Multi-Scale Spatiotemporal Conv-LSTM Network for Video Saliency Detection

Conference Paper · June 2018

DOI: 10.1145/3206025.3206052

Yi Tang
Beijing University of Posts and Telecommunications
9 PUBLICATIONS **496** CITATIONS

Wenbin Zou
Shenzhen University
108 PUBLICATIONS **2,553** CITATIONS

Zhi Jin
Sun Yat-Sen University
66 PUBLICATIONS **886** CITATIONS

Multi-Scale Spatiotemporal Conv-LSTM Network for Video Saliency Detection

Yi Tang
College of Information Engineering, Shenzhen University
Shenzhen, Guangzhou, P.R.China
yitang@szu.edu.cn

Wenbin Zou*
College of Information Engineering, Shenzhen University
Shenzhen, Guangzhou, P.R.China
zouszu@sina.com

Zhi Jin
College of Information Engineering, Shenzhen University
Shenzhen, Guangzhou, P.R.China
jinzhi_126@163.com

Xia Li
Chinese University of Hong Kong
Shenzhen, Guangzhou, P.R.China
lixia@cuhk.edu.cn

ABSTRACT

Recently, deep neural networks have been crucial techniques for image salient detection. However, two difficulties prevent the development of deep learning in video saliency detection. The first one is that the traditional static network cannot conduct a robust motion estimation in videos. The other is that the data-driven deep learning is in lack of sufficient manually annotated pixel-wise ground truths for video saliency network training. In this paper, we propose a multi-scale spatiotemporal convolutional LSTM network (MSST-ConvLSTM) to incorporate spatial and temporal cues for video salient objects detection. Furthermore, as manually pixel-wised labeling is very time-consuming, we sign lots of coarse labels, which are mixed with fine labels to train a robust saliency prediction model. Experiments on the widely used challenging benchmark datasets (e.g., FBMS and DAVIS) demonstrate that the proposed approach has competitive performance of video saliency detection compared with the state-of-the-art saliency models.

CCS CONCEPTS

• **Computing methodologies → Interest point and salient region detections**; *Machine learning approaches*; *Neural networks*;

KEYWORDS

Video Saliency, Deep Learning, Spatiotemporal Fusion

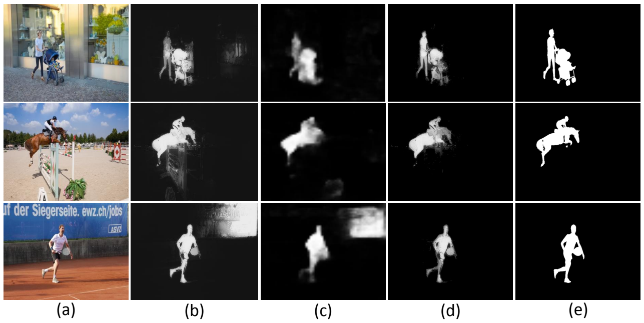

Figure 1: Saliency detection by different deep learning models. (a) Video frame. (b) Saliency maps by image saliency detection [15]. (c) Saliency maps by video saliency detection [31]. (d) Saliency maps by the proposed network. (e) Ground truths.

ACM Reference Format:
Yi Tang, Wenbin Zou, Zhi Jin, and Xia Li. 2018. Multi-Scale Spatiotemporal Conv-LSTM Network for Video Saliency Detection. In *Proceedings of 2018 International Conference on Multimedia Retrieval (ICMR'18)*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3206025.3206052

1 INTRODUCTION

Saliency detection, whose purpose is accurately and uniformly highlighting the regions that draw human attention in an image or video, has become a very active research field in computer vision because of its importance to support high-level computer vision tasks, such as video classification [22, 36], action recognition [6, 19] and content-based image retrieval [33]. According to the type of input data, saliency detection can be classified into image saliency detection and video saliency detection. In this paper, we focus on video saliency detection.

Recently, deep learning models, especially the fully convolutional network (FCN), have widely been employed to detect salient objects and substantially improved the accuracy of saliency detection in static images. However, these deep

[*] Wenbin Zou is the corresponding author.

ICMR'18, June 11–14, 2018, Yokohama, Japan
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5046-4/18/06. . . $15.00
https://doi.org/10.1145/3206025.3206052

learning models cannot adapt directly to the video saliency detection, even if the transfer learning has been introduced to fine-tune the deep networks. The difficulties behind this phenomenon may be two-fold:

One of the difficulties is that the traditional networks in static images cannot capture motion cues in a single frame. These networks learn the features only in an image space, which leads to detect some extra objects in complex video scenes (e.g., Figure.1 (b)). In order to integrate the static and motion cues, [31] tries to train a suitable deep network to complete the task. Their deep model mainly exploits the motion cues between two consecutive frames, but it cannot learn sufficient motion information to completely eliminate non-salient regions in some videos (e.g., Figure.1 (c)).

The other issue is the lack of sufficient pixel-wise labeled video data to train the deep learning network. To obtain robust deep networks, large-scale labeled data are needed. Unfortunately, current datasets for video saliency detection have very limited pixel-wise ground truths. The total number of labeled data in the widely used video datasets (e.g., SegTrackV2 [14], FBMS [1], DAVIS [24]) is less than 5000 frames (including the data for testing). Moreover, the pixel-wise labels of some video sequences are discontinuous. Therefore, insufficient data makes the deep network hardly learn robust motion features, thus reducing the accuracy of video saliency detection.

