Weakly Supervised Salient Object Detection with Spatiotemporal Cascade Neural Networks

Yi Tang, Wenbin Zou, *Member, IEEE*, Zhi Jin, *Member, IEEE*, Yuhuan Chen, Yang Hua, Xia Li

***Abstract*—Recently, deep learning techniques have substantially boosted the performance of salient object detection in still images. However, the salient object detection in videos by using traditional handcrafted features or deep learning features is not fully investigated, probably due to the lack of sufficient manually labeled video data for saliency modeling, especially for the data-driven deep learning. This paper proposes a novel weakly supervised approach to salient object detection in a video, which can learn a robust saliency prediction model by using very limited manually labeled data and a large amount of weakly labeled data that could be easily generated in a supervised approach. Furthermore, we propose a spatiotemporal cascade neural network (SCNN) architecture for saliency modeling, in which two fully convolutional networks are cascaded to evaluate visual saliency from both spatial and temporal cues to lead the optimal video saliency prediction. The proposed approach is extensively evaluated on the widely used challenging datasets, and the experiments demonstrate that our proposed approach substantially outperforms the state-of-the-art salient object detection models.**

***Index Terms*—Video saliency, weakly supervised learning, spatiotemporal prior fusion, cascade fully convolutional network**

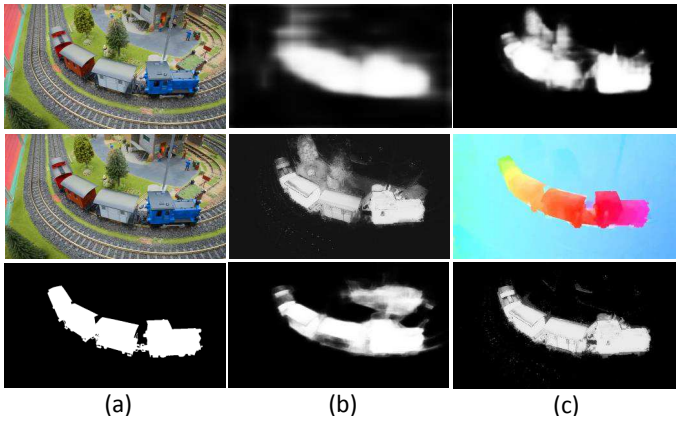

Fig. 1. Salient object detection in dynamic scenes by using different models. (a) Paired frames and the ground truth of the second frame. (b) Saliency maps for the second frame generated by the previous DSMT model [1] (top), DCL model [2] (middle) and UCF model [4] (bottom). (c) Results generated by our proposed SCNN model, including spatial prior (top), optical flow (middle) and the saliency map for the second frame (bottom).

I. Introduction

SALIENT object detection, which aims to identify the objects or regions that are noticeable and mostly attract human attention in an image/video, has become a research focus of computer vision for decades. It is generally as a preprocessing step to support high-level computer vision tasks, such as object segmentation, object recognition, object tracking and content-based video compression. A number of approaches have been proposed to detect salient objects. The recent approaches based on deep Convolutional Neural Networks (CNNs), e.g., [1]–[3], have substantially improved the performance of salient object detection in still images. However, these deep learning models trained by still images may not perform well in some dynamic video scenes.

One of the key issues is the difficulty in eliminating the interference of relatively complex background regions, which may be unmoving or moving in a video. The image saliency approaches take each video frame as a still image and perform saliency detection one by one, without considering the motion of objects. Therefore, those background regions, which are salient in a still image, are easily highlighted in the generated saliency map (See the examples in Fig. 1). However, the motion is the most important cue to attract human attention and these regions may not be salient in a video. Before the usage of neural networks, such motion information is introduced into video saliency approaches by graphics models [5], [6]. These approaches firstly generate an initial saliency map from global motion clues [5] or gradient flow field [6], and then exploit an energy function with spatiotemporal constraint to estimate the final saliency in video sequences. Due to their employment of the handcrafted features in these methods, the salient objects are difficult to be detected in complex video scenes.

Recently, deep learning models are employed into the video saliency prediction. As far as we know, a robust deep learning model needs to be trained by the large-scale labeled pixel-wise video frames. Unfortunately, the pixel-wise labeling is very time-consuming and needs huge human resources. It should be noted the fact that current datasets for video salient object

This work is supported by the NSFC Project under Grant 61771321, Grant 61472257 and Grant 61701313, in part by the Natural Science Foundation of Shenzhen under Grant KQJSCX20170327151357330, Grant JCYJ20170818091621856, and Grant JSGG20170822153717702, and in part by the Interdisciplinary Innovation Team of Shenzhen University. (*Corresponding author: Wenbin Zou*).

Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen are with the Shenzhen Key Laboratory of Advanced Telecommunication and Information Processing, College of Information Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen key laboratory of Advanced Machine Learning and Applications, Shenzhen 518060, China.(e-mail:yitang@szu.edu.cn; zouszu@sina.com; chenyuhuan126@163.com; jinzhi_126@163.com; lixia@cuhk.edu.cn).

Yang Hua is with the EEECS/ECIT Queen's University Belfast, United Kingdom (e-mail:Y.Hua@qub.ac.uk).

Xia Li is with the Chinese University of Hong Kong, Shenzhen 518172, China (e-mail:lixia@cuhk.edu.cn).

detection have very limited manually labeled ground truth. For instance, a total number of labeled data are no more than 7000 frames in the widely used video datasets, including SegTrackV2 [7], FBMS [8], VSB100 [9] and DAVIS [10], where some of them only label a small part of frames in a video sequence. In [11], an approach by synthesizing video data from still frames has been proposed to generate large-scale simulated video data and the corresponding pixel-wise annotations. The other methods introduce weakly supervised learning to train networks by image-level labels [12]–[14]. These labels, that indicate the presence/absence or specific category of objects in an image, are easier to collect than the pixel-wise ones. However, as for the deep models of saliency detection, the pixel-wise labels are more suitable to train the network.

Bearing in mind the issues aforementioned, we, on the one hand, aim to find a solution to learn a salient object detection model in a weakly supervised approach, that trains a saliency model by using limited manually labeled ground truth and huge weakly pixel-wise labeled data which are generated in a fusing saliency maps way. On the other hand, we expect to propose a deep neural network architecture which can learn spatial and temporal cues to identify salient objects in a video.

Therefore, we propose a spatiotemporal cascade neural network (SCNN) architecture, which utilizes spatial and temporal priors to model visual saliency in videos. Moveover, to overcome the problem of the lack of pixel-wisely labeled data, we propose a weakly supervised learning strategy to train deep neural networks. In summary, the contributions of this paper are as follows:

1. We propose a spatiotemporal cascade neural network (SCNN) architecture, which leverages two fully convolutional neural networks to evaluate visual saliency from both spatial and temporal cues.
2. We introduce an weakly supervised approach, which takes advantage of large-scale weakly labeled data for saliency model learning. These weakly labeled data not only complement manually labeled ones, but also enable to achieve obvious performance improvement for salient object detection.
3. We present a novel approach to extract the motion information of salient object from optical flow fields, which is able to be effectively incorporated into the proposed SCNN framework.
4. We demonstrate that our proposed approach substantially outperforms the state-of-the-art salient object detection models.

II. Related works

In this section, we review related works in spatiotemporal saliency models, CNN-based saliency methods and relevant approaches by using weak supervision.

Spatiotemporal saliency models. Over the recent decades, a variety of techniques and theories have be exploited to detect salient objects in still images, such as spatial prior [15], low-rank matrix recovery [16], regional contrast [17], graphical modeling [18], and information theory [19].

