Overview of deep-learning based methods for salient object detection in videos

Qiong Wang, Lu Zhang, Yan Li, Kidiyo Kpalma

Highlights

- This paper provides an overview of recent deep-learning based methods for salient object detection in videos;
- A classification of the state-of-the-art methods and their frameworks is provided;
- The performance of state-of-the-art methods is further analysed through experimental comparison on different public datasets and ablation study on the impact of variants.

Overview of deep-learning based methods for salient object detection in videos

Qiong WANG[a,b], Lu ZHANG[b], Yan LI[c,*], Kidiyo KPALMA[b]

[a] *Zhejiang University of Technology, 310023 Hangzhou, China*
[b] *Univ Rennes, INSA Rennes, CNRS, IETR (Institut d'Electronique et de Télécommunication de Rennes) - UMR 6164, F-35000 Rennes, France*
[c] *Université libre de Bruxelles, Belgium.*

Abstract

Video salient object detection is a challenging and important problem in computer vision domain. In recent years, deep-learning based methods have contributed to significant improvements in this domain. This paper provides an overview of recent developments in this domain and compares the corresponding methods up to date, including 1) classification of the state-of-the-art methods and their frameworks; 2) summary of the benchmark datasets and commonly used evaluation metrics; 3) experimental comparison of the performances of the state-of-the-art methods; 4) suggestions of some promising future works for unsolved challenges.

Keywords: deep-learning, salient object detection, video

1. Introduction

Salient Object Detection (SOD) in videos aims at locating primary foregrounds mostly attracting the human attention in each frame. Its output is a saliency map for each frame, where the pixel value indicates the probability of the corresponding pixel belonging to a salient object [1, 2]. The higher the value, the higher the saliency. The SOD is popularly used in applications where the task is driven by the human attention, such as image segmentation [3], im-

*Corresponding author
Email address: liyanxian19@gmail.com (Yan LI)

age change detection [4], autonomous driving [5], autonomous facial expression recognition [6], etc.

Recently, several researchers tend to solve the problems of SOD in videos using deep-learning based methods, which largely improves the performance of both the accuracy and the efficiency. However, there is few related survey. Table 1 lists the most relevant works, from which we can see that former works mainly focus on traditional methods for images [7, 8, 9]. Among the recent works related to deep-learning based methods, the survey presented in [10] is only for images; and the benchmark [11] only compares deep-learning based methods proposed for images with traditional methods proposed for videos. The survey of existing deep-learning based methods for salient object detection in videos is less explored.

Table 1: Comparison of the existing survey/benchmark for Salient Object Detection

	Year	Benchmark	Survey	Traditional	Deep-learning	Video	Image
[7]	2014	×	✓	✓	×	×	✓
[8]	2014	×	✓	✓	×	×	✓
[9]	2015	✓	×	✓	×	×	✓
[10]	2018	×	✓	×	✓	×	✓
[11]	2018	✓	×	✓	✓	✓	✓

This paper has two main motivations:

- Recently, deep learning-based video SOD has achieved high performing results in this research field but there are still several challenging research directions that need to be explored, so it is interesting to have a general idea about the existing methods, which may pave the way for future works.
- To serve the research community, it is necessary to present a global assessment of state-of-the-art methods with common metrics and comprehensive datasets. To further understand algorithms, it is attractive to make analyses of the strength and weakness of each method, and conduct the ablation study to offer insights into the impact of different components.

The remaining of this paper is organized as follows. Section 2 gives an

classification of deep-learning based methods for SOD in videos. It details the framework of each of the representative methods. Section 3 introduces popular used benchmark datasets and evaluation metrics, then gives experimental comparison of these methods, presents the ablation study and discusses promising future works. Section 4 concludes the paper.

2. Classification of the state-of-the-art methods

Deep-learning based methods for video SOD gain great research interests, and some methods are proposed. However, there still lack sufficient methods for comprehensive analysis. Inspired by [11], the inherently correlated tasks like video foreground object segmentation, moving object segmentation and image SOD are considered for analysis and comparison in this work.

According to the common concepts used in deep learning methods, firstly, the global framework for each method is described in 2.1, then the deep network in each method is analyzed in 2.2, and finally an overview of the categorization of methods is shown at a functional level in 2.3. As a matter of convenience, the describled methods, are denoted as SCOMd [12], NRF [13], DHSNet [14], OSVOS [15], NLDF [16], LMP [17], SFCN [18], SegFlow [19], LVO [20], WSS [21], SCNN [22], DSS [23], SPD [24], AFNet [25] and CPD [26].

2.1. Analysis of the frameworks

According to the involved tasks, these frameworks can be divided into two categories: single-task and multi-task. According to the domain of detection, these frameworks can be classified into 1) Spatial; 2) Temporal; 3) or Spatio-temporal.

2.1.1. Single-task vs Multi-task

The single-task framework is designed just for the SOD task, while the multi-task framework not only predicts the salient objects, but also evaluates other tasks. It exploits the connections between the SOD task and other highly related tasks (such as image classification, optical flow, edge detection and etc.), and

then improves the SOD performance by making use of the deep representation from these tasks.

Specifically, the WSS proposes a network which has two subnetworks: one is designed for classification and the other is designed for SOD. Both subnetworks share convolutional layers firstly and then are separated on the top of the shared layers, as shown in Fig. 1 (a). The SegFlow proposes a network which also consists of two subnetworks: the segmentation subnetwork and the flow subnetwork. A bi-directional feature propagation is built between these two networks as shown in Fig. 1 (b).

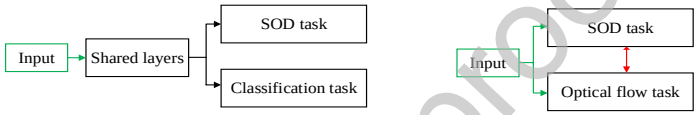


Figure 1: Multi-tasks models: the left one is the WSS and the right one is the SegFlow.

The OSVOS proposes two fully convolutional networks (FCNs) with the same architecture. The first FCN is used as a foreground branch and the second FCN is employed as a edge detection branch. The output of the first FCN is optimized by combining with that from the second FCN. The NLDF adds the boundary loss term to design extra constraints to saliency prediction. The AFNet applies a boundary-enhanced Euclidean loss to overcome blurred saliency boundaries. The SPD proposes joint training with the edge detection task. In the training procedure, the images from the edge detection and salient object detection dataset are inputted alternatively. Note that in the published codes, the OSVOS dose not contain the boundary snapping branch and the SPD does not contain the joint edge training. We only focus on their SOD task in the following part.

2.1.2. Domain of used features

The DHSNet, the DSS, the SPD, the NLDF, the OSVOS, the WSS, the AFNet and the CPD design networks to predict the salient object from the spatial domain, while the LMP detects motion patterns in videos with a motion

pattern network from the temporal domain, as in Fig. 2.

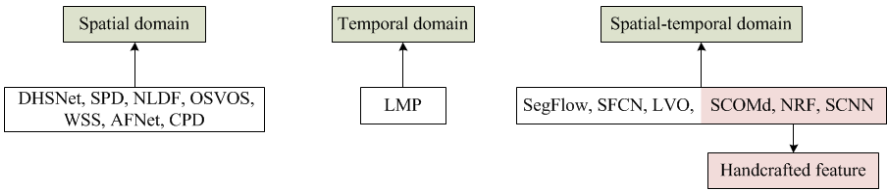


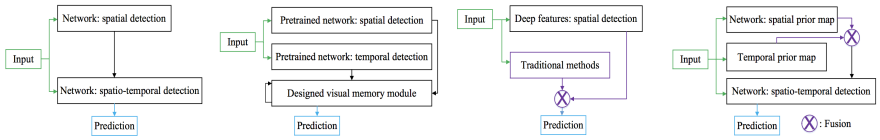
Figure 2: Classification based on the domain of detection.

The SegFlow, the SFCN, the LVO, the SCOMd, the NRF, and the SCNN estimate the salient object in a video sequence from the spatio-temporal domain. Among them, the SegFlow and the SFCN design the networks to learn deep features, while the LVO extracts deep features in spatial and temporal domains from pretrained networks, and builds a visual memory module to get the prediction, as in Fig. 3.

The SCOMd, the NRF and the SCNN combine handcrafted features for detection. The SCOMd uses a pretrained network to get deep spatial features and formulates the detection as energy minimization using a spatio-temporal constrained optimization model. In the NRF, the authors firstly obtain the initial salient object and background estimation with a proposed network, and then construct a neighborhood reversible flow to propagate salient object and background along the most reliable inter-frame correspondences. The SCNN firstly employs the proposed network to get a spatial prior map, secondly uses a graph-based algorithm to get superpixels on the optical flow map, and extractes deep features from a pretrained network for each superpixel to generate the temporal prior map, thirdly combines these two prior maps to be a spatio-temporal prior map which guides the proposed network to generate the spatio-temporal saliency map. At last, the output saliency map is optimized by a conditional random field (CRF) model.

