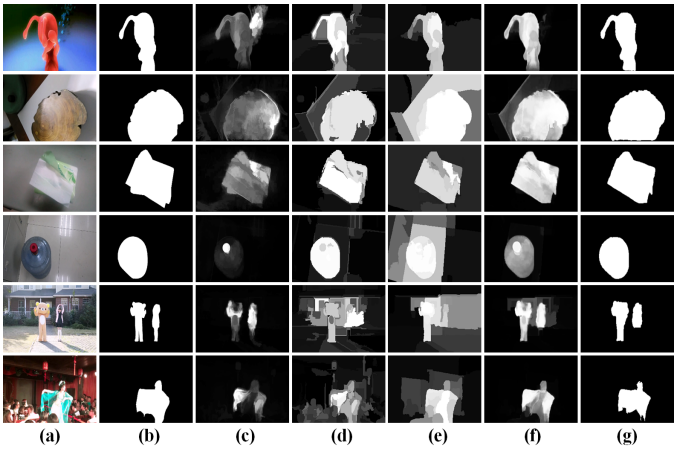

Fig. 8: Saliency cues and the estimated saliency maps. (a) Frames, (b) ground-truth, (c) pixel-based saliency, (d) superpixel-based saliency, (e) object-based saliency, (f) initial saliency maps obtained by the saliency-guided stacked autoencoder, (g) final saliency maps obtained after post-processing.

where $\delta(p \in \mathcal{O})$ is an indicator function which equals to 1 if $p \in \mathcal{O}$ and 0 otherwise. $\mathbb{O}$ is the set of objects used for computing the object-based saliency maps, and we set $\|\mathbb{O}\| = 50$ in 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 successfully pop-out the whole salient objects but often contain the background regions near them.

C. Learning a Stacked Autoencoder for Video-based SOD

Given the saliency cues, we have to estimate a non-negative saliency score for each pixel, which, statistically, has positive correlations with the saliency cues. Moreover, as stated in many previous works [22], [32], [33], 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. Correlated pixels in adjacent frames should have similar saliency scores so that salient objects can consistently pop-out throughout a video.

To develop a model under such constraints, we propose to train a stacked autoencoder that takes the saliency cues at a pixel and its spatiotemporal neighbors as the input so that the spatial smoothness and temporal consistency of predicted saliency scores can be guaranteed to some extent. Considering the computational efficiency, for each pixel we adopt its eight spatial neighbors and only the temporal neighbor in the subsequent frame with respect to the optical flow precomputed in extracting saliency cues. In this manner, a pixel can be 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**, denoted as $\{\mathbf{x}_n^1\}_{n=1}^N$. With these data, we train a stacked autoencoder with T encoding layers and the same number of decoding layers with logistic sigmoid transfer functions. In the training process, the tth encoding layer f_t, $t \in \{1, \ldots, T\}$ and its corresponding decoding layer $\hat{f}_t$ is 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 a ℓ-2 regularization term that penalizes the ℓ-2 norm of weights in the encoding and decoding layers (we empirically set $\lambda_w = 0.001$). Ω_s is a sparsity regularizer that 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 (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

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

where $\mathcal{N}(\cdot)$ indicates the normalization operation that enforces each dimension of the input data that enters a encoding layer falls in the same dynamic range of $[-1, 1]$. In this study, we set $T = 4$ encoding layers with $15, 7, 3, 1$ neurons at each layer, and each layer is trained with 100 epochs. Note that the Tth layer contains only one neuron, and by using its output scores 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 $\{\mathbf{x}_n^{T+1}\}_{n=1}^N$ is tightly correlated with the input saliency cues $\{\mathbf{x}_n^0\}_{n=1}^N$, and the 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 $\mathbf{v}_p$ that contains the saliency cues from p and its spatiotemporal neighbors, can be computed as

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

After computing a saliency score with (8) for each pixel, 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 impressive in highlighting salient objects and suppressing distractors. To further pop-out salient objects and suppress distractors, we conduct three post-processing operations, including:

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 [35].
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 three post-processing operations will generate a compact and precise salient map for each frame. Note that the center-biased re-weighting and the spatial smoothing operations, which are popular in existing models, are not adopted in this study. This is because the autoencoder, which is unsupervisedly learned over a large-scale dataset, already has the capability to accurately detect various types of salient objects despite their positions and sizes.