Based on the aforementioned issues above, we, on the one hand, try to find a desirable deep network architecture which can incorporate spatial and temporal cues to accurately detect salient objects in video sequences. On the other hand, we expect to obtain adequate pixel-wise labeled video data to train the proposed network. Hence, we propose a multi-scale spatiotemporal convolutional LSTM network (MSST-ConvLSTM) for video saliency detection and train this network with a amount of coarse pixel-wise labels. The proposed network consists of two FCN-based sub-networks, spatial sub-network and temporal sub-network. First, we exploit the spatial sub-network, whose input is a static frame, to extract spatial features. Second, we extract the motion features from two aspects. One is the optical flow. A motion prior map generated by optical flow is fed into the temporal sub-network to extract temporal features. The other is the proposed MSST-ConvLSTM. The multi-scale feature maps of several consecutive frames are produced from both sub-networks, and then the recurrent-based model exploits these feature maps to further extract the sequential motion features. Third, the spatial and temporal features are effectively fused at the top of the proposed network to generate fine-grained video saliency prediction (e.g., Figure.1 (d)). Besides, we introduce a coarse labeling strategy, which fuses saliency maps from different saliency models, and then manually erases the error detection regions from the fusing saliency maps. Through this strategy, a number of coarse pixel-wise labels are rapidly generated to support the training of the proposed network.

To summarize, the main contributions of this paper are as follows:

1. We propose a multi-scale spatiotemporal convolutional LSTM network for video saliency detection. This architecture not only retains original spatial cue, but also effectively integrates temporal cue from optical flow map and the structure of LSTM.
2. Following the labeling criteria of [1], we choose some videos in FBMS dataset and label the frames in a coarse pixel-wise way, which makes our deep network learn more robust salient features.
3. We demonstrate that our proposed approach substantially outperforms the state-of-the-art salient object detection models.

2 RELATED WORKS

Over the recent years, a large number of approaches have been proposed in the field of saliency detection. Especially, with the raise of deep learning architecture, the performance of saliency detection obtains a significant improvement. In this section, we give a brief overview of recent saliency detection works in two categories: traditional modeling-based approaches and deep learning-based approaches.

2.1 Modeling-based saliency detection

Saliency detection can be further divided into image saliency and video saliency according to their inputs. The traditional modeling-based approaches in image saliency detection contain low-rank matrix recovery [39], spatial prior [32], regional contrast [38, 40], and graphical modeling [4]. These models are built in the image space and cannot be directly adapted to video saliency. Therefore, some spatiotemporal fusing models [34, 37] have been proposed to incorporate the spacial and temporal cues. In [2], the global motion clues are firstly exploited to generate an initial saliency, and then the low-rank coherency clues are employed to guide the spatial-temporal saliency diffusion. By incorporating the spatial transition matrix and the temporal restarting distribution, Kim et al. [12] propose a novel saliency detection algorithm based on the random walk to estimate saliency distribution. In [25], a statistical framework is proposed for saliency prediction by using motion, illumination and color information. Moreover, as the estimation of moving pixels, optical flow has been widely used in video saliency detection [29, 30]. Although the computation of optical flow is time-consuming, it can extract robust motion features to detect salient objects. Meanwhile, some other methods have been proposed to adopt some optimization mechanisms, such as center-surrounding hypothesis [9], motion attention cue [8], motion boundaries [21], and nonparametric kernel density feature [26].

2.2 Deep learning-based saliency detection

Compared with traditional modeling-based methods, the development of deep learning-based methods can be classified into two phases. In the first phase, convolutional neural network (CNN) is exploited as a feature extractor. The handcrafted features are replaced by the deep features from CNNs

