Learning to Explore Intrinsic Saliency for Stereoscopic Video

Qiudan Zhang[1,2] Xu Wang[1*] Shiqi Wang[2] Shikai Li[1] Sam Kwong[2] Jianmin Jiang[1]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Department of Computer Science, City University of Hong Kong

Abstract

The human visual system excels at biasing the stereoscopic visual signals by the attention mechanisms. Traditional methods relying on the low-level features and depth relevant information for stereoscopic video saliency prediction have fundamental limitations. For example, it is cumbersome to model the interactions between multiple visual cues including spatial, temporal, and depth information as a result of the sophistication. In this paper, we argue that the high-level features are crucial and resort to the deep learning framework to learn the saliency map of stereoscopic videos. Driven by spatio-temporal coherence from consecutive frames, the model first imitates the mechanism of saliency by taking advantage of the 3D convolutional neural network. Subsequently, the saliency originated from the intrinsic depth is derived based on the correlations between left and right views in a data-driven manner. Finally, a Convolutional Long Short-Term Memory (Conv-LSTM) based fusion network is developed to model the instantaneous interactions between spatio-temporal and depth attributes, such that the ultimate stereoscopic saliency maps over time are produced. Moreover, we establish a new large-scale stereoscopic video saliency dataset (SVS) including 175 stereoscopic video sequences and their fixation density annotations, aiming to comprehensively study the intrinsic attributes for stereoscopic video saliency detection. Extensive experiments show that our proposed model can achieve superior performance compared to the state-of-the-art methods on the newly built dataset for stereoscopic videos.

1. Introduction

In recent years, we have witnessed the strong growth of 3D content and fast development of 3D display technologies, such that the automatic prediction of saliency on stereoscopic videos has become ever important. Stereoscopic video saliency prediction (see Fig. 1), which attempts to distinguish salient regions or objects in diverse, dynamic and immersive scenes, is challenging yet rewarding. For example, it can serve as the perceptual preprocessing for numerous stereoscopic applications, such as stereoscopic video coding, quality assessment, medical image analysis and robot vision. However, the saliency prediction of stereoscopic videos is quite complex as non-intuitive interactions between the video pair and depth clues are involved, making it difficult to directly apply the traditional 2D saliency prediction algorithm.

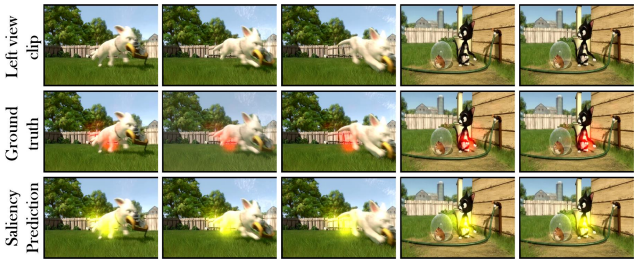

Figure 1. Examples of left view clip (including five consecutive frames), ground truth and their saliency results obtained with our model on the newly built SVS dataset.

With a variety of deep learning architectures available, significant advances have been achieved in static image saliency prediction due to the successful deployment of convolutional neural networks (ConvNet) [5, 15]. The features learned based on ConvNet are commendable for saliency inference, as they are capable of mining underlying cues and heuristic semantic priors to better interpret the image content compared to traditional hand-crafted features. Nevertheless, it is not feasible to infer dynamic video saliency by straightforwardly applying static image saliency models, especially for stereoscopic videos. In particular, since videos are composed of consecutive frames, temporal coherence is an important element in stereoscopic video saliency prediction. While recent studies [36, 9, 27] mostly adopt optical flow to explore saliency in temporal domain for dynamic scenes, it will induce non-salient features and lead to more time consumed during inferring. Moreover, depth serves as an important component for holistic stereoscopic video

*Corresponding author. wangxu@szu.edu.cn

perception, and in essence depth cue is another significant ingredient which can influence locations of fixation points during the 3D rendering.

A series of 3D saliency detection models have also been proposed by considering low-level features and depth relevant information. Fang *et al.* [10] proposed a visual attention model for stereoscopic video based on the Gestalt theory, in which feature contrast and motion contrast were calculated to estimate spatial and temporal saliency, respectively. However, the ignorance of high-level semantic features leads to the limited success of the above method. Moreover, how to determine the interactions between spatial, temporal and depth cues is also non-trivial and challenging for stereoscopic video saliency prediction.

In this paper, we propose a saliency prediction model for stereoscopic video by resorting to deep learning, which infers stereoscopic saliency by automatically exploring saliency-related features in terms of spatio-temporal coherence and intrinsic depth. The obtained saliency distributions on spatio-temporal and depth cues are ultimately combined based on a Conv-LSTM fusion network to produce final saliency maps. To facilitate the training and evaluation on the proposed method, a large dataset with 175 videos, in which multiple visual stimulus are contained, is built. Experimental results show that our proposed model has significantly outperformed the state-of-art saliency models over the created dataset. The contributions of this work are as follows,

- We propose a new deep learning based attention model for stereoscopic videos by singling out the contributions of spatio-temporal and depth cues, in an effort to explore the intrinsic stereoscopic video saliency with a data-driven strategy. The high-level semantic features are learned correspondingly, and the final attention model is developed based on pyramid spatio-temporal saliency prediction, intrinsic-depth saliency estimation and Conv-LSTM based fusion.
- We create a new challenging dataset, stereoscopic video saliency dataset (SVS), for the further research and evaluation towards stereoscopic video saliency estimation. This dataset contains both natural and synthetic scenes and will be made publicly available. Our proposed stereoscopic video saliency model has been validated using this new dataset, showing competitive performance.
- We carry out analyses to investigate the influence of image content, temporal characteristics and depth cues on the stereoscopic video saliency prediction. We believe these analyses are capable of providing useful insights to facilitate the future research of comprehensive attention models for stereoscopic video saliency estimation.