V. Experiments

In this Section, we compare the Saliency-guided Stacked Autoencoder (SSA) with 24 state-of-the-arts on **VOS** (see Table I). The main objectives are two-fold: 1) validate the effectiveness of **SSA**, and 2) provide a comprehensive benchmark to reveal the key challenges in video-based SOD. The rest of this section will first introduce the experimental settings and then discuss the results.

A. Settings

As shown in Table I, 24 state-of-the-art models are tested on **VOS** (19 image-based and 5 video-based). Similar to image-based SOD, we 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

TABLE I: State-of-the-arts for benchmarking ([I] for image-based model and [V] for video-based model)

Model	Pub. & Year	Model	Pub. & Year
CB [40]	BMVC 2011 [I]	**RC** [41]	CVPR 2011 [I]
ULR [42]	CVPR 2012 [I]	**LMLC** [43]	TIP 2013 [I]
DRFI [3]	CVPR 2013 [I]	**GMR** [10]	CVPR 2013 [I]
HS [11]	CVPR 2013 [I]	**PCA** [44]	CVPR 2013 [I]
CHM [45]	ICCV 2013 [I]	**DSR** [46]	ICCV 2013 [I]
MC [37]	ICCV 2013 [I]	**HDCT** [47]	CVPR 2014 [I]
RBD [48]	CVPR 2014 [I]	**BL** [49]	CVPR 2015 [I]
BSCA [50]	CVPR 2015 [I]	**GP** [51]	ICCV 2015 [I]
MB [35]	ICCV 2015 [I]	**MB+** [35]	ICCV 2015 [I]
SMD [52]	PAMI 2016 [I]		
SIV [19]	ECCV 2010 [V]	**FST** [32]	ICCV 2013 [V]
NLC [53]	BMVC 2014 [V]	**SAG** [33]	CVPR 2015 [V]
GF [22]	TIP 2015 [V]		

TABLE II: Performance of our approach and 24 state-of-the-arts on **VOS** and its two subsets **VOS-E** and **VOS-N**. Bold and underline indicate the best and the second best performance, respectively.

Models	VOS-E				VOS-N				VOS			
	MAP	MAR	F_β	MAE	MAP	MAR	F_β	MAE	MAP	MAR	F_β	MAE
CB [40]	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 [41]	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 [42]	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 [43]	0.687	0.736	0.697	0.154	0.408	0.501	0.426	0.262	0.543	0.615	0.558	0.210
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
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
PCA [44]	0.750	0.725	0.744	0.143	0.420	0.696	0.462	0.142	0.580	0.710	0.606	0.143
CHM [45]	0.756	0.765	0.758	0.124	0.409	0.611	0.443	0.186	0.578	0.685	0.599	0.156
DSR [46]	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 [37]	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 [47]	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 [48]	0.799	0.782	0.795	0.091	0.516	0.709	0.550	0.145	0.653	0.745	0.672	0.119
BL [49]	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 [50]	0.766	0.758	0.764	0.133	0.457	0.663	0.493	0.195	0.607	0.709	0.628	0.165
GP [51]	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 [35]	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+ [35]	0.803	0.792	0.801	0.096	0.492	**0.754**	0.535	0.162	0.643	0.772	0.669	0.130
SMD [52]	0.811	0.789	0.806	0.096	0.528	0.688	0.558	0.148	0.665	0.737	0.681	0.123
SIV [19]	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 [32]	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* [53]	0.462	0.446	0.458	0.199	0.611	0.604	0.609	0.138	0.531	0.520	0.528	0.171
SAG [33]	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 [22]	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	**0.873**	0.788	**0.852**	**0.061**	**0.647**	0.686	**0.656**	**0.096**	**0.756**	0.736	**0.752**	**0.079**