2.2. Analysis of the networks

In this part, according to the common concepts used in deep learning, we analyze the networks designated in representative methods from aspects of the



(a) (b) (c) (d)

Figure 3: Models detected from the spatio-temporal domain: (a) the SFCN, (b) the LVO, (c) the SCOMd and the NRF, (d) the SCNN.

architecture and training details.

2.2.1. Architecture

The architecture of the designed networks can be divided into side-fusion network and bottom-up/top-down network.

The side-fusion network aggregates multi-layer responses of the backbone network (i.e. an existing trained model with published weights). In the SCNN, the OSVOS and the SegFlow, feature maps from various layers of the backbone are up-sampled and summed together. The SCNN considers responses from 4th and 5th layers. The SegFlow mainly uses that from 3rd to 5th layers. While, the OSVOS adopts all layers for predicting the final output. Feature maps obtained from each layer are fused into a single output and short connections are added from the low-level layer to the high-level layer. The DSS adds multiple short connections from deeper side outputs to the shallower ones. The NLDF fuses multi-level features to generate a local map, then integrates the local map with the global map got by the top layer of the backbone to obtain the final prediction.

The bottom-up/top-down network generates hierarchical features layer by layer. In the WSS, the SFCN, the LMP, the DHSNet, the SPD, the AFNet and the CPD, the rich and detailed low-level representations are incorporated into the coarse-level semantic representations, which benefits the high-level features with finer details. The DHSNet uses recurrent convolutional layers (RCL) that can incorporate recurrent connections into each convolutional layer in the decoder. For refining the high-level representations, pooling-based modules are

adopted in the SPD and the NRF. The SPD builds pyramid pooling module (PPM), and the NRF uses three parallel modules with "à trous" pyramid pooling (ASPP). Attentive maps are added in the AFNet and the CPD. The AFNet builds Attentive Feedback Modules to guide the boundary-aware learning phase, and the CPD generates the attention map for refining high-level features. Besides, the designed "visual memory module" in the LVO is realized with the convolutional recurrent unit. The classification of architectures can be found in Fig. 4.

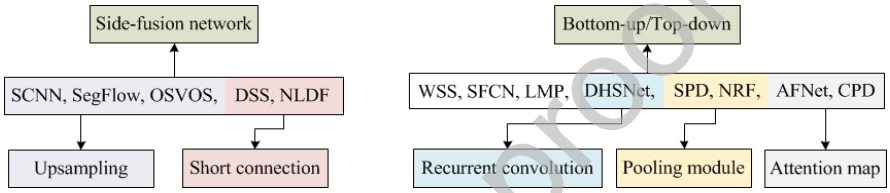


Figure 4: Classification of the architectures.

2.2.2. Training details

The training details of networks are introduced from aspects of the strategy, the backbone, the training dataset and the loss function.

The strategy to employ CNNs on salient object detection can be divided into "Off-the-shelf CNN features" (without retraining the CNN) and "Multi-stage/end-to-end trained".

In "Off-the-shelf CNN features", the used deep representations are directly extracted from pretrained deep networks. Thus, this is a simple way to directly use these deep representations for further researches. The SCOMd and the SCNN extract deep features from image SOD networks (built on VGGNet and AlexNet respectively, and pretrained on MSRA-B dataset), while the LVO gets deep spatial features using a semantic segmentation network (built on VGGNet and pretrained on PASCAL VOC 2012 dataset) and obtains temporal features from a pretrained moving object segmentation network (pretrained on FlyingThings3D dataset).

In "Multi-stage/end-to-end trained", methods usually get more efficient deep representations through their own training phase, where the inputs-outputs relationship is learned by their designed deep architectures. "Multi-stage trained" models are with intermediate supervision to ones trained end-to-end.

Specifically, the networks designed in the DHSNet, the DSS, the SPD, the NLDF, the OSVOS, the NRF, the LMP, the LVO, the AFNet and the CPD are end-to-end trained, while the WSS, the SegFlow, the SCNN and the SFCN are multi-stage trained. The WSS jointly trains the network for the foreground and image-level tag prediction to produce the initial saliency map, which is then used to fine-tune the foreground branch. The SegFlow uses an iterative training between the segmentation task and optical flow task. The SCNN firstly trains the network to get a spatial prior map, and then uses a fine-tuning strategy to generate the spatio-temporal saliency map with the guidance of spatio-temporal prior map. The SFCN uses the proposed network for spatial saliency detection with the input of each frame (the generated spatial saliency map is denoted as the SFCNs), and uses the same network for spatio-temporal saliency detection with the input of adjacent frame pairs and the detected spatial-temporal saliency results.

The backbone is commenly used to build networks in most methods. Image classification networks (e.g. VGGNet and ResNet) and the optical flow network (e.g. FlowNetS), trained on large-scale datasets, have a strong ability to learn both low-level and high-level features. The SegFlow initializes segmentation branch and optical flow branch using the weights from ResNet-101 and FlowNetS respectively. The DHSNet, the DSS, the SPD, the NLDF, the OSVOS, the WSS, the SFCN, the SCNN, the NRF, the AFNet and the CPD adopt the VGG16 as the backbone.

The training dataset is used for networks to learn deep representations. According to their utilization degree of the labeled datasets, the models can be further divided into supervised and weakly-supervised models. Supervised models need training datasets with accurate ground truth, while weakly-supervised models train the network without requiring all training datasets to have accurate

annotations.

For weakly-supervised models, in the WSS, the image-level annotations (ImageNet dataset) are used as weakly labeled datas, based on the assumption that image-level tags can provide the classes of the dominant objects which can be regarded as the salient foregrounds. Sometimes, pseudo pixel-level labels are used. In the SCNN, saliency maps generated from existing image saliency detection are used, while the WSS adopts its initial saliency maps for training.

For the supervised datasets: Image-based SOD datasets (e.g. MSRA-B, MSRA10K, DUT-OMRON, HKU-IS and CSSD) are commenly used in the DSS, the NLDF, the SCNN, the DHSNet, the SFCN and the NRF, and video object segmentation datasets (e.g. SegTrackV2, DAVIS 2016) are used in most methods (the SFCN, the LVO, the SegFlow, the SCNN and the OSVOS); image object segmentation datasets (e.g. DUTS) are used in the SPD, the AFNet and the CPD; moving object segmentation datasets (e.g. FBMS is used in methods SFCN and SCNN; optical flow datasets (FlyingThings3D) are used in the LMP; and datasets (MPI Sintel, KITTI, Scene Flow) are used in the SegFlow.

Besides, due to the limitation of existing datasets, some methods generate new video datasets. In the SFCN, the authors create synthesized video dataset from two large image saliency datasets (MSRA10K and DUT-OMRON), and in the LVO, the authors create training sequences from DAVIS 2016 dataset, which simulate cases where the object stops moving.

The loss function is used to compute the error between the result and the ground truth. During the training phase, a network learns all the parameters via minimizing errors. The "cross entropy" is commonly used for methods DHSNet, SegFlow, LMP, LVO, NRF, DSS, SPD, WSS, NLDF, AFNet and CPD. Given the generated saliency map S and the ground truth G, the cross entropy loss P is given by Eq (1).

$$\mathrm{P} = -\sum_{i=1}^{h1\times w1} (g_i \log s_i + (1-g_i)\log(1-s_i)) \tag{1}$$

where h_1 is the frame height, w_1 is the frame width, $g_i \in G$ and $s_i \in S$. Since the numbers of salient and non-salient pixels are not balanced, the "balanced

cross entropy", given by Eq (2), is more commonly used for methods OSVOS, SCNN and SFCN.

$$\mathrm{P} = -\sum_{i=1}^{h1\times w1}((1-\alpha)g_i \log s_i + \alpha(1-g_i)\log(1-s_i)) \tag{2}$$

where α is the ratio of the number of salient pixels in ground truth G over that of all pixels in G. Besides, the NLDF adds a boundary Intersection over Union (IOU) loss, given by Eq (3), for SOD.

$$\mathrm{IOUloss} = 1 - \frac{2\,|G_b \bigcap S_b|}{|G_b| + |S_b|} \tag{3}$$

G_b and S_b are contours pixels of G and S respectively, which are obtained using the magnitude of Sobel operator followed by a tanh activation. The AFNet adds Euclidean loss for enhancing boundary. The SegFlow uses endpoint error (EPE) loss to optimize the optical flow branch. In order to prevent learning high responses at all locations, the WSS applies sparse regularization on the generated saliency map ($\|S\|_1$) to reduce background noise during pre-training phases. To keep more details of the information, the DSS and the OSVOS add the side-out supervision for each side output of the backbone. The AFNet and the DHSNet supervise the intermediate maps by the ground truth.

