This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

Improving Video Saliency Detection via Localized Estimation and Spatiotemporal Refinement

Xiaofei Zhou, Zhi Liu, *Senior Member, IEEE*, Chen Gong, *Member, IEEE*, and Wei Liu

***Abstract*—Video saliency detection aims to pop out the most salient regions in every frame of a video. Up to now, many efforts have been made from various aspects for video saliency detection. Unfortunately, the existing video saliency models are very likely to fail in challenging videos with complicated motions and complex scenes. Therefore, in this paper we propose a novel framework to improve the saliency detection results generated by existing video saliency models. The proposed framework consists of three key steps including localized estimation, spatiotemporal refinement, and saliency update. Specifically, the initial saliency map of each frame in a video is first generated by using an existing saliency model. Then, by considering the temporal consistency and strong correlation among adjacent frames, the localized estimation models, which are generated by training the random forest regressor within a local temporal window, are employed to generate the temporary saliency map. Finally, by taking the appearance and motion information of salient objects into consideration, the spatiotemporal refinement step is deployed to further improve the temporary saliency map and generate the final saliency map. Further, such improved saliency map is then utilized to update the initial saliency map and provide reliable cues for saliency detection in the next frame. The experimental results on four challenging datasets demonstrate that the proposed framework is able to consistently and significantly improve the saliency detection performance of various video saliency models, thereby achieves the state-of-the-art performance.**

***Index Terms*—video saliency, localized estimation, local temporal window, spatiotemporal refinement, saliency update.**

I. Introduction

SALIENCY detection has become a booming research topic in recent years. The inherent visual attention mechanism in human visual system is deployed to computationally identify the salient objects in the complicated scenes. Up to now, numerous saliency models have been proposed for static images, and have been intensively used in various applications such as object detection and segmentation [1]–[10], content-aware image/video retargeting [11]–[13], image/video quality assessment [14], and content-based image/video compression [15], [16]. However, there are relatively few researches investigating the video saliency. Therefore, in this paper we focus on detecting the salient object in a given video.

X. Zhou and Z. Liu are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China, and School of Communication and Information Engineering, Shanghai University, Shanghai, China (email: zxforchid@outlook.com; liuzhisjtu@163.com).

C. Gong is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China (email: chen.gong@njust.edu.cn).

W. Liu is with the Key Laboratory of Ministry of Education for System Control and Information Processing, Shanghai Jiao Tong University, Shanghai, China (email: liuwei.1989@sjtu.edu.cn).

Video saliency detection differs from the traditional image saliency detection majorly in the introduction of temporal information apart from the spatial information inherited by an image. Therefore, to simultaneously cope with the temporal information and spatial information of a video, many prior works have been done from various aspects such as the center-surround scheme [17]–[22], information theory [23]–[25],control theory [26], [27], frequency domain analysis [15], [28], machine learning [29]–[33], information fusion [34]–[41], and regional saliency assessment [42]–[45]. The above saliency models can obtain satisfactory results to some degree, however their performances will degrade in dealing with the unconstrained videos with complicated motion and complex scenes such as fast motion, dynamic background, nonlinear deformation, and occlusion, etc. Concretely, the existing video saliency models are insufficient to uniformly highlight salient objects with well-defined boundaries and meanwhile suppress irrelevant background regions.

To elevate the performance of saliency detection in videos, this paper proposes a novel framework to effectively improve the saliency detection results in unconstrained videos generated by any existing saliency model. The advantages of our framework are twofold. First, to preserve the global shape of salient object in video, we design a localized estimation step based on bootstrap learning [46] within a local temporal window, in which temporal consistency and strong correlation among frames are considered. As a result, the saliency map will generally highlight most part of salient object throughout the video. Second, to refine the estimation result with well-defined boundaries, we devise a spatiotemporal refinement step which takes the appearance and motion cues of potential salient objects into consideration simultaneously. Consequently, the obtained saliency map will be more accurate.

Our framework detects salient objects in a video frame by frame, and it consists of three key steps, *i.e.* localized estimation, spatiotemporal refinement, and saliency update, as shown in Fig. 1. Specifically, for a given video, the initial saliency maps of all frames are generated via using an existing saliency model. Then, the localized estimation models are generated by training the random forest regressor within a local temporal window centered on the current frame. The estimated saliency maps, which are produced by the localized estimation models, are combined with the initial saliency map to yield the temporary saliency map. Finally, by deploying the appearance and motion information of salient object, the spatiotemporal refinement is performed to generate the final saliency map.

Overall, our main contributions are summarized as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

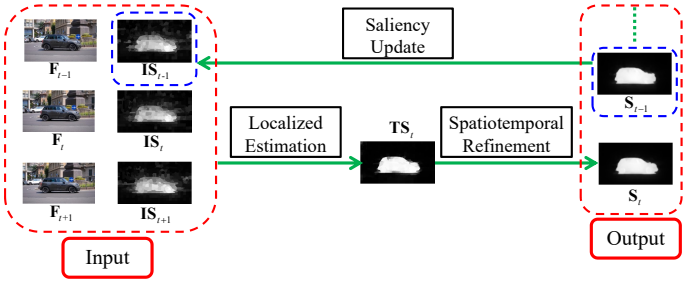


Fig. 1: Illustration of the proposed framework.

1) We propose a novel framework, which consists of three key steps including localized estimation, spatiotemporal refinement and saliency update, for boosting the saliency detection results in unconstrained videos.
2) The proposed localized estimation method reasonably exploits the temporal consistency and strong correlation among adjacent frames, where a local temporal window based estimators are employed to highlight the global shape of salient objects in each current frame.
3) The proposed spatiotemporal refinement method simultaneously incorporates the appearance and motion information of salient objects to effectively highlight the salient objects with well-defined boundaries and achieve more accurate results.
4) We tested our framework with several state-of-the-art video saliency models on four public video datasets, and the results firmly demonstrate the effectiveness and superiority of our framework.

The rest of this paper is organized as follows. The related works are reviewed in Section II. The proposed framework is described in Section III. Experimental results and the related analyses are presented in Section IV. Finally, we conclude this paper in Section V.

II. Related Works

The saliency detection for still images has been studied for decades, during which numerous effective models have been proposed via bottom-up or top-down strategy [47], [48]. For bottom-up models, the pioneering work was done by Itti *et al.*, who proposed the well-known center-surround saliency model [49]. In this model, luminance, color and orientation across multiple scales are employed to compute the center-surround difference. Similar to [49], a global contrast saliency model is proposed in [6], where the global region contrast in entire image and the spatial relationship across different image regions are deployed to detect the salient object. As for the top-down strategy, it is usually task and knowledge driven. For example, in [1], the conditional random field is used to integrate multiple features and generate the saliency map. In [50], the discriminative features of each region are mapped to saliency score using random forest regression. More recently, deep learning based saliency models such as [51]–[54] push forward the progress of saliency detection for still images. Besides, some prior works have been done to perform saliency detection based on existing saliency models, such as bootstrap learning based models [3], [55] and optimization-based saliency prediction [56], which is a similar and related work for the spatiotemporal refinement in our framework. Different from [56], which processes only still color images at single scale, our spatiotemporal refinement method incorporates motion information and operates at multiple scales. Generally speaking, such aforementioned efforts focus on image saliency detection, so they are inappropriate to conduct video saliency detection.

Since the proposed framework focuses on saliency detection in videos, next we will review some representative video saliency models. Roughly speaking, the existing models are based on center-surround scheme, information theory, control theory, machine learning, or information fusion, etc.

The well-known center-surround scheme in [49] has been exploited by numerous video saliency models and interpreted as the feature difference by defining various mathematical principles. The surprise model [17] incorporates multiple features including color, luminance, orientation, flicker and motion energy, to compute the feature difference and generate the saliency map. Akin to [49], in [19], the Kullback-Leibler divergence on dynamic texture feature is used to compute the video saliency based on the discriminant center-surround hypothesis [18]. Besides, the feature difference has also been formulated as local regression kernel based self-resemblance [20], earth movers distance [21], or directional coherence [22].

Based on the information theory, the video saliency is characterized by different models such as self-information [23], minimum conditional entropy [24], and incremental coding length [25]. As for the control theory, a linear dynamic system [26], [27] is used to discriminate the salient object from dynamic scenes. Besides, the frequency domain analysis is also used for video saliency detection, such as the phase spectrum of quaternion Fourier transforms [15] and temporal spectral residual [28].

Machine learning methods have also been widely used in video saliency detection. For example, probabilistic multi-task learning [29], support vector machine with Gaussian kernels [30], and support vector regression [31] are utilized to predict fixations on videos. Besides, the one-class support vector machine is performed on object trajectories and the video saliency is determined based on the diffusion results of such trajectories [32]. Sparse representation is also employed in video saliency detection [33], in which the video saliency detection is formulated as a problem of regularized feature reconstruction.

Considering the difference between spatial and temporal information possessed by a video, some models first generate spatial saliency map and motion saliency map, respectively, and then adopt certain fusion schemes to combine such two saliency maps into the final saliency map. The examples include intra-map and inter-map competition based fusion [34], mean/maximum value based combination [35], linear summation with location prior [36], and weighted linear summation [37]. In addition, Fang *et al.* [38], [39] propose to use the parameterized normalization or sum/product fusion to effectively combine spatial and motion saliency. In [40], the conditional random field is leveraged to integrate spatial and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

motion information. More recently, in [41], the color-based saliency is fused with global motion cues in a batch-wise manner.