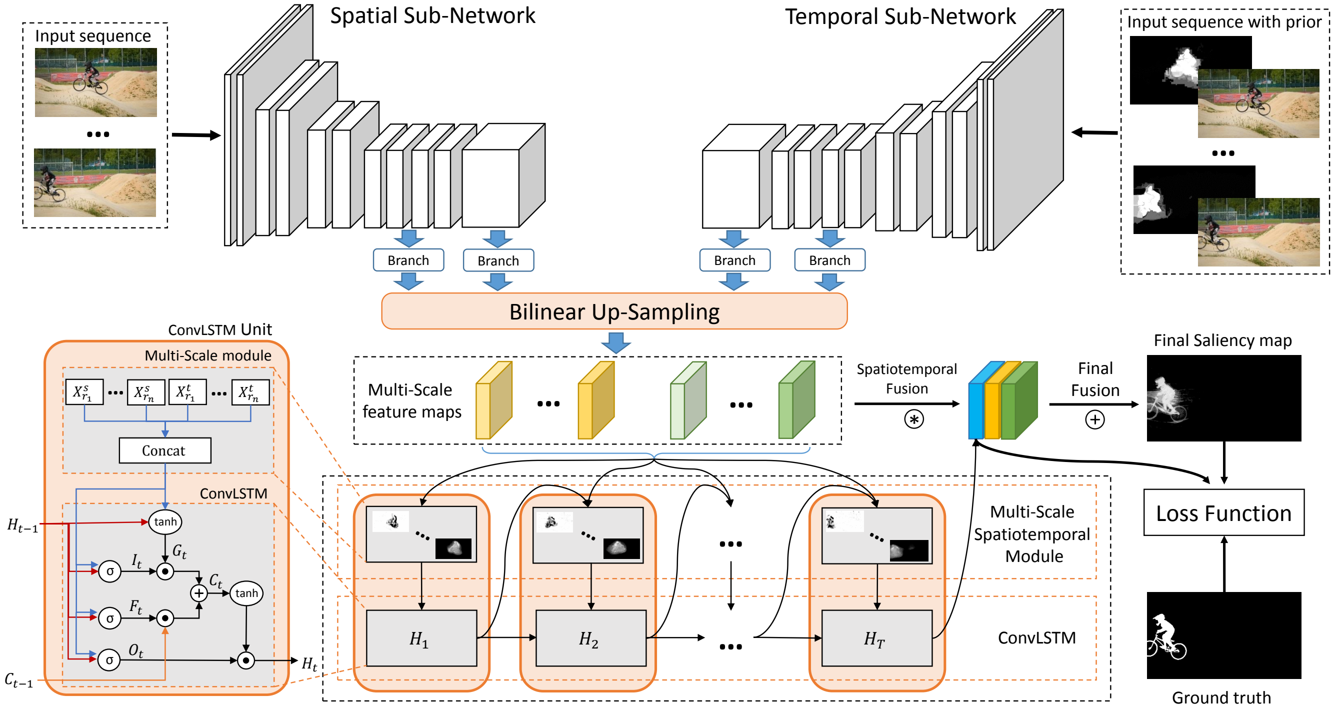

Figure 2: The framework of the proposed network. The multi-scale feature maps are generated from spatial sub-network and temporal sub-network, which are trained by video frames and video frames with the corresponding motion prior, respectively. Along with the modified ConvLSTM units to refine motion information, all of feature maps are fused to generate saliency map.

to detect salient objects in an image. In [16], the multi-scale deep features are combined with multi-region decomposition to detect salient regions. In [35], the deep features of image superpixels are extracted to generate saliency maps. Besides, the integration of region proposal and deep features are adopted for local estimation and global search [27]. In the second phase, an end-to-end FCN [20] has been employed into saliency detection field. Based on the FCN, a series of approaches have been proposed. In order to exploit multi-level feature maps, DHSN [18] proposes a hierarchical FCN to estimate visual saliency. RFCN [28] incorporates saliency prior knowledge, recurrent neural network (RNN) and FCN to perform a full saliency prediction. DCL [15] considers the reduction of receptive field with excessive pooling layers, so the dilated convolutional operation [3] is introduced into FCN. DSMT [17] designs a multi-task convolutional neural network to optimize FCN. Considering a more suitable up-sampling structure, DSS [10] introduces a short connection into FCN.

Due to the shortage of consecutive pixel-wise ground truths, the approaches of video saliency detection by deep learning are not much. Recently, with a novel data augmentation techniques, SFCN [31] proposes an effective deep learning architecture, which fuses spatial and temporal saliency stimuti well. However, their extracted motion information from two consecutive frames is not robust enough. We expect to design more suitable network structure to further extract the motion features.

3 MSST-CONVLSTM FOR VIDEO SALIENCY DETECTION

As shown in Figure.2, the proposed framework is composed of three components, spatial sub-network, temporal sub-network and ConvLSTM units. The spatial sub-network is a modified VGG-based FCN, which introduces dilated convolutional layers in the last two convolutional blocks. This sub-network processes static frames to extract the feature maps with spatial information. The temporal sub-network has almost the same structure with the spatial one. The only difference is that a four-channel image, which consists of a RGB frame and corresponding motion prior (described in section 3.1), is fed into temporal sub-network to generate the feature maps with temporal information. We design serval branches after pooling layers to produce multi-scale feature maps in both sub-networks. Then, the feature maps are fed into the ConvLSTM units to further refine motion information. At last, the feature maps from spatial sub-network, temporal sub-network and ConvLSTM units are fused at the top of network through a convolutional layer (1 × 1 kernel) to produce the final saliency maps.

3.1 The generation of motion prior

In the temporal sub-network, we aim to obtain robust motion features from the video sequences. Instead of solely relying on optical flow map, we feed a four-channel image into the temporal sub-network, which can ensure that the sub-network learns robust motion information even if the optical flow map is serious fault. In the generation of motion prior, we firstly conduct $\mathcal{M}$ segmentations on the optical flow map by using [7] to obtain multi-scale superpixels. Secondly, the deep features [16] are extracted for each superpixel $r_i^j (j = 1, 2, ..., \mathcal{M})$, where i denotes the superpixel index in the j-level segmentation. Thirdly, a binary classifier based on a three-layer neural network (300 neurons) is trained for salient superpixel prediction. Finally, a linear fusion is employed to generate the motion prior $\mathcal{P}_m$ from different segmentation levels as below:

$$\mathcal{P}_m(r_i) = \frac{1}{\mathcal{M}} \sum_{j=1}^{\mathcal{M}} \mathcal{S}^j \left(\mathcal{D}(r_i^j)\right) \tag{1}$$

where $\mathcal{D}(\cdot)$ and $\mathcal{S}^j(\cdot)$ represent the deep features of the superpixel r_i^j and saliency value predicted by the classifier, respectively.