While spatial information has been extensively investigated for still images, video salient object detection models need to integrate both spatial and temporal information. A dynamic fusion model by combining spatial and temporal saliency maps has been proposed in [20]. In [21], Gao et al. propose a novel framework based on center-surrounding hypothesis to predict salient objects from multi-scale handcrafted features of color, orientation and luminance. Then, by extending to the center-surrounding hypothesis, a discriminant saliency model is proposed in [22], where dynamic spatiotemporal textures are employed for saliency detection in the video sequence. Rahtu et al. propose a novel statistical framework [23] for saliency prediction by fusing motion, illumination and color information. In [24], an adaptive fusion method is proposed to integrate pixel-level spatial and temporal saliency maps by using color and motion handcrafted features in superpixel level. By fusing spatial edges and temporal motion boundaries from continuous optical flow maps, Wang et al. [6] use a geodesic model to detect salient regions in video sequences. Kalboussi et al. [25] produce the dynamic map and static map by exploiting dense motion estimation and spatial edges detection, respectively. Then, the flood fill algorithm is introduced to fuse the maps for saliency prediction. In addition, recent works about video saliency employ motion attention cue [26], nonparametric kernel density feature [27], motion continuity [28], low-rank coherency diffusion [5] and gradient flow field [29] to fuse spatial and temporal saliency maps.

CNN-based saliency methods. The above methods utilize handcrafted features and optimization models. However, with the resurgence of the neural network, especially the appearance of CNN, the saliency detection field has a breakthrough. The development of saliency detection with CNN consists of two stages. The first stage is mainly using deep features to take the place of handcrafted features in saliency detection models. In [30], Zhao et al. treat image patches based on superpixels as CNN input and extract their corresponding deep features to complete saliency detection. Besides, in [31], the combination of region proposals and deep features are used for local estimation and global search in saliency detection. In [32], the deep features are extended to multi-scale deep features, which combine with multi-level region decomposition to generate saliency maps. The second stage is directly generating pixel-wise saliency maps in an end-to-end Fully Convolutional Network (FCN) [33]. Recent work [34] modifies FCN and proposes a deep hierarchical network. In [3], [35], Recurrent Neural Network (RNN) is composed with FCN to perform a full saliency prediction. In [2], pooling layers in FCN are decreased to make prediction map denser and branches of convolutional layers are increased to generate multi-scale saliency maps. Meanwhile, multi-task learning is employed to optimize the FCN in [1]. The works of video saliency detection by deep learning are not very much, but recent work [11] has succeeded to fuse spatial and temporal saliency stimuli via FCN.

Weakly Supervised Learning. Recently, weakly supervised learning methods have been introduced into many areas such as object detection [36], [37], object localization [38], [39], semantic segmentation [40], [41]. In [42], the saliency detec-

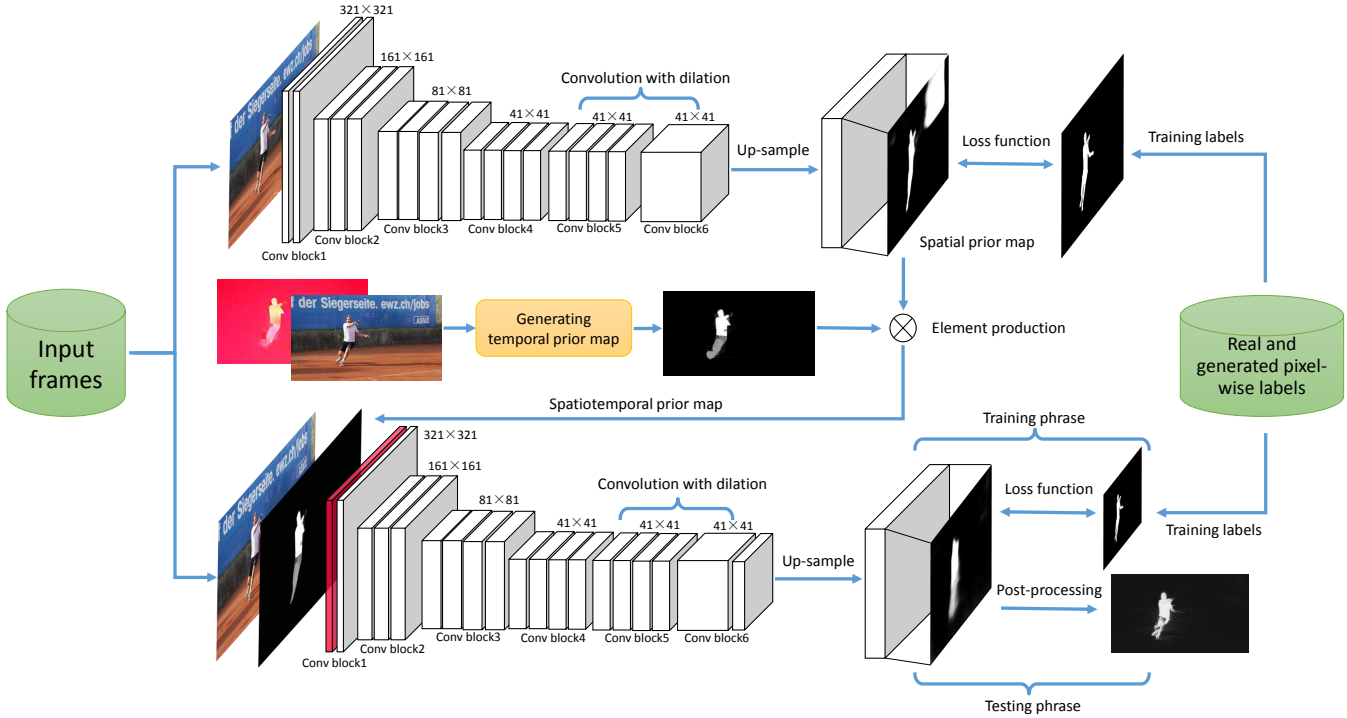

Fig. 2. A schematic diagram of the proposed SCNN framework. A fully convolutional network (FCN) is proposed to generate a spatial prior map which is combined with a temporal prior map, evaluated from optical flow fields, as the input of the second FCN having the same architecture as the first one. A post-processing operation is followed to generate the final saliency map. The network parameters are learned through a weakly supervised approach.

tion is applied for weakly-supervised object detection via a self-paced curriculum learning regime. Lai et al. [43] integrates saliency information into an end-to-end neural network to perform weakly supervised object detection. In [44], a weakly supervised image parsing method is proposed by using saliency results to guide the dictionary learning. In the area of saliency detection, Cholakkal et al. [12] firstly propose a weakly supervised top-down saliency approach by exploiting the backtracking ScSPM image classifier. Then, they extend the approach by combining a selected saliency map from the fast bottom-up saliency approaches to generate the final map in [45]. After that, a two-stage weakly supervised network [14] is proposed for saliency prediction. The network is firstly pre-trained with image-level tags, and then self-trained by using estimated pixel-level labels. In [46], an image-level classifier and a pixel-level map generator are composed to conduct saliency detection.

III. The Proposed Model

Fig.2 illustrates the framework of our proposed model. Given a video frame, a spatial prior map, on one hand, is generated through a fully convolutional network (FCN). On the other hand, a temporal prior map is obtained by evaluating visual saliency from optical flow fields. Furthermore, the spatial prior map and the temporal prior map are combined to generate a spatiotemporal prior map to guide the second FCN for saliency prediction. The generated saliency map is refined further through a model of conditional random field. The connected two FCNs for visual saliency prediction from both spatial and temporal evaluation is called spatiotemporal cascade neural network (SCNN). The SCNN parameters are learned by our proposed weakly supervised approach.

In the following, we present firstly the details of SCNN. Then, we show how to generate the spatiotemporal prior map. After that, we introduce the weakly supervised learning approach. Finally, the CRF for saliency refinement is given.

