Video Saliency Detection via Sparsity-based Reconstruction and Propagation

Runmin Cong, Jianjun Lei, *Senior Member, IEEE,* Huazhu Fu, *Senior Member, IEEE,* Fatih Porikli, *Fellow, IEEE,* Qingming Huang, *Fellow, IEEE,* and Chunping Hou

***Abstract*—Video saliency detection aims to continuously discover the motion-related salient objects from the video sequences. Since it needs to consider the spatial and temporal constraints jointly, video saliency detection is more challenging than image saliency detection. In this paper, we propose a new method to detect the salient objects in video based on sparse reconstruction and propagation. With the assistance of novel static and motion priors, a single-frame saliency model is firstly designed to represent the spatial saliency in each individual frame via the sparsity-based reconstruction. Then, through a progressive sparsity-based propagation, the sequential correspondence in the temporal space is captured to produce the inter-frame saliency map. Finally, these two maps are incorporated into a global optimization model to achieve spatiotemporal smoothness and global consistency of the salient object in the whole video. Experiments on three large-scale video saliency datasets demonstrate that the proposed method outperforms the state-of-the-art algorithms both qualitatively and quantitatively.**

***Index Terms*—Video saliency detection, sparse reconstruction, color and motion prior, forward-backward propagation, global optimization.**

I. Introduction

THE visual attention mechanism is remarkably effective in perceiving the contents and selecting the salient regions from the complex scenes. Imitating this, visual saliency techniques have facilitated a broad range of computer vision tasks such as image segmentation [1], foreground annotation [2], thumbnail creation [3], photo cropping [4], image enhancement [5], [6], and quality assessment [7], [8].

In the past few decades, saliency detection for static image has gained much attention and achieved encouraging performances on the public benchmarks. The saliency detection models can be divided into two categories, *i.e.*, data-driven bottom-up model and task-driven top-down model. In early, saliency detection methods focus on bottom-up visual attention mechanisms by using various low-level features, priors, and techniques, such as compactness prior [9], background prior [10], sparse coding [11]–[13], random walks [14], hierarchical decomposition [15], cellular automata [16], color transform [17], metric learning [18], and Bayesian framework [19]. More recently, the top-down saliency models using the deep learning have achieved remarkable performance [20]–[25].

By contrast, video saliency detection still remains as a relatively challenging and emerging issue. Different from the image saliency detection, video saliency detection aims to continuously locate the motion-related salient object from the video sequences by considering both the spatial and temporal information jointly, where the spatial information represents the intra-frame saliency in the individual frame, and the temporal information provides the inter-frame constraints and motion cues. Moreover, the salient objects in video are continuous in time axis and consistent among different frames, and the motion information is essential to distinguish the salient object from a complex scene.

In video data, the moving objects often attract more attention than the static ones. However, not all moving objects are salient targets and need to be further discriminated by the surrounding regions and adjacent frames. Therefore, how to make full use of the motion information to highlight the salient regions and suppress the backgrounds is essential to video saliency detection. Some motion-based features, such as optical flow contrast and optical flow gradient, have been utilized to separate the foreground regions from the video directly [26], [27]. Nevertheless, these methods are fragile due to the noises and moving backgrounds. In this paper, we introduce the motion compactness and motion uniqueness as motion cues to improve the motion saliency measurement. The motion compactness describes the distribution of the optical flow, and the motion uniqueness represents the appearance characteristics of the motion amplitude information.

Exhibiting robustness to noise, sparsity-based techniques have been demonstrated to yield discriminative representations that have potential to improve the performances in a variety of inference tasks. In addition, several saliency detection methods [11]–[13] construct sparse models from the image and report satisfactory results against complex backgrounds. Li *et al.* [11] computed the saliency via the multi-scale dense and sparse reconstruction, and the reconstruction errors are used

Manuscript received Mar. 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61722112, Grant 61520106002, Grant 61731003, Grant 61332016, Grant 61620106009, Grant U1636214, Grant 61602344, in part by the National Key Research and Development Program of China under Grant 2017YFB1002900, and in part by the Key Research Program of Frontier Sciences, CAS under Grant QYZDJ-SSW-SYS013. (*Corresponding author: Jianjun Lei*)

R. Cong is with the Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: rmcong@126.com).

J. Lei and C. Hou are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: jjlei@tju.edu.cn; hcp@tju.edu.cn).

H. Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: hzfu@ieee.org).

F. Porikli is with the Research School of Engineering, Australian National University, Canberra, ACT 0200, Australia (e-mail: fatih.porikli@anu.edu.au).

Q. Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@ucas.ac.cn).

to represent the saliency. In [12], a weighted sparse coding framework on different data inputs is proposed to locate the salient objects. Recently, Yuan *et al.* [13] combined the deep neural network (DNN) and dense and sparse labeling (DSL) framework for saliency detection. By contrast, only a few studies [28], [29] employed the sparse representations to achieve video saliency detection. However, these methods only use sparse representations to capture the spatial information from individual frames, thus do not generalize well on the temporal space. To address this, we develop a progressive sparse propagation framework with the forward-backward strategy to model the inter-frame correspondence and generate the inter-frame saliency map in the spatiotemporal space. For the forward pass, the previous frame is utilized to build the forward dictionary and reconstruct the current frame. On the contrary, the backward pass processes the video from the last frame to the first frame, and the current frame is reconstructed by the backward dictionary constructed by the latter frame. Through the bidirectional propagation processes, the inter-frame relationship is exploited and the inter-frame saliency is achieved.

Generally, spatiotemporal consistency should be considered in video saliency models to achieve more homogeneous result, *i.e.*, the saliency value of the salient region or background should not change drastically along the time axis. Moreover, in the most of existing methods, the input video is processed frame by frame without considering a global measure across the whole video sequence. In this way, the saliency result can only guarantee local consistency rather than global consistency. Therefore, we propose a global optimization scheme based on energy function to obtain more homogeneous and consistent saliency result, which includes unary data term, spatiotemporal smooth term, spatial incompatibility term, and global consistency term.

In summary, we present a video saliency detection method based on sparse reconstruction and propagation, considering the spatiotemporal priors and global consistency. The main contributions are summarized as follows:

(1) A novel sparsity-based saliency reconstruction is introduced to generate single-frame saliency map, making the best use of the static and motion priors. The motion priors are defined as motion compactness cue and motion uniqueness cue.
(2) A new and efficient sparsity-based saliency propagation is presented to capture the correspondence in the temporal space and produce inter-frame saliency map. The salient object is sequentially reconstructed by the forward and backward dictionaries.
(3) To attain the global consistency of the salient object in the whole video, a global optimization model, which integrates unary data term, spatiotemporal smooth term, spatial incompatibility term, and global consistency term, is formulated.
(4) Extensive experimental evaluations and ablation studies show that the proposed method achieves a superior performance, outperforming the current state-of-the-art approaches.

The remainder of the paper is organized as follow. Section II presents a review of the related work. Section III details the proposed sparse-based video saliency framework. The experimental comparisons and analyses are presented in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

In this section, we review the related work in image saliency detection and video saliency detection.

A. Image Saliency Detection

The last decade have witnessed the significant progress towards image saliency detection, and a number of methods have been presented [9]–[25]. Zhou *et al.* [9] integrated two complementary cues including compactness and local contrast to detect the salient regions. Li *et al.* [15] proposed a saliency detection algorithm by using reconstruction errors derived from the dense reconstruction and sparse reconstruction. Li *et al.* [21] proposed an end-to-end deep network, which consists of a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. Liu and Han [23] proposed a deep hierarchical saliency network, which integrates a CNN over the global view producing the global saliency map and a hierarchical recurrent CNN recovering the image details. Hou *et al.* [24] developed a deeply supervised saliency detection network, which introduces a series of short connections between shallower and deeper side-output layers to highlight the entire salient object and accurately locate the boundary. Moreover, some works focus on performance evaluation for saliency detection task, such as S-measure [30] and E-measure [31].