2.3. An overview of the categorization

In order to better determine the difference between these multiple approaches, we compare 15 algorithms at a functional level. The details can be found in Fig. 5 and Fig. 6.

3. Experimental evaluation

This section firstly reviews the most popular datasets and metrics in video SOD, and secondly assesses the performance of the methods introduced in Section 2.

Domain	Methods	Multi-task*	Deep					
			Architecture	Strategy	Backbone	Training dataset		Loss
						Weakly-supervised	Supervised	
Spatial	SPD	Edge detection	Bottom-up/Top-down	End-to-end	VGG16		DUTS	Cross-Entropy
	OSVOS	Edge detection	Side-fusion	End-to-end	VGG16		DAVIS 2016	Balanced cross-entropy
	NLDF	Boundary detection	Side-fusion	End-to-end	VGG16		MSRA-B	Cross-Entropy, Intersection over Union
	DSS		Side-fusion	End-to-end	VGG16		MSRA-B	Cross-Entropy
	DHSNet		Bottom-up/Top-down	End-to-end	VGG16		MSRA10K,DUT-OMRON	Cross-Entropy
	WSS	Image classification	Bottom-up/Top-down	Multi-stage	VGG16	Pseudo pixel-level label, ImageNet		Cross-Entropy
	AFNet	Boundary detection	Bottom-up/Top-down	End-to-end	VGG16		DUTS	Cross-Entropy, Euclidean
	CPD		Bottom-up/Top-down	End-to-end	VGG16		DUTS	Cross-Entropy

Domain	Methods	Multi-task*	Deep					
			Architecture	Strategy	Backbone	Training dataset		Loss
						Weakly-supervised	Supervised	
Temporal	LMP		Bottom-up/Top-down	End-to-end			FlyingThings3D	Cross-Entropy

Domain	Methods	Multi-task*	Deep						Handcrafted
			Architecture	Strategy	Backbone	Training dataset		Loss	
						Weakly-supervised	Supervised		
Spatio-temporal	SCOMd		*Image SOD	Off-the-shelf	VGG16		MSRA-B		Yes
	NRF		Bottom-up/Top-down	End-to-end	VGG16		HKU-IS,MSRA10K,CSSD,DUT-OMRON	Cross-Entropy	Yes
	SCNN		*Image SOD	Off-the-shelf	Alexnet		MSRA-B		Yes
			Side-fusion	Multi-stage	VGG16	Pseudo pixel-level label	MSRA10K,SegTrackV2,FBMS	Balanced cross-entropy	Yes
	LVO		*Semantic segmentation	Off-the-shelf	VGG16		PASCAL VOC 2012		
			*Moving object segmentation	Off-the-shelf			FlyingThings3D		
			Convolutional gated units	End-to-end			DAVIS 2016, Synthetic dataset from DAVIS 2016	Cross-Entropy	
	SegFlow	Optical flow estimation	Side-fusion	Multi-stage	ResNet-101, FlowNetS		DAVIS 2016,MPI Sintel,KITTI,Scene Flow	Cross-Entropy,Endpoint error	
	SFCN		Bottom-up/Top-down	Multi-stage	VGG16		MSRA10K,SegTrackV2,DUT-OMRON,FBMS, Synthetic dataset from MSRA10K and DUT-OMRON	Balanced cross-entropy	

*: only the tasks that different from salient object detection are listed.

Figure 5: Algorithms comparison at a functional level.

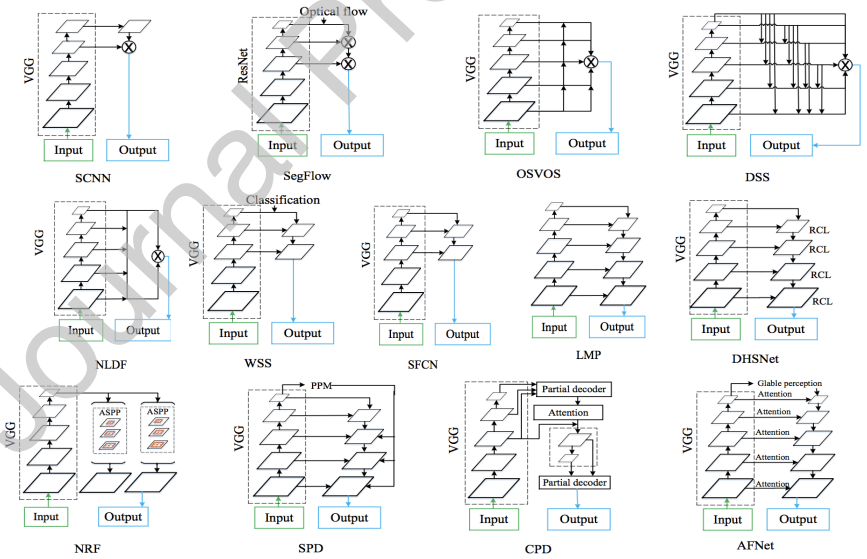


Figure 6: Examples of the architectures. RCL: recurrent convolutional layers, ASPP: "à trous" pyramid pooling, PPM: pyramid pooling module.

3.1. Benchmark datasets and evaluation metrics

Benchmark datasets: the VOS [11] dataset is a recently published large dataset for SOD in videos, which is based on human eye fixation. These videos are grouped into two subsets: VOS-E and VOS-N. Due to the limited number of large-scale datasets designed for SOD in videos, existing methods usually use other datasets from highly related domains like the dataset hereafter. The Freiburg-Berkeley Motion Segmentation (FBMS) dataset [11] is designed for moving object segmentation. Moving objects attract large attention and thus can be regarded as salient objects in videos. As in the methods [12, 22, 18], we also use the 30 test videos with the provided ground truth. The DAVIS 2016 dataset [27] is a popular video dataset for video foreground segmentation. It is divided into two splits: the training (30 sequences) part used for training only and the validation (20 sequences) part for the inference. Though DAVIS 2016-val dataset is designed for video foreground segmentation, it is also widely used for SOD in videos, because of their foreground properties (most of the objects in the video sequences have distinct colors, which can be regarded as salient objects). DAVIS 2017-val [28] is mainly an extension of DAVIS 2016-val dataset (10 new video sequences), which also used for inference in this work.

Similar to [27, 11], we make comparisons of benchmark datasets from aspects of dataset statistics, salient object categories and video attributes. From Table 2, we can observe that the VOS dataset is the largest dataset, while the FBMS is with sparsely-sampled annotated frames.

Table 2: Dataset statistics.

	VOS	VOS-E	VOS-N	FBMS	DAVIS 2016-val	DAVIS 2017-val
#Sequence	200	97	103	30	20	30
#Frame	116103	49206	66897	13860	1376	1999
#Ground truth	7467	3236	4231	720	1376	1999
Resolution (in pixel)	[408,800]	[408,800]	[448,800]	[350,960]	[480,854]	[480,854]
Year	2018	2018	2018	2014	2016	2017

Salient object categories are compared to explore the content diversity. In

FBMS, the humans, animals, vehicles are evenly distributed. DAVIS 2016-val and DAVIS 2017-val datasets consist of more objects and actions. For the creation of VOS dataset, videos are collected by volunteers from video-sharing websites without giving any instructions on the video contents, which significantly increases the object diversities and shape complexities. The area ratio distribution of salient objects per dataset is demonstrated in Fig. 7. Small and medium salient objects are uniformly distributed in all datasets, and large objects mainly appear in VOS datasets.

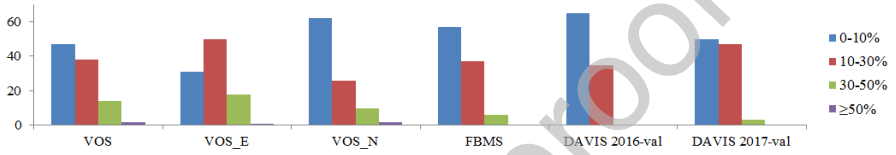


Figure 7: (Better viewed in color) Histogram of the area ratio of salient objects per dataset: the x axis represents the bins regarding average area ratio of salient objects per frame in one video sequence, and y axis is the percentage of total video sequences.

Video attributes, representing specific situations, are important to influence the video salient object detection. The VOS-E dataset contains obvious salient objects with slow camera motion, while the VOS-N dataset presents multiple complex scenes, highly dynamic objects and motion blur. The FBMS mainly provides challenges cases such as fast motion and occlusion. DAVIS 2016-val and DAVIS 2017-val datasets provide multiple balanced video attributes such as appearance change, camera-shake, background cluster, out-of-view [27], deformation [27], etc.