2. Related Work

In the literature, various biological characteristics inspired visual saliency computational models have been proposed for 2D or 3D images. Inspired by the behavior and neuronal architecture of the visual system of primates, the pioneer saliency model proposed by Itti and Koch *et al.* [17] calculated saliency map from multi-scale center-surround feature contrast based on the underlying features of an image (e.g., color, luminance, texture and orientation). Harel *et al.* [13] built a complete bottom-up saliency model by employing the dissimilarity measure based on graph theory to evaluate saliency across distinct feature activation maps. Goferman *et al.* [12] designed a context-aware saliency detection model based on four visual saliency principles, such that the significant regions of scenes can be detected to represent the saliency. Hou *et al.* [14] further proposed a spectral residual based visual saliency detection model that constructs saliency maps by the log-spectrum of image.

The key for further improving the performance of static visual saliency model is to extract meaningful features to capture the attention relevant information. The superiority of deep neural network in terms of feature extraction provides new opportunities, and several learning based static visual saliency detection models were proposed to locate human eye fixations. For example, Vig *et al.* [30] attempted to build a visual saliency detection model based upon ensembles of deep neural network. Later, Kümmerer *et al.* [22] developed a saliency model relying on extracted deep learning features. In [15], the gap between model prediction and human behavior was narrowed by a deep neural network (DNN) based saliency prediction method, which was achieved by fine-tuning the DNNs model with respect to an objective function based upon the saliency evaluation metrics and integrating information at different image scales. Subsequently, Cornia *et al.* [5] proposed a saliency attentive model for fixations prediction on natural images. In this study, an attentive Conv-LSTM model was designed to sequentially enhance saliency prediction.

For video saliency prediction, previous research works focus on exploiting the saliency relevant feature representations from spatial and temporal perspectives. For instance, Tu *et al.* [29] proposed a video saliency detection method in the compression domain on the basic of discrete cosine transformation (DCT) coefficients and motion information. In [29], H.264/AVC video bitstream was used to extract the corresponding information. Xu *et al.* [34] designed a learning based video saliency model that utilizes the support vector machine to fuse spatio-temporal features extracted by High Efficiency Video Coding (HEVC). Kalboussi *et al.* [20] introduced a video saliency model that integrates a static map and dynamic map by the Gestalt principle of figure-ground segregation. In that study, a dense optical flow was used to represent the motion information. In order

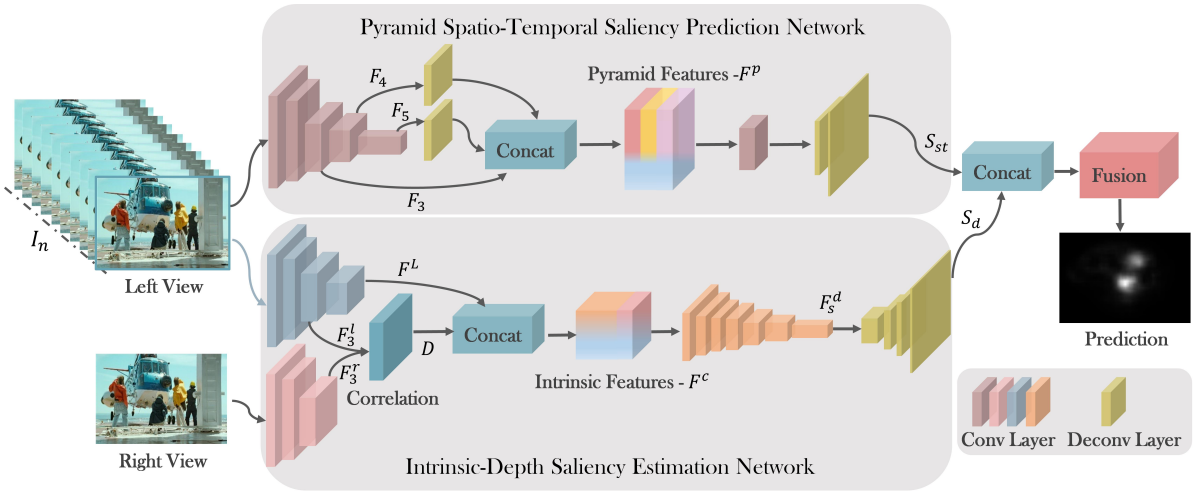

Figure 2. The overall architecture of the proposed stereoscopic video saliency prediction model.

to avoid the time-consuming optical flow calculation, Wang *et al.* [32] proposed a video saliency model based on fully convolutional networks that is capable of directly producing the spatio-temporal saliency inference by incorporating the spatial saliency estimation into the dynamic saliency model.