A. The spatiotemporal cascaded neural network

The proposed SCNN consists of two FCNs having the same architecture extended from VGG network [48], which contains five convolutional blocks, each of which has several convolutional layers. In order to generate feature maps rather than feature vectors, the last two fully connected layers need to be transferred into convolutional layers with 1×1 kernel like [33]. Therefore, the FCN in our SCNN contains six convolutional blocks. Besides, in order to recover the scale of feature maps, an up-sampling layer is added at the top of FCN to generate a saliency map with a resolution of the input video frame. In the FCNs, each convolutional operation can be formulated as follow:

$$f(\mathcal{X}; \mathcal{W}, b) = \sigma(\mathcal{W} * \mathcal{X} + b) \tag{1}$$

where $f(\cdot)$ denotes the generated feature map by a convolutional operation; $\mathcal{X}$ is the input and contains three channels tensors ($\mathcal{X} \in \mathcal{R}^{h*w*c}$); b is a bias term; $\mathcal{W}$ is a set of kernel parameters; $\sigma(\cdot)$ denotes the activation function, which is Rectified Linear Unit (ReLU) in our experiments.

The original VGG network contains five max-pooling layers. The scale of feature maps is reduced twice after each max-pooling layer, which largely reduces the scale of the feature

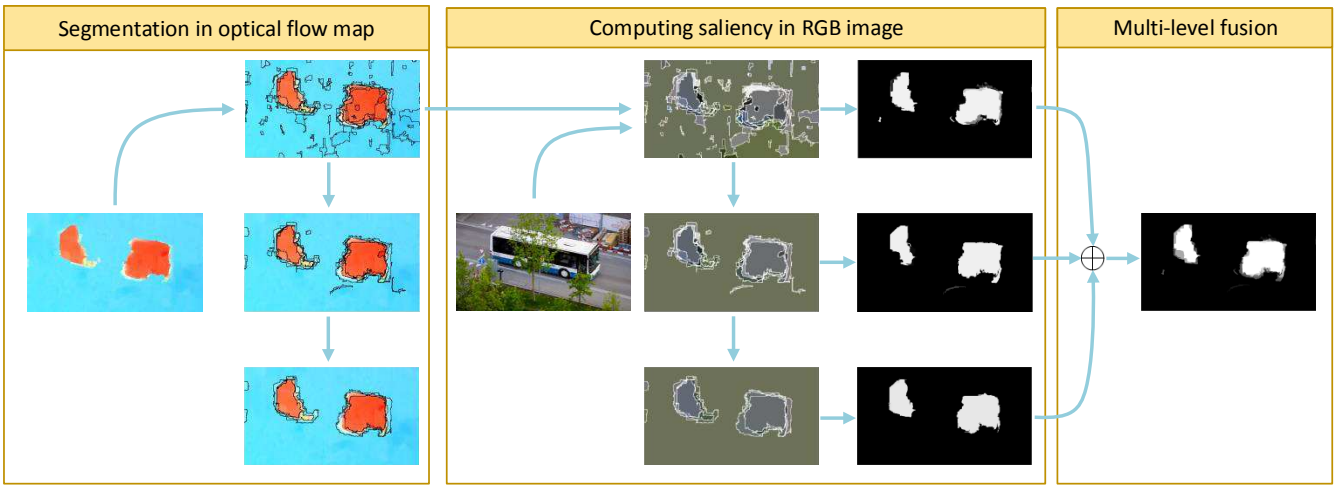

Fig. 3. The pipeline of the generation of motion prior map. Superpixels are obtained by [47] in the left block. The middle block shows the multi-level saliency maps are computed by deep features. In the right block, the fusing saliency maps are generated by linear addition.

maps from the last convolutional layer. After the operation of up-sampling, the feature maps are very coarse and the spatial context information is also lost, which is not conducive to generate the final saliency map. To keep the scale of final feature map suitable, a padding operation of 100 pixels is added in the Conv block1 in the first convolutional layer in [33] (See Fig.2). Although it can increase the scale of feature map, it brings a lot of useless information as well. In this paper, we remove the last two max-pooling layers so that the feature maps after the Conv block3 have enough scale to retain dense features. However, this operation also changes the receptive fields of convolution and makes the original parameters of kernels not suitable for new convolutional layers. Therefore, the convolution layers with dilation [49], that add holes into convolutional kernels, are employed to keep receptive fields and spatial context information.

In our SCNN, the multi-layer up-sampling is used for resizing the feature maps. In [2] [33], the up-sampling operation is embedded after every convolutional block. However, through our experiments, we find that the up-sampled feature maps from the first four max-pooling layers have less effect on the final saliency map. To simplify the structure of the network and speed up the process, we only up-sample the feature maps from the last two convolutional blocks, and then sum them up element-wisely to obtain the final feature map.

The generated feature map from the first FCN is called spatial prior map and combined with the temporal prior map evaluated from optical flow fields to obtain a spatiotemporal prior map, which is presented in detail in Section III-B. Then, the spatiotemporal prior is exploited to guide the second FCN learning for saliency prediction. Due to the embedding of the spatiotemporal prior map, the convolutional operation in the first layer of the second FCN is modified as follow:

$$\begin{aligned} f(\mathcal{X},\mathcal{P};\mathcal{W}_1,\mathcal{W}_2,b) &= \sigma(\mathcal{W}_1 * \mathcal{X} + \mathcal{W}_2 * \mathcal{P} + b) \\ &= \sigma([\mathcal{W}_1\ \mathcal{W}_2] * [\mathcal{X}\ \mathcal{P}]^T + b) \end{aligned} \tag{2}$$

where $\mathcal{P}$ denotes and the spatiotemporal prior map; $\mathcal{W}_1$ is a set of kernel parameters for the input frame $\mathcal{X}$ whereas $\mathcal{W}_2$ is the one for the spatiotemporal prior map $\mathcal{P}$; b is a bias term.

At the top of SCNN, a loss function is applied to compute the errors between the final feature map $\mathcal{S} \in [0,1]^{h*w*1}$ and the pixel-wise labeled $\mathcal{G} \in [0,1]^{h*w*1}$, where h and w denote the height and width of an input video frame, respectively. Considering the unbalance between the number of salient pixels and that of non-salient pixels, we exploit a modified cross-entropy loss function as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{S},\mathcal{G}) = &-\alpha \sum_{i=1}^{h*w} g_i \log P(s_i = 1|\mathcal{X}_i,\mathcal{W}) \\ &-(1-\alpha)\sum_{i=1}^{h*w}(1-g_i)\log P(s_i = 0|\mathcal{X}_i,\mathcal{W}) \end{aligned} \tag{3}$$

where $s_i \in \mathcal{S}$ and $g_i \in \mathcal{G}$ denote the saliency value and the label of ground truth for a pixel $\mathcal{X}_i$, respectively; α denotes the balance factor and is set as the ratio of background pixels in the ground truth $\mathcal{G}$.

B. Generation of spatiotemporal prior map

In our SCNN framework, a spatiotemporal prior map $\mathcal{P}$, which is a combination of the spatial prior map $\mathcal{P}_s$ and the temporal prior map $\mathcal{P}_t$, is proposed to guide the second FCN learning for salient object prediction. As the guidance seed, we emphasize the precision rather than recall; i.e., we do not expect to highlight the whole salient object in the spatiotemporal prior map $\mathcal{P}$, but those salient regions highlighted should be reliable. Therefore, the element-wise production is adopted to fuse the spatial and temporal prior maps, i.e.

$$\mathcal{P} = \mathcal{P}_s \otimes \mathcal{P}_t \tag{4}$$

where $\otimes$ denotes the operation of element-wise production. Such an operation highlights the shared salient regions in

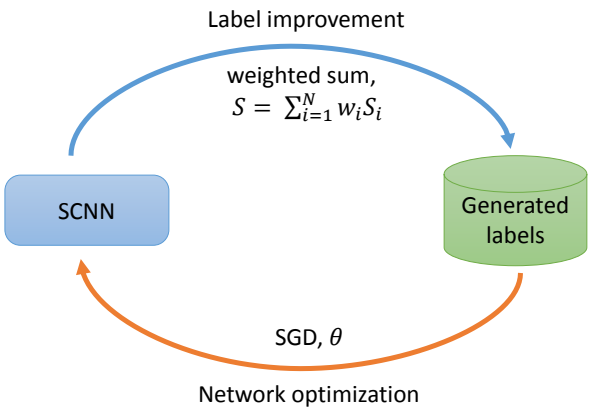

Fig. 4. Weakly supervised learning approach generates weak labels increasingly and learns SCNN parameters iteratively with the augmented data.