Evaluation metrics: for salient object detection, various metrics are used to measure the similarity between the generated saliency map S and the ground truth G.

- Precision, Recall, F_β: an adaptive threshold T is used for binarizing S to a mask M:

$$\text{Precision} = \frac{|M \bigcap G|}{|M|}, \text{Recall} = \frac{|M \bigcap G|}{|G|}, F_\beta = \frac{(1+\beta^2) \times (\text{Precision} \times \text{Recall})}{(\beta^2 \times \text{Precision} + \text{Recall})} \tag{4}$$

F_β comprehensively considers both Precision and Recall and is computed as the weighted harmonic mean of Precision and Recall. β^2 is set to 0.3 (commonly used to weight precision more than recall as proposed by Achanta *et al.* [29]). F_2 (β=2) which weights recall more than precision, and F_1 (β=1) that weights recall and precision equally are also used. The threshold T is set to be the minimum value between T_α' and T_α in our experiments as in NRF.

$$T'_\alpha = max(S(i)) \quad 1 \leq i \leq h1 \times w1, \quad T_\alpha = \frac{2}{h1 \times w1} \sum_{i=1}^{h1 \times w1} S(i) \tag{5}$$

where h_1 is the frame height, w_1 is the frame width. A higher F_β means a better performance.

- P-R curve [8]: S is converted to a binary mask M via a threshold that varies from 0 to 255. For each threshold, a pair of (Precision, Recall) values are computed which are used for plotting P-R curve. The curve closest to the upper right corner (1.0, 1.0) corresponds to the best performance.

- Mean Absolute Error (MAE): computed as the average absolute difference between all pixels in S and G. It considers the true negative saliency assignment, i.e., the pixel correctly masked as non-salient [8]. A smaller MAE value means a higher similarity and a better performance.

$$MAE = \frac{1}{h1 \times w1} \sum_{i=1}^{h1 \times w1} |G(i) - S(i)| \tag{6}$$

For video SOD evaluation, the metrics values are firstly computed over each video, and secondly computed the mean values over all videos in each dataset.

3.2. Experimental comparison and results analysis

In this part, large-scale datasets (including FBMS, VOS-E, VOS-N, VOS, DAVIS 2016-val and DAVIS-2017-val) are used. Metrics (including MAE, Recall, Precision, F_β, F_1, F_2 and P-R curve) are used to evaluate saliency methods (SCOMd, SFCN, SFCNs, DHSNet, NLDF, WSS, DSS, SPD and SCNN) and

metrics (including MAE, Recall, Precision, F_β, F_1 and F_2) are used to evaluate segmentation methods (LMP, LVO, SegFlow, NRF and OSVOS).

For methods SCOMd and SCNN, without published source codes, the results (only for FBMS and DAVIS 2016-val datasets) are those reported by the authors. For other methods, applied to all datasets, the results are generated using the provided source codes. When the authors give their results, we just report these results even if they provide their code. For network inputs of the methods LMP and LVO, the computer flow vector is generated by the method proposed by Tripathi *et al.* [30].

3.2.1. Performance on the VOS-E dataset

Fig. 8 shows the performance on the VOS-E dataset. The methods DHSNet, NLDF, NRF, SFCN, SFCNs, WSS, DSS and SPD, based on backbone networks, all get high Precision, high Recall and high F_β scores. The DHSNet and the SPD also get the best P-R curve, and the NRF and the SPD get the best MAE value. Most of these methods only detect the salient object from spatial domain, which shows that spatial saliency detection has a good performance for SOD on video dataset with slow camera motions.

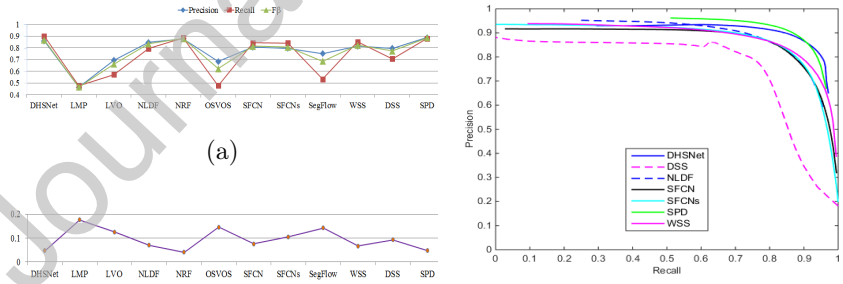


Figure 8: (Better viewed in color) Performances on the VOS-E dataset: (a) $F_\beta\uparrow$, Precision↑, Recall↑, (b) MAE↓, (c) P-R curve. ↑ means the higher the better and ↓ means the lower the better.

3.2.2. Performance on the FBMS dataset

Fig. 9 presents the performances on FBMS dataset. The SPD gets best

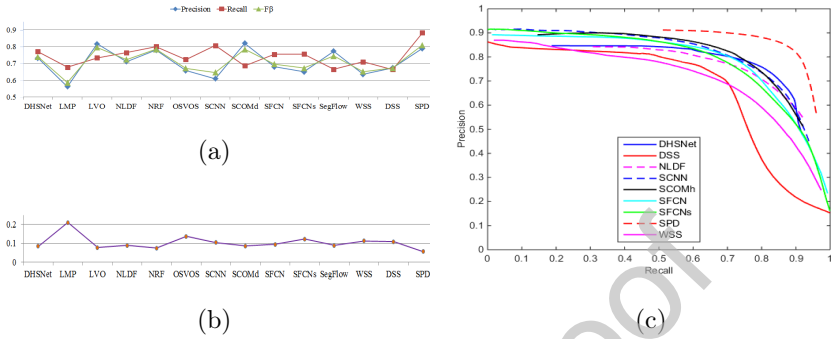


Figure 9: (Better viewed in color) Performances on the FBMS dataset: (a) $F_\beta\uparrow$, Precision↑, Recall↑, (b) MAE↓, (c) P-R curve.

scores on all metrics, which further verifies the effectiveness of the SPD. The SCNN gets high Recall score, and the SCOMd gets high Precision score, and the LVO gets high F_β score, and the SegFlow gets low MAE value. They not only detect the salient object from spatial domain, but also from temporal domain or fused spatio-temporal domain, which indicates that the temporal detection plays a significant role for SOD on video dataset with highly dynamic foreground objects.

3.2.3. Performance on the VOS-N and VOS dataset

Fig. 10 and Fig. 11 show the performances on the VOS-N and the VOS datasets respectively.

Salient objects in these two datasets are obtained according to the saliency fixation, which is similar with that in image SOD datasets. That may explain why the methods (e.g. DHSNet, NRF, NLDF, SFCN, SFCNs, SPD and WSS) trained from image SOD datasets get better Recall scores than others.

3.2.4. Performance on the DAVIS 2016-val and DAVIS 2017-val dataset

Fig. 12 shows the performances on the DAVIS 2016-val dataset. Fig. 13 shows the performances on the DAVIS 2017-val dataset.

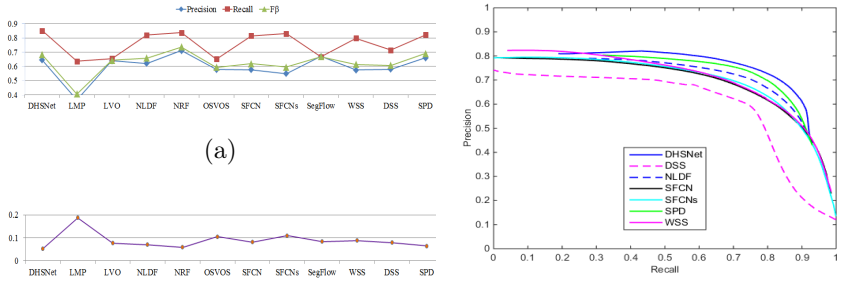


Figure 10: (Better viewed in color) Performances on the VOS-N dataset: (a) $F_\beta\uparrow$, Precision↑, Recall↑, (b) MAE↓, (c) P-R curve.

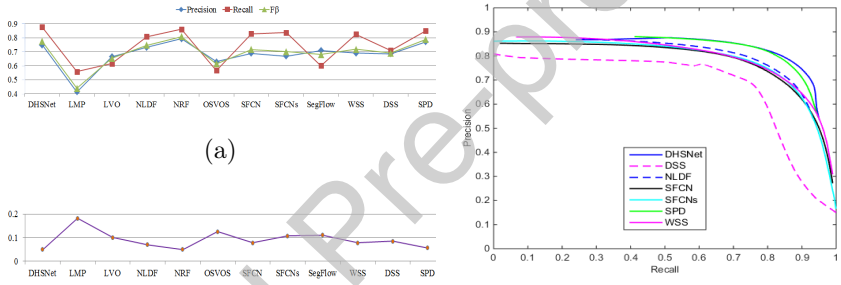


Figure 11: (Better viewed in color) Performances on the VOS dataset: (a) $F_\beta\uparrow$, Precision↑, Recall↑, (b) MAE↓, (c) P-R curve.