3.2 The proposed architecture

The VGG-based FCN contains five pooling layers, which gradually increase the receptive field and separate the network into six convolutional blocks. In order to retain fine semantic context, the last pooling layer is removed and the dilated convolutional layers are introduced into the last two convolutional blocks in modified FCN. Meanwhile, we add several branches after these pooling layers. Each of the branches contains a 128-channel convolutional layer (3 × 3 kernel) and a 1-channel convolutional layer (1 × 1 kernel). With the bilinear up-sampling layers, the multi-scale feature maps are generated from these branches in both spatial and temporal sub-networks. These feature maps are fed into the ConvLSTM units to further refine the motion information.

The bottom-left part of Figure.2 shows the specific structure of ConvLSTM unit, which is composed of a multi-scale spatiotemporal module and a ConvLSTM module. Assume I^s and I^t denote the inputs of both sub-networks, the processing steps of the MSST-ConvLSTM network can be demonstrated as follows:

$$X^{\tau} = \mathbf{Cat}(X^s_{r_1}, ..., X^s_{r_n}, X^t_{r_1}, ..., X^t_{r_n}) \tag{2}$$

where

$$\begin{aligned} X^s_{r_i} &= \mathcal{U}^s_{r_i}(\mathcal{F}^s_{r_i}(I^s; \theta^s_{r_i})) \\ X^t_{r_i} &= \mathcal{U}^t_{r_i}(\mathcal{F}^t_{r_i}(I^t; \theta^t_{r_i})), i \in \{1, 2, ..., n\} \end{aligned} \tag{3}$$

The spatial feature maps $X^s_{r_i}$ and temporal feature maps $X^t_{r_i}$ at different scale r_i are extracted through a series of convolutional operations $\mathcal{F}^s_{r_i}(\cdot)$ and $\mathcal{F}^t_{r_i}(\cdot)$, and up-sampling operations $\mathcal{U}^s_{r_i}(\cdot)$ and $\mathcal{U}^s_{r_i}(\cdot)$. Then, a cross-channel concatenation operation **Cat** is introduced to stack the feature maps.

After that, the sequence of concatenated feature maps $X^1, ..., X^T$ is fed into the ConvLSTM $\mathcal{M}$, where T is the timesteps, which is set to four in our experiment; θ_m is the parameters. As we select a many-to-one model, the output of these operations is a 2-d refined spatiotemporal feature map O^T:

$$O^T = \mathcal{M}(X^1, ..., X^T; \theta_m) \tag{4}$$

At the top of the proposed network, the feature maps from spatial sub-network, temporal sub-network and ConvLSTM units are fused to produce the final prediction map. To retain relative semantic context, a hierarchical structure is exploited to fuse the spatial, temporal and historical semantic information:

$$\mathcal{S} = W^s * X^s + W^t * X^t + O^T \tag{5}$$

where

$$\begin{aligned} X^s &= \mathbf{Cat}(X^s_{r_1}, ..., X^s_{r_n}) \\ X^t &= \mathbf{Cat}(X^t_{r_1}, ..., X^t_{r_n}), i \in \{1, 2, ..., n\} \end{aligned} \tag{6}$$

X^s and X^t are concatenated multi-scale feature maps from spatial and temporal sub-network, respectively; W^s and W^t are their corresponding convolutional parameters; O^T is the spatiotemporal feature map from ConvLSTM units; $\mathcal{S}$ denotes the final prediction map. We firstly use two convolutional layers (1 × 1 kernel) to fuse the feature maps from spatial and temporal sub-network. Then, a element-wise sum is employed to fuse all of the feature maps.

To optimize the whole network, a loss function is applied to compute the errors between the final prediction map $\mathcal{S} \in [0, 1]^{h*w*1}$ and the pixel-wise label $\mathcal{G} \in [0, 1]^{h*w*1}$, where h and w represent the height and width of the input video frame, respectively. Considering the unbalance between salient pixels and non-salient pixels, we employ a weighted sigmoid cross-entropy loss function $\mathcal{L}_f$ as follow:

$$\begin{aligned} \mathcal{L}_f(\mathcal{S}, \mathcal{G}) = &- \alpha \sum_{i=1}^{h*w} g_i \log P(s_i = 1 | I_i^s, I_i^t; \mathcal{W}) \\ &- (1 - \alpha) \sum_{i=1}^{h*w} (1 - g_i) \log P(s_i = 0 | I_i^s, I_i^t; \mathcal{W}) \end{aligned} \tag{7}$$

where $s_i \in \mathcal{S}$ and $g_i \in \mathcal{G}$ demonstrate the saliency value and the label of ground truth for a pixel, respectively; $\mathcal{W}$ is the parameter of the proposed network; $P(\cdot|\cdot)$ denotes the confidence score of the prediction; α is the balance factor.