In addition, as the extended studies, many models focus on extracting the salient object from the RGBD images [32]–[39] or image group [40]–[46]. Feng *et al.* [33] proposed a Local Background Enclosure (LBE) measure to directly capture salient structure from depth map. Considering the negative influence of poor depth map, Cong *et al.* [34] proposed a depth confidence measure to evaluate the quality of depth map, and combined with multiple cues to achieve RGBD saliency detection. Chen *et al.* [38] presented a multi-scale multi-path fusion network for RGB-D saliency detection, which advances the traditional two-stream fusion architecture. Chen *et al.* [39] proposed a three-stream attention-aware multi-modal fusion network for RGBD saliency detection, which integrates the cross-modal distillation stream and the channel-wise attention mechanism. In [40], a cluster-based co-saliency detection algorithm for multiple images is proposed, which integrates the contrast, spatial, and corresponding cues. With the FCN framework, Wei *et al.* [42] proposed an end-to-end group-wise deep co-saliency detection model via the collaborative learning structure with convolution-deconvolution. Han *et al.* [43] introduced metric learning into co-saliency detection, which jointly learns discriminative feature representation and co-salient object detector via a new objective function. Cong *et al.* [44] proposed a co-saliency detection method for RGBD images by using the multi-constraint feature matching and cross label propagation.

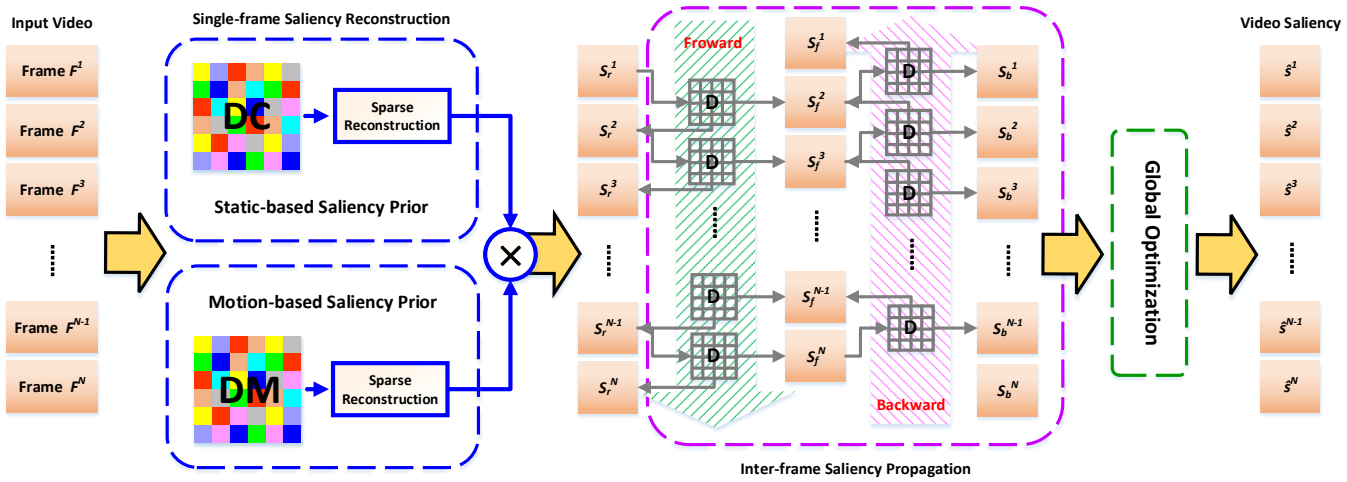


Fig. 1. The flowchart of the proposed video saliency detection framework.

However, these methods mainly focus on capturing saliency attribute from a single image or noncontinuous images, and do not directly extend to video saliency detection.

B. Video Saliency Detection

In video sequences, moving objects generally draw more attention than static ones, even if the static objects appear more salient in a single frame. Inspired by biological mechanisms of motion-based perceptual grouping, Mahadevan *et al.* [47] proposed a spatiotemporal saliency method based on a center-surround framework. Fang *et al.* [48] proposed a video saliency model based on feature contrast in compressed domain, where four features are extracted from the discrete cosine transform coefficients and motion vectors in video bitstream. In [26], the temporal saliency and spatial saliency were adaptively fused to generate spatiotemporal saliency map, which incorporates superpixel-level motion distinctiveness, global contrast, and spatial sparsity. Wang *et al.* [27] presented a novel spatiotemporal saliency model based on the gradient flow field and energy optimization. Kim *et al.* [49] presented a robust saliency detection algorithm for video, which integrates spatiotemporal information through the random walk with restart framework. Wang *et al.* [50] integrated the spatiotemporal edge cue and geodesic distance to obtain accurate spatiotemporal saliency result. In [51], the color saliency and motion saliency are fused in a batch-wise way, and the temporal smoothness is guaranteed through low-rank coherency diffusion. Liu *et al.* [52] presented a spatiotemporal saliency model for unconstrained videos, which utilizes superpixel-level graph-based motion saliency and spatiotemporal propagation. Xi *et al.* [53] proposed a graph-based video saliency detection approach by combining the temporal background prior and spatial background prior simultaneously.

It is worth mentioning that deep learning has been successfully applied to the video saliency detection [54]–[59]. Le *et al.* [54] proposed an end-to-end deeply supervised 3D recurrent fully convolutional network (DSRFCN3D) for salient object detection in video, which contains an encoder network, a decoder network, and a refinement mechanism. Wang *et al.* [56] proposed an effective and efficient video saliency framework by using deep learning, which consists of two modules, *i.e.*, the static saliency model and dynamic saliency model. Moreover, the saliency result from the static saliency network is directly incorporated in the dynamic saliency network to produce final spatiotemporal saliency inference with less computation load. In addition, Li *et al.* [58] proposed a very large video saliency detection dataset and an unsupervised approach for video salient object detection by using saliency-guided stacked autoencoders.

Compared with image saliency detection, video saliency detection is still an emerging topic that needs to be further investigated. Most of the existing methods usually devote to setting up a direct fusion scheme through integrating image cues with motion cues, and lacking of a holistic framework to fully explore the intra-frame and inter-frame information jointly.

III. Proposed Method

Motivated by the inherent aspects of salient objects in video, three progressive steps are proposed to achieve video saliency detection, *i.e.*, single-frame saliency reconstruction, inter-frame saliency propagation, and global optimization. The flowchart is shown in Fig. 1. First, with the intuition that salient objects in the video should be salient in each individual frame, thus, a single-frame saliency model is designed to capture the spatial saliency using sparse reconstruction with the static and motion saliency priors. Then, an inter-frame saliency propagation with forward-backward strategy is utilized to model the sequential correspondence in the temporal space and generate the inter-frame saliency map. Finally, a global optimization model is designed to guarantee the global consistency of the salient object across the whole video and achieve more homogeneous saliency result. We explain each of these steps next.

A. Single-frame Saliency Reconstruction

For video saliency detection, the detected object should be salient with respect to the background and underlying motion in each frame. To this end, a sparse reconstruction model with two saliency priors is used to detect the salient object in each individual frame. The first one is the static saliency prior, which utilizes three color saliency cues to construct a color-based reconstruction dictionary (DC). The second one is the motion-based saliency prior, which integrates the motion uniqueness cue and motion compactness cue to build a motion-based dictionary (DM).

Given a video sequence $\mathbf{F} = \{F^t\}_{t=1}^{N}$ including N frames F^t, we firstly derive some homogeneous superpixels $\mathbf{R}^t = \{r_k^t\}_{k=1}^{N^t}$ using SLIC algorithm [60] for each frame F^t, where N^t is the number of superpixels. In addition, the large displacement optical flow [61] is calculated to represent the pixel-level motion vector. The motion vector $\boldsymbol{v}_k^t$ of superpixel r_k^t is defined as the mean value of pixel-level motion vector in the superpixel.

*1) **Static-based Saliency Prior**:* The static-based saliency prior measures the static saliency in each frame by incorporating the background dictionary into a sparse representation framework. Three color-based cues, including background cue, compactness cue, and uniqueness cue, are integrated to select the background seeds and build the dictionary for reconstruction.

Background Cue. It is generally accepted that in video production, the important objects are close to the image center rather than the boundaries, which is a natural response of the cameraman operating the imaging system. Thus, the superpixels located at the image boundaries are more likely to be the background seeds, and this observation has been applied to many saliency detection models [10], [11]. In our work, the superpixels along the image boundaries are selected as the background candidate set Φ_{SB}^t that represents the spatial location attribute of the background regions.