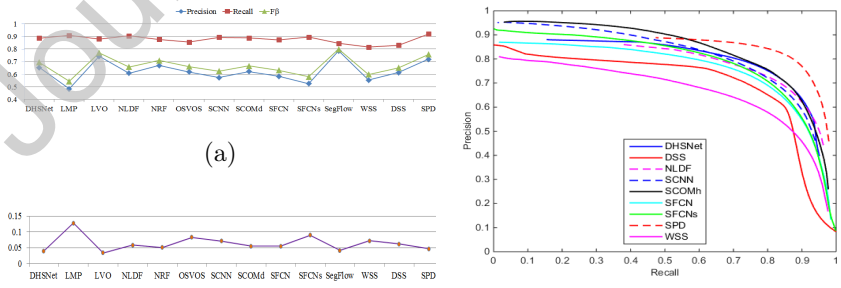


Figure 12: (Better viewed in color) Performances on the DAVIS-2016-val dataset: (a) $F_\beta\uparrow$, Precision↑, Recall↑, (b) MAE↓, (c) P-R curve.

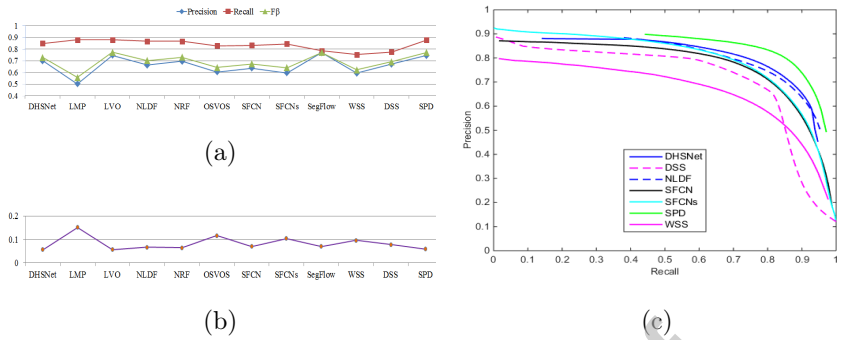


Figure 13: (Better viewed in color) Performances on the DAVIS-2017-val dataset: (a) $F_\beta\uparrow$, Precision↑, Recall↑, (b) MAE↓, (c) P-R curve

The methods that detect saliency from two domains (e.g. the LVO, the NRF, the SegFlow) perform better than those only from one domain (e.g. the LMP, the OSVOS, the WSS), which shows that saliency from two domains is more efficient for SOD on complex videos datasets. Weakly supervised methods (e.g. the SCNN and the WSS) get a little lower recall and F_β values. The methods (e.g. the LVO and the SegFlow) are trained from object segmentation datasets only, which shows the effectiveness of using the training datasets from closely related domains. All methods achieve high Recall scores, which shows that salient objects in these datasets are easy to be detected. Besides, if we compare the SFCNs with the SFCN, we can find that they use the same deep-learning network but with different training datasets. The input of the former one is each frame with provided ground truth, while the input of the later one is the video sequence and the detection results from the SFCNs. Thus, the SFCN refines the output of the SFCNs, by learning more deep features from the temporal domain. If we compare the LMP and the LVO, we can find that the LVO uses the same saliency detection from temporal domain as the LMP but with extra deep spatial saliency information, and deep fused spatio-temporal features. It helps the LVO to achieve a much better performance than the LMP, which also further prove that saliency detection from two domains is significant for SOD in videos.

3.2.5. Global performance on various datasets

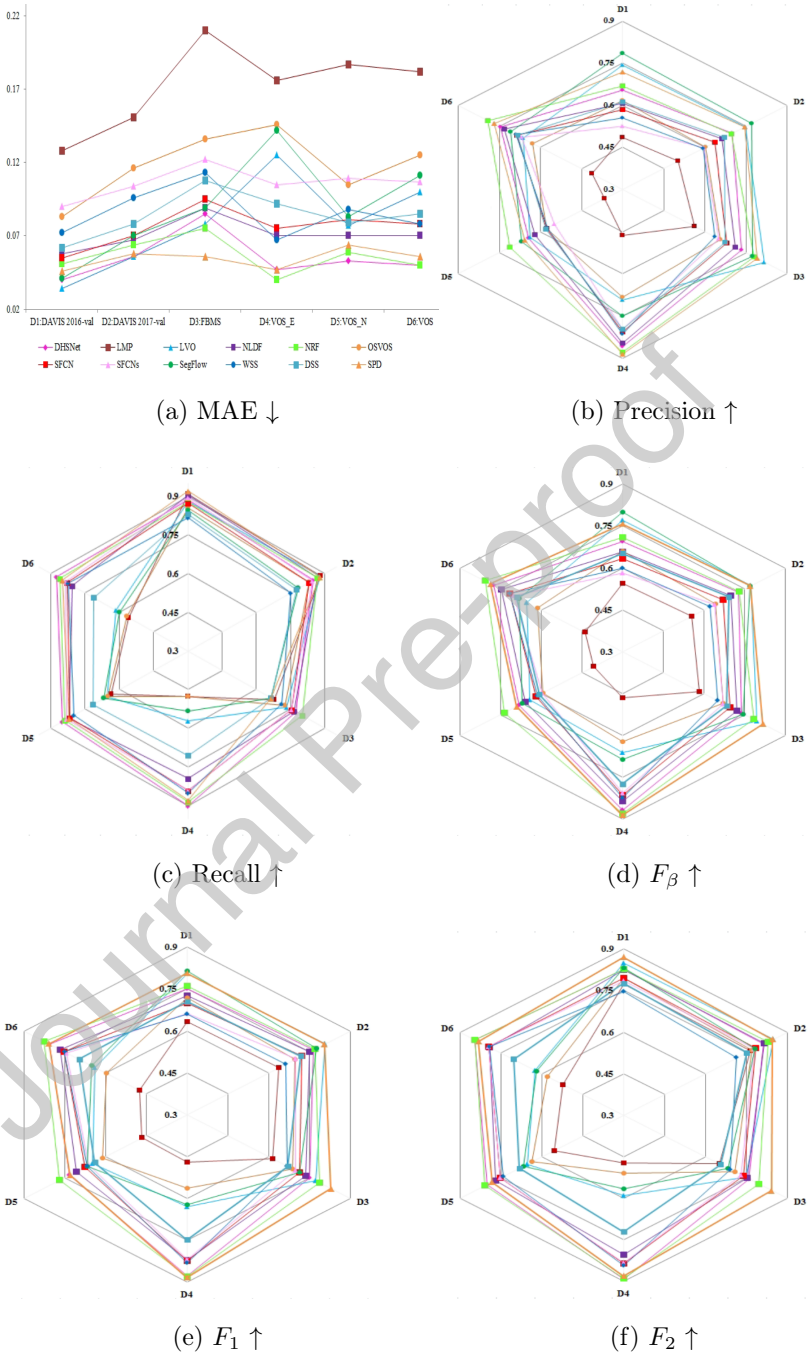
In order to catch the global view of the performance of a method on various datasets, the following Fig. 14 (a) shows the comparative results of the methods for MAE metric on 6 datasets. As can be seen on this figure, methods perform worse on dataset FBMS.

Fig. 14 (b-f) shows the comparative results of the methods for Precision, Recall, F_β, F_2 and F_1 metrics on different datasets. In each figure, the radar chart contains various closed curves, where each curve shows the performance of a method on the datasets. The area of the closed curve can reflect the performance of the method on the whole datasets. The larger the area the better the performance. Table 3 shows the detailed areas of these curves (corresponding to the methods) in Fig. 14 (b-f) respectively.