With the development of stereoscopic display technologies, it is necessary to study visual saliency in binocular domain. For example, Bruce *et al.* [2] proposed a stereo saliency model by expanding the existing 2D attention model to the binocular domain. Region-of-interest (ROI) extraction method was proposed for adaptive rendering as well [4]. Chamaret *et al.* used the disparity information to weight 2D saliency map for producing saliency map of stereoscopic image [4]. In addition, Potapova *et al.* [25] proposed a stereoscopic saliency detection model by integrating the top-down cues into the bottom-up saliency detection model. Eye tracking experiments were also performed on 2D and 3D images for depth saliency analysis in [23], in which stereoscopic saliency map was computed by extending the previous 2D attention models. More specifically, stereo saliency of an image region was computed based on the distance between its perceived location and the comfort zone. Fang *et al.* [8] proposed to calculate the contrast among the color, intensity, texture, and depth features to produce stereoscopic saliency maps. Zhang *et al.* [35] proposed a deep learning based visual saliency model for 3D image. In that study, color and depth features were extracted by a pretrained convolutional neural network model to infer saliency value of regions. Kim *et al.* [21] proposed a stereoscopic video saliency model that generates the final saliency map by involving low-level features, motion and depth attributes as well as the high level scene type.

In essence, automatically predicting the saliency for stereoscopic videos is a very challenging task, especially when considering the sophisticated interactions between multiple clues such as spatial, temporal and depth information. The traditional hand-crafted features adopted in the existing methods largely limit the accuracy of saliency prediction due to the absence of semantic information. In this paper, we single out the important contributions of stereoscopic cues and develop a learning based stereoscopic video saliency model based on 2D and 3D convolutional neural networks, leading to enhanced saliency oriented spatio-temporal and depth representations. A Conv-LSTM based fusion network is designed to finally produce ultimate saliency map by exploring the high-level feature representations, leading to superior performance compared to the state-of-the-art methods.

3. The Proposed Model

As shown in Fig. 2, the architecture of our proposed model consists of three modules, including pyramid spatio-temporal coherence based saliency prediction, intrinsic-depth saliency estimation and Conv-LSTM based fusion. More specifically, the pyramid spatio-temporal coherence based saliency prediction module first generates a series of temporally consistent saliency maps from consecutive frames based on 3D ConvNet. Subsequently, the intrinsic-depth saliency estimation module is applied to obtain the depth guided saliency map between left and right views based on the deep ConvNet. Finally, the acquired spatio-temporal and depth coherence saliency maps are consecutively fed into the Conv-LSTM based fusion network for the stereoscopic saliency inference.

3.1. Pyramid Spatio-Temporal Saliency Prediction

Herein, to construct the efficient saliency detection model for stereoscopic videos, we propose a pyramid

spatio-temporal saliency prediction network termed as PySTSP-Net to exploit the spatio-temporal intrinsic coherence. The overall architecture is shown in Fig. 2. In particular, our proposed PySTSP-Net module is based on the 3D ConvNet [28] that enhances the saliency related feature representations among multiple contiguous frames from both spatial and temporal perspectives, composing of the input layer, feature encoding and decoding layers evolving from 3D residual blocks. The feature encoding layers consist of one 3D convolution block with the kernel size $7 \times 7 \times 3$ and four 3D residual blocks, whereas the feature decoding layers are composed of two 3D deconvolution blocks. Meanwhile, the kernel size of four 3D residual blocks is $3 \times 3 \times 3$, as suggested in [26] that stacked smaller kernels can achieve better classification performance compared with larger kernels. Each convolution layer is followed with the batch normalization layer and scale layer to speed up the training convergence process [16].

Compare to the traditional 2D convolutional neural network for spatial feature extraction, the distinct difference of 3D convolutional neural network is to add an extra temporal dimension in both convolution kernel and input tensor. In this paper, we pack n consecutive frames ($I_n = \{i_1, ..., i_n\}$) as the input of PySTSP-Net, which represents the frame index, height and width of the video frame and the number of channels (denoted as $t \times h \times w \times c$). The extra temporal dimension in the 3D convolution kernel enables to capture temporally consistent visual representations from a segment of n consecutive frames rather than a single video frame. As such, the j-th feature map in the i-th 3D convolution layer is given by

$$F_{i,j} = \sigma\left(BN_{(\gamma,\beta)}\left(\sum_m w_{i,j,m} * f_{(i-1),m} + b_{i,j}\right)\right), \tag{1}$$

where $w_{i,j,m}$ and $b_{i,j}$ denote the weight and bias of 3D convolution kernel connected to the m-th feature map in the previous convolution layer, and $BN_{\gamma,\beta}$ indicates the batch normalization [16] with the trainable parameters γ and β. σ represents the nonlinear activation layer.

According to the cognitive research [33], human visual system follows the coarse to fine strategy in terms of the scale when viewing natural scenes. After resizing the features to the same scale by the 3D deconvolution operation, we cascade the output features of the third, fourth and fifth 3D residual blocks to construct multi-scale pyramid features in the PySTSP-Net module,

$$F^p = [F_3, F_4 \uparrow_2, F_5 \uparrow_4], \tag{2}$$

where F_3, F_4 and F_5 represent output features of the third, fourth and fifth 3D residual blocks, respectively. Moreover, $\uparrow_x$ refers to the x times up-sampling operation and F^p denotes the multi-scale pyramid features. $[\cdot]$ indicates the concatenation operation.