Static Compactness Cue. The salient regions incline to have a small spatial variance, whereas the backgrounds usually have a high spatial variance since their superpixels are often distributed over the entire image. Therefore, the compactness cue is introduced to describe the spatial distribution of the background regions. Following the DCLC method [9], the spatial variance of superpixel r_k^t is calculated by:

$$v_s(r_k^t) = \frac{\sum_{l=1}^{N^t} a_{kl}^t \cdot n_l^t \cdot \|\boldsymbol{p}_l^t - \boldsymbol{u}_k^t\|_2}{\sum_{l=1}^{N^t} a_{kl}^t \cdot n_l^t} \tag{1}$$

where n_l^t represents the number of pixels that belong to the superpixel r_l^t, $\boldsymbol{p}_l^t$ denotes the spatial coordinates of superpixel r_l^t, $\boldsymbol{u}_k^t$ is the spatial mean, $a_{kl}^t = \exp(-\|\boldsymbol{lc}_k^t - \boldsymbol{lc}_l^t\|_2/\sigma^2)$ denotes the color similarity between two superpixels, $\boldsymbol{lc}_k^t$ is the mean Lab color value of superpixel r_k^t, and σ^2 is a parameter to control degree of the similarity, which is set to 0.1 as suggested in [14]. Then, the top Q_1 superpixels with larger spatial variances are selected as the compactness-based background candidate set Φ_{SC}^t.

Static Uniqueness Cue. The third cue represents the global appearance of the background regions in which the salient object shows different properties in appearance compared with the background. In our work, a cluster-based method is proposed to define the uniqueness cue. First, *K-means++* clustering [62] is used to group the superpixels into K clusters $\{C_i^t\}_{i=1}^{K}$ with cluster centers $\{c_i^t\}_{i=1}^{K}$, where the cluster number is set to 20 in the experiments. Then, two clusters with the largest Euclidean distance are selected by:

$$\{C_p^t, C_q^t\} = \underset{m,n\in\{1,\ldots,K\}}{\arg\max} E_d(c_m^t, c_n^t) \cdot e^{-|v_s(C_m^t) - v_s(C_n^t)|} \tag{2}$$

where $E_d(c_m^t, c_n^t)$ is the Euclidean distance between the two cluster centers, and $v_s(C_m^t)$ denotes the mean spatial variance of the cluster C_m^t. The selected two clusters correspond to one foreground cluster and one background cluster. Finally, a decision scheme considering the spatial variance and background probability is designed to determine the uniqueness-based background candidate set Φ_{SU}^t as:

$$\Phi_{SU}^t = \begin{cases} \{C_p^t\}, & \text{if } [v_s(C_p^t) > v_s(C_q^t)] \cap [P_b(C_p^t) > P_b(C_q^t)] \\ \{C_q^t\}, & \text{if } [v_s(C_p^t) \le v_s(C_q^t)] \cap [P_b(C_p^t) \le P_b(C_q^t)] \\ \varnothing, & \text{otherwise} \end{cases} \tag{3}$$

where $P_b(C_p^t)$ is the mean background probability of the cluster C_p^t by using the method in [10].

Static-based Saliency Reconstruction. The final background set is obtained by combining all background candidates as $\Phi_{CB}^t = \Phi_{SB}^t \cup \Phi_{SC}^t \cup \Phi_{SU}^t$. Then, three types of features considering the color components, spatial location, and texture distribution are used to describe each superpixel. The color features in different color spaces are the intuitive representation of the superpixel, which is denoted as $\boldsymbol{c} = [R, G, B, L, a, b, H, S, V]$. The position coordinates benefit for depicting the spatial relationship of the superpixel, which is represented as $\boldsymbol{p} = [x, y]$. The texton histogram $\boldsymbol{t}$ describes the local texture information of the superpixel [63]. Similar to [11], [12], [35], all these hand-crafted features are firstly normalized to [0, 1], and then concatenated into a feature vector to represent the superpixel r_k^t, which is denoted as $\boldsymbol{x}_k^t = [\boldsymbol{c}_k^t, \boldsymbol{p}_k^t, \boldsymbol{t}_k^t]^T$. The background dictionary $\mathbf{D}_B^t$ is constructed by the feature representations of the stacking background seeds in Φ_{CB}^t.

Based on the assumption that reconstruction error should be different for foreground and background through a sparse reconstruction model, the image saliency can be measured by the reconstruction error [11]. Each superpixel r_k^t is encoded by:

$$\alpha_k^{t*} = \underset{\alpha_k^t}{\arg\min} \|\boldsymbol{x}_k^t - \mathbf{D}_B^t \cdot \alpha_k^t\|_2^2 + \lambda \cdot \|\alpha_k^t\|_1 \tag{4}$$

where α_k^{t*} is the optimal sparse coefficient for superpixel r_k^t, $\mathbf{D}_B^t$ denotes the background dictionary for frame F^t, $\boldsymbol{x}_k^t$ is the feature representation of superpixel r_k^t, λ is set to 0.01 as suggested in [11], and $\|\cdot\|_1$, $\|\cdot\|_2$ indicate the ℓ_1-norm and ℓ_2-norm functions, respectively.

For the sparse reconstruction with a background dictionary, the salient region will have a large reconstruction error, while the reconstruction error of the background region should be

(a) (b) (c)

Fig. 2. Optical flow data of different video frames. (a) RGB image. (b) Optical flow map. (c) The magnitude of optical flow data.

small. Thus, the saliency of superpixel r_k^t can be measured by the reconstruction error ε_k^t:

$$S_s(r_k^t) = \varepsilon_k^t = \|\boldsymbol{x}_k^t - \mathbf{D}_B^t \cdot \alpha_k^{t*}\|_2^2 \tag{5}$$

where $S_s(r_k^t)$ denotes the static saliency value of superpixel r_k^t via the reconstruction error ε_k^t.

*2) **Motion-based Saliency Prior**:* Moving target attracts more attention in visual perception, thus, we introduce a motion-based saliency prior to represent the salient object from the perspective of motion space. An example of the optical flow data is shown in Fig. 2, where the spatial distribution of moving object is more concentrated than the background regions in the optical flow data. In addition, the moving object is often different from the background regions in terms of the magnitude of optical flow (MOF), which is consistent with the uniqueness cue in the color space. Based on these observations, we extend the color-related cues to the motion field and determine the background seeds for dictionary construction.

Motion Compactness Cue. Our intuition is that, in the whole video sequences, the spatial location distribution of moving object is more concentrated and compact in the optical flow field, whereas the background is distributed over the entire image. Therefore, we introduce a "motion compactness" cue to describe the distribution of the optical flow data and determine the background candidates. Similar to the color-based spatial variance, the motion-based spatial variance is defined as:

$$v_m(r_k^t) = \frac{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t \cdot \|\boldsymbol{p}_l^t - \widetilde{\boldsymbol{u}}_k^t\|_2}{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t} \tag{6}$$

where $m_{kl}^t = \exp(-\|\boldsymbol{v}_k^t - \boldsymbol{v}_l^t\|_2/\sigma^2)$ is the motion similarity between two superpixels, $\boldsymbol{v}_k^t$ denotes the optical flow vector of superpixel r_k^t, σ^2 is a constant parameter, n_l^t represents the number of pixels that belong to superpixel r_l^t, $\boldsymbol{p}_l^t = [x_l^t, y_l^t]$ is the centroid coordinates of superpixel r_l^t, and $\widetilde{\boldsymbol{u}}_k^t = [ux_k^t, uy_k^t]$ represents the spatial mean in optical flow field, which is defined as:

$$\begin{cases} ux_k^t = \frac{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t \cdot x_l^t}{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t} \\ uy_k^t = \frac{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t \cdot y_l^t}{\sum_{l=1}^{N^t} m_{kl}^t \cdot n_l^t} \end{cases} \tag{7}$$

where a larger $v_m(r_k^t)$ indicates that the distribution of superpixel r_k^t in optical flow field is more dispersed, thus, the background probability of the superpixel is greater. Then, the top Q_1 superpixels with larger motion-based spatial variance are composed to the background candidate set Φ_{MC}^t.

Motion Uniqueness Cue. In general, the moving target exhibits different motion appearance compared with the background regions in the MOF data. Therefore, we define a "motion uniqueness" cue in the MOF field by calculating the global contrast of each superpixel:

$$u_m(r_k^t) = \sum_{k=1,k\neq l}^{N^t} |M_f(r_k^t) - M_f(r_l^t)| \cdot e^{-E_d(\boldsymbol{p}_k^t, \boldsymbol{p}_l^t)/\sigma^2} \tag{8}$$

where $u_m(r_k^t)$ is the motion-based uniqueness measure of superpixel r_k^t, $M_f(r_k^t)$ denotes the MOF value of the superpixel r_k^t, and $E_d(\cdot)$ is the Euclidean distance function between two superpixels, which emphasizes the effect of closer superpixels. The smaller the uniqueness value is, the greater background probability of the superpixel achieves. Thus, we select the top Q_1 superpixels with smaller motion-based uniqueness value to build the background candidate set Φ_{MU}^t.