Recently, more efforts have been made on video saliency detection. For example, in [57], the spatial transition matrix and the temporal restarting distribution are systematically unified to compute the video saliency. Besides, there are also some literatures working on the segmented regions/superpixels. In [42], superpixel-level motion distinctiveness, global contrast, as well as spatial sparsity are first used to measure spatial and temporal saliency, and then they are fused via an adaptive scheme to generate the final saliency map. In [43], a superpixel-level graph based motion saliency measurement is leveraged to generate the initial saliency map, and then bidirectional temporal propagation and two-phase spatial propagation are performed successively to generate the final saliency map. In [44], the intra-frame boundary information together with the inter-frame motion information are first employed to construct the gradient flow field, and then the local and global contrast mechanism is deployed to obtain the coarse saliency cues. Such coarse saliency cues are finally improved by the energy optimization method, yielding the refined saliency map. In [45], spatial edges and temporal motion boundaries are exploited to generate the initial saliency map based on the geodesic distance over an intra-frame graph, and then an inter-frame graph is constructed to generate the final saliency map. Some more recent and prominent approaches such as spatiotemporal background priors based video saliency model [58] and video quantum cuts [59] have also been proposed, which achieved a very encouraging performance.

All the aforementioned saliency models, *i.e.* existing models for video saliency detection, can generate visually promising results in some cases, however the performance will degrade in dealing with complicated scenarios such as fast motion, cluttered background, deformation and so on. For the purpose of boosting saliency maps generated by existing video saliency models, we present a novel framework, which combines three key steps including localized estimation, spatiotemporal refinement and saliency update in an effective way.

III. Proposed Framework

This section details our proposed video saliency detection framework.

A. Architecture Overview

The main architecture of our proposed framework is illustrated in Fig. 1, which consists of three steps including localized estimation, spatiotemporal refinement and saliency update. For saliency computation of each current frame $\mathbf{F}_t$, a local temporal window $\mathrm{WT}_t = \{\mathbf{F}_{t-1}, \mathbf{F}_t, \mathbf{F}_{t+1}\}$ is established centered on $\mathbf{F}_t$, where $\mathbf{F}_{t-1}$ is the previous frame and $\mathbf{F}_{t+1}$ is the next frame. The initial saliency maps $\mathbf{IS}_{t-1}$, $\mathbf{IS}_t$ and $\mathbf{IS}_{t+1}$ of the three frames in WT_t can be generated by any existing video saliency model, and could serve as the input of our framework without the saliency update. However, since the final saliency map $\mathbf{S}_{t-1}$ of the previous frame $\mathbf{F}_{t-1}$ is actually available, $\mathbf{IS}_{t-1}$ is updated by $\mathbf{S}_{t-1}$ in our framework with the update step, *i.e.* "Saliency Update", as shown in Fig. 1. Furthermore, in Section IV-C, we will evaluate the performance of our framework with/without update and demonstrate the contribution of saliency update. With $\mathbf{S}_{t-1}$, $\mathbf{IS}_t$ and $\mathbf{IS}_{t+1}$ as the input, the temporary saliency map $\mathbf{TS}_t$ is generated via the localized estimation step (Section III-B). Then, $\mathbf{TS}_t$ is further improved by the spatiotemporal refinement step (Section III-C) to obtain a more precise result, namely the final saliency map $\mathbf{S}_t$, for the current frame $\mathbf{F}_t$. For saliency computation of the next frame $\mathbf{F}_{t+1}$, the initial saliency map $\mathbf{IS}_t$ is also updated by $\mathbf{S}_t$. The above process iterates until all the frames in a video have been processed. In this way, our framework detects the salient objects frame by frame in a video, and operates on a local temporal window centered on each current frame. Besides, it should be noted that the saliency computation for the first frame is performed on a local temporal window, which only contains the first frame and the second frame, while for the last frame, its local temporal window only contains the penultimate frame and the last frame.

TABLE I: Features extracted for each superpixel

Feature Description	Notation	Dimension
Color Features		
The average of RGB values	$x_1 \sim x_3$	3
The variances of RGB values	$x_4 \sim x_6$	3
The average of CIELab values	$x_7 \sim x_9$	3
The variances of CIELab values	$x_{10} \sim x_{12}$	3
Texture Features		
The average of LBP values	x_{13}	1
The variances of LBP values	x_{14}	1
Location Features		
The average of normalized x coordinates	x_{15}	1
The average of normalized y coordinates	x_{16}	1
Motion Features		
The average of motion amplitude values	x_{17}	1
The variances of motion amplitude values	x_{18}	1
The average of motion orientation values	x_{19}	1
The variances of motion orientation values	x_{20}	1

In our method, we follow the recent works [42]–[45] and segment every frame $\mathbf{F}_t$ $(t = 1, 2, \cdots)$ into some perceptually homogenous superpixels $\{sp_t^i\}_{i=1}^{n_t}$ (n_t is the number of generated superpixels) via the simple linear iterative clustering (SLIC) algorithm [60]. Salient objects are likely to appear at different scales, so we generate three layers of superpixel with different granularities in our implementation with $n_t = 350, 400, 450$, respectively. In order to guarantee the temporal consistency of saliency maps, in Section III-B, the local temporal window based estimation models are designed to incorporate and exploit the temporal consistency and strong correlation among temporally adjacent frames. The above models complement to each other and will be used for predicting the saliency map of current frame. Furthermore, the individual performances at certain scales will be discussed in Section IV-C. Note that the notation without superscript denotes all superpixels at certain scale in $\mathbf{F}_t$.

In our work, four kinds of features are extracted on every superpixel. Firstly, color features are extracted from RGB

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

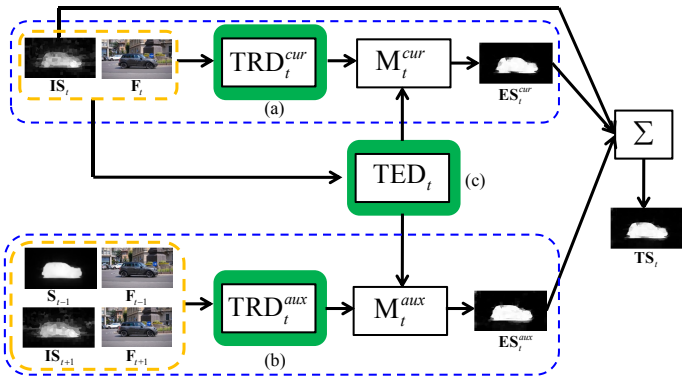


Fig. 2: Illustration of the localized estimation step. (a) and (b) denote the training data $\{\mathrm{TRD}_t^{cur}, \mathrm{TRD}_t^{aux}\}$ collected from the current frame $\mathbf{F}_t$ and the adjacent two frames $\{\mathbf{F}_{t-1}, \mathbf{F}_{t+1}\}$, respectively, (c) represents the test data TED_t extracted from the current frame $\mathbf{F}_t$.

and CIELab color spaces. Secondly, we employ LBP [61] to characterize the texture of image regions. Thirdly, horizontal and vertical locations of superpixels are used to specify the spatial information of superpixels. Lastly, the motion information which is an important cue for video processing is also incorporated in our work. For the current frame $\mathbf{F}_t$, its pixel-level motion vector field $\mathbf{MVF}_{t,t+1}$ with respect to the next frame $\mathbf{F}_{t+1}$ is calculated using the method of large displacement optical flow (LDOF) [62]. The motion feature of each superpixel is then computed based on the amplitudes and orientations of pixels in $\mathbf{MVF}_{t,t+1}$.

Table I gives the detailed information about the adopted features. For each superpixel at each scale, a 20-dimensional feature vector $\mathbf{x} = [x_1, x_2, \cdots, x_{20}]$ is obtained by concatenating all the features mentioned above.

B. Localized Estimation

To exploited the temporal consistent and strong correlation among temporally adjacent frames, we propose a novel localized estimation method to obtain temporary saliency map, as shown in Fig. 2. In the following, we will provide a detailed description of localized estimation step.

For the current frame $\mathbf{F}_t$, the local temporal window $\mathrm{WT}_t = \{\mathbf{F}_{t-1}, \mathbf{F}_t, \mathbf{F}_{t+1}\}$ contains the previous frame $\mathbf{F}_{t-1}$ with its final saliency map $\mathbf{S}_{t-1}$, the current frame $\mathbf{F}_t$ with its initial saliency map $\mathbf{IS}_t$, and the subsequent frame $\mathbf{F}_{t+1}$ with its initial saliency map $\mathbf{IS}_{t+1}$. Therefore, the current estimation model M_t^{cur} is learned from current frame $\mathbf{F}_t$, and the auxiliary estimation model M_t^{aux} is learned from adjacent two frames $\{\mathbf{F}_{t-1}, \mathbf{F}_{t+1}\}$. Such two estimation models are denoted as the localized estimation models. Both models are built via the random forest regressor. Here, we take $\mathbf{F}_t$ as an instance. A binary mask $\mathbf{BM}_t$ is first obtained using the Otsu's method [63] on $\mathbf{IS}_t$. The confidence score is then obtained to select the reliable training samples for the random forest regressor, namely:

$$CS_t^i = \frac{\left|sp_t^i \cap \mathbf{BM}_t\right|}{\left|sp_t^i\right|}, \tag{1}$$

where CS_t^i denotes the confidence score measuring the percentage of the pixels in the superpixel sp_t^i at one scale that belong to the salient object, and "$|.|$" denotes the number of pixels in the corresponding region. Then, we can compute the saliency score A_t^i of the superpixel sp_t^i as:

$$\mathrm{A}_t^i = \begin{cases} 1 & CS_t^i \geq q_h \\ 0 & CS_t^i \leq q_l \end{cases}, \tag{2}$$

which means that sp_t^i is treated as a positive sample if CS_t^i is not less than the upper threshold q_h , so the corresponding saliency score A_t^i is set to 1. If CS_t^i is not larger than q_l , sp_t^i is treated as a negative sample and the corresponding saliency score A_t^i should be 0. To obtain confident samples, here we set the upper threshold q_h and the lower threshold q_l as 0.8 and 0, respectively. By this way, we can obtain the training data $\mathrm{TRD}_t^{cur} = \left\{\left(\mathbf{x}_t^1, \mathrm{A}_t^1\right), \left(\mathbf{x}_t^2, \mathrm{A}_t^2\right), \cdots, \left(\mathbf{x}_t^Q, \mathrm{A}_t^Q\right)\right\}$ from three scales in current frame, which consists of totally Q confident samples.