Finally, the multi-scale pyramid features are further fed into feature decoding layers that consist of two 3D deconvolution blocks, in which the spatio-temporal coherence guided saliency map can be accurately predicted by evaluating local, neighboring and background representations. The deconvolution layer has recently been adopted to reconstruct features into pixel space in image processing like semantic segmentation [24], and meanwhile it also plays an important role in upsampling the obtained features to the original image size. As such, we employ the 3D deconvolution layers to reconstruct saliency maps by altering the stride in spatial and temporal dimensions. The relevance between parameters of 3D deconvolution layer and upsampling factor k is defined as follows,

$$k_s = k_d = k * 2 - k\%2, s_s = t_s = k, \tag{3}$$

$$p = (k-1)/2, \tag{4}$$

where k_s and k_d are the kernel size and depth. The parameters s_s and t_s refer to spatial stride and temporal stride in 3D deconvolution layer, and p denotes additional padding added to the feature maps. Finally, the reconstructed n continuous saliency maps $S_{st} = \{S^1_{st}, \cdots, S^n_{st}\}$ can be obtained after the two 3D deconvolution blocks.

3.2. Saliency Estimation Based on Intrinsic-Depth

The binocular depth is an important cue in determining the saliency of stereoscopic videos. However, inferring precise depth information based on stereoscopic pairs is inherently a non-trivial task. In this work, we design an intrinsic-depth saliency estimation network (IDSE-Net), which serves to automatically explore the depth-oriented saliency between left and right views from the stereoscopic video. As shown in Fig. 2, our proposed IDSE-Net consists of 2D convolution based feature extraction, displacement correlation calculation between left and right views, and depth-oriented saliency reconstruction modules.

The IDSE-Net in Fig. 2 first produces meaningful features of left and right views separately by three convolution blocks, each of which consists of one convolution layer followed by a relu activation layer. The feature maps in the i-th convolution layer is given by

$$F^l_i = \max(0, w_i * f^l_{(i-1)} + b_i), \tag{5}$$

$$F^r_i = \max(0, w_i * f^r_{(i-1)} + b_i), \tag{6}$$

where F^l_i and F^r_i represent the feature maps of left and right views. The parameters w_i and b_i denote the weight vector and bias of the convolution kernel.

The displacement between left and right views is evaluated based on two feature vectors F^l_i and F^r_i with a specifically designed correlation layer. This resembles the stereo

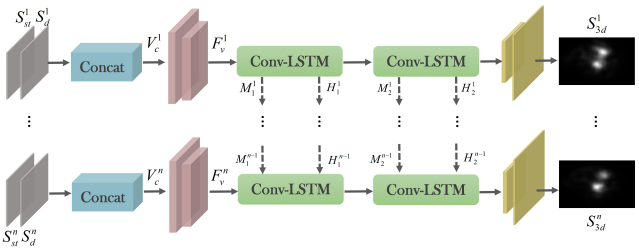

Figure 3. Illustration of the Conv-LSTM based fusion network.

matching which aims to effectively identify the corresponding pixel pairs. The introduced correlation layer is inspired by [6], aiming to discover displacement D by performing multiplicative path comparisons between two feature maps,

$$D = c\left(F_3^l, F_3^r\right), \tag{7}$$

where $c(\cdot)$ indicates the correlation operation [6]. Moreover, the spatial features F^L of the left view are further concatenated with D, and the concatenated feature vector F^c is then fed into a serials of convolution layers with the kernel size 3×3. As such, the depth-oriented saliency features F_s^d can be obtained. This process can be formulated as follows,

$$F^c = [F^L, D], \tag{8}$$

$$F_s^d = \max(0, w_i * F_{i-1}^c + b_i), \tag{9}$$

where F^c denotes the concatenated spatial and depth features. Finally, the depth-oriented saliency map S_d can be reconstructed by the deconvolution layers.

3.3. Conv-LSTM Based Fusion

The obtained saliency maps from PySTSP-Net and IDSE-Net are characterized by the individual components in the stereoscopic video. However, the spatial, temporal and depth information can jointly determine the human eye fixation locations with varying degrees [10]. Hence, fusing the obtained saliency maps S_d and S_{st} is a crucial process to achieve the ultimate saliency prediction. Moreover, due to the dynamic transitions of attention across continuous video frames [19], considering the dynamic coherence when merging the obtained saliency maps S_d and S_{st} with the feature contrast is a meaningful exploration. As such, we develop a Conv-LSTM based fusion network with the target of learning to produce the final stereoscopic saliency maps of a video clip. The architecture is illustrated in Fig. 3. More specifically, the saliency maps S_d^n and S_{st}^n for n-th frame are fed into two convolution blocks after cascading into a vector V_c^n. Each of the convolution block is composed of a convolution layer with the kernel size 3×3, a batch normalization layer, a relu layer and a max pooling layer. The numbers of output channels are set to 256 and 128. Subsequently, the feature vector F_v^n is treated as the input of Conv-LSTM layers to generate the final saliency map. The long-short term correlations between the input feature vectors are acquired through the memory cells (M_1^{n-1}, M_2^{n-1}) and the hidden states (H_1^{n-1}, H_2^{n-1}) of the two Conv-LSTM layers at the last frame. The LSTM cells at n-th frame is given by,