* The executable of NLC only output valid results on 156 videos (83 from **VOS-E** and 73 from **VOS-N**). This may be caused by the fact that some videos in **VOS** contain thousands of frames, which are difficult to be processed by approaches that consume too much time and memory.

and G, which directly reflects the visual difference [8], [54]. Moreover, the Recall and Precision scores can be computed by converting S into a binary mask M and compare it with G:

$$\begin{aligned} \text{Recall} &= \frac{\#(\text{Non-zeros in } M \cap G)}{\#(\text{Non-zeros in } G)}, \\ \text{Precision} &= \frac{\#(\text{Non-zeros in } M \cap G)}{\#(\text{Non-zeros in } M)}, \end{aligned} \tag{9}$$

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 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. After that, 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 performance over all videos. Correspondingly, F_β is computed by fusing MAR and MAP to quantize the overall performance of a model:

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

where we set $\beta^2 = 0.3$ as most of existing image-based models did in the performance evaluation.

Another problem in assessing models with MAP, MAR and F_β is how to turn a gray-scale saliency map S into a binary mask M. Similar to image-based SOD, we enumerate the fixed thresholds from 0 to 255 and compute MAP and MAR of a model at each threshold so as to generate a Precision-Recall curve. Moreover, the adaptive threshold proposed in [2], which are computed as twice the average values of S, is used to generate a binary mask from each saliency map. Considering that such adaptive threshold may sometimes exceed the maximal saliency value of S if there exists a very large salient object, we re-set this threshold to the maximal saliency value in this case. In this manner, the unique MAR, MAP and F_β scores can be generated to measure the overall performance of a model.

B. Results

The performance scores of **SSA** and the other 24 state-of-the-arts over **VOS-E**, **VOS-N** and **VOS** are illustrated in Table II, and some representative results from the best models are shown in Fig. 9. Moreover, we also provide the Precision-Recall curves of all the 25 approaches in Fig. 10. With Table II and Fig. 9-10, we conduct several comparisons and discussions, including:

1) Comparisons between SSA and the other 24 models. From Table II, we can see that **SSA** achieves the best F_β and MAE scores over **VOS** and its two subsets. On **VOS**, **SSA** has $F_\beta = 0.752$ and MAE $= 0.079$, while the best model among the other 24 state-of-the-arts, **FST**, only has $F_\beta = 0.718$ and MAE $= 0.097$. The larger F_β and the smaller MAE indicate that **SSA** outperforms **FST** and the other 23 models, no matter how the saliency maps are compared (*i.e.*, direct comparison in computing MAE or binarized comparison in compute F_β).

One more thing that worth mentioning is that on **VOS** and its two subsets, **SSA** always has the best Precision

(MAP $= 0.756$ on **VOS**), while its MAR scores are relatively lower than other models. Actually, it is widely recognized that a high precision is much more difficult to obtain than a high recall in both image-based and video-based SOD, and the most commonly used trade-off is to gain a remarkable increase in precision along with the decrease in recall as small as possible. That is why the computation of F_β in this work and almost all the image-based models emphasize more on precision than recall. Although a higher recall usually leads to a better subjective impression in qualitative comparisons, the overall performance, especially the F_β score, may be not very satisfactory due to the emphasis of precision in computing F_β.