Then, a random forest regressor is exploited to obtain the estimation model M_t^{cur} on the training data TRD_t^{cur}. Next, for the test data $\mathrm{TED}_t = \{\mathrm{TED}_t^s\}_{s=1}^3$ (*i.e.* all the superpixels at three scales in current frame $\mathbf{F}_t$), the estimated saliency map $\mathbf{ES}_t^{cur}$ is computed by:

$$\mathbf{ES}_t^{cur} = \frac{1}{3}\sum_{s=1}^{3} \mathrm{M}_t^{cur}\left(\mathrm{TED}_t^s\right). \tag{3}$$

By incorporating the temporal correlation, an auxiliary estimation model M_t^{aux} is constructed based on the other two frames $\{\mathbf{F}_{t-1}, \mathbf{F}_{t+1}\}$. Akin to the generation of estimated saliency map $\mathbf{ES}_t^{cur}$, we can obtain the corresponding training data TRD_t^{aux}, estimation model M_t^{aux}, and the estimated saliency map $\mathbf{ES}_t^{aux}$ that is defined as:

$$\mathbf{ES}_t^{aux} = \frac{1}{3}\sum_{s=1}^{3} \mathrm{M}_t^{aux}\left(\mathrm{TED}_t^s\right). \tag{4}$$

The estimated saliency maps $\mathbf{ES}_t^{cur}$ and $\mathbf{ES}_t^{aux}$ are shown in Fig. 2. Compared with the initial saliency map $\mathbf{IS}_t$, the estimated saliency maps $\mathbf{ES}_t^{cur}$ and $\mathbf{ES}_t^{aux}$ highlight most part of the car more effectively. Besides, due to the difference of training data, $\mathbf{ES}_t^{cur}$ pays more attention to discriminate the object details (*e.g.* tyre), while $\mathbf{ES}_t^{aux}$ is able to highlight the car more uniformly.

Lastly, by integrating the initial saliency map $\mathbf{IS}_t$ and the predicted saliency maps $\mathbf{ES}_t^{cur}$ and $\mathbf{ES}_t^{aux}$, the temporary saliency map $\mathbf{TS}_t$ is generated as:

$$\mathbf{TS}_t = \mathbf{IS}_t + \mathbf{ES}_t^{cur} + \mathbf{ES}_t^{aux}. \tag{5}$$

As a result, the temporary saliency map $\mathbf{TS}_t$ well preserves the details of the car as well as uniformly highlights the entire car region due to the combination of such saliency maps.

C. Spatiotemporal Refinement

To further improve the performance of temporary saliency map output by the localized estimation step, we propose a spatiotemporal refinement method to generate the final saliency map with well-defined boundaries for current frame (see

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

Fig. 3). To this end, the appearance and motion information of salient object are used simultaneously. Specifically, the current frame $\mathbf{F}_t$ is first fed into "ST-BP", which is a spatiotemporal background probability computation model that will be later introduced, to yield the background probability map $\mathbf{BPM}_t$. Then, the temporary saliency map $\mathbf{TS}_t$ and the background probability map $\mathbf{BPM}_t$ are together fed into "ST-SOP" and graph cut (GC) method [55], in which "ST-SOP" denotes a spatiotemporal saliency optimization model that will be detailed below. In this way, we obtain the final saliency map $\mathbf{S}_t$ for the current frame $\mathbf{F}_t$.

(a) Spatiotemporal background probability (ST-BP). As mentioned above, a background probability computation model is employed by us to infer the potential background priors in spatiotemporal domain. Concretely, we use the strategy proposed in [56], which models the boundary connectivity $BondCon(.)$ as the degree that a superpixel is connected to image boundaries. For a superpixel sp_t^i at certain scale in $\mathbf{F}_t$, the $BondCon(.)$ is defined as:

$$BondCon\left(sp_t^i\right) = \frac{Len_{bnd}\left(sp_t^i\right)}{\sqrt{Area\left(sp_t^i\right)}}, \tag{6}$$

where $Area\left(sp_t^i\right)$ measures the soft area of the region that sp_t^i belongs to, and $Len_{bnd}\left(sp_t^i\right)$ defines the length along the boundary of the region that sp_t^i is located. Such two terms are computed by:

$$Area\left(sp_t^i\right) = \sum_{j=1}^{n_t} \exp\left(-\frac{d_{geo}^2\left(sp_t^i, sp_t^j\right)}{2\sigma_{geo}^2}\right), \tag{7}$$

$$Len_{bnd}\left(sp_t^i\right) = \sum_{j=1}^{n_t}\left[\exp\left(-\frac{d_{geo}^2\left(sp_t^i, sp_t^j\right)}{2\sigma_{geo}^2}\right)\cdot\delta\left(sp_t^i \in Bnd\right)\right], \tag{8}$$

where Bnd denotes the set of image boundary superpixels and $\delta(.)$ is equal to 1 for superpixel on the image boundary and 0 otherwise. The geodesic distance $d_{geo}\left(sp_t^i, sp_t^j\right)$ between any two superpixels is defined as the accumulated edge weights, which contains two kinds of geodesic distances $d_{geo,C}\left(sp_t^i, sp_t^j\right)$ and $d_{geo,M}\left(sp_t^i, sp_t^j\right)$ that are computed as:

$$d_{geo,C}\left(sp_t^i, sp_t^j\right) = \min_{sp_t^1=sp_t^i, sp_t^2, \cdots, sp_t^{n_t}=sp_t^j} \sum_{k=1}^{n_t-1} d_C\left(sp_t^k, sp_t^{k+1}\right), \tag{9}$$

$$d_{geo,M}\left(sp_t^i, sp_t^j\right) = \min_{sp_t^1=sp_t^i, sp_t^2, \cdots, sp_t^{n_t}=sp_t^j} \sum_{k=1}^{n_t-1} d_M\left(sp_t^k, sp_t^{k+1}\right), \tag{10}$$

where d_C and d_M are the Euclidean distance between any two adjacent superpixels using color and motion feature, respectively. Here, the color feature is the average of CIELab values of pixels in each superpixel, *i.e.* $[x_7, x_8, x_9]$, and the motion feature is the mean value of motion amplitude and orientation values of pixels in each superpixel, *i.e.* $[x_{17}, x_{19}]$. Besides, the σ_{geo} in Eqs. (7) and (8) is computed as the mean of all distances between any two adjacent superpixels in above two feature spaces, respectively.

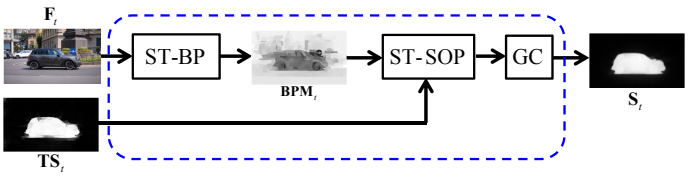


Fig. 3: Illustration of the spatiotemporal refinement step. "ST-BP" denotes spatiotemporal background probability computation model, "ST-SOP" represents spatiotemporal saliency optimization model, and "GC" stands for graph cut.

Then, according to above explanations, we compute the background probability map $\mathbf{BPM}_t$ at each scale as:

$$\mathbf{BPM}_t = \mathbf{BPM}_t^C + \mathbf{BPM}_t^M, \tag{11}$$

where $\mathbf{BPM}_t^C$ and $\mathbf{BPM}_t^M$ are the background probability maps generated by using color feature and motion feature, respectively. The uniform computation equation is then defined as:

$$\mathbf{BPM}_t^{*,i} = 1 - \exp\left(-\frac{BondCon^2\left(sp_t^i\right)}{2\sigma_{bondcon}^2}\right), \tag{12}$$

where the superscript "*" can be "C" or "M" that refers to the color feature or motion feature, respectively. The normalization term $\sigma_{bondcon}$ is also set to the mean value of all the boundary connectivity values $Len_{bnd}(.)$.

(b) Spatiotemporal saliency optimization (ST-SOP). spatiotemporal saliency optimization model is utilized to establish the final saliency map $\mathbf{S}_t$ based on the obtained background probability map $\mathbf{BPM}_t$ and the temporary saliency map $\mathbf{TS}_t$. Concretely, the optimization model is designed to assign the salient object region value 1 and the background region value 0. The objective function at each scale is expressed as:

$$\mathbf{S}_t = \arg\min_{\mathbf{S}_t^i}\left[\sum_{i=1}^{n_t}\mathbf{BPM}_t^i\cdot\left(\mathbf{S}_t^i\right)^2 + \sum_{i=1}^{n_t}\mathbf{TS}_t^i\cdot\left(\mathbf{S}_t^i-1\right)^2 + \sum_{i,j} w_t^{ij}\left(\mathbf{S}_t^i-\mathbf{S}_t^j\right)^2\right]. \tag{13}$$

Eq. (13) is conducted on all three scales (*i.e.* n_t = 350, 400 and 450), therefore the optimization result is the average of the outputs of all three scales, namely $\mathbf{S}_t := \left(\sum_{n_t=350,400,450}\mathbf{S}_t\left(n_t\right)\right)\Big/3$. The smoothness term $\sum_{i,j} w_t^{ij}\left(\mathbf{S}_t^i-\mathbf{S}_t^j\right)^2$ encourages the adjacent superpixels to obtain similar saliency values. For a pair of adjacent super-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

pixels , the weight w_t^{ij} is calculated as:

$$w_t^{ij} = \exp\left(-\frac{\left(d_C\left(sp_t^i, sp_t^j\right)\right)^2}{2\sigma_C^2}\right) + \exp\left(-\frac{\left(d_M\left(sp_t^i, sp_t^j\right)\right)^2}{2\sigma_M^2}\right) + \mu, \tag{14}$$

where the normalization terms σ_C and σ_M are computed as the mean value of all the distances between any two adjacent superpixels in color and motion feature spaces, respectively. The trade-off parameter μ is empirically set to 0.1. The objective function Eq. (13) can be easily solved by least-square regression.