To speed up the training of the ConvLSTM units, we set another auxiliary loss function $\mathcal{L}_r$ after the ConvLSTM units. This loss function is also a weighted sigmoid cross-entropy loss, but the parameters do not include the fusing operations. Therefore, the final joint loss function for the whole network is defined as below:

$$\mathcal{L} = \mathcal{L}_f(\mathcal{S}, \mathcal{G}) + \beta \mathcal{L}_r(\mathcal{R}, \mathcal{G}) \tag{8}$$

where $\mathcal{R}$ is the feature map from ConvLSTM units; β is the loss weight. Here, it is set to 0.1.

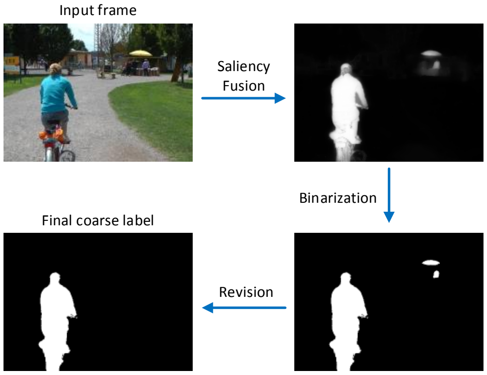

Figure 3: The generation of coarse pixel-wise labels.

3.3 The coarse pixel-wise labeling

Due to the shortage of pixel-wise labeled video data, the proposed network is hard to train. The best solution is labeling sufficient and fine pixel-wise labels, but it is very time-consuming. Therefore, instead of the fine labels, we sign video data with coarse pixel-wise labels, which retain the rough outline but ignore the details of salient objects. The generation pipeline is displayed in Figure.3. Firstly, given the video sequences in the training set of FBMS dataset, the fused saliency maps of video frames are generated by a weighted linear combination [11] of three image saliency detection methods (i.e., DCL [15], RFCN [28], DSMT [17]), which are the recent deep learning models and achieve competitive performance in image saliency detection. Although some background regions are highlighted in the complex scenes, these approaches are able to detect the main salient objects. Besides, we employ the weighted linear combination to fuse these saliency maps, which can make these saliency maps compensate for each other and generate more accurate ones. Then, with the fusing saliency maps, the binary pixel-wise labels are generated by Otsu thresholding method. Finally, we manually erase the error detection regions and produce the final coarse pixel-wise labels. In our approach, we totally obtain 3,326 coarse pixel-wise labels, which are mixed with the original labeled ones to train the proposed network.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. Three datasets, SegTrackV2 [14], FBMS [1], and DAVIS [24], are used for evaluating our experiments.

In our experiments, the training data (totally 51 sequences, 6,098 frames) includes the half of DAVIS, the training set of FBMS (with generated coarse pixel-wise labels) and all of SegTrackV2 videos. The testing set is the remaining video sequences of DAVIS and the testing set of FBMS (with ground truths).

Evaluation criteria. In our experiments, the standard precision-recall (PR) curve and mean absolute error (MAE) are adopted. In PR curve, the binary mask is converted from the saliency map with a integer threshold. Then, the precision and recall is computed by comparing the binary mask against the ground-truth. In addition, MAE is introduced as a complemental measure. MAE measures the average absolute difference between a saliency map and its corresponding ground truth in pixel level.

Implementation. To better fuse spatial and temporal cues and obtain robust feature maps, the network training follows next several settings:

- Before the network training, both spatial and temporal sub-networks are initialized by the same pre-trained VGG-based FCN from ImageNet [5]. In order to match the input of the temporal sub-network, the first convolutional layer is re-initialized by Gaussian distribution.
- In the proposed network, we design several branches to extract the multi-scale feature maps. Specifically, there are two branches in each sub-network, one is after the fourth pooling layers, the other is after the last convolutional layer.
- As the application of FCN, the input size of images is arbitrary. To make the network to learn more fine-grained features, two resolution images (321 × 321 and 512 × 512) are used for training.
- In the testing stage, to improve spatial coherence, a dense CRF [3] is employed as a post-processing method to refine the generated saliency maps.
- In the entire training process, the Adaptive Moment Estimation (Adam) [13] is chosen to optimize the network. The initial learning rate is 10^{-3} for the proposed network training.

The PC configuration is an Intel(R) i7-5820 CPU (3.3GHz), 64G RAM and a Nvidia Geforce TITAN X GPU. The proposed method costs 2.9 seconds to process a sequence (4 frames with 512 × 512 resolution). The network forwarding operation takes 2.1 seconds and the post-processing costs 0.8 seconds.

4.2 Comparison to the State-of-the-art Methods

In this section, 6 video saliency detection methods and 8 image saliency detection methods are compared with the proposed approach. The video saliency detection methods include ST [37], CS [8], SS [25], CG [30], SA [29] and SFC-N [31]. The image saliency detection methods include SF [23], MDF [16], DCL [15], RFCN [28], DSMT [17], LEGS [27], DSS [10] and DHSN [18].

As shown in Figure.4, the proposed method outperforms the existing approaches in term of both PR curve and MAE criteria. Notice that the PR curves demonstrate that the proposed method achieves the competitive performance. At the same time, the MAEs are decreased to 8.2% and 4.5% on the FBMS and DAVIS, respectively. The proposed method obtains the lowest MAE among all the state-of-the-art methods.