Table 3: Area of each method in the Fig 14 (b-f). (The best score is in **bold**)

Metric↑	DHSNet	LMP	LVO	NLDF	NRF	OSVOS	SFCN	SFCNs	SegFlow	WSS	DSS	SPD
Precision	1.3505	0.5650	1.3388	1.2562	1.4768	1.0244	1.1332	1.0282	1.4354	1.0685	1.1595	**1.5021**
Recall	**1.8987**	1.2482	1.3665	1.7740	1.8966	1.2212	1.7656	1.8061	1.2142	1.6278	1.3974	1.8075
F_β	1.4477	0.6519	1.3356	1.3436	1.5561	1.0447	1.2384	1.1471	1.3690	1.1635	1.2104	**1.5989**
F_1	1.5803	0.7899	1.3412	1.4657	1.6564	1.0862	1.3868	1.3205	1.3052	1.2951	1.2590	**1.7063**
F_2	1.7573	1.0184	1.3516	1.6341	1.7997	1.1563	1.5938	1.5775	1.2463	1.4790	1.3285	**1.8441**

Fig. 14 (b), (d), (e) and (f) show that methods achieve highest Precision, F_β, F_1 and F_2 scores on VOS-E dataset, which is reasonable since VOS-E dataset contains slow camera motion. Fig. 14 (c) presents that methods achieve lowest Recall scores on FBMS dataset. We can learn that the deep-learning technique provides a poorly ability to detect moving salient objects from dynamic background. From Table 3, one can observe that the DHSNet gets good Recall score and the SPD obtains good Precision, F_β, F_1 and F_2 scores, while the LMP performs not very well. We can firstly find that the end-to-end trained networks, the DHSNet and the SPD, are efficient to learn and detect the salient object. We secondly observe that though temporal saliency is significant, saliency information only detected from the temporal domain is not enough.



(a) MAE ↓ (b) Precision ↑

(c) Recall ↑ (d) F_β ↑

(e) F_1 ↑ (f) F_2 ↑

Figure 14: (Better viewed in color) Global performance on various datasets.

3.2.6. Least and most difficult scenes

To achieve a more in-depth analysis, it would be interesting to explore least and most difficult scenes for the compared methods.

Least difficult scenes: Fig. 15 presents the video sequences in which all compared algorithms achieve similar high accuracies according to two metrics (MAE < 0.5 and $F_\beta > 0.5$). The metric threshold is chosen to 0.5 for the sake of the balance between a high accuracy and a considerable quantity of video sequences. The compared algorithms find similar solutions for saliency accuracy on these sequences. The background are mainly simple and static, and the camera motion is slow. Salient objects are almost belonging to the humans, animals and vehicles categories, which mostly appear in the training sets. Whereas, some tricky video attributes, e.g., background cluster, deformation, fast motion, out-of-view, are more or less found.

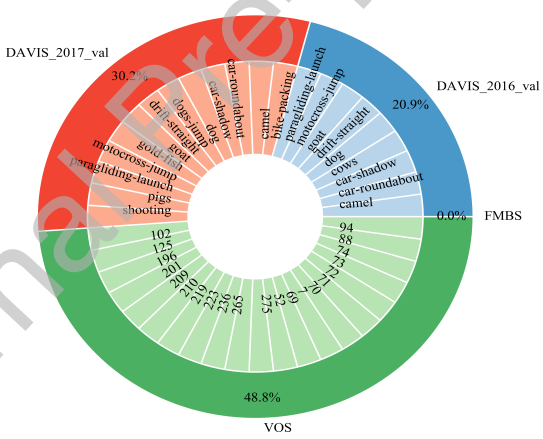


Figure 15: (Better viewed in color) Examples of video sequences (with detailed name) that all compared algorithms achieve similar high accuracies.

Most difficult scenes: Fig. 16 summaries the number of failure video sequences of each method, which are selected using two metrics (MAE > 0.5 or $F_\beta < 0.5$). For the failure sequences, we seek for the breakdown factors of each method on aspects of "false positives (FP)" vs "false negative (FN)", as is shown in Fig 17. The same saliency threshold (in Eq (5)) is adopted. The LMP,

the LVO and the SegFlow wrongly predict more non-salient pixels to be salient. The DHSNet and the NRF detect the main body of the salient object, and miss the smallest part of the salient object. The DHSNet and the SPD, keeping the small number of FP and FN pixels, perform better than others.

Figure 16: (Better viewed in color) The number of failure video sequences of each method in four datasets (MAE > 0.5 or $F_\beta < 0.5$).

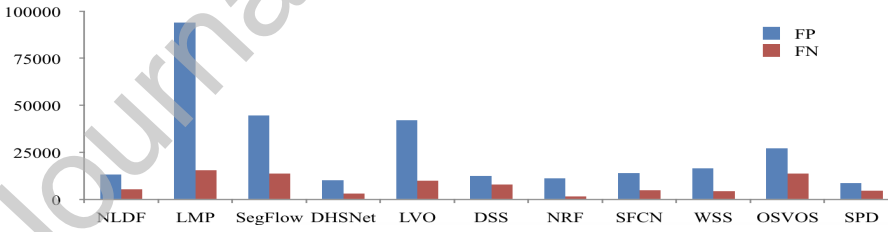


Figure 17: (Better viewed in color) The average number of pixels belong to "False positives (FP)" or "false negative (FN)" per frame.

3.2.7. Speed performance

For different models, the training time (obtained from the published paper or provided by authors) is listed in Table 4.

A PC with a NVIDIA 1080 GPU is used for testing the speed of the methods on the DAVIS-2016-val dataset. For different models (except SCOMd and SCNN with unpublished codes), the average run-time is listed in Table 4. We can observe that the DSS has the least computation costs, which is similar to that of the OSVOS, the SFCN, the DHSNet and the NLDF. Methods SegFlow, NRF, LMP and LVO are much more time-consuming.

Table 4: Training time in hours, average run time in seconds (per frame) of the compared models. (The best run time score is in **bold**, "-" indicates that the time is not available.)

Methods	DHSNet	LMP	LVO	NLDF	NRF	OSVOS	SFCN	SegFlow	WSS	SCNN	DSS	SPD
Train (h)	17.7	<24	<24	9	15	16	40	>30	>30	12	8	9
Test(s)↓	0.069	0.2	0.42	0.091	0.297	0.072	0.072	0.174	0.067	-	**0.056**	0.092

3.3. Ablation study

To better analyse and understand algorithms, we try to retrain advantageous and representative networks in Section 3.3.1. Ablation experiments related to domain shifts influences are performed in Section 3.3.2.

3.3.1. Retrain advantageous and representative networks

The representative networks, i.e., the OSVOS, the DSS and the SPD, are selected. The training codes are provided by the authors and these networks are with various architectures. The OSVOS fuses each side output of the backbone network together through upsampling; the DSS adds multiple short connections from deeper side outputs to the shallower ones; and the SPD adds short connections from the encoder features to the mirror decoder features, and refines the high-level semantic features by adding pyramid pooling module (PPM).

We try to retrain these networks onto the same training set, and break apart these algorithms in terms of crucial variants, such as data augmentation, short connections, loss function, etc. We evaluate the retrained models onto the same test set to conclude the effectiveness of components used in training.

We adopt the DAVIS 2016-train dataset for training salient object detection and the DAVIS 2016-val for performance evaluation. All networks are imple-

mented using PyTorch. VGGNet is chosen to be the backbone and experiments are trained for 50 epochs under the default hyper-parameter settings.

As is shown in Table 5, we conduct different variants, and the corresponding results are summarized.

Table 5: Comparisons of the performance under different settings. CE: cross entropy, BCE: balanced cross entropy, HF: horizontal-flipping, R: resizing, SC: short connections, SS: side output supervision, PPM: pyramid pooling module. ("x" indicates that the method is not based on corresponding technique, "~~~" indicates that the score is better than that of the baseline $OSVOS_1$, DSS_1 or SPD_1).

Model		Pretrain	Loss		Deformation	Augmentation		Metrics		
			CE	BCE		HF	R	Precision↑	Recall↑	MAE↓
OSVOS	1	✓	x	✓	x	x	x	0.664	0.771	0.119
	2	✓	x	✓	x	✓	x	0.653	0.785	0.115
	3	✓	x	✓	x	x	✓	0.663	0.655	0.103
	4	x	x	✓	x	x	x	0.484	0.614	0.136
DSS	1	✓	✓	x	x	x	x	0.517	0.808	0.061
	2	✓	✓	x	x	✓	x	0.466	0.645	0.074
	3	✓	✓	x	x	x	✓	0.631	0.839	0.050
	4	x	✓	x	x	x	x	0.304	0.439	0.083
	5	✓	x	✓	x	x	x	0.492	0.790	0.068
	6	✓	✓	x	-SC	x	x	0.425	0.871	0.101
	7	✓	✓	x	-SS	x	x	0.467	0.560	0.206
SPD	1	✓	✓	x	x	x	x	0.644	0.813	0.054
	2	✓	✓	x	x	✓	x	0.460	0.842	0.076
	3	✓	✓	x	x	x	✓	0.571	0.551	0.081
	4	x	✓	x	x	x	x	0.581	0.708	0.062
	5	✓	x	✓	x	x	x	0.688	0.832	0.042
	6	✓	✓	x	-SC	x	x	0.695	0.799	0.046
	7	✓	✓	x	-SC -PPM	x	x	0.620	0.782	0.063
	8	✓	✓	x	-PPM	x	x	0.478	0.722	0.086

- We retrain the three networks from scratch without loading pretrained weights from ImageNet. It can be observed from Table 5 that the performance of the retrained networks ($OSVOS_4$, DSS_4 and SPD_4) with random initialized weights are worse than that of the baselines ($OSVOS_1$, the DSS_1 and the SPD_1), which explains that initializing the VGG network with pretrained weights from the classification task could help to achieve performance gains for salient object detection. Whereas, the performance

might become competitive if more and clean samples are used for training from scratch [31].