Finally, by feeding the output of objective function Eq. (13) into a graph cut (GC) based refinement method [55], a salient object mask $\mathbf{SM}_t$ can be generated. By combining $\mathbf{SM}_t$ with $\mathbf{S}_t$, the final saliency map $\mathbf{S}_t$ is computed as $\mathbf{S}_t := (\mathbf{SM}_t + \mathbf{S}_t)/2$ as shown in Fig. 3. Compared with the initial saliency map $\mathbf{IS}_t$, the temporary saliency map $\mathbf{TS}_t$ better highlights the salient object and suppresses background regions. Furthermore, by fusing $\mathbf{TS}_t$ and the background probability map $\mathbf{BPM}_t$, the final saliency map $\mathbf{S}_t$ renders the salient object more uniformly and completely with well-defined boundaries.

IV. Experimental Results

In this section, we performed comprehensive experiments on four public video datasets including SegTrackV2 [64], UVSD [43], DAVIS [65] and ViSal [44]. First, the video datasets and experimental settings are detailed in Section IV-A. Then, the comprehensive comparison results over the aforementioned four video datasets are provided in Section IV-B. Some validation experiments are performed in Section IV-C. In Section IV-D, some failure cases are presented, and finally, the computation issue of our framework is discussed in Section IV-E.

A. Datasets and Experimental Settings

The four typical video saliency datasets with manually annotated binary ground truths are employed for evaluation. The first dataset SegTrackV2 consists of 14 videos with challenging circumstances such as appearance change, motion blur, occlusion, complex deformation and so on. The second dataset UVSD contains a total of 18 challenging videos with complicated motions and complex scenes. The third dataset DAVIS is a recent dataset for video object segmentation, which contains 50 high-quality videos with different motions of human, animal and vehicle in challenging circumstances. As for the fourth dataset, ViSal contains 17 challenging video sequences such as complex color distributions, camera motion, rapid topology changes and so on.

We applied our framework with five state-of-the-art saliency models including SGSP [43], GD [45], MC [51], CVS [44], and RWRV [57]. Therefore, the comparison is performed between the original saliency models (*i.e.* SGSP, GD, MC, CVS and RWRV), and the corresponding improved version (denoted as SGSP*, GD*, MC*, CVS* and RWRV*) based on the proposed framework. The saliency model MC is designed for image saliency detection while the rest four models are used for video saliency detection. For a fair comparison, the source codes of SGSP, GD, MC, CVS and RWRV are directly provided by their authors, and the saliency maps generated by different models are normalized into the same resolution as original videos with pixel value ranging from 0 to 255.

B. Performance Comparison

1) Qualitative Evaluation: Figs. 4-7 provide the qualitative evaluation between SGSP, GD, MC, CVS, RWRV and the corresponding improved SGSP*, GD*, MC*, CVS*, RWRV* on SegTrackV2, UVSD, DAVIS and ViSal, respectively. From Figs. 4(k), 5(k), 6(k) and 7(k), it can be observed that the video saliency model RWRV that works in a patches/volumes way only highlights the regions around the boundaries of salient object or falsely highlights some background regions. The other three video saliency models including SGSP, GD and CVS perform better than RWRV and achieve the decent visual effect to some degree, but the results are not sufficiently good on the four challenging datasets. The reason behind this lies in the heavy dependence of motion information on the construction of basic saliency cue in such models. As for the image saliency model MC, its performance is insufficient for the dynamic scenes due to the lack of temporal information. However, it shows competitive performance with the aforementioned three video saliency models, and this indicates the power of deep learning for saliency detection.

Compared with the original saliency models including SGSP, GD, MC, CVS and RWRV, it can be seen that the improved version (*i.e.* SGSP*, GD*, MC*, CVS* and RWRV*) render the better performance. The saliency maps improved by our framework highlight the salient objects more completely and meanwhile suppress background regions more effectively in most test examples, as shown in Figs. 4(d, f, h, j, l), 5(d, f, h, j, l), 6(d, f, h, j, l) and 7(d, f, h, j, l). In particular, the improved saliency maps exhibit more promising visual results on some challenging scenarios such as camera motion and appearance change in Figs. 4 and 7, cluttered background, low resolution, fast motion, shape complexity in Fig. 5, and edge ambiguity, heterogeneous object and complex shapes of objects in Fig. 6. Furthermore, for some examples including the two examples in Fig. 5 and the bottom example in Fig. 7, there are two objects in each video frame such as horse and horseman in the top example of Fig. 5. It can be observed, when most parts of salient objects can be highlighted by the original saliency models, as shown by the top example in Fig. 5(e), the improved version can highlight the salient objects more completely, as shown in Fig. 5(f). Even though the salient objects cannot be highlighted well by existing saliency model, such as the top example in Fig. 5(k), the improved saliency maps produced by our framework can not only uniformly pop out the entire salient objects, but also effectively suppress the background noise.

In general, the performance boosting of the original saliency models can be attributed to the following three aspects. First,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

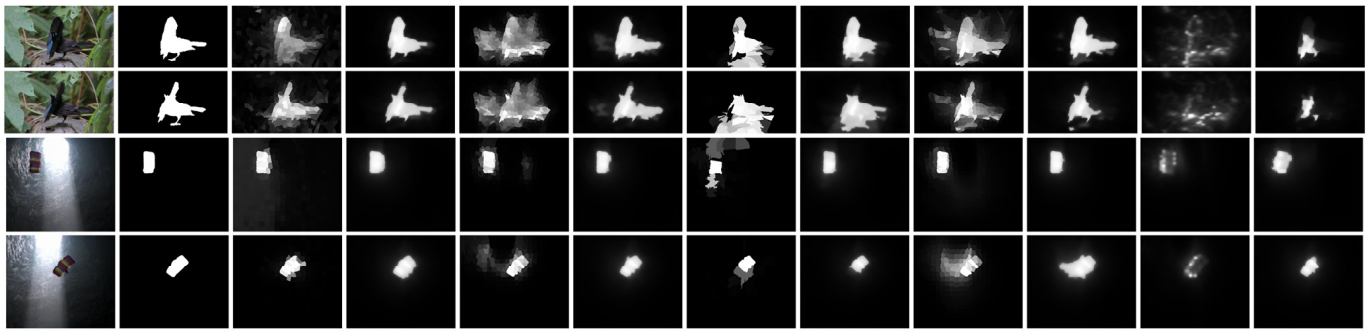


Fig. 4: Qualitative comparison between the original saliency maps and the final saliency maps improved by our framework on the SegTrackV2 dataset. (a): Input video frames, (b): binary ground truths, (c): SGSP, (d): SGSP*, (e): GD, (f): GD*, (g): MC, (h): MC*, (i): CVS, (j): CVS*, (k): RWRV, (l): RWRV*.

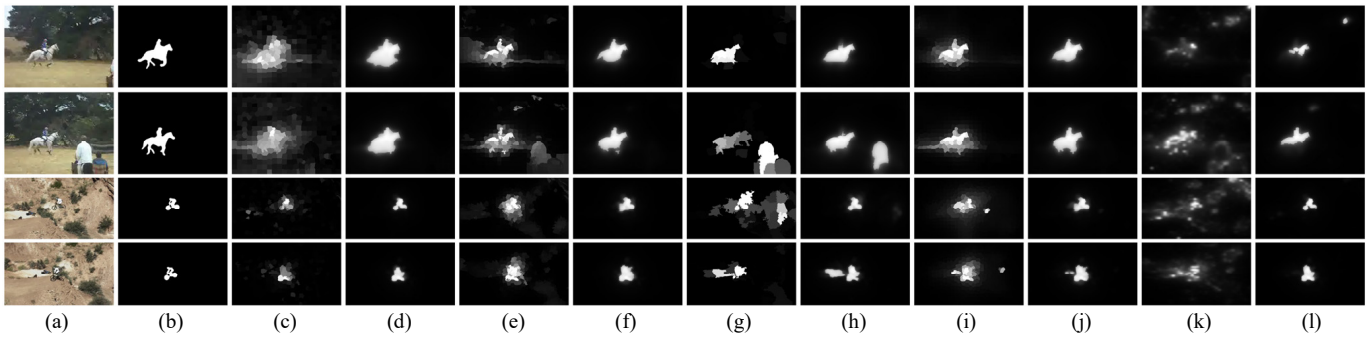


Fig. 5: Qualitative comparison between the original saliency maps and the final saliency maps improved by our framework on the UVSD dataset. (a): Input video frames, (b): binary ground truths, (c): SGSP, (d): SGSP*, (e): GD, (f): GD*, (g): MC, (h): MC*, (i): CVS, (j): CVS*, (k): RWRV, (l): RWRV*.