Algorithm 1 Parallel iteration strategy in training process

Input: the initial network parameters θ_0 and network input $\mathcal{X}$, the initial weak pixel-wise label $\mathcal{G}_0$ by fusing saliency maps $\mathcal{S}_i$ with initial weights w_0, the number of epoch β;
Output: final network parameters θ_β
1: **for** $t = 1, 2, ..., \beta$ **do**
Network optimization:
2: Use weak pixel-wise label $\mathcal{G}_{t-1}$ and input images $\mathcal{X}$ to compute network loss by Eq.3;
3: Update network parameters θ_t by using SGD;
Weak label improvement:
4: Add saliency map generated by our SCNN and update weights w_t by Eq.8;
5: Generate binary weak labels $\mathcal{G}_t$ through segmenting the fused saliency map $\mathcal{S}$ in Eq.7 by Otsu threshold.
6: **end for**

both the spatial and temporal prior maps and suppresses those regions being salient in one prior map only.

As mentioned in the previous subsection, the spatial prior map $\mathcal{P}_s$ is generated from the first FCN. Specifically, the FCN takes a frame as the input and produces the corresponding feature maps. Then, the feature map of the last convolutional layer is activated by a sigmoid function to generate the spatial prior map $\mathcal{P}_s$ as follows

$$\mathcal{P}_s = \Psi\Big(\mathcal{U}_s\big(\mathcal{F}_s(\mathcal{X};\theta)\big)\Big) \tag{5}$$

where $\Psi(\cdot)$ denotes sigmoid operation; $\mathcal{U}_s(\cdot)$ denotes the upsampling operation; $\mathcal{F}_s(\cdot)$ represents the convolution operation with the parameters θ. In our experiments, the bilinear upsampling is used to ensure the spatial prior map $\mathcal{P}_s$ having a resolution of the input video frame.

To generate the temporal prior map $\mathcal{P}_t$, we propose a novel approach to evaluate visual saliency from optical flow fields. As illustrated in Fig.3, we firstly perform $\mathcal{M}$ segmentations on the optical flow map by using graph-based algorithm [47] with different parameters to generate multi-scale superpixels. Then we extract deep features [32] from RGB image (video frame) for each superpixel $r_i^j (j = 1, 2, ..., \mathcal{M})$, where i denotes the superpixel index in the j-level segmentation, to train a binary classifier of the three-layer neural network for salient superpixel prediction. Finally, the saliency maps from different segmentation levels are linearly fused to generate the temporal prior map $\mathcal{P}_t$ as follows:

$$\mathcal{P}_t(r_i) = \frac{1}{\mathcal{M}} \sum_{j=1}^{\mathcal{M}} \mathcal{S}^j\big(\mathcal{D}(r_i^j)\big) \tag{6}$$

where $\mathcal{D}(\cdot)$ denotes the deep features of the superpixel r_i^j; $\mathcal{S}^j(\cdot)$ denotes the saliency value predicted by the binary classifier. the segmentation level $\mathcal{M}$ is set to 3, which is a balance setting between the accuracy and the processing time.

C. Weakly supervised learning approach

Deep learning has demonstrated its success in various tasks. However, it is a data-driven approach and needs large-scale data with ground truth to learn network parameters. Although it is easy to obtain video data, the pixel-wisely labeling for the ground truth is very time consuming, that is why some datasets labeled a part of video frames only for the performance evaluation of salient object detection. Therefore, we propose a weakly supervised approach which learns network parameters by generating weakly labeled data increasingly along with the model learning iteration. Our hypothesis is that if we assign weak labels for some frames unlabeled in the training set and then make the labels stronger during training processing, these frames and weak labels are also useful to learn network parameters. To this end, we use a method fusing saliency maps to generate the pixel-wise labels and propose an interactive iteration approach to complete our weakly supervised learning.

As illustrated in the Fig.4, the proposed weakly supervised learning approach consists of two components, including network optimization for SCNN and labeling improvement for the generated weak labels. Given some labeled data available, the SCNN model is optimized, through the stochastic gradient descent (SGD) method, and generates a saliency map for the input video frame. The generated saliency map is combined with other saliency maps, obtained by existing salient object detection models, to generate weak labels. Interactively, these weak labels are included in the training data for the next iteration of SCNN model optimization. In this way, the two components boost each other interactively and iteratively until the convergence is reached.

To fuse the SCNN generated saliency with other saliency maps generated by several existing saliency models, we exploit weighted linear combination, i.e.,

$$\mathcal{S} = \sum_{i=1}^{N} w_i * \mathcal{S}_i \tag{7}$$

where $\mathcal{S}_i (i = 1, \cdots, N)$ denotes the saliency map generated by SCNN or existing saliency models thus N models used in total. The weights $\{w_1, w_2, ..., w_N\}$ of combination are obtained by the quadratic programming through minimizing

the following objective function

$$w^* = \arg\min_{w_i} ||\mathcal{G} - \sum_{i=1}^{N} w_i * \mathcal{S}_i|| \tag{8}$$

where $\mathcal{G}$ denotes the available manually labeled ground truth and w^* are the optimal weights. In other words, we use the data with ground truth to learn the weights in Eq.8, and these weights are used to fuse those saliency maps without ground truth. With the fused saliency map $\mathcal{S}$, we generate binary labels through Otsu thresholding method. Compared with other methods such as mean thresholding, P-Tile thresholding, etc [50], Ostu is stable, effective and able to produce automatically binary images. Hence, we choose it to generate the binary labels. The overall algorithm of the weakly supervised learning approach is summarized in Algorithm 1. It should be noted that the weak pixel-wise labels by fusing saliency maps are used to optimize the proposed network during the training process. In the testing process, all of the fusing saliency maps are removed in the network. Then, the trained SCNN produces directly the saliency prediction.

In the implementation, three existing saliency models [1]–[3] are introduced to generate fusing saliency maps. These three models are recent deep models and achieve competitive performance in image saliency detection. Although their saliency maps exist some flaws in video datasets, the main salient regions can be detected. Besides, by using the weighted linear combination, the saliency maps can compensate for each other. Thus we can produce more accurate weak pixel-wise labels and exploit them to support the network training.

D. Pixel-wise saliency refinement

The SCNN is already able to detect salient regions in frames, but saliency labels may be too coarse. Therefore, the conditional random field (CRF) framework is exploited to refine saliency pixel-wisely. Specifically, a CRF energy function is defined respect to saliency labels (as random variables), and its minimum leads to the optimal saliency labeling. Following the previous work [51], the CRF energy function is defined as follows:

$$\mathcal{E}(s_i, s_j) = -\sum_i \log \Theta(s_i) + \sum_{i,j} \varphi_{ij}(s_i, s_j) \tag{9}$$

where s_i and s_j denote the pixels in a saliency map $\mathcal{S}$, respectively. The first is a unary potential, where $\Theta(s_i)$ is defined as the saliency value s_i. The second is a pairwise potential and expanded as follows:

$$\begin{aligned}\varphi_{ij}(s_i, s_j) = \mu(s_i, s_j)[\omega_1 * \exp(-\frac{||p_i - p_j||^2}{2\delta_\alpha^2} - \frac{||I_i - I_j||^2}{2\delta_\beta^2}) \\ + \omega_2 * \exp(-\frac{||p_i - p_j||^2}{2\delta_\gamma^2})]\end{aligned} \tag{10}$$

where $\mu_{ij}(s_i, s_j) = 1$ if $s_i \neq s_j$, otherwise 0. The first term is appearance kernel, which promotes the adjacent pixels with similar color appearance to be assigned with the same label. The second term is smoothness kernel, whose purpose is to eliminate small isolated regions. In this equation, p_i, p_j denote coordinates of pixels, and I_i, I_j denote intensity of pixels. The parameters $\delta_\alpha, \delta_\beta$ control the impact of spatial distance, whereas δ_γ determines the impact of intensity contrast; ω_1 and ω_2 are the weights of the two kernels.