Motion-based Saliency Reconstruction. The final motion-related background set is determined by combining two background candidate sets, as $\Phi_{MB}^t = \Phi_{MC}^t \cup \Phi_{MU}^t$. For the motion-based sparse reconstruction, the motion feature is necessarily introduced to represent the motion cue. Furthermore, in order to guarantee the robustness of the feature representation, the basic color components are also embedded into the feature pool. Each superpixel is represented as a 12-dimensional feature vector $\boldsymbol{x}_k^t = [\boldsymbol{c}_k^t, \boldsymbol{m}_k^t]^T$, where $\boldsymbol{c}$ is the 9-dimensional color feature, and $\boldsymbol{m}$ denotes the 3-dimensional motion feature involving the components and magnitude of optical flow data. Then, the feature representation of each motion-related background seed is used to construct the background dictionary for frame F^t as $\mathbf{D}_B^t$. At last, as same as the static-based saliency reconstruction in Eqs. (4)-(5), the motion saliency of each superpixel is represented by the reconstruction error, which is denoted as $S_m(r_k^t)$.

*3) **Single-frame Saliency Map**:* The static saliency and motion saliency aim to discover the salient object from different feature domains. We integrate these two saliency maps to produce the single-frame saliency map as:

$$S_r(r_k^t) = S_s(r_k^t) \cdot S_m(r_k^t). \tag{9}$$

B. Inter-frame Saliency Propagation

The sequential relationship across the time axis is crucial to video saliency detection. The salient object in an individual frame should be further discriminated by using the inter-frame information. Considering the high consistency and smoothness of the salient object in appearances and views between two adjacent frames, the previous frame can be employed to build a foreground dictionary and reconstruct the current frame in a forward way. Likewise, the current frame can be reconstructed by the next frame in a backward propagation manner. Therefore, a spatiotemporal saliency model is established via sparse propagation with a forward-backward strategy to smooth the salient object and suppress the background.

For the inter-frame sparsity-based saliency propagation, the directly adjacent frames are most relevant to the current frame,

which benefits for capturing the common attributes of the salient objects. The proposed forward-backward propagation strategy is a heuristic method for inter-frame relationship abstraction in a progressive manner. The forward saliency and backward saliency are progressively correlated, where the forward saliency result is embedded into the feature pool to construct the dictionary and conduct the backward propagation. Through the bidirectional propagation processes, the exploitation of inter-frame relationship becomes more comprehensive and accurate.

*1) **Forward Propagation**:* In the forward propagation, the current frame is reconstructed by a foreground dictionary derived from the previous frame, and the video is sequentially processed from the first frame to the last frame.

First, top Q_2 superpixels with larger single-frame saliency values in frame F^{t-1} are selected as the foreground seeds in the forward pass. Then, using the spatiotemporal features, each superpixel is represented as $\boldsymbol{x}_k^t = [\boldsymbol{c}_k^t, \boldsymbol{p}_k^t, \boldsymbol{t}_k^t, \boldsymbol{m}_k^t, S_r(r_k^t)]^T$, where $\boldsymbol{c}$ represents the 9-dimensional color feature, $\boldsymbol{p}$ is the 2-dimensional spatial coordinates, $\boldsymbol{t}$ is the 15-dimensional texton histogram, $\boldsymbol{m}$ denotes the 3-dimensional motion feature vector, and S_r is the single-frame saliency value. The feature representations of all foreground seeds from frame F^{t-1} are stacked to construct the forward foreground dictionary for frame F^t, which is denoted as $\mathbf{D}_F^{t-1}$.

Each superpixel in the current frame F^t is reconstructed by the forward foreground dictionary $\mathbf{D}_F^{t-1}$ through the sparse framework, and the reconstruction error $\overline{\varepsilon_k^t}$ is calculated to measure the forward saliency of superpixel r_k^t. Since the foreground dictionary is used for sparse reconstruction, the reconstruction error of the foreground regions should be small, and the background regions have a large reconstruction error. In other words, the superpixel with smaller reconstruction error should be assigned to a greater saliency value, and vice versa. Therefore, following [12], the forward saliency of superpixel r_k^t is measured by an exponential function of reconstruction error:

$$S_f(r_k^t) = \exp(-\overline{\varepsilon_k^t}/\sigma^2) = \exp(-\|\boldsymbol{x}_k^t - \mathbf{D}_F^{t-1} \cdot \overline{\alpha_k^{t*}}\|_2^2/\sigma^2) \tag{10}$$

where $S_f(r_k^t)$ is the saliency value in the forward pass, $\sigma^2 = 0.1$ is a weighted parameter, and $\overline{\alpha_k^{t*}}$ denotes the optimal sparse coefficient obtained by solving Eq. (4) with the forward foreground dictionary $\mathbf{D}_F^{t-1}$.

*2) **Backward Propagation**:* The forward propagation captures the pre-order inter-frame relationship. Similarly, a backward pass is further carried out, which processes the video from the last frame to the first frame in a post-order way. The backward pass is the same as the forward pass, except for the foreground dictionary construction.

In the backward propagation, the single-frame saliency and forward saliency are combined to determine the foreground seeds. First, top $Q_2/2$ superpixels with larger saliency values in the single-frame and forward saliency models are selected, respectively. Then, the union of these superpixels are determined as the final foreground seeds in the backward pass. Different from the forward pass, the forward saliency S_f is added into the feature pool, which is denoted as $\boldsymbol{x}_k^t = [\boldsymbol{c}_k^t, \boldsymbol{p}_k^t, \boldsymbol{t}_k^t, \boldsymbol{m}_k^t, S_r(r_k^t), S_f(r_k^t)]^T$. Finally, the backward reconstruction error $\widetilde{\varepsilon_k^t}$ is used to define the backward saliency:

$$S_b(r_k^t) = \exp(-\widetilde{\varepsilon_k^t}/\sigma^2) = \exp(-\|\boldsymbol{x}_k^t - \mathbf{D}_F^{t+1} \cdot \widetilde{\alpha_k^{t*}}\|_2^2/\sigma^2) \tag{11}$$

where $\widetilde{\alpha_k^{t*}}$ denotes the optimal sparse coefficient obtained by solving Eq. (4) with the backward foreground dictionary $\mathbf{D}_F^{t+1}$.

C. Global Optimization

In order to achieve superior and globally consistent saliency map, we propose an efficient optimization model with an energy function that consists of four complementary terms.

Unary Data Term. This term encourages the similarity between the final saliency map and initial saliency map, which is defined as:

$$E_u = \sum_k (\widehat{s_k^t} - s_k^t)^2 \tag{12}$$

where $\widehat{s_k^t}$ represents the final optimized saliency value of superpixel r_k^t, and $s_k^t = S_r(r_k^t) + S_f(r_k^t) + S_b(r_k^t)$ is the initial saliency value combining three obtained saliency maps.

Spatiotemporal Smooth Term. This term favors that all the similar and spatiotemporally adjacent superpixels across the whole video should be assigned to consistent saliency scores, which is calculated by:

$$E_s = \sum_{(k,l)\in\Omega_{st}} \omega_{kl} \cdot (\widehat{s_k^t} - \widehat{s_l^t})^2 \tag{13}$$

where $\omega_{kl} = \exp(-\|\boldsymbol{lc}_k - \boldsymbol{lc}_l\|_2/\sigma^2)$ is the Lab color similarity between superpixels r_k and r_l, $\boldsymbol{lc}_k$ is the mean Lab color value of superpixel r_k, and $\Omega_{st} = \Omega_s \cup \Omega_t$ is the spatiotemporal adjacent set. Ω_s is the spatially adjacent set in one frame, which is defined as:

$$\Omega_s = \{(r_m^t, r_n^t) |\ r_m^t \text{ and } r_n^t \text{ are spatially adjacent in } F^t\} \tag{14}$$

Following the settings in [27], the temporally adjacent set Ω_t is represented as:

$$\Omega_t = \{(r_m^t, r_n^{t'}) |\ \|\boldsymbol{p}_m^t - \boldsymbol{p}_n^{t'}\| \le 800 \ \&\ |t - t'| = 1\} \tag{15}$$

where $\boldsymbol{p}_m^t$ is the spatial coordinates of superpixel r_m^t, and t' denotes the frame index.