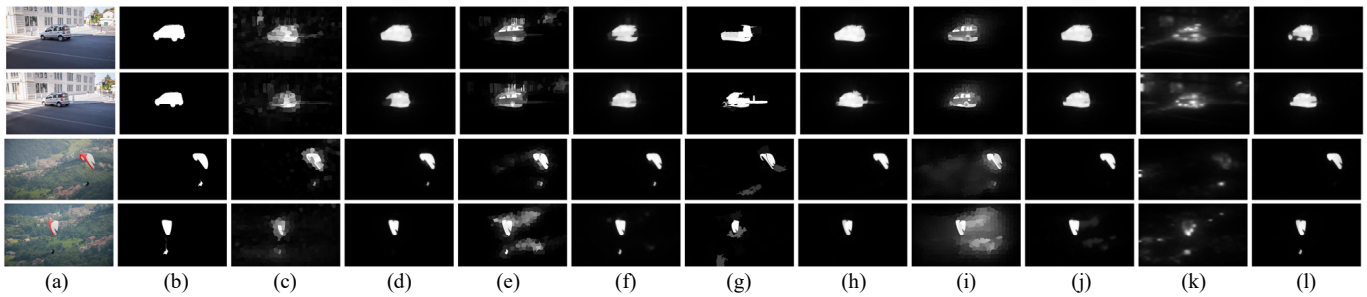
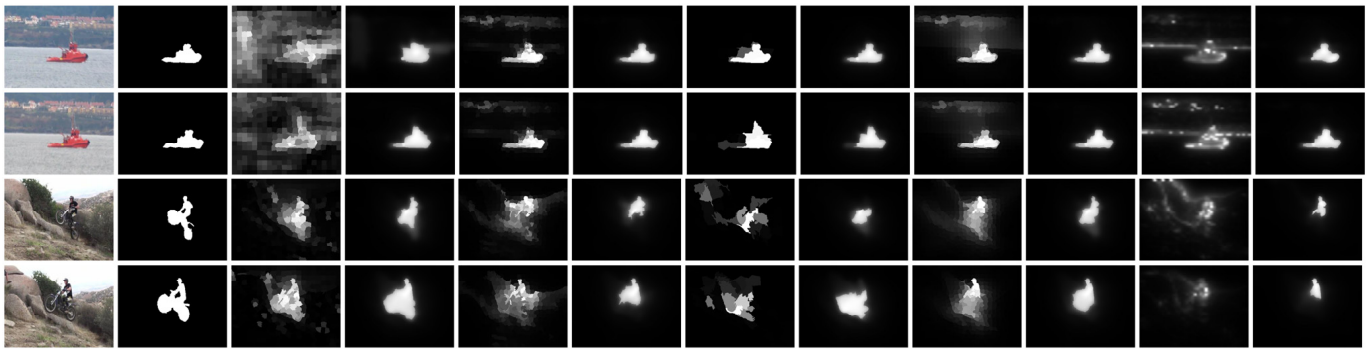


Fig. 6: Qualitative comparison between the original saliency maps and the final saliency maps improved by our framework on the DAVIS dataset. (a): Input video frames, (b): binary ground truths, (c): SGSP, (d): SGSP*, (e): GD, (f): GD*, (g): MC, (h): MC*, (i): CVS, (j): CVS*, (k): RWRV, (l): RWRV*.

to consider the temporal correlation and consistency, two local temporal window based estimation models are leveraged in our framework. Specifically, one of the two models is trained on the current frame, and the other one is trained on other frames in the local temporal window. Therefore, such two estimation models complement to each other to achieve satisfactory performance. Second, the output of estimation, *i.e.* temporary saliency map, is further improved via the refinement step that incorporates the appearance and motion information simultaneously. As a result, more accurate results can be generated. Finally, by updating the initial saliency map of current frame with the obtained final saliency map, more reliable samples can be collected for processing the subsequent frames.

2) Quantitative Evaluation: To objectively evaluate the saliency detection performances of different models, we adopt three widely used performance measures, including precision-recall (PR) curve, F-measure curve, and mean absolute error (MAE). Specifically, the precision-recall (PR) curves plot the trade-off between precision and recall achieved by an algorithm. Precision corresponds to the ratio of salient pixels correctly assigned, while recall denotes the percentage of detected salient pixels in relation to the salient pixels in ground truth. F-measure is then defined as the weighted harmonic mean of precision and recall for a comprehensive evaluation,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 7: Qualitative comparison between the original saliency maps and the final saliency maps improved by our framework on the ViSal dataset. (a): Input video frames, (b): binary ground truths, (c): SGSP, (d): SGSP*, (e): GD, (f): GD*, (g): MC, (h): MC*, (i): CVS, (j): CVS*, (k): RWRV, (l): RWRV*.

which has the following form:

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{15}$$

where β^2 is set to 0.3 as suggested in [6] and [55]. To plot the curve, the saliency maps are binarized with thresholds ranging from 0 to 255, and then 256 pairs of precision-recall combination and F-measure against thresholds are generated. Different from F-measure, MAE provides a more balanced comparison between the binary ground truth $\mathbf{GT}$ and the continuous saliency map $\mathbf{S}$, which is defined as:

$$\mathrm{MAE} = \frac{1}{W * H} \sum_{i=1}^{W*H} \left|\mathbf{S}(i) - \mathbf{GT}(i)\right|, \tag{16}$$

where W and H denote the width and height of video frame, respectively. In the computation of MAE, *i.e.* Eq. (16), $\mathbf{S}$ and $\mathbf{GT}$ are normalized to [0,1] for all the compared methods.

The PR curves of the original video saliency models and our improved versions on the four datasets are plotted in Fig. 8(a). It can be seen that the improved versions consistently outperform the corresponding original models on all datasets. In terms of F-measure curves and MAE values shown in Figs. 8(b) and (c), the improved version (*i.e.* SGSP*, GD*, MC*, CVS* and RWRV*) achieves better performance with a noticeable margin than the corresponding original saliency models (*i.e.* SGSP, GD, MC, CVS and RWRV) on all datasets. Overall, the PR curves, F-measure curves and MAE values shown in Fig. 8 convincingly demonstrate the capability of our framework to improve the performance of various video saliency models across diverse challenging videos

C. Validation of the Proposed Framework

1) Component-wise analysis: In this subsection, we first study the contribution of each step in our framework. Then, we make an analysis for the multiscale strategy adopted by our framework. Lastly, we explore the influence of neighboring frames used in our framework.

To demonstrate that all the critical steps (including localized estimation, spatiotemporal refinement, and saliency update) in our framework are beneficial for improving the saliency detection performance, quantitative comparisons are performed on the DAVIS dataset to show the contribution of each of these critical steps. We use SGSP to do such a comparison as this method achieves the best performance among the aforementioned five state-of-the-art models. Specifically, we present the initial saliency map (denoted as "initial") generated by the original SGSP, final saliency map with update (denoted as "wup-F"), temporary saliency map with update (denoted as "wup-T"), final saliency map without update (denoted as "woup-F"), and the temporary saliency map without update (denoted as "woup-T"). Fig. 9 shows the performances with the above five different settings, *i.e.* initial, wup-F, wup-T, woup-F and woup-T. It can be seen that wup-F performs best among all the compared settings in terms of PR curves, F-measure curves and MAE values. Besides, the superiority of wup-F over woup-F and the superiority of wup-T over woup-T can be easily identified from all the evaluations metrics. This clearly demonstrates the rationality and effectiveness of the proposed update operation (in Fig. 1). Furthermore, we can also observe that the performance gains of wup-F over wup-T and woup-F over woup-T, which clearly demonstrate the effectiveness of the proposed spatiotemporal refinement step (in Section III-C). It can also be observed that the saliency maps including wup-T, wup-F, woup-T and woup-F all perform better than the initial saliency map generated by SGSP, and this clearly reflects the effectiveness of the proposed localized estimation step (in Section III-B). Therefore, the PR curves, F-measure curves and MAE values in Fig. 9 reveal that every step in the proposed framework contributes to enhance the saliency detection performance.

As for the multiscale strategy adopted by our framework, the performance evaluation under different scales is conducted. Quantitative comparisons are performed on the DAVIS dataset based on SGSP, as shown in Fig. 10. Here, the results at different scales including 350, 400 and 450 are denoted as S350, S400 and S450, respectively. It can be observed that our framework, *i.e.* SGSP* with multiple scales, performs best in terms of PR curves, F-measure curves and MAE values, and this demonstrates the rationality and effectiveness of multiscale strategy adopted in our framework. As for the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