IV. Experiments

In this section, we present firstly the most commonly used benchmark datasets and evaluation criteria for salient object detection. Moreover, the implementation of our approach is introduced in detail. Then, we compare our proposed SCNN with the state-of-the-art saliency detection models, and analyze the effect of each module. Finally, the runtime complexity is reported.

A. Datasets and performance evaluation criteria

We perform experiments on four benchmark datasets including MSRA10k [52], SegtrackV2 [7], FBMS [8] and DAVIS [10].

MSRA10K contains 10k images from diverse scenes, such as person, animals, plants, traffic signs, etc. This dataset is widely used in image saliency detection.

SegtrackV2 contains 14 video sequences including 1,066 frames in total. Thus each sequence contains 100 frames appropriately. Each frame is manually labeled for salient objects.

FBMS has 59 video sequences including 13,960 frames in total. This dataset is divided into a training set (with 29 video sequences) and a test set (with 30 video sequences). The ground truths are incomplete and discontinuous. For example, there are 7,306 frames in the test set, but only 720 frames have their corresponding ground truths.

DAVIS consists of 50 video sequences containing 3,455 frames in total, and each frame is pixel-wisely labeled. This dataset contains a diversity of difficult scenes, such multiple objects with occlusion, appearance variation, motion blurred and low contrast, that makes it challenging for salient object detection.

In our experiments, we train our SCNN model with all images in MSRA10K, SeqtrackV2 and the training set of FBMS, and test the performance with the testing set of FBMS and the DAVIS dataset.

As for evaluation criteria, the standard precision-recall (PR) curve is adopted. In computing PR curve, each saliency map is normalized into the range of [0, 255]. Each integer within this range is used as a threshold for segmentation to generate a binary mask to compute precision and recall by comparing against the ground truth. Furthermore, the mean absolute error (MAE) (See Eq.11) is also used to measure the average prediction error between saliency maps and ground truths, which is defined as:

$$MAE = \frac{1}{|\mathcal{S}|} \sum_i |\mathcal{S}(p_i) - \mathcal{G}(p_i)| \tag{11}$$

where p_i denotes any pixel in a frame; $\mathcal{S}$ and $\mathcal{G}$ denote the generated saliency map and its corresponding ground truth, respectively.

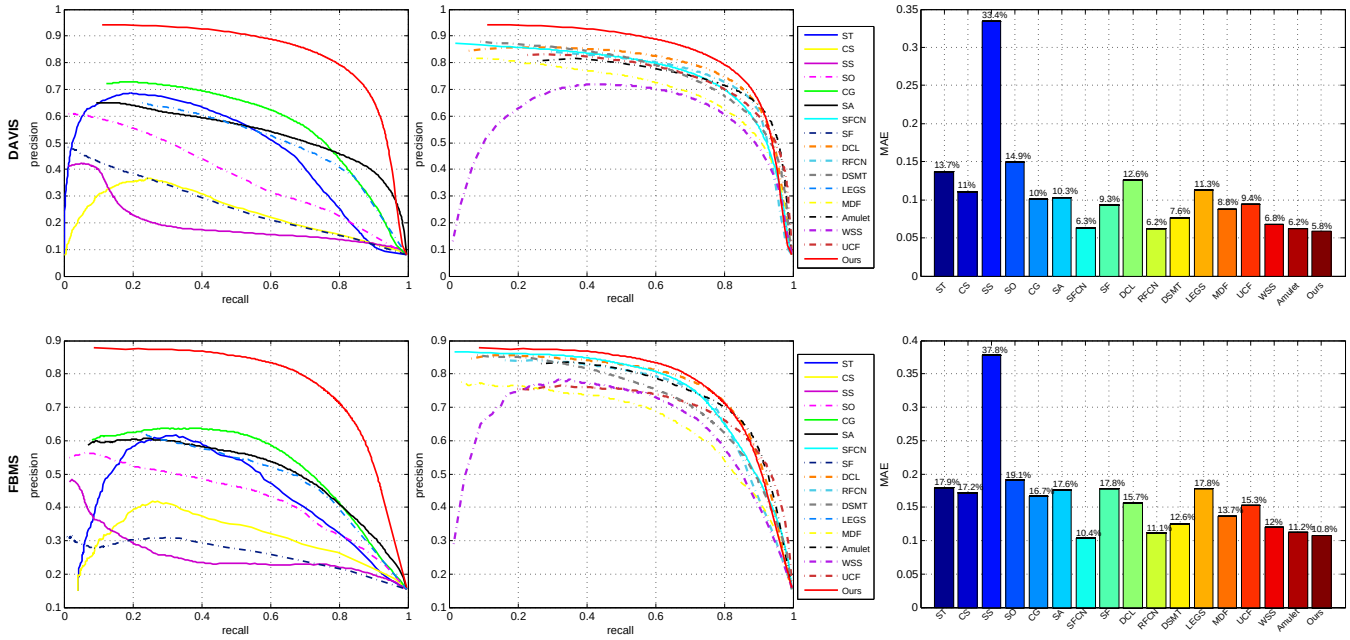

Fig. 5. Comparison with 16 different saliency detection methods including 6 video saliency detection methods (solid lines) and 10 image saliency detection methods (dashed lines) by using DAVIS dataset (top) and FBMS dataset (bottom). The left two columns are PR curves and the right one is MAEs of different methods.

B. Implementation

The proposed SCNN has mainly been implemented with Caffe library [53] and its MATLAB API. As an auxiliary, deep learning toolbox [54] in MATLAB is used at the same time.

For the network training, we start firstly with the pre-trained VGGNet [55], learned from ImageNet dataset [56], and transfer its fully connected layers into fully convolutional layers. Secondly, MSRA10K is used for fine-tuning the FCN. Then, two pre-trained FCN are connected to form the SCNN. The spatiotermporal prior map fusing the spatial prior map generated from the first FCN and temporal prior map evaluated from optical flow field is fed into the second FCN. Since the spatiotermporal prior is added to guide the network learning, the first convolutional layer in the second FCN is not suitable. Therefore, the parameters of this layer are re-initialized with four channels by Xavier function. Thirdly, we use the SegtrackV2 and the training set of FBMS to fine-tune the SCNN. Lastly, in the inference phase, the saliency maps generated from SCNN are refined further by CRF.

In the whole training process, stochastic gradient descent (SGD) is used to update the parameters. The initial learning rate is 10^{-2} and 10^{-10} for the two-phrase fine-tuning, respectively. The weight decay is set to 0.005 and momentum is 0.9.

C. Comparison to the state-of-the-art saliency models

For performance comparison, we compare the proposed approach with six state-of-the-art video saliency approaches and ten image saliency detection approaches both qualitatively and quantitatively. The video saliency approaches we compared are space-time saliency detection (ST) [57], cluster-based co-saliency method (CS) [26], segmenting saliency detection (SS) [23], consistent gradient based saliency (CG) [29], saliency-aware method (SA) [6] and video salient object detection via fully convolutional networks (SFCN) [11]. The compared image saliency detection models include eight deep learning based models: deep contrast learning (DCL) [2], recurrent fully convolutional network (RFCN) [3], deep saliency multi-task (DSMT) [1], local estimation and global search (LEGS) [31], visual saliency on multi-scale deep features (MDF) [32], aggregating multi-level convolutional features (Amulet) [58], saliency detection with image-level supervision (WSS) [14], learning uncertain convolutional features (UCF) [4] and two models based on handcrafted features: robust background detection (SO) [59], saliency filters (SF) [15].