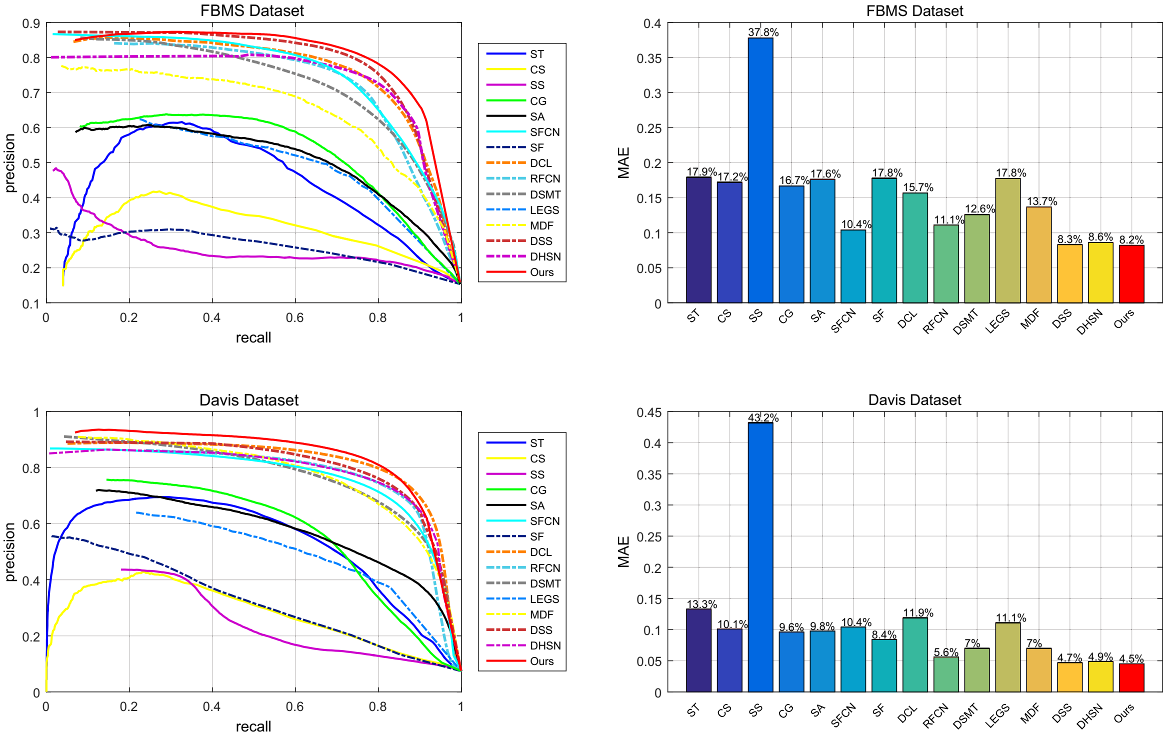

Figure 4: Comparison with 14 different saliency detection methods including 6 video saliency detection methods (solid lines) and 8 image saliency detection methods (dashed lines) by using FBMS dataset (top) and DAVIS dataset (bottom). The left two figures are PR curves and the right two are MAEs of different methods.

In qualitative comparison, Figure.5 shows the saliency maps generated by different saliency detection models. Based on these examples, serval observations are obtained as follows:

Multiple salient objects: In the *horse04* sequence (the third row), there are three salient objects in the ground truth. Previous models cannot highlight all of them. Our model by fusing spatial and temporal features successfully detects all of the salient objects in this video.

Tiny salient objects: For some video scenes containing tiny saliency (e.g., the fifth row), previous models can detect the moving soccer ball, but some other regions also are detected at the same time. On the contrary, the proposed approach with multi-scale spatiotemporal features is able to pop out the tiny object from background regions.

Motion features: Our network effectively extracts motion features from two aspects (optical flow and RNNs). These features are robust enough to make the moving objects more salient and eliminate extra background regions (e.g., the bench in *lucia* sequence).

4.3 Ablation Studies

The proposed deep learning architecture consists of spatial sub-network, temporal sub-network and MSST-ConvLSTM. In order to validate the effectiveness of these components, we report the saliency performance by different configurations of network.

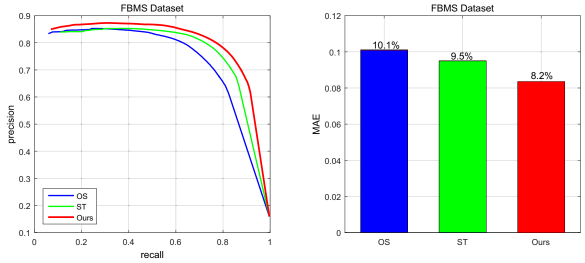

Figure 6: PR curves and MAEs by using different configurations on FBMS dataset.

- *OS*: This architecture only contains a spatial sub-network, which is extended by modified VGGNet. We train this network only with video frames and do not introduce any temporal information.
- *ST*: This network combines spatial and temporal sub-networks. The feature maps generated from both sub-networks are integrated into the final saliency map at the top of the network.
- *MSST-ConvLSTM*: Based on the integration of spatial and temporal sub-network, the proposed MSST-ConvLSTM architecture is introduced to make the network further learn motion features.

