



Received August 23, 2019, accepted September 24, 2019, date of publication October 9, 2019, date of current version October 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946479

Real-Time Video Saliency Prediction Via 3D Residual Convolutional Neural Network

ZHENHAO SUN^{1,2}, XU WANG^{1,2} , (Member, IEEE), QIUDAN ZHANG³,
AND JIANMIN JIANG^{1,2} 

¹College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

²Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China

³Department of Computer Science, City University of Hong Kong, Hong Kong

Corresponding author: Xu Wang (wangxu@szu.edu.cn)


This work was supported in part by the National Natural Science Foundation of China under Grant 61871270 and Grant 61620106008, in part by the Guangdong Natural Science Foundation of China under Grant 2016A030310058, in part by the Shenzhen Commission for Scientific Research and Innovations under the Grant JCYJ20160226191842793, in part by the Natural Science Foundation of SZU under Grant 827000144, and in part by the National Engineering Laboratory for Big Data System Computing Technology of China.

ABSTRACT Attention is a fundamental attribute of human visual system that plays important roles in many visual perception tasks. The key issue of video saliency lies in how to efficiently exploit the temporal information. Instead of singling out the temporal saliency maps, we propose a real-time end-to-end video saliency prediction model via 3D residual convolutional neural network (3D-ResNet), which incorporates the prediction of spatial and temporal saliency maps into one single process. In particular, a multi-scale feature representation scheme is employed to further boost the model performance. Besides, a frame skipping strategy is proposed for speeding up the saliency map inference process. Moreover, a new challenging eye tracking database with 220 video clips is established to facilitate the research of video saliency prediction. Extensive experimental results show our model outperforms the state-of-the-art methods over the eye fixation datasets in terms of both prediction accuracy and inference speed.

INDEX TERMS Video saliency prediction, eye fixation dataset, 3D residual convolutional neural network.

I. INTRODUCTION

Saliency originates from the fact that the cone and rod photoreceptors are not uniformly distributed on the retina, such that more attention is allocated to the attractive regions [1]–[3]. Numerous applications such as action recognition [4], robotic navigation [5], video compression [6] and semantic segmentation [3] benefit from robust saliency prediction. In particular, a rapid and robust saliency model can be applied as a pre-processing tool to efficiently extract region-of-interest (ROI) and distinguish the areas which are less important. Recent years have witnessed a dramatic performance improvement on image saliency prediction, especially after the introduction of deep learning [7], [8]. Inspired by the successful practice of convolutional neural network (ConvNet) [9] in computer vision tasks, many researchers employed ConvNet as a generic feature extractor, which is able to extract more sophisticated and descriptive features compared to traditional hand-crafted features.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang .

ConvNet-based models have also been successfully incorporated to automatically learn spatial feature to predict image saliency maps. For instance, the static saliency model in [7] was built on two typical ConvNet architectures including VGG [10] and ResNet [11].

Though the performance of static image saliency prediction approaches saturated to some extent [12], there is an increasing demand for video saliency prediction due to the numerous applications of the video big data. A video is generally composed of consecutive frames such that the static image model can be directly utilized in the video task. Since one key ingredient of videos is the temporal correlation, it is imperative to properly model the dynamical temporal properties in video saliency prediction. Currently, many methods have been developed to extract temporal related features to improve the performance of video saliency prediction [13]–[15]. In the pioneering work, Itti *et al.* [16] modeled the motion flow as the absolute difference of adjacent frames on both spatial and Gabor domains. Fang *et al.* [17] employed the optical flow between two consecutive frames for spatio-temporal feature extraction. However, estimating

optical flow is time consuming, which becomes the bottleneck in the inference stage. On the other hand, most of the models use a frame-by-frame inference manner, which is also cumbersome in real application scenarios.