Spatial Incompatibility Term. Inspired by the related work [64], the distributions of the salient and background regions should have high probabilities at mutually exclusive domains. Thus, the spatial incompatibility term enforces that the same region should not have high foreground and background probabilities simultaneously, which is represented as:

$$E_i = \sum_{(k,l)\in\Omega_s} \omega_{kl} \cdot \widehat{s_k^t} \cdot \widehat{s_l^t} \tag{16}$$

When a highly probable salient region is surrounded by unlikely background neighbors, the spatial incompatibility energy is reduced. Therefore, for a low spatial incompatibility energy, the foreground and the background should form their own dominant regions.

Global Consistency Term. Video saliency detection aims to continuously locate the motion-related salient object from

the video sequences. In other words, the salient objects in video should be salient with distinct motion patterns in each individual frame, and appear in most of the frames. Therefore, the salient objects will not change throughout the whole video sequence, and need to be consistently highlighted. However, the input video is processed frame by frame in most of the existing methods, which ignores the global property across the whole video sequence. In this way, the saliency result only guarantees the local consistency rather than global consistency. In our work, the global consistency term is proposed to improve the consistency from the global perspective, which imposes the appearance of salient object approximate to a global video foreground model. Moreover, even for some occlusion regions that occur in the video, the global consistency term does not highlight them due to the introduction of global appearance similarity. In other words, the global consistency term only improves the global consistency of non-occluded salient objects, which can be used to process the long video sequences. This term is described as:

$$E_g = \sum_k \kappa_k \cdot \widehat{s_k^t}^2 \tag{17}$$

where $\kappa_k = \chi^2(\boldsymbol{h}_k, \boldsymbol{h}_{vs})$ is the chi-square distance of Lab color histograms between the superpixel and video foreground model. The top 10 superpixels with larger initial saliency value in each frame are extracted as the foreground samples to represent the foreground distribution of the whole video.

To sum up, the energy function is defined as follows:

$$E = \eta_1 \cdot E_u + \eta_2 \cdot E_s + \eta_3 \cdot E_i + \eta_4 \cdot E_g \tag{18}$$

where η_i is the weighting parameter for balancing the relative influence of different components. Following [27], the weighting parameter η_1 for unary data term is set to 0.5 to constrain the updating change not to be large, and other weighting parameters are set to 1 with equal contribution.

Let $\boldsymbol{s} = [s_k]_{N_a \times 1}$, and $\widehat{\boldsymbol{s}} = [\widehat{s_k}]_{N_a \times 1}$, where $N_a = \sum_{i=1}^{N} N^i$ is the total number of superpixels in the whole video. The energy function can be rewritten as the following matrix form:

$$\begin{aligned}\mathbf{E} =& \eta_1 \cdot (\widehat{\boldsymbol{s}} - \boldsymbol{s})^T \cdot (\widehat{\boldsymbol{s}} - \boldsymbol{s}) + \eta_2 \cdot \widehat{\boldsymbol{s}}^T \cdot (\mathbf{D}_{st} - \mathbf{W}_{st}) \cdot \widehat{\boldsymbol{s}} \\ &+ \eta_3 \cdot \widehat{\boldsymbol{s}}^T \cdot \mathbf{W}_s \cdot \widehat{\boldsymbol{s}} + \eta_4 \cdot \widehat{\boldsymbol{s}}^T \cdot \mathbf{K} \cdot \widehat{\boldsymbol{s}}\end{aligned} \tag{19}$$

where $\mathbf{W}_{st} = [\omega_{kl}]_{N_a \times N_a}^{(k,l) \in \Omega_{st}}$ is the spatiotemporal color similarity matrix, $\mathbf{D}_{st} = diag(d_1, d_2, \ldots, d_{N_a})$ denotes the degree matrix, $d_i = \sum_{j=1,(i,j)\in\Omega_{st}}^{N_a} \omega_{ij}$, $\mathbf{W}_s = [\omega_{kl}]_{N_a \times N_a}^{(k,l) \in \Omega_s}$ is the spatial color similarity matrix, and $\mathbf{K} = diag(\kappa_1, \kappa_2, \ldots, \kappa_{N_a})$ is the difference matrix between the superpixels and global foreground model.

Combining these four quadratic function terms, the energy function is a convex function, which can be solved by setting its derivative with respect to $\widehat{\boldsymbol{s}}$ to be 0. The transformation formula is represented as:

$$\eta_1 \cdot (\widehat{\boldsymbol{s}} - \boldsymbol{s}) + \eta_2 \cdot (\mathbf{D}_{st} - \mathbf{W}_{st}) \cdot \widehat{\boldsymbol{s}} + \eta_3 \cdot \mathbf{W}_s \cdot \widehat{\boldsymbol{s}} + \eta_4 \cdot \mathbf{K} \cdot \widehat{\boldsymbol{s}} = 0 \tag{20}$$

Then, the solution is obtained by:

$$\widehat{\boldsymbol{s}} = [\eta_1 \cdot \mathbf{I} + \eta_2 \cdot (\mathbf{D}_{st} - \mathbf{W}_{st}) + \eta_3 \cdot \mathbf{W}_s + \eta_4 \cdot \mathbf{K}]^{-1} \cdot (\eta_1 \cdot \boldsymbol{s}) \tag{21}$$

where $\mathbf{I}$ is an identity matrix with the size of $N_a \times N_a$.

IV. Experiments

A. Experimental Settings

We evaluate the proposed approach on the SegTrackV1 dataset [65], DAVIS dataset [66], and ViSal [27] dataset. The SegTrackV1 dataset includes 6 videos, and the pixel-level ground truth for each frame is available. Five videos in the dataset except for the penguin video are used for evaluation since the ground truth for this video is not usable. The DAVIS dataset contains 50 video sequences with the high quality resolution of 854×480 and full HD 1080p, and the fully-annotated pixel-level ground truth for each frame is provided. It is a challenging dataset because the scenes span multiple occurrences of occlusions, motion-blur, and appearance changes. In our work, the 854×480 resolution video sequences are utilized. As a specially designed dataset for video saliency detection, ViSal dataset consists of 17 challenging video sequences with manually annotated ground truth containing complex color distributions, highly cluttered background, various motion patterns, rapid topology changes, and camera motion. In experiments, the number of superpixels for each frame is set to 500, and the number of seeds are set to $\{Q_1, Q_2\} = \{250, 50\}$. The project, including the codes, results, and demos, is available on our website.[1]

Comparing the thresholding saliency map against the ground truth, the precision and recall scores are achieved, and the Precision-Recall (P-R) curve can be drawn. As an overall performance measurement, F-measure [67] is defined as:

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{22}$$

where β^2 is set to 0.3 that weighs the precision more than recall as suggested in [68], [69].

In addition, Mean Absolute Error (MAE) [27], [56] is introduced as a complementary measure, which is calculated as:

$$MAE = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} |S(x, y) - G(x, y)| \tag{23}$$

where S denotes the binary saliency map, G is the ground truth map, w and h represent the width and height of the image, respectively.

B. Comparison with State-of-the-art Methods

We compare the proposed method with 15 state-of-the-art methods, including 6 static image saliency methods for each frame (DSR [11], DCLC [9], HS [15], BSCA [16], RRWR [14], and HDCT [17]), 2 co-saliency detection methods for each video (CCS [40] and SCS [41]), and 7 video saliency detection methods (SP [26], CVS [27], RWRV [49], SG [50], SGSP [52], STBP [53], and VFCN [56]), where VFCN is a deep learning based video saliency detection method. All the compared methods are implemented by the source codes or released results provided by the authors. The qualitative comparison of different methods on three datasets are illustrated in Fig. 3, and the quantitative evaluation results are reported

[1] https://rmcong.github.io/proj_video_sal_SRP_tip.html

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/TIP.2019.2910377, IEEE Transactions on Image Processing