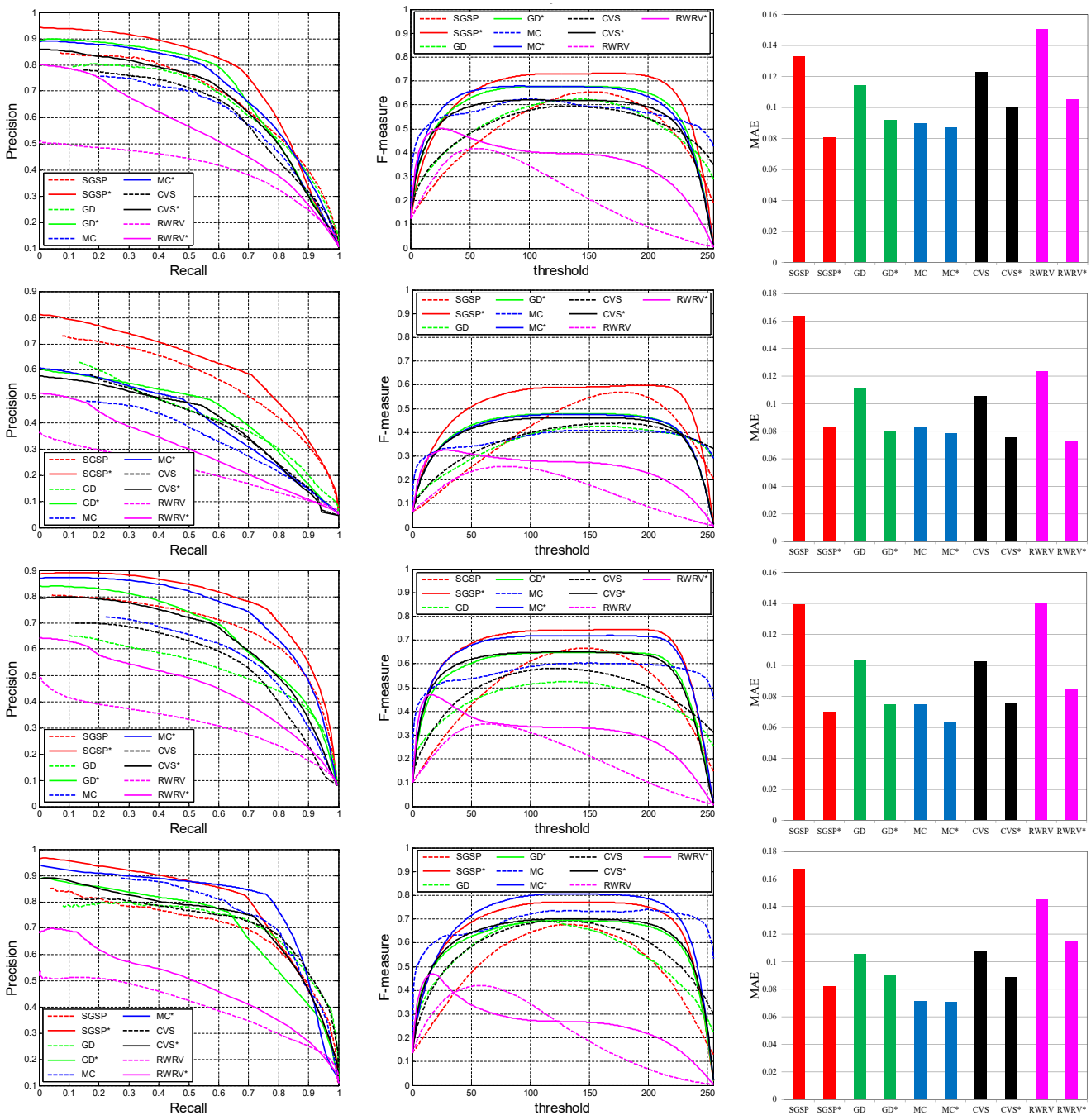
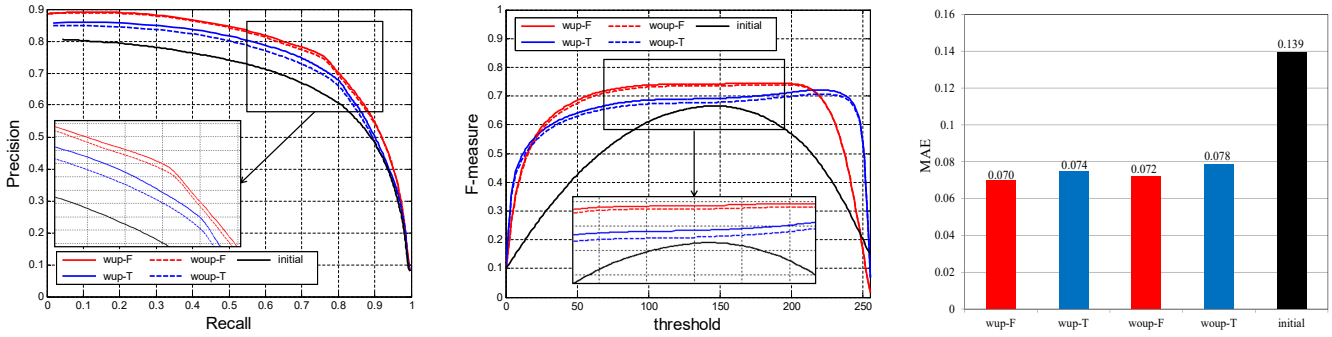


Fig. 8: (better viewed in color) Quantitative evaluation of different saliency models: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE values. From top to down, each row shows the results on the SegTrackV2 dataset, the UVSD dataset, the DAVIS dataset and the ViSal dataset, respectively.

results at different scales including S350, S400 and S450, they also achieve a competitive performance in terms of all the three metrics. Furthermore, we can also see that the results at all the three scales perform better than SGSP, and this indicates the effectiveness of our framework again.

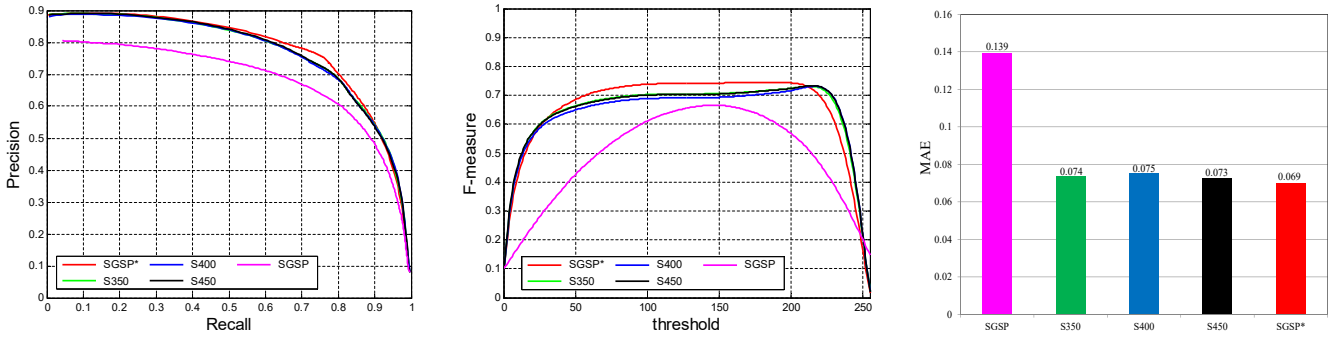
In order to explore the influence of neighboring frames used in our framework, the performance comparison of our framework with different numbers of neighboring frames is performed on the DAVIS dataset using the original model SGSP, and the results are shown in Fig. 11. The three improved versions obtained by setting the neighbor size of forward and backward frames to 1, 3 and 5 are represented as SGSP*1, SGSP*3 and SGSP*5, respectively. It can be seen from Fig. 11 that the performances of the three improved versions are all better than the original model SGSP in terms of PR curves, F-measure curves and MAE values. Further, we can see that the three improved versions achieve almost the same performance in terms of all the three metrics. This

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia



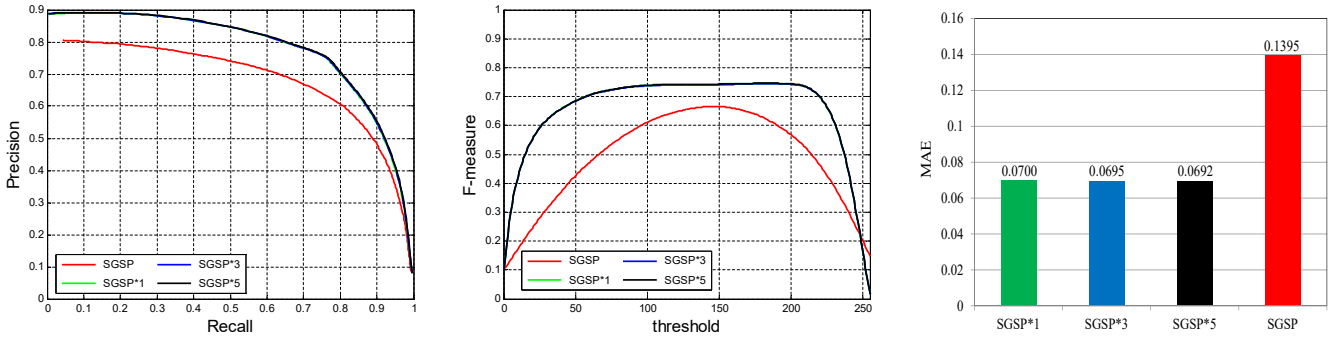
(a) (b) (c)

Fig. 9: (better viewed in color) Component-wise efficacy of our framework: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE values.



(a) (b) (c)

Fig. 10: (better viewed in color) Effects at different scales of our framework: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE values.



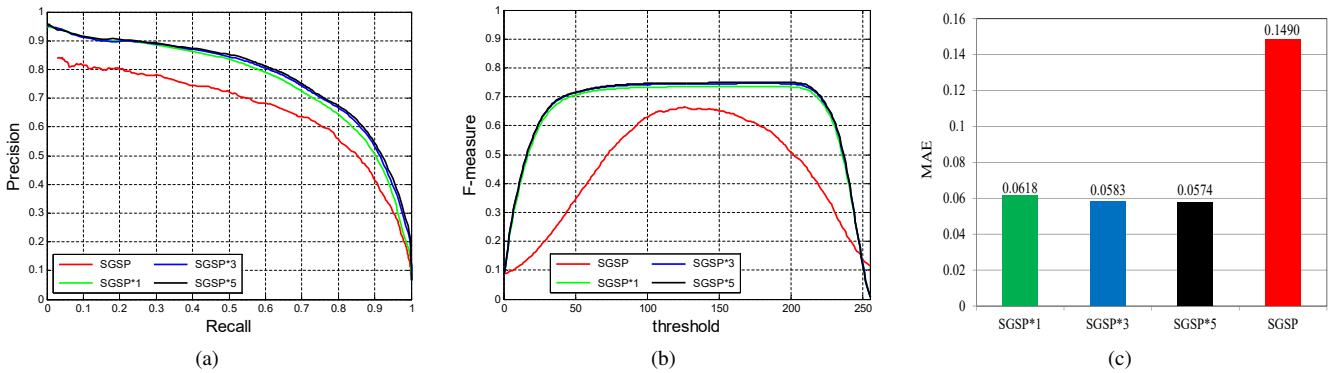
(a) (b) (c)

Fig. 11: (better viewed in color) Performances with different numbers of neighboring frames used in our framework: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE values.

indicates the effectiveness and robustness of our framework. We performed the experiments on a PC with Intel Core i7-4790K 4GHz CPU and 32GB RAM. The average processing time per frame with the video resolution 320×240 is 5.928 seconds using SGSP*1, 6.706 seconds using SGSP*3 and 7.630 seconds using SGSP*5, respectively. It can be seen that with more neighboring frames, the corresponding computation cost increases.