To address the above-mentioned issues, this paper proposes a low complexity end-to-end method for video saliency prediction via 3D residual convolutional neural network (3D-ResNet) structures, which is efficient for spatio-temporal feature representation learning. Instead of extracting the temporal information straightforwardly, our model takes the video clip as the input, making it possible to generate a volume of saliency maps in a single forward pass, which significantly improves the inference speed. Moreover, a skipping frame strategy is also introduced to further boost the computational efficiency. Compared with the popular two-stream network structure and optical flow based architecture used in [17]–[19] and [20], the proposed model incorporates the temporal representation into spatial feature extraction as an integrated process and directly produces spatio-temporal saliency results. As such, there are two advantages of the proposed scheme. First, the time-consuming optical flow or other motion feature extractors are avoided, which significantly reduces the computational complexity. Second, in contrast to previous approaches [18], [20], the proposed model extracts the spatio-temporal feature directly from a video clip, such that the saliency-relevant spatio-temporal feature representation can be learned in an end-to-end manner.

Furthermore, to facilitate the training and evaluation on the proposed model, an large-scale eye fixation datasets with 220 5-second videos composed of 29,850 frames is constructed. Experimental results show that our proposed model has significantly outperformed the state-of-art video saliency models over the existing datasets and the proposed dataset. In summary, the main contributions of this paper are as follows,

- We propose a new end-to-end video saliency prediction model base on 3D-ResNet for data-driven spatio-temporal feature representation. A multi-scale feature representation scheme and a frame skipping strategy are proposed in this model to improve the prediction accuracy and inference speed. The experimental results show that the average inference speed of our model is 60 ~ 107 fps, which is more than 3 ~ 7 times faster than the existing deep learning based saliency prediction models such as OM-LSTM [21].
- We create a new challenging eye fixation video dataset, termed as VideoSet-Gaze (VSG), for the further research and evaluation towards video saliency prediction. The VSG dataset contains static close shot, slow and rapid motion sports scene, movie scene and synthetic video, which is publicly available.¹ Our proposed 3D-ResNet based video saliency model has been validated using this new dataset, showing competitive performance.

- We carry out analyses to investigate the influence of video content and temporal characteristics on video saliency prediction. These analyses provide useful insights to facilitate the future research of comprehensive models for video saliency prediction.

The rest of this paper is organized as follows. Section II reviews the related works of video saliency detection models. The proposed 3D-ResNet based video saliency prediction is discussed in section III. Section IV introduces the proposed video eye fixation dataset VSG. Extensive experimental results are provided in Section V. Finally, Section VI concludes this paper.

II. RELATED WORKS

A. VIDEO SALIENCY PREDICTION

Many state-of-the-art static saliency models utilize top-down or bottom-up modeling schemes, which perform well in predicting saliency on images [22], [23], especially after the introducing of deep neural networks. However, dynamic saliency prediction for video is much more complicated. Since the rapidly moving object of a dynamic scene may extract human attention, the eye fixation areas will be significantly different [24] when a frame is displayed in a static or dynamic way. Thus, the key issue of video saliency detection task is to jointly model the spatio-temporal cues. According to the method of feature extraction, existing video saliency prediction methods can be classified into heuristic methods (or hand-crafted feature based methods) and learning based methods.

1) HEURISTIC DYNAMIC METHOD

Inspired by the feature integration theory, a typical video saliency detection model consists of feature extraction module and fusion module. The feature extraction module generates feature maps for saliency related cues. The fusion module combines the feature maps to obtain the final saliency map. In addition to utilizing the spatial features [25] for static saliency prediction, Itti *et al.* extracted the temporal differences between frames for the dynamic task. By computing the difference between the luminance vector and the spatially-shifted Gabor pyramids of two adjacent frames, they yielded flicker pyramid and motion pyramid as temporal feature maps [16]. Nguyen *et al.* [24] used a spatial feature extraction scheme which is similar with Itti's model. The main difference is that in Nguyen's model, the optical flow fields were applied for dense temporal feature representation instead of flicker feature.