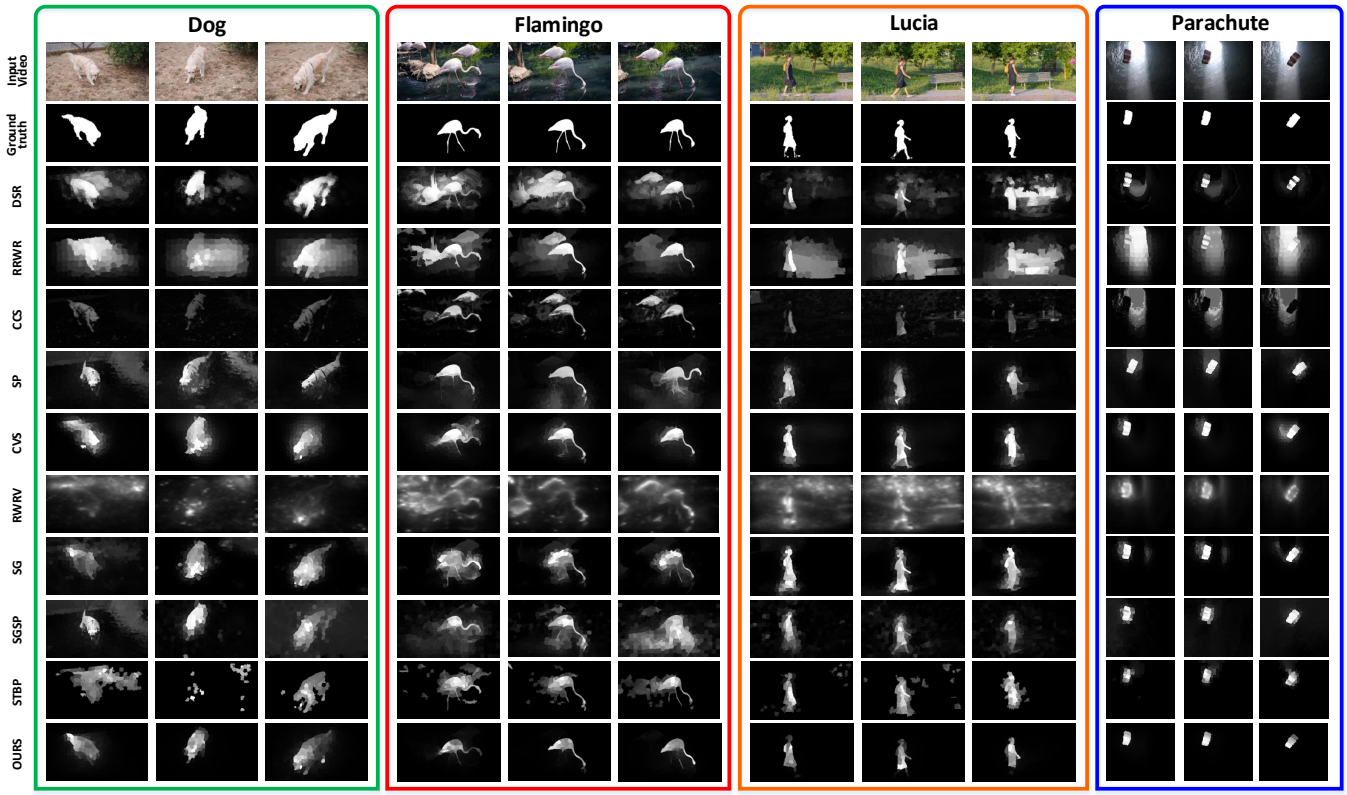


Fig. 3. Visual examples of different saliency detection methods.

in Fig. 4 and Table. I, including the P-R curves, F-measures, and MAE scores.

Visual results of different methods are shown in Fig. 3. For the image saliency model (*e.g.*, DSR and RRWR), it is difficult to extract the salient object completely and accurately from a complex scene due to the lack of motion perception and inter-frame information. For example, in the *Flamingo* video, two birds are both detected as the salient objects by the DSR and RRWR methods. In fact, only the front one is the unique salient object in the whole video. In other words, it is insufficient to directly use the static saliency model to detect the salient object in video. In the *Dog* video, the salient object and the background have the similar color appearance, which lead to some backgrounds are wrongly detected as foregrounds by RRWR method. In the *Lucia* video, the bench is relatively static compared to the moving human, and should not be detected as the salient object in the video. However, the image saliency models fail to effectively suppress these regions without considering the motion constraints. In the *Parachute* video, some backgrounds are wrongly highlighted by the image saliency models due to the strong luminance. For the co-saliency detection model, benefiting from the introduction of inter-image correspondence, some backgrounds are effectively suppressed, such as the trees and lawns in the *Lucia* video. However, some foregrounds are missed through the CCS model, such as the salient objects in the *Parachute* video. Moreover, for the co-saliency model, it is difficult to distinguish motion related salient object from all foreground objects, such as the *Flamingo* video. Without introducing the motion cue, the back of the bird is wrongly retained by CCS method. By contrast, the video saliency detection methods produce better results.

Our method achieves the best and most consistent performance compared with other methods. The salient objects are accurately and completely detected from some challenging videos, such as *Flamingo* video. Note that, other video saliency detection models either cannot exactly locate the salient object (such as RWRV) or cannot effectively suppress the background regions (such as SG and SGSP). For example, in the *Dog* video, the salient object is not accurately and completely detected by the RWRV and STBP methods. In addition, some video saliency models fail to discover the salient object accurately from the clustered backgrounds, such as the SG and STBP models in the *Flamingo* video. The SGSP method induces many false positives in the background regions, and cannot locate the front bird perfectly. In the *Lucia* video, compared with other video saliency methods, superior performance in shape preserving and pinpointing is achieved through our method.

The P-R curves are shown in Fig. 4. As visible, our method achieves the highest precision of the whole P-R curves on these three datasets with remarkable performance gain. In particular, on the DAVIS dataset, the proposed SRP method achieves better performance than the deep learning based video saliency method (*i.e.*, VFCN). The F-measure and MAE scores are reported in Table I. From the table, it can be seen that the proposed method obtains the highest F-measure on these three datasets and the minimum MAE score on the ViSal

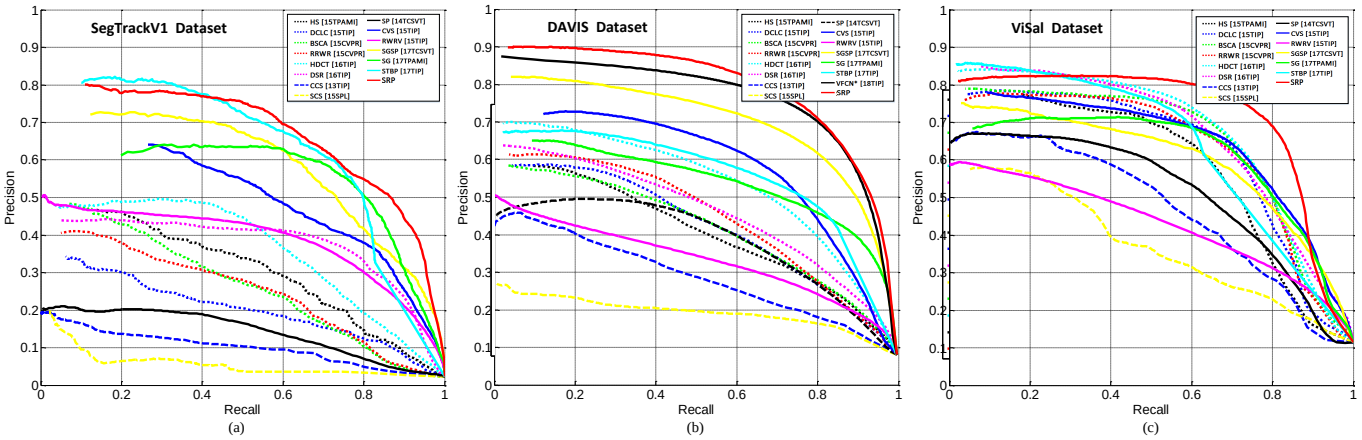


Fig. 4. P-R curves of different methods on three datasets. (a). SegTrackV1 dataset. (b). DAVIS dataset. (c). ViSal dataset.

TABLE I
QUANTITATIVE COMPARISONS WITH DIFFERENT METHODS ON THREE DATASETS. THE BEST RESULTS ARE BOLDED.