2) Comparisons between image-based and video-based models. Beyond analyzing the best models, another issue that worth discussing is the performance of the image-based and video-based models. Interestingly, video-based models like **GF** and **SAG** may sometimes perform even worse than the image-based models (*e.g.*, **SMD**, **RBD** and **MB+**). This may be caused by two reasons. First, the impact of incorporating temporal information in visual saliency computation is not always positive. In some videos, the salient objects, as assumed by many video-based models, have specific motion patterns that are remarkably different from the distractors (*e.g.*, the dancing bear/girl in the second row of Fig. 4). However, such an assumption may not always hold in 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 fine-tuned on small datasets, which may be somehow "over-fitting" to specific video scenarios. Given an unknown scenario contained in **VOS**, these parameters often 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 the energy-based optimization.

3) Failure cases. Although **SSA** achieves the best performance, we can see that its F_β score is still far from perfect. On **VOS-E** that contains only simple videos with nearly static salient objects and distractors as well as slow camera motion, **SSA** only reaches a F_β score of 0.852, while the performance score drops sharply to 0.656 on **VOS-N**. This implies that the videos from the real-world scenarios are much more challenging than the videos taken in the laboratory environment. Actually, this is also the main reason that prevents the usage of existing SOD models in other applications.

To validate this point, we illustrate in Fig. 11 two representative scenarios that **SSA** fails, which actually 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 in some later frames the pen is correctly detected, it is difficult to transfer such correct results to the frames far away. This indicates that the local spatiotemporal correspondences between pixels used by **SSA** is still insufficient to handle more challenging scenarios, and a salient object should be detected by computing saliency from the global perspective as well.

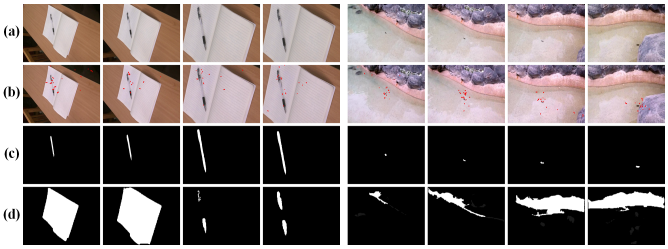

Fig. 11: Failure cases. (a) Frames, (b) the fixations received in 30ms after a keyframe is displayed, (c) binary masks of salient objects and (d) the estimated saliency maps of **SSA**.

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 proposed **VOS** dataset. Actually, the scenarios in the realistic videos from **VOS**, which are mainly taken by non-professional photographers, 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 video, even though it is very small. The existence of such scenarios in **VOS** increase the difficulties to transfer the knowledge learned on existing image datasets (*e.g.*, the regression model learned in **DRFI**) 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 performance in processing real-world videos.

4) Discussion and conclusion. 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 perfect without fully utilizing the temporal information from both local and global perspective. Second, there exist some inherent correlations 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 have better capability to process real-world scenarios.

VI. Conclusion

Salient object Detection is a hot topic in the area of computer vision. In the past five years, dozens of innovative models have been proposed for detecting salient objects in images, which gradually evolve from heuristic to learning-based due to the presence of large-scale image datasets. However, the problem of video-based SOD has not been sufficiently explored since there lacks a large-scale video dataset. Actually, the most challenging part in building such a dataset is to

provide a reasonable and unambiguous definition of salient objects from the spatiotemporal perspective.

To address this problem, this paper proposes **VOS**, a large and realistic 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 a saliency-guided stacked autoencoder for video-based SOD, which, together with 24 state-of-the-art models, are compared over **VOS** to show the challenges of video-based SOD as well as its differences and correlations with image-based SOD. We find that **VOS** is very challenging for containing a large amount of realistic videos, and its subset **VOS-E** serves as a good baseline to extend existing image-based models to the spatiotemporal domain. Moreover, its subset **VOS-N** covers many real-world scenarios that can help the deployment of better algorithms. This dataset can be very helpful for the area of video-based SOD, and the proposed saliency-guided stacked autoencoder can be used 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

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

[4] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *CVPR*, 2015, pp. 1884–1892.

[5] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[6] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[7] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 414–429.

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

[9] MSRA10K and THUR15K, "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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[11] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[12] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," in *IEEE Transactions on Image Processing*, 2014.

