Two-Stream Recurrent Convolutional Neural Networks for Video Saliency Estimation

Xiao Wei[1,2], Li Song[1,2], Rong Xie[1,2], Wenjun Zhang[1,2]
[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2] Cooperative Medianet Innovation Center,
Shanghai, China
{smile_wx94, song_li, xierong, zhangwenjun}@sjtu.edu.cn

***Abstract*—Recently, research has emphasized the need for video saliency estimation since its application covers a large domain. Traditional saliency prediction methods for video based on hand-crafted visual features lead to slow speed and ineffective results. In this paper, we propose a real-time end-to-end saliency estimation model combining two-stream convolutional neural networks from global-view to local-view. In global view, the temporal stream CNN extracts the inter-frame features from optical flow map, and spatial stream CNN extracts the intra-frame information. In local view, we adopt the recurrent connnections to refine the local details through correcting the saliency map step by step. We test our model TSRCNN on three datasets in video saliency estimation, and it shows not only exceedingly commendable performance but almostly real-time GPU processing time of 0.088s compared to other state-of-art methods.**

Keywords—Saliency estimation; Video processing; Optical flow; CNN; Recurrent connections

I. INTRODUCTION

Recently, the late-model convolutional neural networks and feasible online datasets provide saliency detection with rapid development. Under the circumstances, several directions in computer vision can incorporate saliency models, such as semantic segmentation[1], object detection[2], object proposals[4], image clustering and retrieval[5], cognitive saliency applications such as image captioning and high-level image understanding[3], video compression[6] and summarization[7]. Broadly speaking, visual saliency approaches can be roughly divided into two information-processing categories: top-down mechanisms and bottom-up mechanisms. According to the application, it can be further divided into image saliency and video saliency estimation. While estimating saliency map in dynamic scenes constitutes a great challenge for scholars and researchers compared to handling the same task in still images.

Traditional saliency models [8], [9], [10], [11], [12], [13] usually take the spatial characteristics into account, such as intensity,color and orientation. In addition, some researchers observe human's eye fixtion to track the interesting area and use human eye-gaze or annotations as ground truth for training. For video saliency estimation [14], [15], [16] it need to focus on motion cues, since the pixels around the optical flow field change abruptly drawing more attention by people.

There is an obvious tendency in applying deep learning to still image saliency detection. These methods [14], [15], [16] employ neural networks for feature extraction and outerperform contrast to traditional methods in most of the benchmark datasets. In this paper, our contributions are as follows. First, inspired by the success of deep learning methods, we propose a novel two-stream architechture which combines convolutional neural networks to extract the effective features. Second, we design this framework from global estimation to local refinement by uniting the recurrent connections. Finally, due to the whole end-to-end framework, we achieve almostly real-time processing speed satisfing the requirement in video application.

II. RELATED WORKS

The present studies for attention model can be divided into two aspects along application scenarios: image saliency and video saliency, since our works are derived from static image methods and improved to video sequences. Image saliency estimation focuses on salient area detection and eye fixation prediction. More specifically, a bottom-up visual saliency model called Graph-Based Visual Saliency was proposed by Schölkopf et al. [8], consisted of two steps: first forming activation maps on certain feature channels, and then normalizing them. Hou and Zhang[9] extracted the spectral residual of an image in spectral domain and construct the corresponding saliency map in spatial domain. Yang et al.[10] considered a two-stage scheme with the background and foreground queries for ranking, which incoprated local grouping cues and boundary priors. Since in ImageNet 2012 deep learning methods outperformed traditional object recognition algorithms, a large number of scholars began to use neural networks for saliency detection. Zhao et al. [11] introduced a multi-context deep learning method for saliency detection to solve problems in low-contrast background with confusing visual appearance, which took both global and local context into account. Furthermore, Wang et al. [12] integrated local estimation and global search for saliency detection by using two deep neural networks, to learn local patch features and predict the saliency score of each object region based on