	SegTrackV1 Dataset		DAVIS Dataset		ViSal Dataset	
	F-measure	MAE	F-measure	MAE	F-measure	MAE
DCLC [9]	0.2755	0.1496	0.4783	0.1350	0.6700	0.1265
DSR [11]	0.4445	0.1305	0.4972	0.1303	0.6923	0.1061
RRWR [14]	0.3267	0.1963	0.5089	0.1693	0.6707	0.1690
HS [15]	0.3821	0.3142	0.4523	0.2505	0.6442	0.2019
BSCA [16]	0.3579	0.2366	0.4680	0.1957	0.6949	0.1703
HDCT [17]	0.4681	0.1268	0.5664	0.1346	0.7047	0.1282
CCS [40]	0.1486	0.1437	0.3476	0.1510	0.5317	0.1427
SCS [41]	0.1137	0.2664	0.2307	0.2567	0.4384	0.2523
SP [26]	0.2159	0.1195	0.4616	0.1430	0.5723	0.1510
CVS [27]	0.5370	0.1085	0.6212	0.1004	0.6676	0.1139
RWRV [49]	0.4458	0.1511	0.3776	0.2001	0.4662	0.1903
SG [50]	0.6218	0.0810	0.5553	0.1034	0.6640	0.1129
SGSP [52]	0.6275	0.1258	0.6911	0.1374	0.6226	0.1772
STBP [53]	0.6583	**0.0342**	0.5848	0.1015	0.6815	0.0987
VFCN* [56]	–	–	0.7488	**0.0588**	–	–
SRP	**0.6830**	0.0949	**0.7652**	0.0688	**0.7517**	**0.0924**

dataset. The proposed method achieves the second and third places in term of MAE score on the DAVIS and SegTrackV1 datasets, respectively. In addition, the performance gains of our method against others are more remarkable. Compared with the second best method in terms of F-measure, the percentage gain of our method reaches 3.7% on the SegTrackV1 dataset, 2.2% on the DAVIS dataset, and 6.7% on the ViSal dataset. Moreover, the proposed unsupervised method is superior to the deep learning based VFCN method, and the percentage gain of F-measure achieves 2.2% on the DAVIS dataset. All the quantitative measures demonstrate the effectiveness of the proposed method.

C. Ablation Study

We comprehensively evaluate each main component (single-frame saliency reconstruction integrating the static saliency and motion saliency, inter-frame saliency with forward and backward propagations, and global optimization) on the DAVIS dataset, and present the quantitative comparison results in Fig. 5 and Tables II-III.

Compared to the static saliency result, the motion saliency model achieves the higher precision of the P-R curves, and the F-measure is increased by 19.4%, which shows the effectiveness of the motion information in video saliency detection. Through the multiplying combination, the F-measure and MAE of the single-frame saliency model reach 0.7358 and 0.0807, which is better than other existing video saliency models. To fully capture the inter-frame relationship, we propose the sparsity-based propagation with forward-backward strategy. As can be seen, the performance is further improved through the saliency propagation model, and the F-measure reaches 0.7381 after the backward propagation. Considering the spatiotemporal smoothness and global consistency, an optimization model is designed to improve the saliency map,

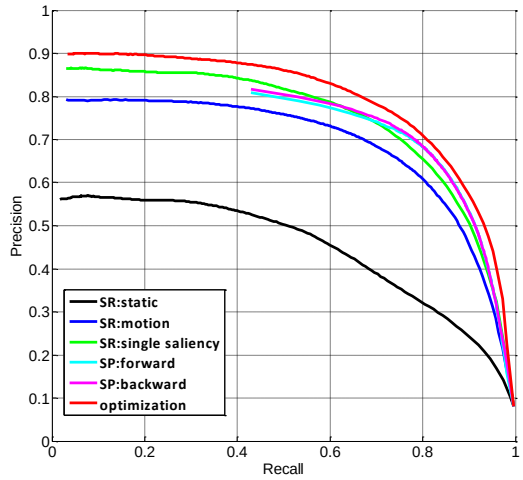


Fig. 5. P-R curves of different modules of the proposed method on the DAVIS dataset.

TABLE II
F-MEASURES OF DIFFERENT MODULES ON THE DAVIS DATASET. SR: SINGLE-FRAME SALIENCY RECONSTRUCTION THAT INTEGRATES THE STATIC AND MOTION SALIENCIES. SP: INTER-FRAME SALIENCY PROPAGATION.

Modules		F-measure	MAE
SR	Static Saliency	0.5029	0.1206
	Motion Saliency	0.6971	0.0807
	Single Saliency	0.7358	0.0712
SP	Forward Propagation	0.7318	0.0924
	Backward Propagation	0.7381	0.0793
Global Optimization		0.7652	0.0688

and the output is regarded as the final video saliency result. From Fig. 5, we can be seen that the optimized result achieves the highest precision of the P-R curves, which is marked by the red line. The same conclusion can be drawn from the F-measure reported in Table II, which demonstrates the rationality and effectiveness of the optimization model. On the whole, the performance is gradually improved through the different modules in our method.

In addition, we conduct an additional comparison experiment with different inter-frame propagation strategies, including a one-step propagation based inter saliency model that integrates two directly adjacent frames to construct the foreground dictionary for inter reconstruction, and the proposed forward-backward propagation based inter saliency model. The quantitative results are shown in Table III. In terms of F-measure, the one-step propagation based inter saliency achieves 0.7306, and the inter saliency with forward-backward propagation strategy reaches 0.7381. In terms of MAE score, the percentage gain of forward-backward propagation strategy achieves 39% against the one-step propagation. All these measures demonstrate the effectiveness of the proposed forward-backward propagation strategy.

TABLE III
COMPARISONS OF THE DIFFERENT INTER SALIENCY WITH DIFFERENT PROPAGATION STRATEGIES ON DAVIS DATASET.

	One-step propagation	Forward-backward propagation
F-measure	0.7306	0.7381
MAE	0.1306	0.0793

D. Parameter Analysis

We comprehensively discuss the influence of different seed numbers, the tendency chart of F-measure on the DAVIS dataset is shown in Fig. 6. Generally, the salient regions in each frame are much smaller than the background regions. To explore the single-frame saliency, some background seeds are selected, and the number is denoted as Q_1. We can choose more background seeds to construct a more complete background dictionary. For the inter-frame saliency propagation, the foreground seeds are determined to propagate the sequential relationship across the time axis in a forward-backward way. The number of foreground seeds is denoted as Q_2. In order to avoid the introduction of interference, the number of foreground seeds should not be too large. In all the experiments, we fixed the ratio of Q_1 to Q_2 as $5:1$. Selecting 100 or 120 background seeds for each frame is too small to completely reconstruct the single-frame saliency and will degenerate the performance. As the seed number increases, the performance becomes better, and the performance reaches optimum when (Q_1, Q_2) is set to $(250, 50)$. Subsequently, the performance begins to drop. The main reason is that too many seeds will introduce some false seed regions and decrease the reconstruction and propagation accuracy. As above, the performance is not highly sensitive to the parameter (Q_1, Q_2), and we set it to $(250, 50)$ in all experiments.

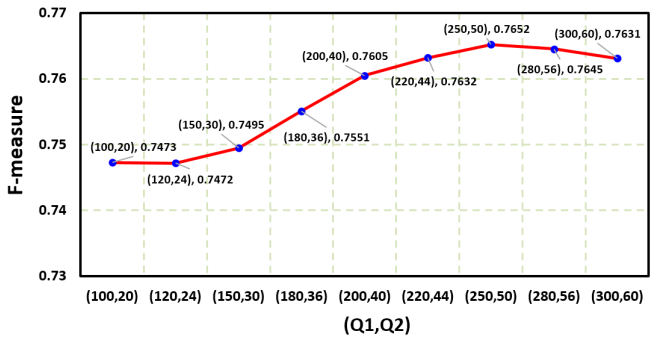


Fig. 6. F-measure of different (Q_1, Q_2) on the DAVIS dataset.

E. Running Time

In this section, we discuss the running time and computational complexity of the proposed method. We implemented the proposed method using MATLAB 2014a on a Quad Core 3.7GHz workstation with 16GB RAM. The running time of different video saliency detection methods on the DAVIS dataset are reported in Table IV. As can be seen, the proposed

TABLE IV
COMPARISONS OF THE AVERAGE RUNNING TIME (SECONDS PER FRAME) ON THE DAVIS DATASET.

Method	SP	CVS	RWRV	SG	SGSP	STBP	SRP
Time	29.81	25.46	36.67	27.20	26.45	174.00	17.03

method takes an average of 17.03 seconds to process one frame with a resolution of 854×480, and ranks the first against other video saliency detection methods. Specifically, for our SRP method, the optical flow calculation costs 65% of the runtime, the single-frame saliency calculation takes 10% of the runtime, the inter-frame saliency costs 22% of the runtime, and the global optimization occupies 3% of runtime. From these statistics, we can see that the global optimization occupies a small part of the computing resources. The time complexity of global optimization is proportional to $o(N_a^2)$. As the number of video frame increases, N_a becomes larger, and the computational complexity will increase. In the future, we can use a faster optical flow method with parallel technique to reduce this cost further.