Fig.5 shows the PR curves and MAEs generated by the proposed SCNN model and the other sixteen state-of-the-art saliency models. Clearly, the proposed SCNN achieves obvious higher performance on each dataset in term of both PR curve and MAE criteria. Notice that the PR curve of our approach outperforms the others by a large margin on the DAVIS dataset and also has an improvement on the FBMS dataset. In the aspect of MAE, the proposed SCNN decreases it to 5.8% and 10.8% on the DAVIS and FBMS datasets, respectively.

Fig.6 shows some saliency maps generated by the top sixteen models on the two datasets. The first four sequences (*bmx-bumps, blackswan, bus, train*) are from DAVIS and last three sequences (*camel01, horse04, goats01*) are from FBMS. Notice that the SA and CG by modeling the motion cues are more difficult to predict saliency than the deep learning based models (e.g. DCL, RFCN, MDF) learned from still images only. The deep learning based models have successfully detected most of the salient regions in video sequences.

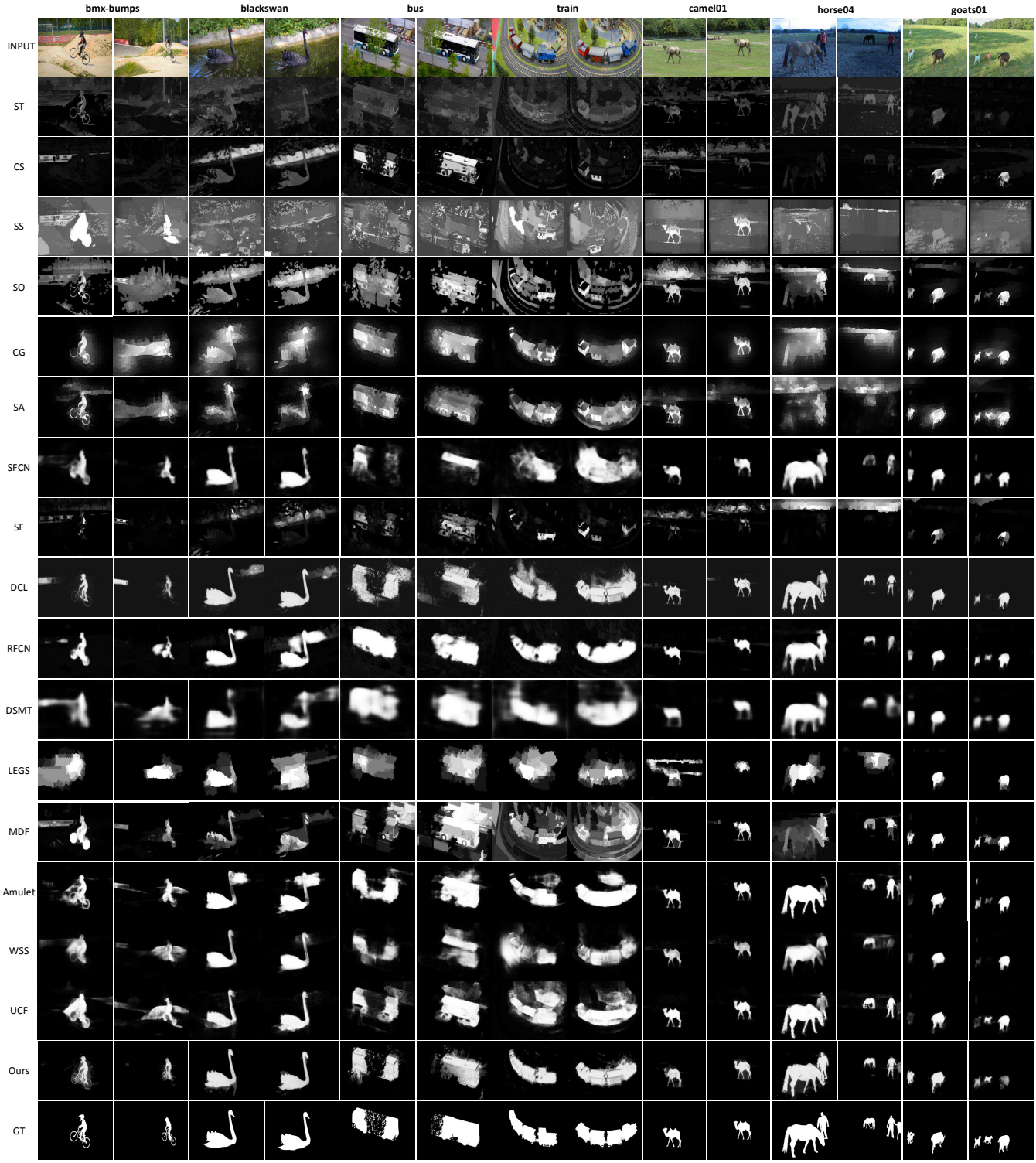

Fig. 6. Saliency maps generated by using part of comparing methods and the proposed approach on DAVIS video sequences (bmx-bumps, blackswan, bus and train), FBMS video sequences (camel01, horses04 and goats01). Qualitatively, our approach removes the inference of still salient region and generates the most similar saliency maps to the ground truths.

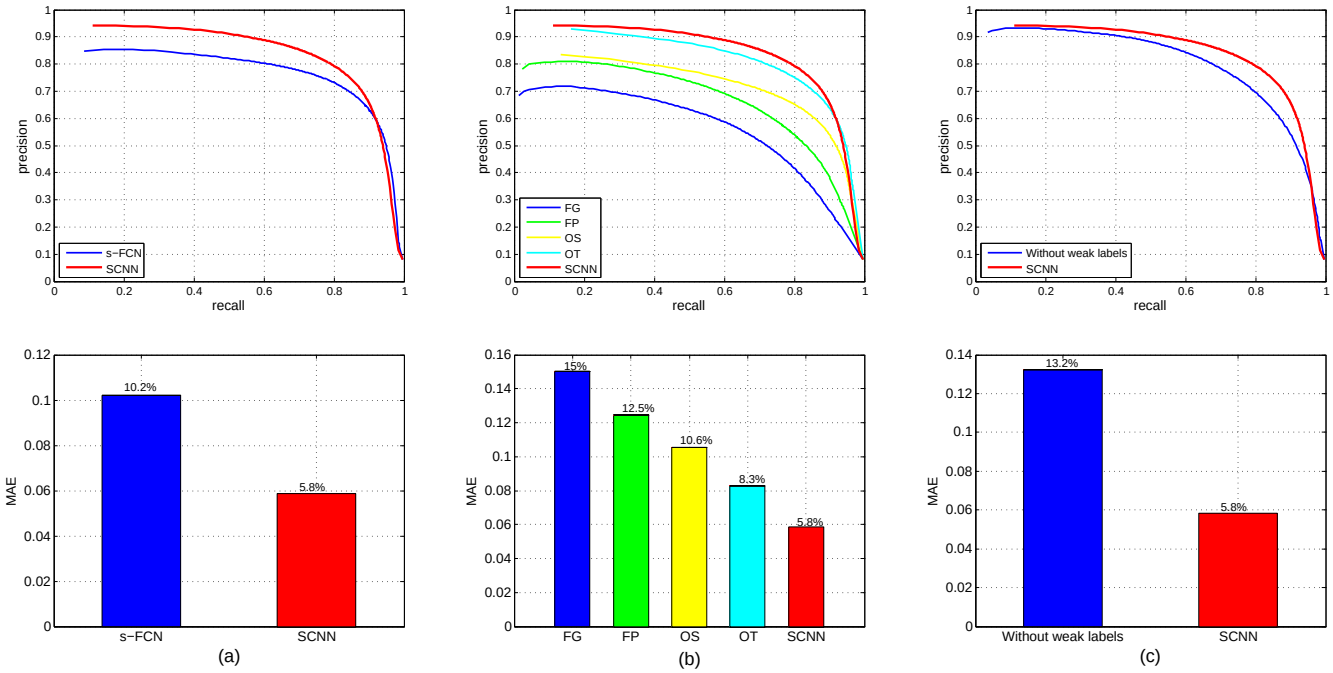

Fig. 7. PR curves (top) and MAEs (bottom) generated by using different configurations on DAVIS dataset. (a) Performance comparison between s-FCN and the proposed SCNN. (b) Performance validation for the spatiotemporal prior. (c) Performance validation for the weakly supervised learning approach.