- We apply data augmentation techniques on-the-fly to the retraining, which might prevent over-fitting the training sets. Experimental results with horizontal-flipping or resizing show that, the DSS_3 increases the accuracy, while the accuracy of the DSS_2 and the SPD_3 decrease.

- We explore the two frequently-used loss in video SOD - cross entropy (CE) and balanced cross entropy (BCE) in ablation experiments. For the OSVOS, we find it fails to work with the CE, which might be caused by the imbalance between salient object and non salient region. Using BCE, the DSS_5 are worse on all metrics.

- For the DSS, without deep supervision for each side of the output, DSS_7 experiences a dramatic drop, only obtaining MAE of 0.206. Without short connections DSS_6 decreases the accuracy. However, the SPD_6 is still with high accuracy without short connections. The results of the SPD_7 and the SPD_8 illustrate that pyramid pooling module contributes most significantly to the performance.

3.3.2. Domain shifts influences

The performance of deep learning-based methods may degrade if the characteristics of the images in the target domain vary from that in the source domain, e.g. the contrast, brightness, etc. It is non-trivial to learn whether deep models are invariant to different domains. Thus, we investigate the influences of the domain shifts on different variants of models by simulating the characteristic/environmental changes in the test images.

The experimental domain shifts are implemented by adding the Gaussian noises and perturbing the contrast and brightness onto the test images. For Gaussian noise, the standard deviation of the noise is progressively increased by 0.02, ranging from 0.0 to 0.16; for brightness, we randomly increase the hue value by 0, 5, 10, 15; for contrast, we randomly set the enhancement value to

0, 30, 60, 90. We conduct experiments on different variants of the OSVOS, the DSS and the SPD, and the experimental results are summarized below.

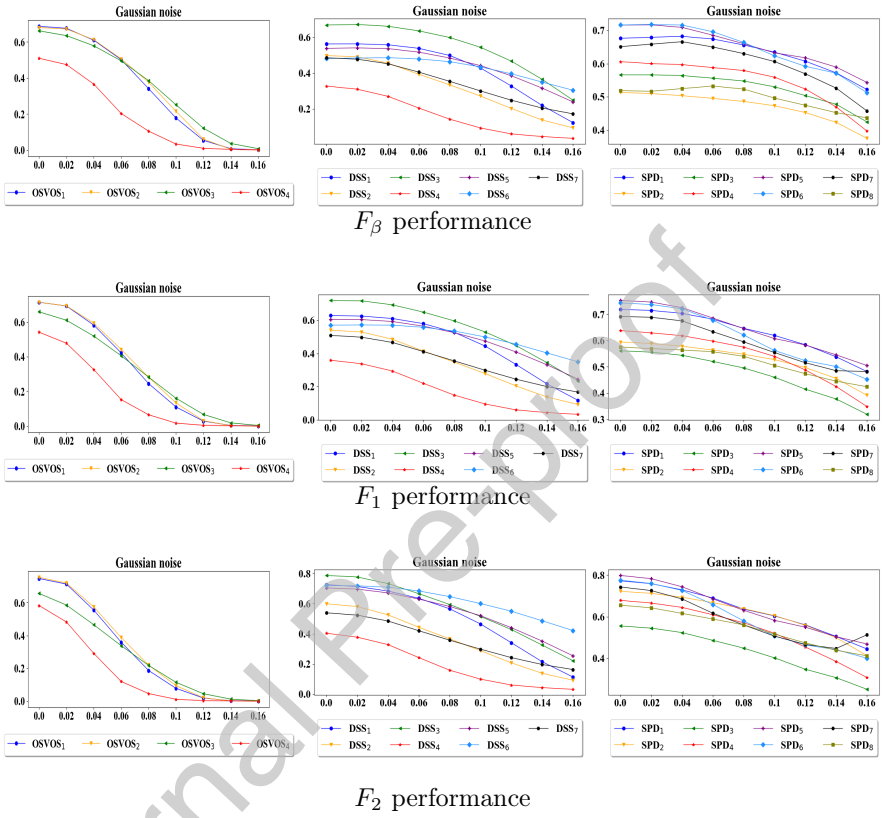


Figure 18: The influences of Gaussian noise. The F_β, F_1, F_2 performance of each model over different standard deviation of the noise which are increased from 0.0 to 0.16.

- For Gaussian noise (as in Fig. 18), the performances of all ablation models degrade with the increased standard deviation of the noise. Compared with the OSVOS and the DSS, the quality of saliency maps estimated by the SPD observe smaller drops in general. With data augmentation techniques, the performance of the OSVOS_3 and the DSS_3 descend a bit slower and less than that of the baselines (OSVOS_1 and DSS_1). When the standard deviation of the noise is large, the DSS_5 using the balanced cross entropy (BCE) performs better than that of using cross entropy (CE)

Figure 19: The influences of random perturbation of brightness and contrast.

loss (DSS_1). Compared with (DSS_1 and SPD_1), the models without side-output supervision (DSS_7) or without Pyramid pooling module (SPD_8) stay less robust against noise.

- Fig. 19 demonstrates the sensitivity of each model to the various choices of brightness and contrast. In the figure, for each model, three pairs of vertical lines are provided to illustrate the F_β, F_1 and F_2 scores over multiple input images. For each pair, the y-axis value of the dot is the mean score, the length of the color line shows the doubled standard deviation, and the top end of the black line is the differences between the maximun and mean metric value, while the bottom end is that between the mamininum and mean metric value. In general, compared with the OSVOS and the SPD, the quality saliency maps estimated by the DSS are more stable with smaller standard deviation. For the random perturbation, these

models are rarely affected by brightness, but relatively unrobust against the contrast effect. Retraining networks from scratch without loading pre-trained weights from ImageNet, the $OSVOS_4$ and the DSS_4 suffer the most with lowest mean metric values. The SPD_8 is more sensitive due to the removed pooling in feature hierarchy.

3.4. Promising future works

Deep-learning based video SOD approaches have greatly improved the accuracy and efficiency for this field. However, there are still some interesting but challenging works to be considered regarding:

-weakly-supervised networks: weakly-supervised models that do not rely on large pixel-wise labels attract much attention in recent years. However, its accuracy is still far from satisfactory. To address theses issues, it may be possible to obtain large amounts of weakly labeled datasets and design weakly supervised triplet ranking loss as in [32]. Mining pseudo ground truth [33] may also be considered and developed to enhance the performance of weakly-supervised video SOD.

-visual attention: attention mechanism introduced in machine translation is recently evolved in closely related vision and video processing tasks, e.g., attention weighted CNN features in video captioning [34], spatial attention, temporal attention and channel-wise attention in visual tracking [35], attentive feedback modules and the attention guidance in image SOD [25, 26], which considerably enhance the accuracy by boosting the representative power of CNNs. It is valuable to put efforts into visual attention mechanism for video SOD to achieve more promising accuracy.

-spatial-temporal saliency learning: existing video SOD methods predominantly rely on spatial features. Other methods that take into consideration temporal features usually use the optical flow information. [36, 37] explore recurrent module and long-short term module, and [38] uses 3D filters. However, the exploration of networks to learn spatio-temporal representations remains limited. Inspired by recent advances in related video tasks, e.g., dense connections for

spatio-temporal interaction in action recognition [39] and space-time memory block in video object segmentation [40], we possibly consider these techniques to avoid the latent dependency issues in video SOD. Learning spatial-temporal features in an end-to-end manner is important for further accuracy improvement.

-knowledge from traditional methods: some deep-learning based SOD models derive their good performance or gains from well-established knowledge of traditional methods. [41, 42] put forward the extraction of contrast information, which is similarly encoded as a contrast layer by [16]. [2] proposes to fuse local and global features for improvements, which is similar to the fusion or guidance modules in recent image SOD networks [43, 44] and the video SOD network [45]. Therefore, it is interesting to explore other knowledge from traditional methods and reformulate them into CNNs to learn more representative features.

4. Conclusion

To the best of our knowledge, this is the first overview of deep learning techniques for video SOD. The classification of the state-of-the-art methods is done regarding the involved tasks and the domain of used features, which presents a clear viewpoint of recent development. Deep networks of representative existing methods are introduced and compared in detail. They are surveyed from three points of view: architectures, training details and results. A comparative summary of methods is presented and their performances on various datasets are discussed. The pros and cons of each method are also pointed out. The various experiments conducted show that the methods DHSNet and SPD produce effective/genetic features with state-of-the-art performance on various tested databases. The representative approaches are selected and broken apart to verify the performance of their components independently. Consequently, our thorough analysis of the methodologies and experiments is presented for readers to quickly grasp these representative state-of-the-art methods.