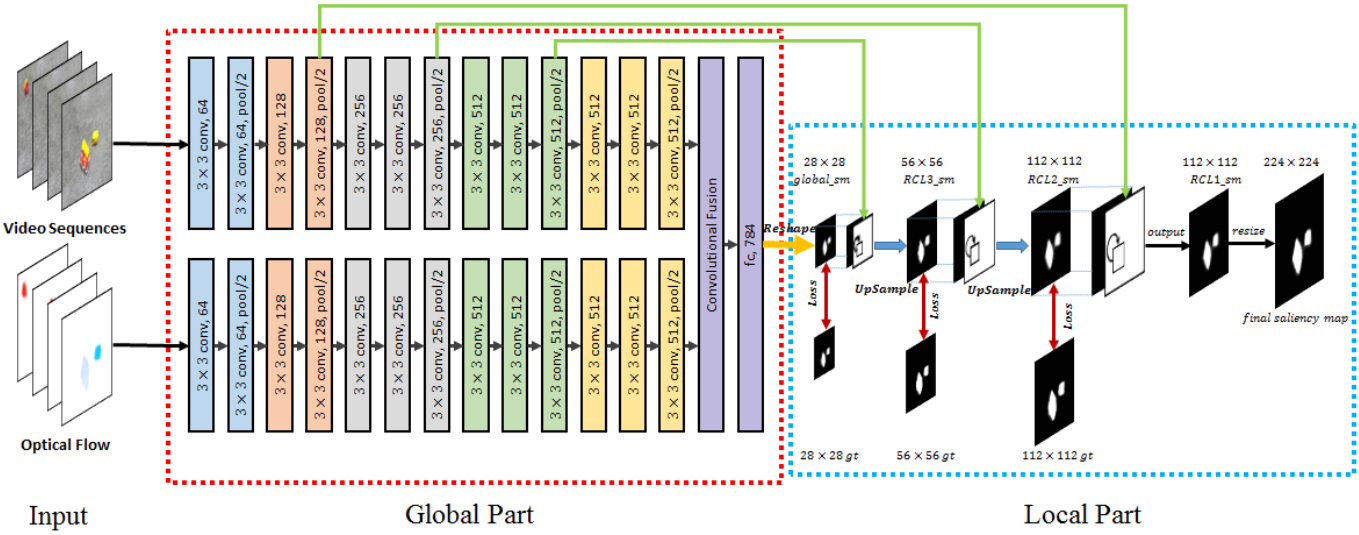


Fig. 1. The framework of our proposed method. Video sequences and optical flow maps are input to global part network simultaneously to produce a coarse saliency map. The local part is built to refine the saliency step by step integrating the recurrent structure.

the global features. Li and Yu[13] discovered a high-quality visual saliency model to learn multiscale features using deep convolutional neural networks and then incoprated a refinement method to enhance the spatial coherence. However, image saliency methods ignore the change of object in scale and missed the motion information, which drives the application in video.

Video saliency detection aims to extract the region attracting people's attention in video sequences. Guo et al. [14] proposed the Phase spectrum of Quaternion Fourier Transform to obtain spatial-temporal saliency map. In [15] Rahtu et al. combined a saliency measure with conditional random field model to segment salient object from images and videos. Considering the consistency of saliency maps, Wang et al. [16] estimated salient regions based on gradient flow field and energy optimization. Nevertheless, these approaches processes the frame sequences frame-by-frame, ignoring the issue of real-time in video processing.

Despite several proposed methods, traditional methods are still inferior to deep learning techniques in speed and accuracy. The amount of video saliency datasets for training neural networks is too poor for training. In this paper, we introduce a novel saliency model learning features in global-view and local-view, with a two-stream recurrent convolutional neural network and few available datasets for training.

III. GLOBAL PART FOR COARSE EXTRACTION

Several experiments have proven that VGGnet performs well in image feature extraction, which is the reason we choose this architecture in this paper. In top stream, we use VGGnet to extract the feature of the current frame. The difference between video and image is that video has inter-frame information which is important to video tasks. We consider to introduce inter-frame information using another CNN stream. We use down stream to extract inter-frame information from optical flows of the current frame. Therefore, we propose two-stream network to extract the inter-frame temporal saliency and intra-frame spatial saliency.

This innovative composition can learn motion information from optical flow map and appearance information through every frame simultaneously. As shown in Figure 1, the top architecture consists of 13 convolutional layers, derived from pretrained VGG 16-layer network [23]. The bottom architecture is similar to the top one. We use the eltwise layer to combine the motion and appearance features extracted from first two stream CNN. The following fully-connected layer and reshape layer are built to generate and reshape corase saliency map for furthur refinement.