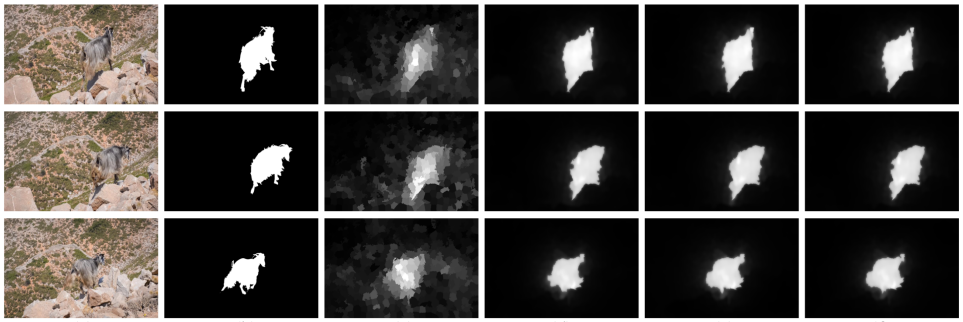
Besides, we also show quantitative and qualitative comparisons on a video which contains a slowly moving object, as shown in Fig. 12 and Fig. 13, respectively. In Fig. 12, we can find that SGSP*5 and SGSP*3 render slightly superior results to SGSP*1 in terms of all the three metrics. Obviously, all the three improved versions outperform the original model SGSP. In Fig. 13, the example shows a slowly moving goat with low contrast to cluttered background. We can see that the results of SGSP*1, SGSP*3 and SGSP*5 are very close to each other, and most parts of the goat are highlighted uniformly, as shown in Fig. 13(d), (e) and (f). As for the results of SGSP shown in Fig. 13(c), the background regions around the goat are falsely highlighted. Therefore, we can conclude that our framework is effective and robust to videos with slowly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia



(a) (b) (c)

Fig. 12: (better viewed in color) Quantitative comparisons with different numbers of neighboring frames on a video which contains a slowly moving object: (a) presents PR curves, (b) presents F-measure curves, and (c) presents MAE values.



(a) (b) (c) (d) (e) (f)

Fig. 13: Qualitative comparisons with different numbers of neighboring frames on a video which contains a slowly moving object. (a) Input video frames; (b) binary ground truths; the original saliency maps generated by using (c) SGSP, and the improved saliency maps generated by using (d) SGSP*1, (e) SGSP*3 and (f) SGSP*5, respectively.

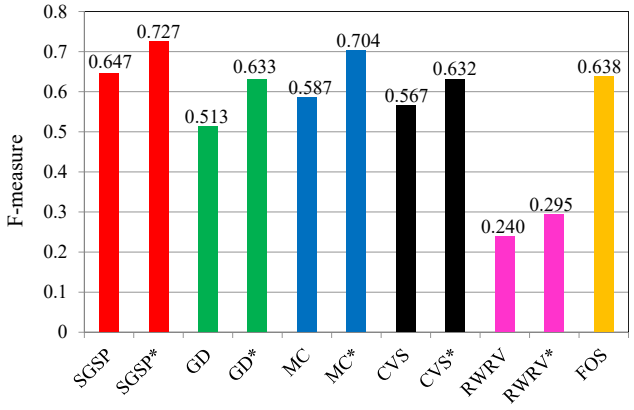


Fig. 14: Quantitative evaluation of video object segmentation using F-measures on the DAVIS dataset.

moving objects. In a word, incorporating more neighboring frames brings slight performance gain for videos with slowly moving objects, but incurs more computation costs. To balance the efficiency and effectiveness, we set the neighbor size of forward and backward frames to one.

2) Evaluation of video object segmentation: We objectively evaluate the quality of video object segmentation using the saliency maps generated by the original saliency models, *i.e.* SGSP, GD, MC, CVS and RWRV, and the corresponding improved versions, *i.e.* SGSP*, GD*, MC*, CVS* and RWRV*. The video object segmentation results are obtained by using the graph cut based segmentation method [66] with the aforementioned ten groups of saliency maps. Furthermore, we also compared with the video object segmentation method FOS [67]. Here, we use the average F-measure defined in Eq. (15) to measure the segmentation quality. As shown in Fig. 14, we can see that the improved saliency maps generated by SGSP*, GD*, MC*, CVS* and RWRV* consistently result in the better segmentation quality compared to the original saliency maps. Concretely, the maximum improvement appears between GD and GD* with an increase of 0.12 on F-measure, from 0.513 to 0.633. The minimum improvement occurs between RWRV and RWRV* with an increase of 0.055 on F-measure, from 0.240 to 0.295. For all the five saliency models, the average improvement on F-measure is 0.087. Besides, among the five saliency models, the video object segmentation with GD, MC and CVS performs worse than FOS, but with the deployment of our framework, the corresponding segmentation with GD*, MC* and CVS* achieves comparable or even better performance than FOS. This clearly demonstrates that our framework can generate the better saliency maps for video object segmentation.

3) Effectiveness of inter-frame interaction: For the purpose of validating the effectiveness of inter-frame interaction using our framework, we present two examples in Fig. 15, where the object is not so salient initially in a particular frame but becomes salient by interacting with other frames. Here, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia

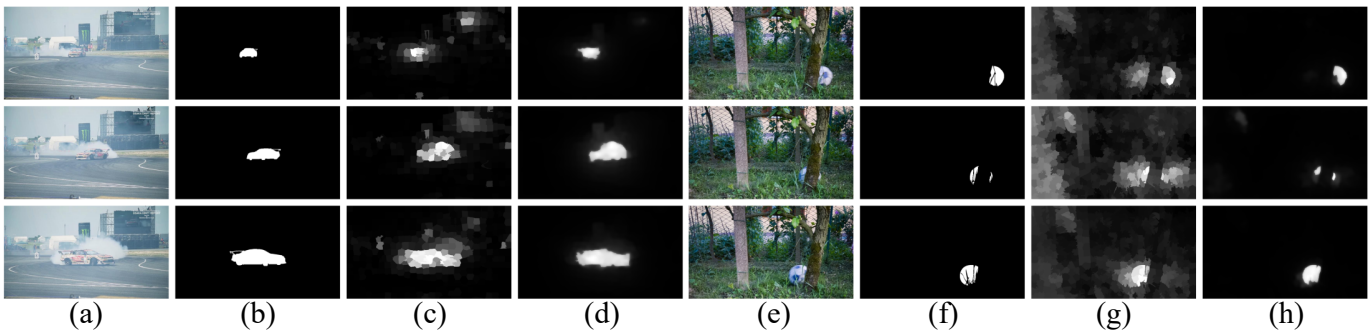


Fig. 15: Examples of saliency maps for some videos where the object is not salient initially but becomes salient by interacting with other frames. (a) and (e): Input video frames; (b) and (f): binary ground truths; (c) and (g): saliency maps generated by SGSP; (d) and (h): saliency maps generated by SGSP*.

saliency maps are generated using the original model SGSP and its improved version SGSP*. In Fig. 15(a), the car moves from far to near, and becomes salient gradually. It can be seen from Fig. 15(c) that the saliency maps generated by SGSP falsely highlight the background regions around the car. In contrast, as shown in Fig. 15(d), due to the inter-frame interaction using our framework, the improved version SGSP* can highlight the car uniformly and suppress background regions effectively. For the second example shown in Fig. 15(e), we can see that the soccer is not so salient in the middle row due to the occlusion of the tree. It can be seen from Fig. 15(g) that some background regions are falsely highlighted and the boundaries of salient object are not well-defined in the saliency maps generated by SGSP. In contrast, as shown in Fig. 15(h), the improved version SGSP* can uniformly highlight the salient object regions with well-defined regions, and also can suppress the background regions more effectively. The reason behind this is that the inter-frame interaction using our framework provides more information about salient objects and background.

D. Failure Examples and Analysis

As aforementioned, our framework can improve the quality of saliency maps generated by the existing video saliency models on both quantitative and qualitative evaluations. However, our framework cannot obtain satisfactory results when dealing with some challenging videos such as the examples shown in Fig. 16. For the example shown in the top two rows, the salient object (*i.e.* the cyclist) is shot from behind with severe camera jitter. As a result, the existing video saliency models (*i.e.* SGSP, GD, MC, CVS and RWRV) cannot locate the salient object. As shown in the top two rows of Fig. 16(c, e, g, i, k), some background regions are mistakenly identified as salient object. Based on such initial saliency maps, the obtained final saliency maps (*i.e.* SGSP*, GD*, MC*, CVS* and RWRV*) also cannot capture the correct salient object as shown in the top two rows of Fig. 16(d, f, h, j, l). In the bottom example, besides the salient object (the singer), the colorful screen content and the audiences also move quickly. It can be seen that the existing saliency models improperly highlight some background regions as shown in the bottom two rows of Fig. 16(c, e, g, i, k). As a result, our final saliency maps are also unable to tackle such challenging videos as revealed by the bottom two rows of Fig. 16(d, f, h, j, l).