$$\begin{aligned}
I_m^n &= \sigma(W_i^h * (H_m^{n-1} \circ Q_i^h) + W_i^f * (F_v^n \circ Q_i^f) + B_i), \\
F_m^n &= \sigma(W_f^h * (H_m^{n-1} \circ Q_f^h) + W_f^f * (F_v^n \circ Q_f^f) + B_f), \\
O_m^n &= \sigma(W_o^h * (H_m^{n-1} \circ Q_o^h) + W_o^f * (F_v^n \circ Q_o^f) + B_o), \\
G_m^n &= \tanh(W_g^h * (H_m^{n-1} \circ Q_g^h) + W_g^f * (F_v^n \circ Q_g^f) + B_g), \\
M_m^n &= F_m^n \circ M_m^{n-1} + I_m^n \circ G_m^n, \\
H_m^n &= O_m^n \circ \tanh(M_m^n),
\end{aligned} \tag{10}$$

where I_m^n, F_m^n and O_m^n denote the gate of input, forget and output for n-th frame at the m-th Conv-LSTM layer. G_m^n, M_m^n and H_m^n are the candidate memory, memory cell and hidden state, respectively. Moreover, $\{Q_i^h, Q_f^h, Q_o^h, Q_g^h\}$ and $\{Q_i^f, Q_f^f, Q_o^f, Q_g^f\}$ are two sets of random masks for the hidden states and input features before the convolution operation [19]. Consequently, the two deconvolution layers with kernel size 4×4 are employed to generate final stereoscopic saliency map S_{3d}^n for n-th frame by reconstructing the output hidden states of the last Conv-LSTM layer.

3.4. Implementation Details

The training of proposed whole framework is not in an end-to-end manner. PySTSPNet and IDSE-Net are separately trained using the eye fixation map as the groundtruth, such that the outputs of these two networks are desired to be saliency maps. The parameters of these two modules are fixed during the training of the Conv-LSTM based fusion network. To train the PySTSP-Net and IDSE-Net, we initialize the kernel parameters of convolution layers by employing the pretrained models in [28, 6], respectively. The inv policy [18] is adopted to control the learning rate while the initialized learning rate is set to be 0.01. We also utilize Adadelta gradient descent (AGD) with momentum 0.9 and a weight decay of 0.0005 to minimize L_1 loss between prediction and eye fixation density map during training process.

Regarding the Conv-LSTM based fusion, the kernel parameters of convolution layers are initialized with a truncated normal distribution. For the training stage, the initialized learning rate is set to 10^{-5}. The Xavier initializer is adopted to initialize the kernel parameters in each LSTM cell, while the memory cells and hidden states are initialized as zeros. The training model is constrained and updated by the minimization of the Kullback-Leibler (KL) divergence based loss function between the prediction and ground truth.

Figure 4. Illustration of complex scenes in our dataset. (a) Cluttered background; (b) Low contrast; (c) Multiple objects.

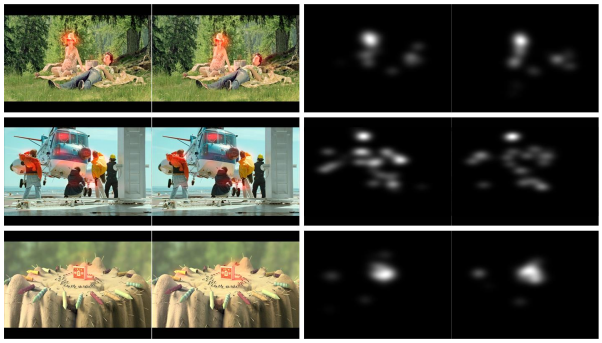

Figure 5. Fixation density annotations from left and right views.

4. Stereoscopic Video Eye Fixation Database

To our best knowledge, only a few eye fixation datasets for stereoscopic videos [11, 7] are public available, with in total 84 video sequences included. However, large-scale eye fixation database for stereoscopic video is critical for learning the meaningful features with diversity of video content. Moreover, the SAVAM dataset [11] only presented the left view to subjects, such that the acquired eye fixation data cannot be adopted to investigate the stereoscopic video saliency. In addition, the dataset in [7] did not provided the left and right gaze point maps that are important for the research of stereoscopic saliency prediction.

In this paper, we have constructed a new challenging dataset termed as the SVS dataset. The SVS dataset includes 175 stereoscopic video sequences with resolution 1920×1080, and each stereoscopic sequence is composed of left and right views. The videos in the dataset span a variety of real-world scenarios ranging from natural to synthetic scenes. We also select 77 video sequences from the dataset in [11] and [7]. During our data collection, the duration of the video is set randomly instead of strictly fixing identical duration for all video sequences. Moreover, the dataset also includes some exceptional circumstances to better reflect the real-world scenes, such as low contrast, multiple objects and cluttered background, and examples are shown in Fig. 4.