Hence, we can describe the whole framework by this procedure: 1) input one frame from video sequences and its corresponding optical flow generated from the current frame and next frame. 2) after feedforwarding frame and optical flow through the top and down VGG nets we can get the corresponding motion and appearance information; 3)Then we use fully-connected and reshape layer to generate and reshape a coarse saliency map of current frame. The coarse saliency map is used as input for the next local part.

IV. LOCAL PART FOR DETAILED CONTRAST

The global part framework focuses on global information to extract the corase salient map. To hierarchically and iteratively refine the salient map details, we incorporate recurrent convolutional layer proposed by [19]. Illustrated in Figure 2, the RCL connections can be unfolded detailedly. The input feature maps can be explained to a 3-dimension structure. Here 3-dimension indicates the width, height and depth of feature maps. Therefore, the unit located in (i,j) of k-th feature maps can be represented as:

$$x_{ijk}(t)=g(f(z_{ijk}(t))) \quad (1)$$

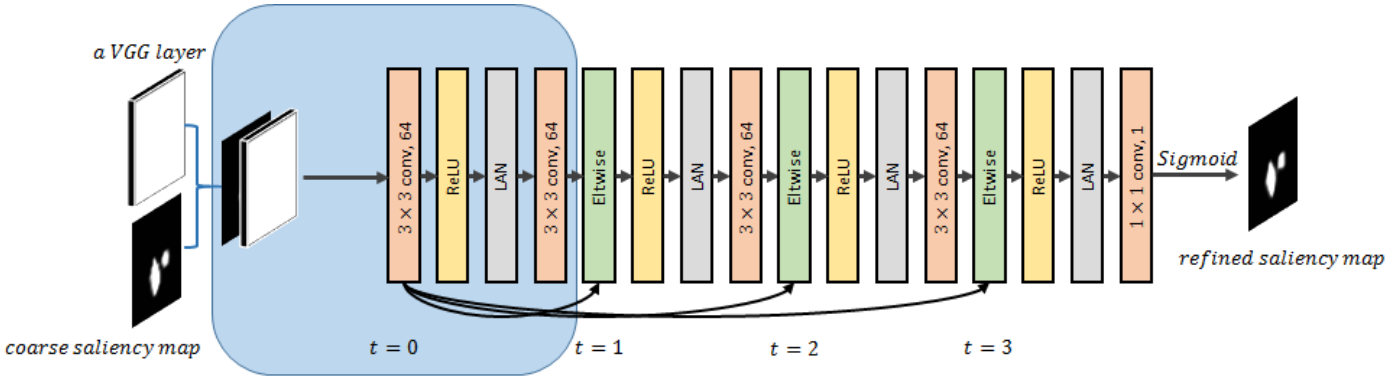


Fig. 2. The extended architechiture of RCL. A VGG layer and coarse saliency map generated from last structure are concatenated and then input to recurrent connections. Finally the refined saliency map is created through a sigmoid layer.

In Eq. (1), $f(\cdot)$ is the Rectified Linear Unit activation function and g(·) is the Local Response Normalization function, $z_{ijk}(t)$ is the input of the unit, which connects the initial feedforward input and last recurrent output. Eq. (2) can explain the architechiture:

$$z_{ijk}(t)=(w_k^f)^T u^{(i,j)}+(w_k^r)^T x^{(i,j)}(t-1)+b_k \quad (2)$$

Here w_k^f and w_k^r defines the feedforward weights and the recurrent weights, t represents the current step time, $u^{(i,j)}$ and $x^{(i,j)}(t-1)$ indicate the feedforward input and last step time recurrent input, b_k is bias value.

We also combine a convolutional layer from the VGG net with a coarse saliency map to generate a finer saliency map. Only convolutional layers in top stream are used for that feature extracted from frame(top stream) is much finer than that extracted from optical flows(down stream).

In this paper, we set time steps to 3 in consideration of the computation complexity, which leads to the depth as 4, and three RCL connections are designed for optimization.