Inspired by the spectral residual method [26], Cui *et al.* proposed a FFT based fast motion saliency detection method [27]. The Fourier transform was applied along video slices on both X-T and Y-T planes to separate the background and foreground objects. After inhibiting noisy candidate by utilizing a threshold, a voting operation was used to obtain the salient motion region. Instead of using traditional transform based method, Guo *et al.* [28] applied the quaternion representation method to model the four

¹https://github.com/sunnycia/video_saliency_scripts

feature channels, including two color channels, one intensity channel, and one motion channel. Then the video saliency was obtained after applying Quaternion Fourier Transform (PQFT). Harel *et al.* proposed a Markov chain Graph-Based Visual Saliency (GBVS) model [29]. Taking advantages of static saliency map generated by GBVS, Zhong *et al.* [18] proposed a modified optical flow method to construct temporal saliency map. Rudoy *et al.* [20] proposed a candidate selection method for saliency prediction, which combines the static saliency model, human face detector and poselets detector to extract static candidates and semantic candidates. Then the most salient region is selected by learning the probability of transition between candidates of continuous frames.

In addition to investigating efficient temporal feature extraction schemes, efforts are also devoted to robust fusion strategy, which attempts to weigh the contributions of spatial and temporal saliency maps. Fang *et al.* proposed an uncertainty weighting based feature fusion scheme [17]. Karpathy *et al.* compared slow fusion, early fusion and late fusion methods when extracting spatio-temporal information from the original frames [30]. The cues from the compression domain can also play important roles in saliency prediction. Based on feature comparison, Fang *et al.* proposed a compression domain video saliency detection method in [31], where luminance, color and texture features are extracted by measuring the residual coefficients of coding unit from intra frame, and spatial saliency map is generated by comparing these features. The motion vector is used to finally generate a motion saliency map, which is combined with the spatial saliency map to fuse the final saliency map. Xu *et al.* [32] found there is consistency between compression domain based features and human fixation ground truth. As such, they utilized the HEVC codec as the spatio-temporal feature extractor and support vector machine is applied for further training.

2) DEEP LEARNING BASED DYNAMIC MODEL

Recently, deep learning based approaches [7], [8] have proven their success on saliency detection task benefiting from existing large-scale eye fixation datasets [33] and sophisticated algorithms. Vig *et al.* [34] firstly proposed a ConvNet-based feature extraction scheme for image saliency prediction. The end-to-end deep learning based image saliency prediction models as proposed in [7], [8] take one single image as input, and predicts the probability map as the saliency confidence. Huang *et al.* [8] proposed a model which takes both low and high-resolution images as input to learn multi-scale spatial features. In [7], deep spatial contextual LSTM based on ResNet [11] achieved the state-of-the-art performance.

Existing learning based dynamic models can be classified into optical-flow based methods and long-term-dependency-based methods. The optical-flow based methods explicitly employ the pre-computed short-term temporal information, mostly optical flow, as the compensation for CNN spatial feature. For example, Bak *et al.* utilized the two-stream

convolution neural network [19] for video saliency prediction. One stream takes the original frames as the input to extract spatial feature and another stream takes the pre-computed optical flow vector as the input for temporal feature extraction. Two different features are then simply concatenated for saliency pooling. The long-term-dependency-based methods mainly employed long short-term memory (LSTM) network to model the temporal dependency among frames. More recently, Jiang *et al.* [21] proposed an object-to-motion structure to model the transition of human eye fixations in the video, where the FlowNet [35] is employed as the backbone for extracting optical flow features.

Due to the limitation of model structure, existing video saliency models mainly take a few video frames as the input. Thus motion information and slow temporal changes may be ignored, which limited the performance. Besides, the frame by frame saliency inference is computationally inefficient, as every single frame in the video has to be fed into the model to obtain the final result. Based on these considerations, an efficient feature representation method which can model video saliency from both spatial and temporal perspectives with low complexity cost is highly desired. As such, feature representation methods in other video analytic tasks are also reviewed in the following subsection.