4.1. Procedures of Data Collection

Each stereoscopic video sequence was displayed on a 27-inch LG screen in 3D left-right pattern with the resolution of 1920×1080. The viewing distance is set to 63cm in our experiment. The Tobii Pro X3-120 is used to capture the eye gaze data and the sampling rate is 120 Hz. The frame rate of each stereoscopic video sequence is 25 fps. Subjects wearing a pair of passive polarized glasses were allowed to view the stereoscopic video sequences. Due to the free-view setting in our experiment, the subjects were allowed to freely move their heads such that normal human viewing behaviors could be better simulated.

All the stereoscopic video sequences were randomly divided into eight groups and presented in a random order for the subjective viewing. Such randomness also ensures that two similar videos will not play continuously to minimize the impact of dependencies. The calibration test was performed before playing each group of video sequences. The subjects were asked to readjust their watching position to guarantee that they can maintain the position stable while watching each group of videos. Each video sequence would provide subjects with a 3-seconds buffer time before playing. During the viewing test, subjects were asked for a 3-minutes break at the beginning of each video group to avoid the feeling of fatigue. There were 28 subjects participated in the experiment ranging from 18-25 years old. Subjects corrected to normal visual acuity were required to keep their glasses clean to ensure the accuracy of gaze data. They have also no experience about saliency prediction and are naive to the purpose of the experiment.

4.2. Data Processing and Outlier Removal

After collecting the eye gaze points for left and right views by the eye tracker, we create two gaze point maps for left and right views by using the coordinates of the eye fixation locations. Nevertheless, it is quite difficult for the saliency method to model the consistency of discrete gaze points. As such, we perform the Gaussian blur on the two gaze point maps to simulate the nonuniform distribution of the photoreceptors on the retina. Finally, we can obtain two fixation density maps as the ground truth for stereoscopic video saliency prediction. Some samples of the stereoscopic video frames and the corresponding fixation density maps are provided in Fig. 5.

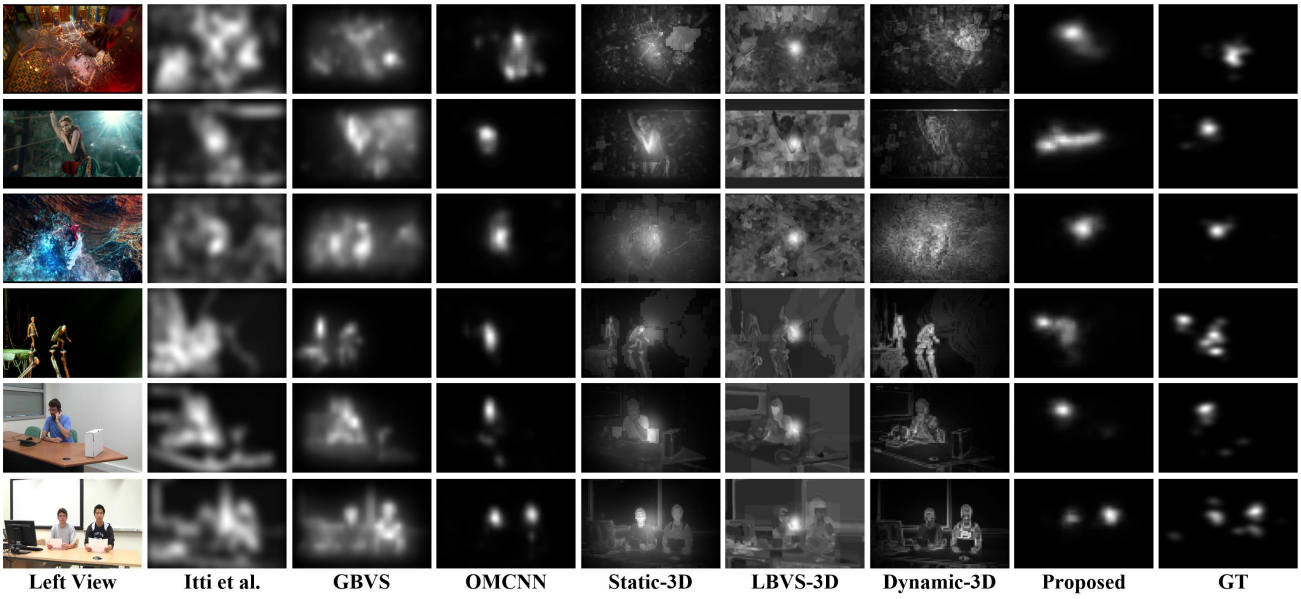

Figure 6. Comparisons of saliency maps generated from the seven different methods. The ground truth is shown in the last column.

5. Experimental Results

5.1. Training and Testing Datasets

Our proposed stereoscopic video saliency model is trained and validated on the newly built SVS dataset. In particular, the SVS dataset is randomly divided into training and testing sets according to the ratio of 9:1. For the PySTSP-Net which deals with one view only, the data from DHF1K [31] are combined with the training set of SVS to jointly improve the diversity of training data, such that 860 video sequences in total are used for training. All of these video sequences are segmented into clips with n consecutive frames, denoted as $I_n = \{i_1, ..., i_n\}$. The spatial resolution of video clips are downsampled into 112×112. We also allow overlaps as the data augmentation method for training. Moreover, extremely long or short duration of video clip may affect the exploration of spatio-temporal coherence. Empirically, the length of video clip is set to 16 in our experiment.