V. EXPERIMENTS

A. Datasets

We test our model on three widly used video segmentation datasets : MCL Dataset[20], SegTrack We v2 Dataset[21], and NTT Dataset[22]Video sequences in MCL dataset have the resolution of and consist of around 800 frames. The binary ground-truth maps have been manually obtained for every 8 frame. SegTrack v2 is a video segmentation dataset with full pixel-level annotations on 14 different kinds of objects at each frame within each video. NTT dataset contains 10 uncompressed AVI clips of natural scenes with 12 fps, including at least one target objects or something others. Length varies 5-10 seconds. Note that Segmented images are provided for almost all the frames exculding first 15 frames.

B. Preprocessing

We wrap video frames into size 224×224 to satisfy the input size of VGGnet. Then we generate the approximate binary image for conforming ground-truth. Optical flow maps are extracted using FlowNet[17] frame by frame.

Since the total number of frames in thest datasets are around 2000, data augmentation is conducted for avoiding over-fitting during training stage. We flip the frame sequences horizontally and crop the top, left, right, bottom and marginal 1/9 portion, which increases the amount of training data 12 times.

C. Implementation details

We randomly select 2000 video frames for training from total 2500 frames,and the rest is used for testing. The RGB mean value is subtracted before to reduce computational complexity. We set the initial learning rate to 0.01 and halve the learning rate every 3000 iterations. The maximum iteration is set to 100000. Besides, we set the momentum and the weight decay factor to 0.9 and 0.0005. It takes about two days for the whole training process. We train and test our TSRCNN using caffe [18] toolbox and Matlab. Two Tesla K20c GPU are used both in training and testing for acceleration.

VI. RESULTS

We compare the performance of TSRCNN with several state-of-art methods including GBVS[8], LEGS[12], MCDL[11], DHS[24], SAG[25], LGFO[16]. We present the extracted saliency maps using different methods in Figure 3. Our network outperforms other algorithms obviously in multi-object and multi-scale. Compared with video methods [8], [16], [25], our results display the outline more clearly and are fitted more acurrately to the ground truth. The image methods reveal poor performance in scale changing. In column 5 and 6 the original frames have similar foreground and background, the experiments verify the robustness and effectiveness of our algorithm.

For quantitative evaluation, the PR curves and F-measure scores are plotted in Figure 4. The precision-recall curves obviously show that our method achieves both high precision and recall. From the right row in Figure 4, TSRCNN wins the highest scores in precision, recall and F-measure. Results