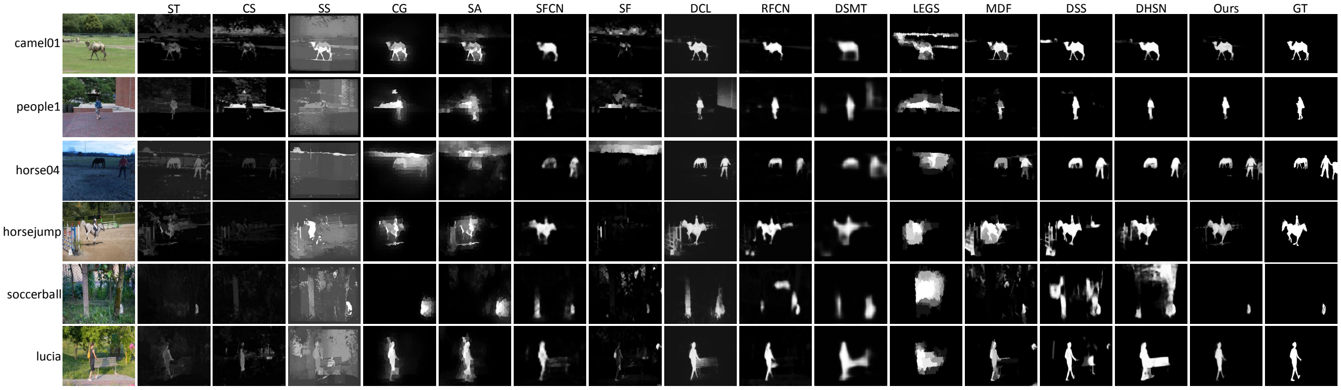

Figure 5: The saliency maps are generated from different saliency detection approaches. Top and bottom three rows are the video frames from FBMS and DAVIS datasets, respectively

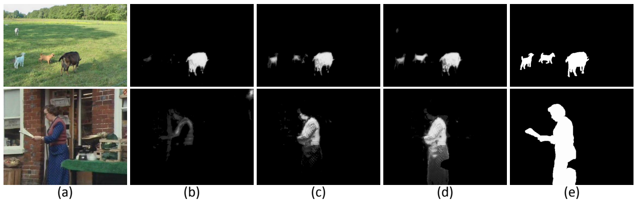

Figure 7: Saliency maps generated by networks with different architectures. (a) Input video frames. (b) Results of only spatial network (c) Results of spatiotemporal network. (d) Results of MSST-ConvLSTM. (e) Ground truths.

As shown in Figure.6, the proposed MSST-ConvLSTM model achieves the best performance in both PR curve and MAE criteria. At the beginning, the network only uses the single frame as training samples, which leads to the failure detection of moving objects (e.g., Figure.7 (b)). With the employment of temporal sub-network, the moving objects are obviously highlighted (e.g., Figure.7 (c)). After the proposed MSST-ConvLSTM step, the saliency maps are closer to the ground truths (e.g., Figure.7 (d))

In order to validate the effectiveness of the generated coarse pixel-wise labels, we report the saliency performance of training the proposed network with and without coarse pixel-wise labels. As the Fig.8 shown, this experiment proves that the generated coarse pixel-wise labels have positive effect on saliency prediction. The training data increases from 2,772 frames to 6,098 frames. Besides, these training data contains many complex video scenes, which can make the network learn more robust salient features, thus improve the accuracy of the saliency detection.

5 CONCLUSION

In this paper, we propose a multi-scale spatiotemporal convolutional LSTM network (MSST-ConvLSTM) for video saliency detection. Our network consists of the complementary components, a spatial sub-network and a temporal sub-network. The network not only commendably retains the spatial features but also learns the temporal features. Moreover, the proposed MSST-ConvLSTM can further fuse the spatial and temporal cues to generate final spatiotemporal saliency maps. Meanwhile, to train the data-driven network, we produce an amount of coarse pixel-wise labels, which effectively improve robustness of the proposed network. Finally, the experimental results on the two datasets, FBMS and DAVIS, demonstrate that the proposed network can outperform the-state-of-the-arts in term of both PR curve and MAE criteria.

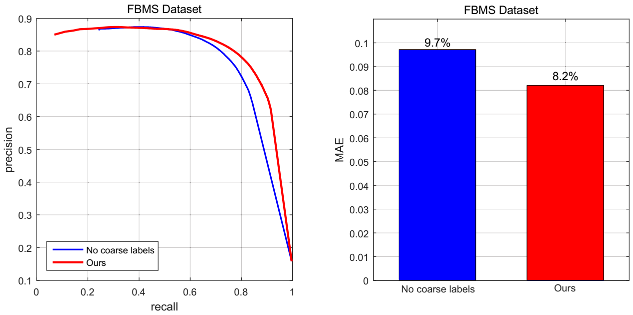

Figure 8: Comparison of PR curves and MAEs with and without coarse pixel-wise labels

6 ACKNOWLEDGEMENT

This work is supported by the NSFC Project under Grant 61771321, Grant 61701313, Grant 6171101606 and Grant 61472257, in part by the Natural Science Foundation of Shenzhen under Grant KQJSCX20170327151357330, Grant JCYJ20170818091621856, Grant JCYJ20160307154003475, Grant JCYJ20160506172651253 and Grant JCYJ20170302145906843, in part by the Interdisciplinary Innovation Team of Shenzhen University, and in part by the Natural Science Foundation of SZU under Grant 827-000152.