Moreover, the extracted training set including 158 video sequences is utilized to train the IDSE-Net. The left and right views are fed into the input layer to estimate the intrinsic depth. For the Conv-LSTM based fusion network, the outputs from PySTSP-Net and IDSE-Net modules are concatenated as the input. It is worth mentioning that the left eye fixation density map serves as the ground-truth during the training stage for all the three modules.

5.2. Comparisons with State-of-the-Art Methods

In this paper, five metrics are employed to measure the accuracy and similarity [3] of saliency detection models, including two variants of area under the ROC curve (AUC) (denoted as AUC_Jud, AUC_Borji), correlation coefficient (CC), similarity metric (SIM) and normalized scan-path saliency (NSS).

To validate the performance of the proposed visual saliency model on stereoscopic videos, we perform the comparisons between our proposed method with six existing state-of-the-art saliency detection models, including Itti *et al.*'s method [17], GBVS [13], OMCNN [19], Static-3D [8], LBVS-3D [1] and Dynamic-3D [10]. Among these models, Itti *et al.*'s method [17] and GBVS [13] are proposed for 2D static images. The method OMCNN [19] focuses on the 2D video. The Static-3D [8] is a saliency model towards 3D stereoscopic images. Besides, LBVS-3D [1] and Dynamic-3D [10] aim to predict saliency distributions for stereoscopic videos.

The comparison results on the SVS dataset are shown in Table 1. The strong competitiveness of our proposed model against the state-of-art saliency prediction approaches is clearly observed. Moreover, Static-3D [8] performs relatively better than traditional methods such as Itti *et al.*'s method [17] and GBVS [13]. This may be explained by the extra depth attribute utilized in Static-3D [8], which also provides useful evidence regarding the necessity of depth information for stereoscopic image and video. In addition, OMCNN [19] is a deep learning based spatio-temporal video saliency detection model, which obtains better performance than other benchmark models except for our proposed model. This might originate from the fact that other benchmark models all exploit hand-crafted features to encode image content without incorporating com-

Model	AUC_Jud ⇑	AUC_Borji ⇑	CC ⇑	SIM ⇑	NSS ⇑
Itti *et al.*'s method [17]	0.7592	0.7518	0.2454	0.2037	1.1173
GBVS[13]	0.8547	0.8268	0.3432	0.2640	1.5990
OMCNN[19]	0.9066	0.8244	0.5184	0.4068	2.6336
Static-3D [8]	0.8743	0.8632	0.3987	0.2568	1.8490
LBVS-3D [1]	0.7376	0.7248	0.2646	0.1784	1.2574
Dynamic-3D [10]	0.8334	0.8066	0.2987	0.2356	1.4004
Proposed-ST	0.8836	0.8088	0.6096	0.4922	3.2075
Proposed-Depth	0.8444	0.7903	0.4767	0.3704	2.2068
Proposed	**0.9201**	**0.8390**	**0.6339**	**0.5171**	**3.2320**

Table 1. Performance evaluations on the SVS dataset.

plex semantic features during the saliency inference. Employing the learning based motion information could also be another reason that explains the superior performance of OMCNN [19]. The results in terms of five evaluation criteria provided in Table 1 show that our proposed model achieves the highest performance, which demonstrates that our learning based saliency method is capable of producing saliency distributions for stereoscopic videos better than other benchmarks.

To better illustrate the advantages of our proposed model, we provide comparison results with the state-of-art saliency models in terms of saliency maps, which are depicted in Fig. 6. In particular, our proposed model accurately predicts the human eye fixation locations than other saliency models. We can also discover that traditional 2D static methods (Itti *et al.*'s method [17] and GBVS [13]) only detect the blurred outline information of salient objects. Besides, we can clearly see that some background information is mistaken as saliency in these two models. For the learning based 2D dynamic method OMCNN [19], the prediction accuracy is still lacking. This also verifies that depth information can affect human eye fixation locations. As shown in Fig. 6, it is obvious that the Static-3D [8], LBVS-3D [1] and Dynamic-3D [10] may treat the background information as salient. By contrast, our proposed model learns to combine spatio-temporal and intrinsic depth saliency distributions to produce more accurate saliency maps than other models.

5.3. Ablation Study

We perform ablation study to evaluate the relative impact of each component in our proposed model. In Table 1, we report five evaluation metrics to estimate these components, including only spatio-temporal component (denoted as Proposed-ST), only depth component (denoted as Proposed-Depth) and our proposed full version model. We can discover that the performance of Proposed-ST is better than Proposed-Depth. As such, color and motion information could play more important roles in predicting the stereoscopic video saliency. However, this does not imply that depth information is incompetent during the stereoscopic video saliency inference process. Compared with the performance of Proposed-ST and Proposed-Depth, our proposed full version model achieves better performance than the models contain only single component. This suggests that color, motion and depth information interact to influence the final prediction performance for stereoscopic video saliency.