[13] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[15] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *ECCV*, 2014, pp. 92–109.

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

[17] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[18] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 594–602.

[19] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision (ECCV)*, 2010.

[20] 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 Transactions on Image Processing*, vol. 22, no. 7, pp. 2600–2610, 2013.

[21] 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 Transactions on Image Processing*, vol. 24, no. 11, pp. 3415–3424, Nov 2015.

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

[23] D. Tsai, M. Flagg, and J. M.Rehg, "Motion coherent tracking with multi-label mrf optimization," *British Machine Vision Conference (BMVC)*, 2010.

[24] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] T. Brox and J. Malik, *Object Segmentation by Long Term Analysis of Point Trajectories*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 282–295.

[26] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

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

[28] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[29] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: Learning to detect a salient object sequence," in *International Conference on Pattern Recognition (ICPR)*, 2008.

[30] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *IEEE International Conference on Multimedia and Expo (ICME)*, June 2009, pp. 638–641.

[31] S. Bin, Y. Li, L. Ma, W. Wu, and Z. Xie, "Temporally coherent video saliency using regional dynamic contrast," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2067–2076, 2013.

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

[33] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3395–3402.

[34] W. C. Chiu and M. Fritz, "Multi-class video co-segmentation with a generative multi-video model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 321–328.

[35] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1404–1412.

[36] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, March 2011.

[37] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[38] J. Pont-Tuset, P. Arbelaez, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[39] J. Li, M. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.

[40] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *British Machine Vision Conference (BMVC)*, 2011.
[41] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
[42] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
[43] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Transactions on Image Processing*, vol. 22, no. 5, 2013.
[44] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1139–1146.
[45] X. Li, Y. Li, C. Shen, A. R. Dick, and A. van den Hengel, "Contextual hypergraph modeling for salient object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3328–3335.
[46] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
[47] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
[48] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
[49] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1884–1892.
[50] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 110–119.
[51] P. Jiang, N. Vasconcelos, and J. Peng, "Generic promotion of diffusion-based salient object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 217–225.
[52] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
[53] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *British Machine Vision Conference (BMVC)*, 2014.
[54] G. Lee, Y. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