However, they highlight some background regions as well. For example, the red brand in *bmx-bumps* sequence and river bank in *blackswan* sequence from the background are highlighted as salient regions. The main reason comes from the lack of motion information for learning these models. Our approach employs the spatiotemporal prior to guide saliency modeling and eliminate the inference of unmoving salient regions effectively, which makes the proposed approach achieve a higher quality of saliency maps. As for the multiple objects, our method also achieves competitive performance. In *horse04* and *goats01* sequences, all of the horses and goats are detected by the SCNN.

D. Validation of the proposed approach with different configurations

We perform several experiments on DAVIS dataset to validate the effectiveness of SCNN framework, the spatiotemporal prior and the weakly supervised learning approach, which demonstrate the contributions of this paper.

To validate the proposed SCNN framework, a single fully convolutional network (s-FCN), which has the same the structure of FCN in our SCNN, is used for generating the saliency maps for input video frames. As shown in the Fig.7 (a) the PR curve of SCNN is obviously higher than that of s-FCN and decreases the MAE from 10.2% to 5.8%, which demonstrates the effectiveness by using two FCNs to model visual saliency from both spatial and temporal cues.

To validate the effectiveness of the spatiotemporal prior, we report saliency performance by replacing it with the following alternative methods while keeping other components of SCNN.

- *FG*: the color optical flow image is converted into gray-scale one which is a typical method representing motion information. Then the gray-scale motion image and the spatial prior map $\mathcal{P}_s$ generated from the first FCN using Eq.(5) are stacked with the input video frame to form a five-channel input of the second FCN.
- *FP*: the temporal prior map $\mathcal{P}_t$ generated with Eq.(6) and the spatial prior map $\mathcal{P}_s$ are stacked with the input video frame to form a five-channel input of the second FCN.
- *OS*: only the spatial prior map $\mathcal{P}_s$ is stacked with the input video frame to form a four-channel input of the

Fig. 8. Saliency maps by different configurations. From top to bottom, they are saliency maps by single FCN (s-FCN), connecting graying optical flow prior map and spatial prior map (FG), connecting spatial prior map and temporal prior map (FP), only using the spatial prior of video frame (OS), only using temporal prior map (OT), saliency maps without weak pixel-wise labels (NA) training and Final saliency map (SCNN) with all of components

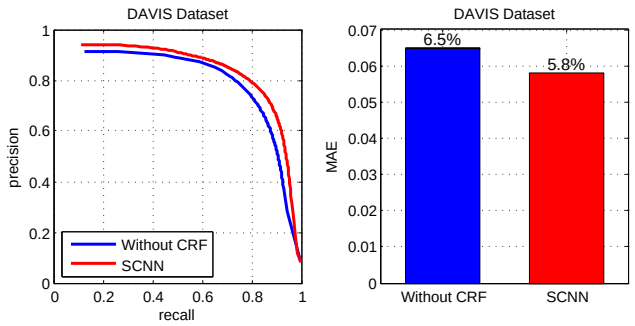

Fig. 9. Comparison between the proposed spatiotemporal cascade neural network and the improvement of the CRF model on DAVIS dataset

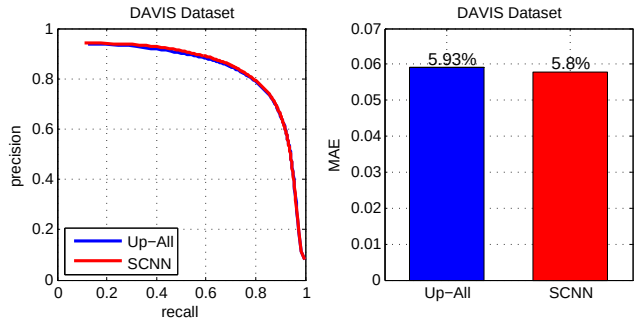

Fig. 10. Comparison of different network configuration by using the multi-layer up-sampling on DAVIS dataset

second FCN.

- *OT*: only the temporal prior map $\mathcal{P}_t$ is stacked with the input video frame to form a four-channel input of the second FCN.

From Fig.7 (b), we can clearly observe that our proposed SCNN guided by the spatiotemporal prior substantially outperforms the four alternative methods mentioned above. In some complex scenes, some background regions are generated in the spatial prior. These background information directly affects the network learning. The element-wise production may remove some regions of salient objects. Besides, due to the accuracy in both spatial prior and temporal prior, the main parts of salient objects can be retained. Therefore, the fusing spatiotemporal prior can guide the SCNN to learn more robust salient features.

The proposed weakly supervised learning approach generates weakly labeled data for training the network. To validate the contribution, in Fig.7(c) we report the saliency performance by using manually labeled ground truths only (without including weak labels) to train our network. Clearly, The proposed SCNN with weakly supervised learning approach achieves notable higher performance. Fig.8 displays the some saliency maps generated by our network with different configurations.

In our approach, we exploit the dense CRF as the post-processing to refine the saliency maps generated from SCNN. In order to validate its effectiveness, we also set a experiment to present the performance of our saliency maps with and without the CRF in DAVIS dataset. The PR curves and MAEs are shown in Fig. 9. We can see that the post-processing indeed enhances the quality of the saliency results from our SCNN, because it can improve the spatial coherence of the generated saliency maps. However, the proposed SCNN also achieves competitive performance.

In the some deep saliency models, the final saliency maps are generated by fusing the up-sampling feature maps from multi-scale convolutional blocks. Through the experiments, we find that the up-sampling feature maps from the first four convolutional blocks have less effect on the final saliency maps in video saliency datasets. Therefore, we try to fuse the up-sampling feature maps from the last two convolutional blocks to generate the final saliency maps. The Fig. 10 shows that the PR curve and MAE of the proposed architecture are slightly better than the ones of integrating the up-sampling feature maps from all of the convolutional blocks.

E. Runtime Analysis

The PC configuration is an Intel(R) i7-5820 CPU (3.3 GHz), a Nvidia Geforce TITAN X GPU (12 GB memory), and 64G RAM. All approaches are run on this PC. Table.I displays the average run time per frame of different methods on DAVIS dataset. Among them, CG, SA, SS and our SCNN by exploiting the optical flow cost much time than others. SCNN$_f$ introduces the FlowNet2.0 [60] to extract optical flow. Compared with traditional method of optical flow extraction [61], FlowNet2.0 is faster by a robust deep learning model. Table. II shows the average run time of each component of the SCNN and SCNN$_f$. We can see that the FlowNet2.0 decreases the computation time from 36.720s to 0.739s and accelerates the proposed method from 38.511s to 2.53s.