6. Conclusion

In this paper, we propose a learning based visual attention model for stereoscopic videos by singling out intrinsic cues in terms of spatio-temporal, intrinsic-depth attributes, as well as the interaction between each other. In order to acquire the saliency from spatio-temporal perspective, we design a pyramid 3D ConvNet to investigate saliency distributions from spatial and temporal feature channels. Implicitly estimating depth indication between left and right views also enables our proposed model to effectively infer saliency influenced by depth information. The ultimate saliency map is predicted by combining the saliency distributions in spatio-temporal and depth cues by a Conv-LSTM based fusion network. Experimental results demonstrate that our proposed model outperforms all existing state-of-art saliency detection algorithms on the newly built SVS dataset.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 61871270, 61672443 and 61620106008, in part by the Hong Kong RGC Early Career Scheme under Grant 9048122 (CityU 21211018), in part by the Guangdong Nature Science Foundation of China under Grant 2016A030310058, in part by the Natural Science Foundation of SZU (grant no. 827000144), and in part by the National Engineering Laboratory for Big Data System Computing Technology of China.

References

[1] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos. A learning-based visual saliency prediction model for stereoscopic 3D video (LBVS-3D). *Multimedia Tools and Applications*, 76(22):23859–23890, 2017. 7, 8

[2] N. D. Bruce and J. K. Tsotsos. An attentional framework for stereo vision. In *Proc. 2nd IEEE Canadian Conf. Comput. Robot Vis.*, pages 88–95, 2005. 3

[3] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):740–757, 2019. 7

[4] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur. Adaptive 3D rendering based on region-of-interest. 7524:75240V, 2010. 3

[5] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Trans. Image Process.*, 27(10):5142–5154, 2018. 1, 2

[6] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. ICCV*, pages 2758–2766, 2015. 5

[7] Y. Fang, J. Wang, J. Li, R. Pépion, and P. Le Callet. An eye tracking database for stereoscopic video. In *Proc. Int. Workshop Quality Multimedia Exper.*, pages 51–52, 2014. 6

[8] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin. Saliency detection for stereoscopic images. *IEEE Trans. Image Process.*, 23(6):2625–2636, 2014. 3, 7, 8

[9] Y. Fang, Z. Wang, W. Lin, and Z. Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE Trans. Image Process.*, 23(9):3910–3921, 2014. 1

[10] Y. Fang, C. Zhang, J. Li, J. Lei, M. P. Da Silva, and P. Le Callet. Visual attention modeling for stereoscopic video: a benchmark and computational model. *IEEE Trans. Image Process.*, 26(10):4684–4696, 2017. 2, 5, 7, 8

[11] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, and F. Alexey. Semiautomatic visual-attention modeling and its application to video compression. In *Proc. ICIP*, pages 1105–1109, 2014. 6

[12] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell*, 34(10):1915–1926, 2012. 2

[13] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Proc. NIPS*, pages 545–552, 2007. 2, 7, 8

[14] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proc. CVPR*, pages 1–8, 2007. 2

[15] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proc. ICCV*, pages 262–270, 2015. 1, 2

[16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. ICML*, pages 448–456, 2015. 4

[17] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998. 2, 7, 8

[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 *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014. 5

[19] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang. Deepvs: A deep learning based video saliency prediction approach. In *Proc. ECCV*, pages 602–617, 2018. 5, 7, 8

[20] R. Kalboussi, M. Abdellaoui, and A. Douik. A spatiotemporal model for video saliency detection. In *Proc. IEEE Conf. IPAS*, pages 1–6, 2016. 2

[21] H. Kim, S. Lee, and A. C. Bovik. Saliency prediction on stereoscopic videos. *IEEE Trans. Image Process.*, 23(4):1476–1490, 2014. 3

[22] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *Proc. ICLR Workshop*, 2015. 2

[23] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. In *Proc. ECCV*, pages 101–115, 2012. 3

[24] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. ICCV*, pages 1520–1528, 2015. 4

[25] E. Potapova, M. Zillich, and M. Vincze. Learning what matters: combining probabilistic models of 2D and 3D saliency cues. *Proc. 8th Int. Comput. Vis. Syst.*, pages 132–142, 2011. 3

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Proc. ICLR*, 2015. 4

[27] Y. Tang, L. Ma, W. Liu, and W. Zheng. Long-Term human motion prediction by modeling motion context and enhancing motion dynamics. *Proc. IJCAI*, pages 935–941, 2018. 1

[28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. ICCV*, pages 4489–4497, 2015. 4, 5

[29] Q. Tu, A. Men, Z. Jiang, F. Ye, and J. Xu. Video saliency detection incorporating temporal information in compressed domain. *Signal Process. Image Commun.*, 38:32–44, 2015. 2

[30] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proc. CVPR*, pages 2798–2805, 2014. 2

[31] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proc. CVPR*, pages 4894–4903, 2018. 7

[32] W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. *IEEE Trans. Image Process.*, 27(1):38–49, 2018. 3

[33] R. Watt. Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *Journal of the Optical Society of America*, 4(10):2006–2021, 1987. 4

[34] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang. Learning to detect video saliency with HEVC features. *IEEE Trans. Image Process.*, 26(1):369–385, 2017. 2

[35] Q. Zhang, X. Wang, J. Jiang, and L. Ma. Deep learning features inspired saliency detection of 3D images. In *Pacific Rim Conference on Multimedia*, pages 580–589, 2016. 3

[36] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *Proc. AAAI*, pages 1063–1069, 2013. 1