V. CONCLUSION

In this paper, a sparsity-based video saliency detection algorithm, which integrates a saliency reconstruction model, a saliency propagation model, and a global optimization model, is proposed. Saliency reconstruction and propagation models leverage on the novel motion priors to discover the salient objects. In addition, their sparse representations not only allow them to extract the salient object from individual frames efficiently, but also capture the inter-frame correspondence along the time axis in a progressive way. Moreover, the performance is further improved by the global optimization model. Our comprehensive analysis demonstrated that the proposed method outperforms the state-of-the-art saliency, co-saliency, and video saliency models. In the future, we plan to incorporate our models into a deep learning framework.

REFERENCES

[1] H. Fu, D. Xu, and S. Lin, "Object-based multiple foreground segmentation in RGBD video," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1418–1427, 2017.

[2] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, "Saliency-aware nonparametric foreground annotation based on weakly labeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1253–1265, 2016.

[3] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Vis. Comput. Graph*, vol. 23, no. 8, pp. 2014–2027, 2017.

[4] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 1–14, 2018.

[5] C. Li, J. Guo, R. Cong, Y. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, 2016.

[6] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, 2019.

[7] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "Learning sparse representation for objective image retargeting quality assessment," *IEEE Transactions Cybernetics*, vol. 48, no. 4, pp. 1276–1289, 2018.

[8] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2035–2048, 2018.

[9] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, "Salient region detection via integrating diffusion-based compactness and local contrast," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3308–3320, 2015.

[10] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014, pp. 2814–2821.

[11] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *ICCV*, 2013, pp. 2976–2983.

[12] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *CVPR*, 2015, pp. 5216–5223.

[13] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Dense and sparse labeling with multi-dimensional features for saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1130–1143, 2018.

[14] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Feng, "Robust saliency detection via regularized random walks ranking," in *CVPR*, 2015, pp. 2710–2717.

[15] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, 2016.

[16] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015, pp. 110–119.

[17] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 9–23, 2015.

[18] C. Zhang, Z. Tao, X. Wei, and X. Cao, "A flexible framework of adaptive method selection for image saliency detection," *Pattern Recognition Lett.*, vol. 63, pp. 66–70, 2015.

[19] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, 2016.

[20] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, 2015.

[21] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.

[22] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016, pp. 660–668.

[23] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.

[24] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017, pp. 5300–5309.

[25] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, 2017, pp. 212–221.

[26] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, 2014.

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

[28] Z. Ren, S. Gao, D. Rajan, L.-T. Chia, and Y. Huang, "Spatiotemporal saliency detection via sparse representation," in *ICME*, 2012, pp. 158–163.

[29] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in *ICASSP*, 2012, pp. 1485–1488.

[30] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017, pp. 4548–4557.

[31] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018, pp. 698–704.

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

[33] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *CVPR*, 2016, pp. 2343–2350.

[34] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, 2016.

[35] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, 2017.

[36] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, 2017.

[37] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *CVPR*, 2018, pp. 3051–3060.

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

[39] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. PP, no. 99, pp. 1–12, 2019.

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

[41] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, 2017.

[42] L. Wei, S. Zhao, O. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *IJCAI*, 2017, pp. 3041–3047.

[43] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based for co-saliency detection framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, 2018.

[44] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, 2018.

[45] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for RGBD images," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 233–246, 2019.

[46] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "HSCS: Hierarchical sparsity based co-saliency detection for RGBD images," *IEEE Trans. Multimedia*, vol. PP, no. 99, pp. 1–12, 2018.

[47] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, 2010.

[48] Y. Fang, W. Lin, Z. Chen, C. Tsai, and C. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, 2014.

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

[50] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *CVPR*, 2015, pp. 3395–3402.

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

[52] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, 2017.

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

[54] T.-N. Le and A. Sugimoto, "Deeply supervised 3D recurrent FCN for salient object detection in videos," in *BMVC*, 2017, pp. 1–13.

[55] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *ECCV*, 2018, pp. 715–731.

[56] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, 2018.

[57] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 985–998, 2019.

[58] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349–364, 2018.

[59] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *CVPR*, 2019, pp. 1–10.

[60] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[61] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, 2011.

[62] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *ACM-SIAM Symp. Discr. Algorithms*, 2007, pp. 1027–1035.

[63] T. Leung and J. Malik, "Recognizing surface using three-dimensional textons," in *ICCV*, 1999, pp. 1010–1017.

[64] W.-D. Jang, C. Lee, and C.-S. Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *CVPR*, 2016, pp. 696–704.

[65] D. Tsai, M. Flagg, and J. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *BMVC*, 2010, pp. 1–11.

[66] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016, pp. 724–732.

[67] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. 99, pp. 1–19, 2018.

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

[69] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *ECCV*, 2018, pp. 1–17.

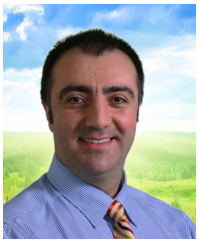
Runmin Cong received the M.S. degree from the Civil Aviation University of China, Tianjin, China, in 2014. He is currently pursuing his Ph.D. degree in information and communication engineering with Tianjin University, Tianjin, China.

He was a visiting student at Nanyang Technological University (NTU), Singapore, from Dec. 2016 to Feb. 2017. Since May 2018, he has been working as a Research Associate at the Department of Computer Science, City University of Hong Kong (CityU), Hong Kong. He is a Reviewer for the IEEE TIP, TMM, and TCSVT, etc. He won the Best Student Paper Runner-Up at IEEE ICME in 2018. His research interests include computer vision, image processing, saliency detection, and 3-D imaging.

Jianjun Lei (M'11-SM'17) received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

He was a visiting researcher at the Department of Electrical Engineering, University of Washington, Seattle, WA, from August 2012 to August 2013. He is currently a Professor at Tianjin University, Tianjin, China. He is on the editorial boards of Neurocomputing and China Communications. His research interests include 3D video processing, virtual reality, and artificial intelligence.

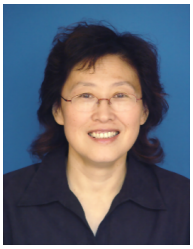
Huazhu Fu (SM'18) received the Ph.D. degree in computer science from Tianjin University, China, in 2013. From 2013 to 2015, he worked as the research fellow at Nanyang Technological University, Singapore. And from 2015 to 2018, he worked as a Research Scientist at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. He is currently the Senior Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, image processing, and medical image analysis. He is the Associate Editor of IEEE Access and BMC Medical Imaging.



Fatih Porikli (M'96-SM'04-F'14) received his Ph.D. degree from New York University, New York, NY, USA, in 2002. He is currently a Professor with the Research School of Engineering, Australian National University, Canberra, ACT, Australia. He is also acting as the Vice President of CBG Device and Hardware at Huawei, Santa Clara, CA, USA. Previously, he led the Computer Vision Research Group at NICTA and served as a Distinguished Research Scientist at Mitsubishi Electric Research Laboratories. Prof. Porikli is the recipient of the R&D 100 Scientist of the Year Award in 2006. He won six best paper awards at premier IEEE conferences and received five other professional prizes. He authored more than 200 publications and invented 73 patents. He is the co-editor of two books. He is currently the Associate Editor for five journals. His research interests include computer vision, deep learning, manifold learning, online learning, and image enhancement with commercial applications in mobile phones, autonomous vehicles, video surveillance, satellite, and medical systems.

Qingming Huang (SM'08-F'18) is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China.

His research areas include multimedia video analysis, image processing, computer vision and pattern recognition He has published more than 400 academic papers in prestigious international journals including IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on Circuits and Systems for Video Tech., etc, and top-level conferences such as ACM Multimedia, ICCV, CVPR, IJCAI, VLDB, etc. He is the associate editor of IEEE Trans. on Circuits and Systems for Video Tech., and Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. on Multimedia, IEEE Trans. on Circuits and Systems for Video Tech., IEEE Trans. on Image Processing, etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, PSIVT, etc.



Chunping Hou received the M.Eng. and Ph.D. degrees, both in electronic engineering, from Tianjin University, Tianjin, China, in 1986 and 1998, respectively.

Since 1986, she has been the faculty of the School of Electronic and Information Engineering, Tianjin University, where she is currently a Full Professor and the Director of the Broadband Wireless Communications and 3D Imaging Institute. Her current research interests include 3D image processing, 3D display, wireless communication, and the design and applications of communication systems.