TABLE I
COMPARISON AVERAGE RUN TIME (SECONDS PER FRAME) ON DAVIS DATASET

Method	SCNN	SCNN$_f$	ST	CS	SS	SO
Time(s)	38.511	2.53	28.193	1.175	37.176	0.671
Method	CG	SA	SFCN	SF	DCL	RFCN
Time(s)	38.075	38.751	0.473	0.842	0.670	4.580
Method	DSMT	LEGS	MDF	Amulet	UCF	WSS
Time(s)	0.14	0.206	11.33	5.299	0.151	0.024

TABLE II
AVERAGE RUN TIME (SECONDS PER FRAME) OF EACH COMPONENT IN THE PROPOSED APPROACH ON DAVIS DATASET

Model	Component	Time (s)	Ratio (%)
SCNN	Optical flow computation	36.720	95.36
	Temporal prior generation	0.823	2.14
	Neural network processing	0.685	1.78
	Saliency refinement	0.283	0.72
	Total	38.511	100
SCNN$_f$	Optical flow computation	0.739	29.2
	Temporal prior generation	0.823	32.5
	Neural network processing	0.685	27.1
	Saliency refinement	0.283	11.2
	Total	2.53	100

V. Conclusion

In this paper, we propose a novel SCNN for salient object detection in a video. The framework integrates the spatial prior of the video frame and the temporal prior based on optical flow, which successfully eliminates unmoving salient region and generates final saliency maps in dynamic scenes. Based on optical flow, we subtly incorporate superpixel segmentation on optical flow map and multi-scale deep features to obtain a high-quality temporal prior map, which can guide the training of SCNN and support accurate inference. Furthermore, facing the shortage of training data, a weak supervised learning strategy is proposed. This method enables our network to learn more information and make saliency maps more accurate. Finally, we performed an extensive evaluation on the widely used FBMS and DAVIS dataset. Experiments denote that the proposed approach substantially outperforms the state-of-the-art video saliency and image saliency models in term of both PR curve and MAE criteria.

VI. Acknowledgment

The authors would like to thank the anonymous reviewers and the Associate Editor for their insightful comments and suggestions, which are greatly helpful for improving the quality of this paper.

References

[1] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

[2] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 478–487.

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

[4] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[5] 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.

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

[7] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2013, pp. 2192–2199.

[8] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proceedings of the European Conference on Computer Vision*. Springer, 2010, pp. 282–295.

[9] F. Galasso, N. Shankar Nagaraja, T. Jimenez Cardenas, T. Brox, and B. Schiele, "A unified video segmentation benchmark: Annotation, metrics and analysis," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2013, pp. 3527–3534.

[10] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 724–732.

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

[12] H. Cholakkal, J. Johnson, and D. Rajan, "Backtracking scspm image classifier for weakly supervised top-down saliency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 5278–5287.

[13] H. Cholakkal, J. Johnson, and D. Rajan, "Weakly supervised top-down salient object detection," *arXiv preprint:1611.05345*, 2016.

[14] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and R. Xiang, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 3796–3805.

[15] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 733–740.

[16] W. Zou, K. Kpalma, Z. Liu, and J. Ronsin, "Segmentation driven low-rank matrix recovery for saliency detection," in *Proceedings of the British Machine Vision on Conference*, 2013, pp. 1–13.

[17] W. Zou and N. Komodakis, "HARF: Hierarchy-associated rich features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 406–414.

[18] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2085–2098, 2015.

[19] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, pp. 32–32, 2008.

[20] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proceedings of the ACM international conference on Multimedia*. ACM, 2006, pp. 815–824.

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

[22] 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.

[23] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proceedings of the European Conference on Computer Vision*. Springer, 2010, pp. 366–379.

[24] 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.

[25] R. Kalboussi, M. Abdellaoui, and A. Douik, "A spatiotemporal model for video saliency detection," in *Proceedings of the Image Processing, Applications and Systems*. IEEE, 2017, pp. 1–6.

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

[27] 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.

[28] R. Dosil, X. R. Fdez-Vidal, and X. M. Pardo, "Motion representation using composite energy features," *Pattern Recognition*, vol. 41, no. 3, pp. 1110–1123, 2008.

[29] 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.

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

[31] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3183–3192.

[32] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3431–3440.

[34] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 678–686.

[35] Y. Tang, X. Wu, and W. Bu, "Deeply-supervised recurrent convolutional neural network for saliency detection," in *Proceedings of the ACM on Multimedia Conference*. ACM, 2016, pp. 397–401.

[36] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 2846–2854.

[37] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. D. Bourdev, "Pronet: Learning to propose object-specific boxes for cascaded neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3485–3493.

[38] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 685–694.

[39] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 189–203, 2016.

[40] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 1796–1804.

[41] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5038–5047.

[42] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," *arXiv preprint:1703.01290*, 2017.

[43] B. Lai and X. Gong, "Saliency guided end-to-end learning for weakly supervised object detection," *arXiv preprint:1706.06768*, 2017.

[44] B. Lai and X. Gong, "Saliency guided dictionary learning for weakly-supervised image parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 3630–3639.

[45] H. Cholakkal, J. Jubin, and D. Rajan, "Weakly supervised top-down salient object detection," *arXiv preprint:1611.05345*, 2016.

[46] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised saliency detection with a category-driven map generator," in *Proceedings of the British Machine Vision Conference*, 2017.

[47] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint:1409.1556*, 2014.

[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint:1412.7062*, 2014.

[50] P. K. Sahoo, S. Soltani, and A. K. C. Wong, "A survey of thresholding techniques," *Computer Vision Graphics and Image Processing*, vol. 41, no. 2, pp. 233–260, 1998.

[51] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proceedings of the Advances in Neural Information Processing Systems*, 2011, pp. 109–117.

[52] 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. 353–67, 2011.

[53] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[54] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," *Technical University of Denmark*, 2012.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1–9.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[57] F. Zhou, S. Bing Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 3358–3365.

[58] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2017.

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

[60] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1647–1655.

[61] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014.

Yi Tang received the B.S. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2012, and the M.S. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the College of Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include saliency detection, deep learning and multimodal analysis in multimedia.

Wenbin Zou received the M.E. degree in software engineering with a specialization in multimedia technology from Peking University, China, in 2010, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2014. From 2014 to 2015, he was a Researcher with the UMR Laboratoire d'informatique Gaspard-Monge, CNRS, and École des Ponts ParisTech, France. Since then, he has been with the faculty of the College of Information Engineering, Shenzhen University, China. His current research interests include saliency detection, object segmentation, and semantic segmentation.

Zhi JIN (M'16-current) received the B.S degree in Telecommunication Engineering from the University of Liverpool (UoL), UK and Xi'an Jiaotong-Liverpool University (XJTLU), China, in 2011. She received the Ph.D. degree from the University of Liverpool, UK in 2016. From 2016, she worked as a Postdoctoral Researcher in Shenzhen University. Her current research interests include 3D video processing, the applications of depth map and image/video quality enhancement.

Yuhuan Chen received the B.S. degree from Gannan Normal University, Ganzhou, China, in 2003, and the M.S. degree from University of Science and Technology of Jiangxi, Ganzhou, China, in 2007. She is currently pursuing the Ph.D. degree with the College of Information Engineering, Shenzhen University, Shenzhen, China. Her current research interests include saliency detection, object segmentation and pattern recognition.

Yang Hua Yang Hua is presently a lecturer in the School of Electronics, Electrical Engineering and Computer Science at the Queens University of Belfast, UK. He received hisPh.D. degree from Universit Grenoble Alpes / Inria Grenoble Rh?ne-Alpes, France, funded byMicrosoft Research C Inria Joint Center. Hewon PASCAL Visual Object Classes (VOC) Challenge Classification Competition in 2010, 2011 and 2012,receptivelyand the Thermal Imagery Visual Object Tracking (VOT-TIR) Competition in 2015.His research interests includemachine learning methods for image and video understanding. Heholds three US patents and one China patent.

Xia Li received her B.S. and M.S. in electronic engineering and SIP (signal and information processing) from Xidian University in 1989 and 1992 respectively. She was later conferred a Ph.D. in Department of information engineering philosophy by the Chinese University of Hong Kong in 1997. She was a former dean of College of Information Engineering at Shenzhen University, a director of Shenzhen Key Laboratory of Advanced Communication and Information Processing. She is currently Associate Vice-President at the Chinese University of Hong Kong, Shenzhen. Her research interests include intelligent computing and its applications, image processing and pattern recognition.