Overall, it can be concluded that our framework depends on the orginal video saliency models, which provide training data for the saliency estimation in our framework. If the original video saliency model fails to offer sufficiently reliable training samples, it is difficult for the final saliency maps improved by our framework to make effective improvements on such challenging videos as shown in Fig. 16.

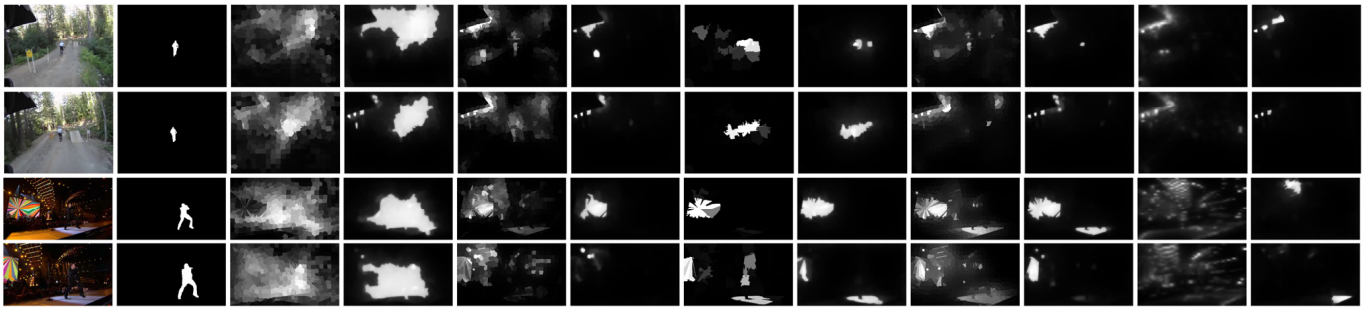
E. Computation Cost

In this section, we report the computation cost of the proposed framework. Our method is implemented on a PC with Intel Core i7-4790K 4GHz CPU and 32GB RAM. Table II gives the average processing time per frame with the video resolution 320×240. Taking SGSP* for example, the average test time for one frame is 5.928 seconds excluding the generation of initial saliency maps by SGSP. Specifically, the optical flow estimation consumes 2.977 seconds, which takes 50.22% of the total processing time. The extraction of features, localized estimation step and spatiotemporal refinement step take 1.143 seconds, 1.378 seconds and 0.430 seconds, respectively, which account for 19.28%, 23.25%, and 7.25% of the total processing time. It can be observed that optical flow estimation and localized estimation are the two most time-consuming components, and they occupy about 73.47% of the total processing time. Thus, the efficiency is one of the limitations of our framework. There are two potential ways to relieve the computational complexity and speed up our algorithm. The first one is to resize every frame of a video to a low resolution for calculating saliency maps, and then resize the obtained final saliency map back to the original resolution. The second one is to use GPU to accelerate the LDOF process.

V. Conclusion

This paper proposed a novel framework to improve saliency detection results generated by existing video saliency models. The framework consists of three key steps including localized estimation, spatiotemporal refinement, and saliency update. Firstly, by considering the temporal consistency and strong

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829605, IEEE Transactions on Multimedia



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 16: Failure examples. (a): Input video frames, (b): binary ground truths, (c): SGSP, (d): SGSP*, (e): GD, (f): GD*, (g): MC, (h): MC*, (i): CVS, (j): CVS*, (k): RWRV, (l): RWRV*.

TABLE II: Processing time of every step for one frame

Component	Optical Flow Estimation (LDOF)	Feature Extraction	Localized Estimation	Spatiotemporal Refinement	Total Time
Time(second)	2.977	1.143	1.378	0.430	5.928

correlation among temporally adjacent frames, a local temporal window based estimation models, *i.e.* localized estimation models, are learned to obtain the temporary saliency map. Such temporary saliency map can preserve the global shape of salient object in a video. Secondly, by incorporating the appearance and motion information simultaneously, a spatiotemporal refinement step is deployed to further improve the temporary saliency map and obtain the final saliency map with well-defined boundaries. Finally, the final saliency map is used to update the initial saliency map of current frame, which provides more reliable information for processing the next frame. Extensive experiments are performed on four challenging public video datasets, and the results show that the proposed framework consistently elevates the performance of the state-of-the-art video saliency models with significant improvements on four datasets.

REFERENCES

[1] 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.

[2] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, "Region diversity maximization for salient object detection," *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 215–218, 2012.

[3] X. Zhou, Z. Liu, G. Sun, L. Ye, and X. Wang, "Improving saliency detection via multiple kernel boosting and adaptive fusion," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 517–521, 2016.

[4] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1937–1952, 2014.

[5] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1275–1289, 2012.

[6] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[7] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Transactions on Multimedia*, 2017, doi:10.1109/TMM.2017.2693022.

[8] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1896–1909, 2016.

[9] L. Chen, J. Shen, W. Wang, and B. Ni, "Video object segmentation via dense trajectories," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2225–2234, 2015.

[10] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *British Machine Vision Conference, BMVC*, 2014, pp. 1–8.

[11] Z. Yuan, T. Lu, Y. Huang, D. Wu, and H. Yu, "Addressing visual consistency in video retargeting: A refined homogeneous approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 6, pp. 890–903, 2012.

[12] H. Du, Z. Liu, J. Jiang, and L. Shen, "Stretchability-aware block scaling for image retargeting," *Journal of Visual Communication and Image Representation*, vol. 24, no. 4, pp. 499–508, 2013.

[13] B. Yan, B. Yuan, and B. Yang, "Effective video retargeting with jittery assessment," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 272–277, 2014.

[14] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[15] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[16] J. Guo, B. Song, and X. Du, "Significance evaluation of video data over media cloud based on compressed sensing," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1297–1304, 2016.

[17] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society Conference on*. IEEE, 2005, pp. 631–637.

[18] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Advances in Neural Information Processing Systems,NIPS*, 2008, pp. 497–504.

[19] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 171–177, 2010.

[20] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, vol. 9, no. 12, pp. 15–15, 2009.

[21] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth mover's distance-based saliency measurement and nonlinear feature combination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 314–328, 2013.

[22] W. Kim and C. Kim, "Spatiotemporal saliency detection using textural contrast and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 646–659, 2014.

[23] C. Liu, P. C. Yuen, and G. Qiu, "Object motion detection using information theoretic spatio-temporal saliency," *Pattern Recognition*, vol. 42, no. 11, pp. 2897–2906, 2009.

[24] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," in *Asian Conference on Computer Vision, ACCV*. Springer, 2009, pp. 246–257.

[25] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems,NIPS*, 2009, pp. 681–688.

[26] V. Gopalakrishnan, D. Rajan, and Y. Hu, "A linear dynamical system framework for salient motion detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 5, pp. 683–692, 2012.

[27] K. Muthuswamy and D. Rajan, "Salient motion detection through state controllability," in *Acoustics, Speech and Signal Processing, ICASSP, IEEE International Conference on*. IEEE, 2012, pp. 1465–1468.

[28] X. Cui, Q. Liu, and D. Metaxas, "Temporal spectral residual: fast motion saliency detection," in *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 617–620.

[29] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *International Journal of Computer Vision*, vol. 90, no. 2, pp. 150–165, 2010.

[30] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "Intrinsic dimensionality predicts the saliency of natural dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1080–1091, 2012.

[31] W.-F. Lee, T.-H. Huang, S.-L. Yeh, and H. H. Chen, "Learning-based prediction of visual attention for video signals," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3028–3038, 2011.

[32] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1336–1349, 2014.

[33] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan, "Regularized feature reconstruction for spatio-temporal saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3120–3132, 2013.

[34] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[35] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 231–243, 2009.

[36] G. Abdollahian, C. M. Taskiran, Z. Pizlo, and E. J. Delp, "Camera motion-based analysis of user generated video," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 28–41, 2010.

[37] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, 2011.

[38] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27–38, 2014.

[39] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[40] 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.

[41] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3156–3170, 2017.

[42] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, 2014.

[43] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, doi:10.1109/TCSVT.2016.2595324.

[44] 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, 2015.

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

[46] B. Kuipers and P. Beeson, "Bootstrap learning for place recognition," in *Association for the Advancement of Artificial Intelligence*. AAAI, 2002, pp. 174–180.

[47] 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.

[48] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[49] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[50] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society Conference on*. IEEE, 2013, pp. 2083–2090.

[51] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society Conference on*. IEEE, 2015, pp. 1265–1274.

[52] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society Conference on*. IEEE, 2015, pp. 5455–5463.

[53] Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with cnns," in *European Conference on Computer Vision, ECCV*. Springer, 2016, pp. 809–825.

[54] L. Wang, L. Wang, H. Lu, P. Zhang, Zhang, and X. R. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision, ECCV*. Springer, 2016, pp. 825–841.

[55] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society Conference on*. IEEE, 2015, pp. 1884–1892.

[56] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society Conference on*. IEEE, 2014, pp. 2814–2821.

[57] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, 2015.

[58] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3425–3436, 2017.

[59] C. Aytekin, H. Possegger, T. Mauthner, S. Kiranyaz, H. Bischof, and M. Gabbouj, "Spatiotemporal saliency estimation by spectral foreground detection," *IEEE Transactions on Multimedia*, 2017, doi: 10.1109/TMM.2017.2713982.

[60] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[61] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognition*, vol. 42, no. 3, pp. 425–436, 2009.

[62] 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, 2011.

[63] N. Otsu, "A thresholding selection method from gray-level histogram," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[64] 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*. IEEE, 2013, pp. 2192–2199.

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

[66] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

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