REFERENCES

[1] Thomas Brox and Jitendra Malik. 2010. Object segmentation by long term analysis of point trajectories. In *Proceedings of the European Conference on Computer Vision*. Springer, 282–295.

[2] Chenglizhao Chen, Shuai Li, Yongguang Wang, Hong Qin, and Aimin Hao. 2017. Video Saliency Detection via Spatial-Temporal Fusion and Low-Rank Coherency Diffusion. *IEEE Transactions on Image Processing* 26, 7 (2017), 3156–3170.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint:1412.7062* (2014).

[4] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2015), 569–582.

[5] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[6] Ionut C. Duta, Bogdan Ionescu, Kiyoharu Aizawa, and Nicu Sebe. 2017. Simple, Efficient and Effective Encodings of Local Deep Features for Video Action Recognition. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 218–225.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. 2004. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision* 59, 2 (2004), 167–181.

[8] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. 2013. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing* 22, 10 (2013), 3766–3778.

[9] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. 2008. The discriminant center-surround hypothesis for bottom-up saliency. In *Proceedings of the Advances in Neural Information Processing Systems*. 497–504.

[10] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Zhuowen Tu, and A Borji. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[11] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. 2013. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2083–2090.

[12] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. 2015. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing* 24, 8 (2015), 2552–2564.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

[14] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. 2013. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2192–2199.

[15] Guanbin Li and Yizhou Yu. 2016. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 478–487.

[16] Guanbin Li and Yizhou Yu. 2016. Visual saliency detection based on multiscale deep CNN features. *IEEE Transactions on Image Processing* 25, 11 (2016), 5012–5024.

[17] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang. 2016. DeepSaliency: Multi-Task deep neural network model for salient object detection. *IEEE Transactions on Image Processing* 25, 8 (2016), 3919–3930.

[18] Nian Liu and Junwei Han. 2016. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 678–686.

[19] Wu Liu, Cheng Zhang, Huadong Ma, and Shuangqun Li. 2018. Learning Efficient Spatial-Temporal Gait Features with Deep Learning for Human Identification. *Neuroinformatics* (2018), 1–15.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3431–3440.

[21] Anestis Papazoglou and Vittorio Ferrari. 2013. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1777–1784.

[22] Yuxin Peng, Yunzhen Zhao, and Junchao Zhang. 2017. Two-stream Collaborative Learning with Spatial-Temporal Attention for Video Classification. *arXiv preprint arXiv:1711.03273* (2017).

[23] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 733–740.

[24] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 724–732.

[25] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. 2010. Segmenting salient objects from images and videos. In *Proceedings of the European Conference on Computer Vision*. Springer, 366–379.

[26] Hae Jong Seo and Peyman Milanfar. 2009. Static and space-time visual saliency detection by self-resemblance. *Journal of vision* 9, 12 (2009), 15–15.

[27] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2015. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3183–3192.

[28] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. 2016. Saliency detection with recurrent fully convolutional networks. In *Proceedings of the European Conference on Computer Vision*. Springer, 825–841.

[29] Wenguan Wang, Jianbing Shen, and Fatih Porikli. 2015. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3395–3402.

[30] Wenguan Wang, Jianbing Shen, and Ling Shao. 2015. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing* 24, 11 (2015), 4185–4196.

[31] Wenguan Wang, Jianbing Shen, and Ling Shao. 2018. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing* 27, 1 (2018), 38–49.

[32] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. 2012. Geodesic saliency using background priors. In *Proceedings of the European Conference on Computer Vision*. Springer, 29–42.

[33] Dayan Wu, Zheng Lin, Bo Li, Mingzhen Ye, and Weiping Wang. 2017. Deep Supervised Hashing for Multi-Label and Large-Scale Image Retrieval. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 150–158.

[34] Yun Zhai and Mubarak Shah. 2006. Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the ACM international conference on Multimedia*. ACM, 815–824.

[35] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1265–1274.

[36] Yunzhen Zhao and Yuxin Peng. 2017. Saliency-guided video classification via adaptively weighted learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, 847–852.

[37] Feng Zhou, Sing Bing Kang, and Michael F Cohen. 2014. Time-mapping using space-time saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3358–3365.

[38] Wenbin Zou and Nikos Komodakis. 2015. HARF: Hierarchy-Associated Rich Features for Salient Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 406–414.

[39] Wenbin Zou, Kidiyo Kpalma, Zhi Liu, and Joseph Ronsin. 2013. Segmentation driven low-rank matrix recovery for saliency detection. In *Proceedings of the British Machine Vision on Conference*. 1–13.

[40] Wenbin Zou, Zhi Liu, Kidiyo Kpalma, Joseph Ronsin, Yong. Zhao, and Nikos Komodakis. 2015. Unsupervised Joint Salient Region Detection and Object Segmentation. *IEEE Transactions on Image Processing* 24, 11 (2015), 3858–3873.