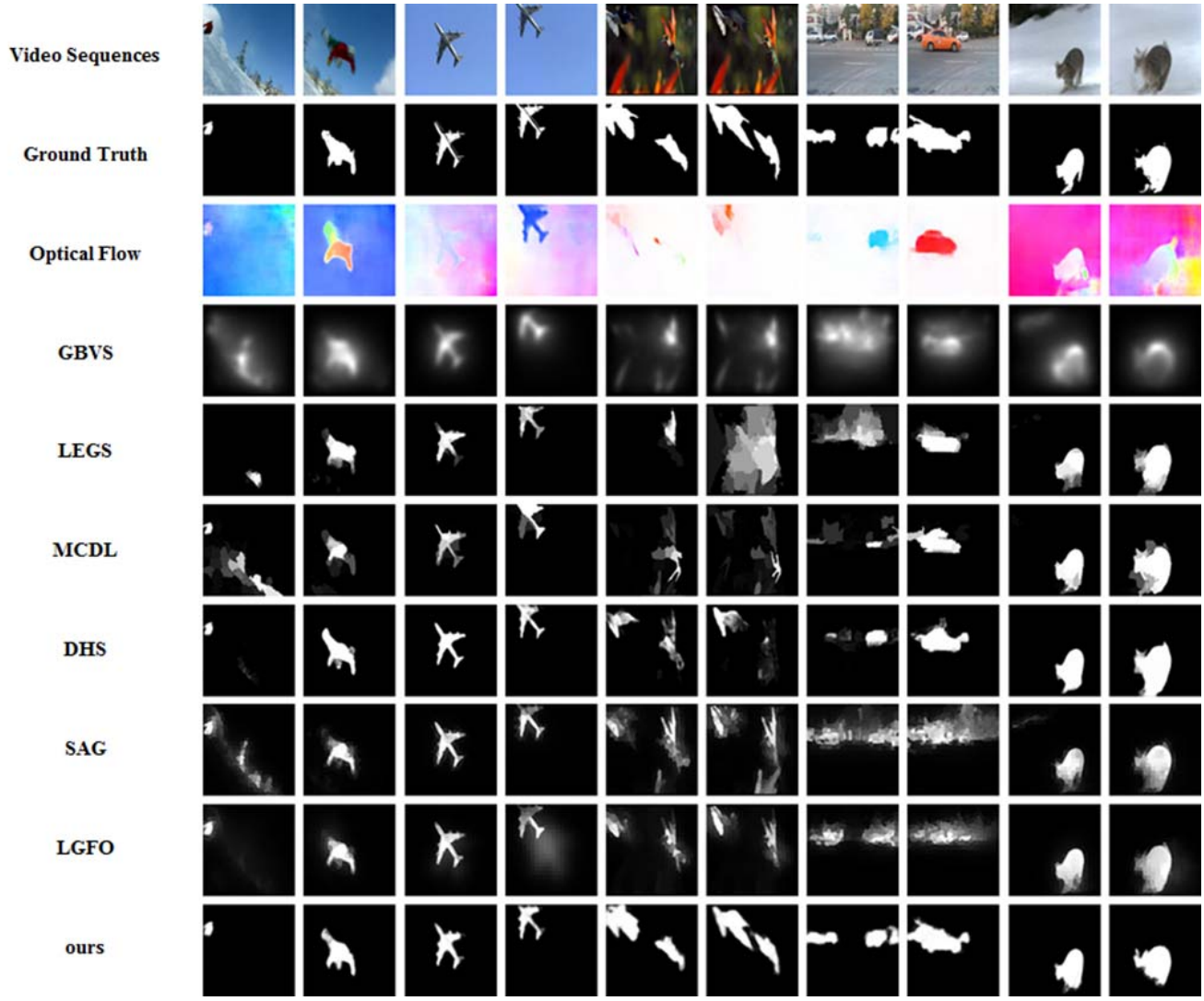
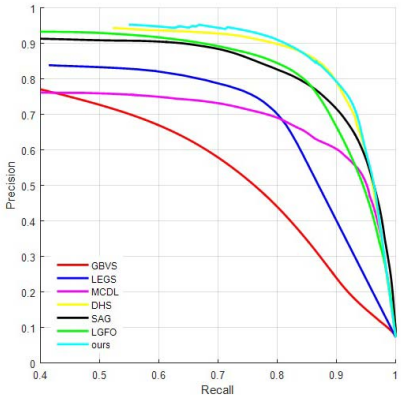
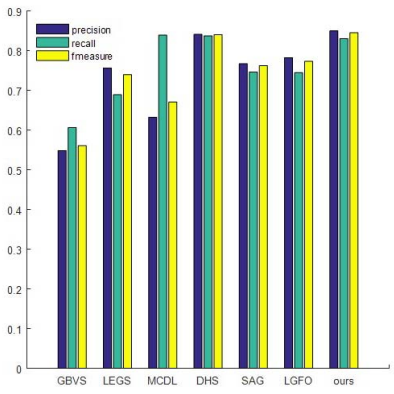


Fig. 3. The comparisons between different methods. Top three rows: input video sequences, ground truth and corresponding input optical flow map. Bottom six rows: the results of seven approaches, GBVS[8], LEGS[12], MCDL[11], DHS[24], SAG[25], LGFO[16].

indicate our network has better performance than other state-of-art methods.We also evaluate the runtimes and amounts of parameters of each method in Table 1. Although parameters of TSRCNN is larger than other deep learning models, our algorithm beats other competitors in terms of performance and speed.



(a)



(b)

Fig. 4. Quantitative evaluation: the (a) is PR curves of 7 different methods, the (b) is the corresponding F-measure scores and average precision and recall.

TABLE I. The runtime and amount of parameters of each method.