Finally, some promising future directions are discussed, in which we suggest

to consider the improvement of weakly-supervised networks for overcoming the drawbacks of fully labeled training sets; the new visual attention-driven models, the explorations of temporal saliency features and spatio-temporal saliency features, and the knowledge from traditional methods for improving accuracy. In general, this survey is expected to pave a way to study the existing deep-learning based video SOD methods and provide promising research directions for future exploration in video SOD.

5. Acknowledgment

The authors would like to thank the pioneer researchers for the available algorithms. The authors would also like to express their sincere appreciation to the reviewers for their comments and suggestions.

Vitae

Qiong Wang received the Ph.D. degree with the "National Institute of Applied Sciences of Rennes", Rennes, France, in 2019. She is currently a faculty of Zhejiang University of Technology. Her current research interests include visual saliency detection, video object segmentation, and deep learning.

Lu Zhang received the Ph.D. degree from University of Angers, Angers, France, in 2012. She is currently an Associate Professor in the "National Institute of Applied Sciences of Rennes" in France. Her research interests include visual saliency detection, multimedia quality assessment, medical imaging and human perception understanding.

Yan LI is currently pursuing the Ph.D. degree with Université libre de Bruxelles, Belgium. His current research interests include depth estimation, light field, and deep learning.

Kidiyo Kpalma received his Ph.D in Image Processing INSA Rennes in 1992. Since 2014, he became Professor at INSA: he teaches Signal and Systems, Signal Processing and DSP. As a member of IETR UMR CNRS 6164,

his research interests include pattern recognition, semantic image segmentation, facial micro-expression and salient object detection.

References

[1] H. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, Journal of vision 9 (12) (2009) 1–27.

[2] Z. Tu, Z. Guo, W. Xie, M. Yan, R. C. Veltkamp, B. Li, J. Yuan, Fusing disparate object signatures for salient object detection in video, Pattern Recognition 72 (2017) 285–299.

[3] X. Zhi, H. Shen, Saliency driven region-edge-based top down level set evolution reveals the asynchronous focus in image segmentation, Pattern Recognition 80 (2018) 241–255.

[4] Y. Zheng, L. Jiao, H. Liu, X. Zhang, B. Hou, S. Wang, Unsupervised saliency-guided SAR image change detection, Pattern Recognition 61 (2017) 309–326.

[5] F. Yan, Autonomous vehicle routing problem solution based on artificial potential field with parallel ant colony optimization (ACO) algorithm, Pattern Recognition Letters 116 (2018) 195–199.

[6] S. Xie, H. Hu, Y. Wu, Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition, Pattern Recognition 92 (2019) 177–191.

[7] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: A benchmark and algorithms, in: Computer Vision - ECCV 2014 - 13th European Conference, 2014, pp. 92–109.

[8] A. Borji, M. Cheng, Q. Hou, H. Jiang, J. Li, Salient object detection: A survey, Computational Visual Media 5 (2) (2019) 117–150.

[9] A. Borji, M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Processing 24 (12) (2015) 5706–5722.

[10] J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: A survey, IEEE Signal Process. Mag. 35 (1) (2018) 84–100.

[11] J. Li, C. Xia, X. Chen, A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection, IEEE Trans. Image Processing 27 (1) (2018) 349–364.

[12] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, N. Komodakis, SCOM: spatiotemporal constrained optimization for salient object detection, IEEE Trans. Image Processing 27 (7) (2018) 3345–3357.

[13] J. Li, A. Zheng, X. Chen, B. Zhou, Primary video object segmentation via complementary cnns and neighborhood reversible flow, in: IEEE International Conference on Computer Vision, ICCV, 2017, pp. 1426–1434.

[14] N. Liu, J. Han, Dhsnet: Deep hierarchical saliency network for salient object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 678–686.

[15] S. Caelles, K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. V. Gool, One-shot video object segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 5320–5329.

[16] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, P. Jodoin, Non-local deep features for salient object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6593–6601.

[17] P. Tokmakov, K. Alahari, C. Schmid, Learning motion patterns in videos, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 531–539.

[18] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Processing 27 (1) (2018) 38–49.

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

[20] P. Tokmakov, C. Schmid, K. Alahari, Learning to segment moving objects, International Journal of Computer Vision 127 (3) (2019) 282–301.

[21] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 3796–3805.

[22] Y. Tang, W. Zou, Z. Jin, Y. Chen, Y. Hua, X. Li, Weakly supervised salient object detection with spatiotemporal cascade neural networks, IEEE Trans. Circuits Syst. Video Techn. 2018.

[23] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, P. H. S. Torr, Deeply supervised salient object detection with short connections, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2019) 815–828.

[24] J. Liu, Q. Hou, M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[25] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 1623–1632.

[26] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 3907–3916.

[27] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. H. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 724–732.

[28] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbelaez, A. Sorkine-Hornung, L. V. Gool, The 2017 DAVIS challenge on video object segmentation, arXiv.

[29] R. Achanta, S. S. Hemami, F. J. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1597–1604.

[30] S. Tripathi, Y. Hwang, S. J. Belongie, T. Q. Nguyen, Improving streaming video segmentation with early and mid-level visual processing, in: IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 477–484.

[31] K. He, R. B. Girshick, P. Dollár, Rethinking imagenet pre-training, in: IEEE International Conference on Computer Vision, ICCV 2019, 2019.

[32] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, J. Sivic, Netvlad: CNN architecture for weakly supervised place recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1437–1451.

[33] Y. Zhang, Y. Bai, M. Ding, Y. Li, B. Ghanem, Weakly-supervised object detection via mining pseudo ground truth bounding-boxes, Pattern Recognition 84 (2018) 68–81.

[34] W. Li, D. Guo, X. Fang, Multimodal architecture for video captioning with memory networks and an attention mechanism, Pattern Recognition Letters 105 (2018) 23–29.

[35] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, H. Lu, Multi attention module for visual tracking, Pattern Recognition 87 (2019) 80–93.

[36] H. Song, W. Wang, S. Zhao, J. Shen, K. Lam, Pyramid dilated deeper convlstm for video salient object detection, in: Computer Vision - ECCV 2018 - 15th European Conference, 2018, pp. 744–760.

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

[38] T. Le, A. Sugimoto, Deeply supervised 3d recurrent FCN for salient object detection in videos, in: British Machine Vision Conference 2017, BMVC, 2017.

[39] W. Hao, Z. Zhang, Spatiotemporal distilled dense-connectivity network for video action recognition, Pattern Recognition 92 (2019) 13–24.

[40] S. W. Oh, J. Lee, N. Xu, S. J. Kim, Video object segmentation using space-time memory networks, in: IEEE International Conference on Computer Vision, ICCV, 2019.

[41] Y. Yan, J. Ren, G. Sun, H. Zhao, J. Han, X. Li, S. Marshall, J. Zhan, Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement, Pattern Recognition 79 (2018) 65–78.

[42] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, S. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.

[43] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection, Pattern Recognition 86 (2019) 376–385.

[44] P. Zhang, W. Liu, Y. Lei, H. Lu, Hyperfusion-net: Hyper-densely reflective feature fusion for salient object detection, Pattern Recognition 93 (2019) 521–533.

[45] T. Le, A. Sugimoto, Video salient object detection using spatiotemporal deep features, IEEE Trans. Image Processing 27 (10) (2018) 5002–5015.

Vitae

Qiong Wang received the Ph.D. degree with the "National Institute of Applied Sciences of Rennes", Rennes, France, in 2019. She is currently a faculty of Zhejiang University of Technology. Her current research interests include visual saliency detection, video object segmentation, and deep learning.

Lu Zhang received the Ph.D. degree from University of Angers, Angers, France, in 2012. She is currently an Associate Professor in the "National Institute of Applied Sciences of Rennes" in France. Her research interests include visual saliency detection, multimedia quality assessment, medical imaging and human perception understanding.

Yan LI is currently pursuing the Ph.D. degree with Université libre de Bruxelles, Belgium. His current research interests include depth estimation, light field, and deep learning.

Kidiyo Kpalma received his Ph.D in Image Processing INSA Rennes in 1992. Since 2014, he became Professor at INSA: he teaches Signal and Systems, Signal Processing and DSP. As a member of IETR UMR CNRS 6164, his research interests include pattern recognition, semantic image segmentation, facial micro-expression and salient object detection.

Declaration of Interest statement

The authors declare that they have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.