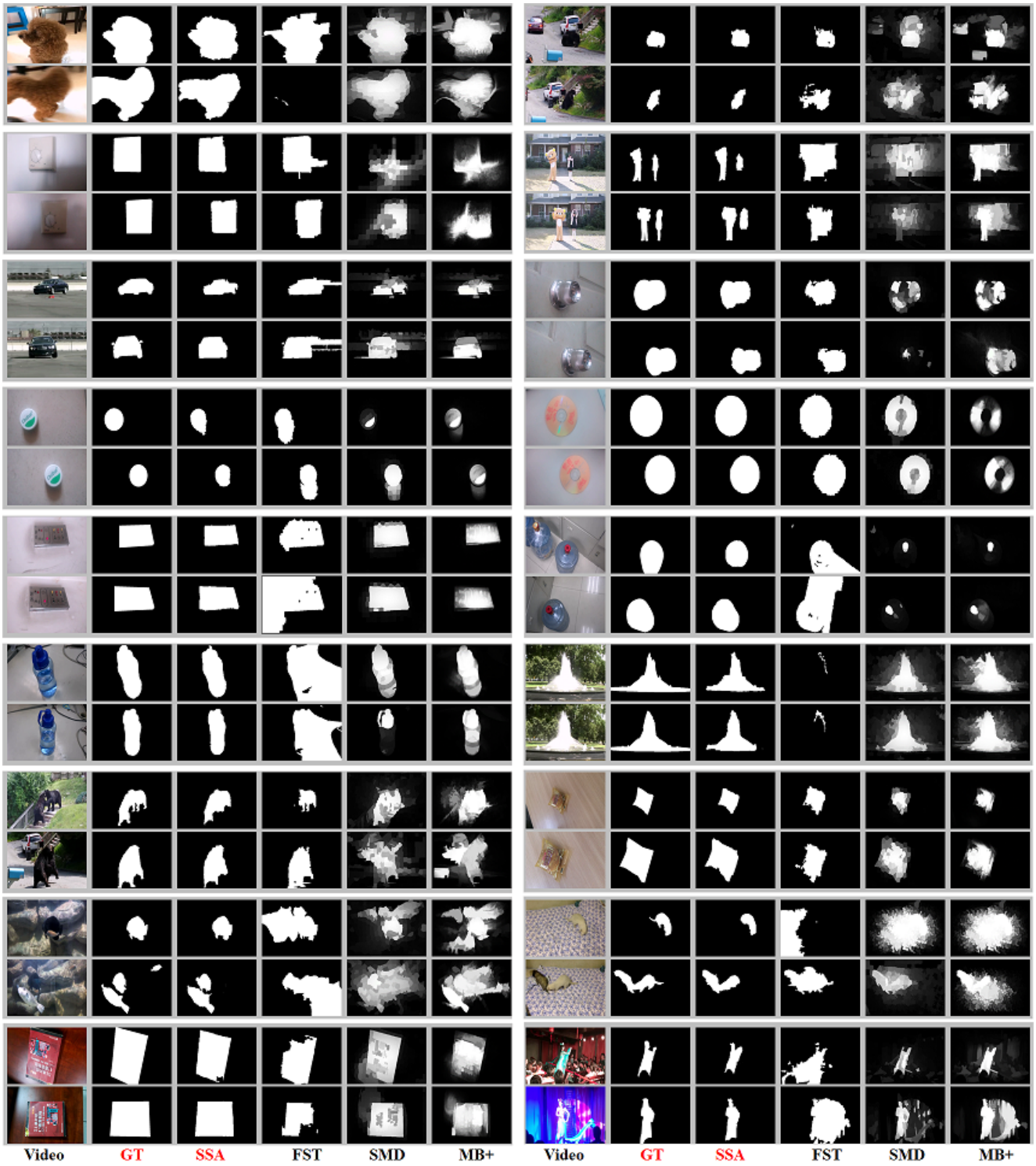

Fig. 9: The representative results of **SSA** and the other 3 models that perform among the best.

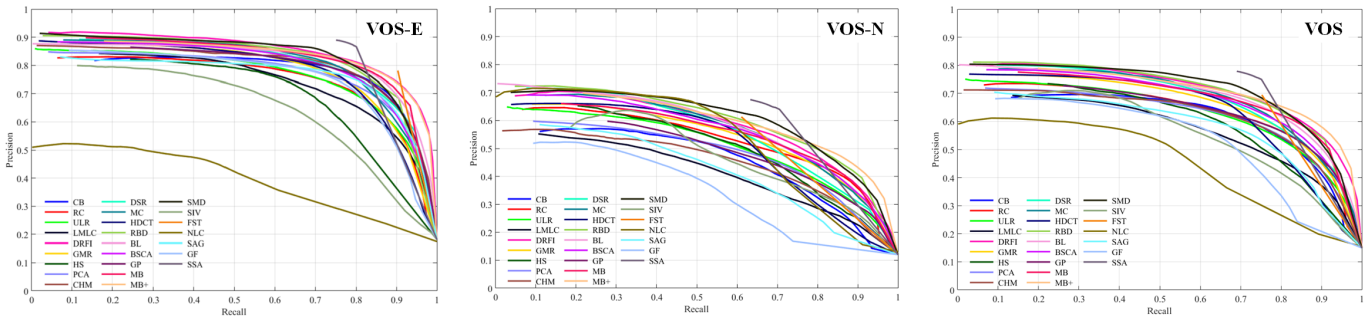

Fig. 10: The Precision-Recall curves of our approach and 24 state-of-the-arts.