	GBVS	LEGS	MCDL	DHS	SAG	LGFO	ours
Runtime(s)	0.882	2.759	1.435	0.082	0.117	0.275	**0.088**
Parameters(MB)	*	73.6	233.1	376.2	*	*	**434.4**

VII. CONCLUSION

We have presented a two-stream architecture incorporating recurrent connections for video saliency estimation from global to local view. In the global part, the convolutional neural network is trained end-to-end to generate corase saliency map. In the local part, recurrent connections are applied to convolutional layer to refine the details of corase saliency maps.

By combining global part and local part, our method can effectively and accurately detect the salient region in videos. The experiment results demonstrate our method outperforms the state-of-art on several saliency detection dataset.

ACKNOWLEDGMENT

This work was supported by NSFC (61521062, 61671296 and U1611461),the 111 Project (B07022 and Sheitc No.150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

REFERENCES

[1] Y. Wei, X. Liang, Y. Chen, X. Shen, M. M. Cheng, J. Feng, Y. Zhao, S. Yan, "STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence , vol. PP, no. 99, pp. 1-1.

[2] W. Qi, M. M. Cheng, A. Borji, H. Lu, L. F. Bai, "SaliencyRank: Two-stage Manifold Ranking for Salient Object Detection", Computational Visual Media, vol. 1, no. 4, pp. 309-320, Dec. 2015.

[3] C. Shen, X. Huang and Q. Zhao, "Predicting Eye Fixations on Webpage With an Ensemble of Early Features and High-Level Representations from Deep Network," in IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2084-2093, Nov. 2015.

[4] A. Borji, M. M. Cheng, H. Jiang and J. Li, "Salient Object Detection: A Benchmark," in IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5706-5722, Dec. 2015.

[5] M. M. Cheng, N. J. Mitra, X. Huang, S. M. Hu, "SalientShape: Group saliency in image collections", The Visual Computer, vol. 30, no. 4, pp. 443-453, 2014.

[6] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," in IEEE Transactions on Image Processing, vol. 19, no. 1, pp. 185-198, Jan. 2010.

[7] W. H. Cheng, C. W. Wang and J. L. Wu, "Video Adaptation for Small Display Based on Content Recomposition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 1, pp. 43-58, Jan. 2007.

[8] B. Schölkopf, J. Platt and T. Hofmann, "Graph-Based Visual Saliency," in Advances in Neural Information Processing Systems, pp. 545-552, 2007.

[9] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[10] C. Yang, L. Zhang, H. Lu, X. Ruan and M. H. Yang, "Saliency Detection via Graph-Based Manifold Ranking," 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166-3173, 2013.

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

[12] L. Wang, H. Lu, X. Ruan and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3183-3192, 2015.

[13] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5455-5463, 2015.

[14] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform," 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2008.

[15] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," European Conference on Computer Vision, pp. 366--379, 2010.

[16] W. Wang, J. Shen and L. Shao, "Consistent Video Saliency Using Local Gradient Flow Optimization and Global Refinement," in IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 4185-4196, Nov. 2015.

[17] P. Fischer, A. Dosovitskiy, E. Ilg, P. Hausser, C. Hazirbas, and V. Golkov, "FlowNet: Learning Optical Flow with Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2758-2766, 2015.

[18] 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 ACM Multimedia, 2014.

[19] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3367-3375, 2015.

[20] S. H. Lee, J. H. Kim, K. P. Choi, J. Y. Sim and C. S. Kim, "Video saliency detection based on spatiotemporal feature learning," 2014 IEEE International Conference on Image Processing (ICIP), pp. 1120-1124, 2014.

[21] F. Li, T. Kim, A. Humayun, D. Tsai and J. M. Rehg, "Video Segmentation by Tracking Many Figure-Ground Segments," 2013 IEEE International Conference on Computer Vision, Sydney, pp. 2192-2199, 2013.

[22] K. Akamine, K. Fukuchi, A. Kimura and S. Takagi, “Fully automatic extraction of salient objects from videos in near real time,” The Computer Journal, vol. 55, no. 1, pp. 3-14, Jan. 2012.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," the International Conference on Learning Representations, 2015.

[24] N. Liu and J. Han, "DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 678-686, 2016.

[25] W. Wang, J. Shen and F. Porikli, "Saliency-aware geodesic video object segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3395-3402, 2015.