B. LEARNING BASED SPATIO-TEMPORAL FEATURE REPRESENTATION

Recently, efforts have been devoted to spatio-temporal feature representation learning in video processing and analytical tasks, such as video segmentation [36], video style transfer [37], video action classification [38] and temporal video localization [36]. Consecutive frames in videos are highly temporally correlated, and the temporal redundancy also appears in the deep feature maps [39]. To reduce the computational cost, Zhu *et al.* [39] proposed a deep feature flow framework which propagates the feature map of previous frames into subsequent inference frames. Ruder *et al.* [37] employed temporary consistency loss to preserve the consistency between adjacent frames and distant frames in video style transfer. Simonyan *et al.* [38] proposed a novel two-stream CNN architecture to extract spatial and temporal feature separately in two CNN branches for action recognition. To explore the fusion of temporal information by adjusting the structure of CNN, Karpathy *et al.* [30] tried four different models, including a baseline model with single-frame and three different methods to process frames in one forward pass for learning temporal information. However, these methods employ 2D convolution kernel, such that the temporal learning capability is constrained.

Due to the efficiency of 3D convolution neural networks (3D ConvNets) on spatio-temporal feature learning, Ji *et al.* [40] introduced the 3D convolution layer into their human action recognition network architecture, which can capture the motion information encoded in multiple adjacent frames. Tran *et al.* [41] proposed the C3D architecture, in which the 2D convolution and 2D pooling of VGG [10]

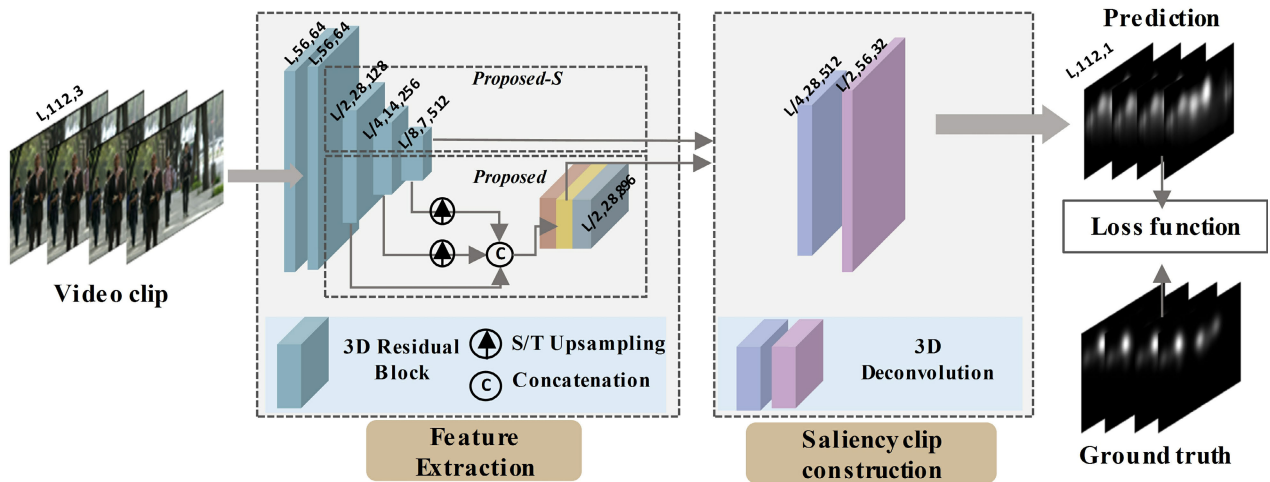


FIGURE 1. Network architecture of proposed saliency clip prediction network.

model were replaced by 3D convolution and 3D pooling, respectively. Carreira *et al.* [42] proposed a two-stream Inflated 3D ConvNet (I3D) which takes both RGB and optic flow volume as inputs. To reduce the computational cost and memory demand of very deep 3D CNN models, Qiu *et al.* [43] proposed the Pseudo-3D (P3D) residual networks with three different bottleneck block designing. Furthermore, Tran *et al.* [44] proposed the ResNet architecture based R(2+1)D model, which decomposed the 3D convolution into a 2D spatial convolution followed by 1D temporal convolution. Since additional non-linear rectification is involved in the (2+1)D operation, the feature representation capacity is significantly improved.

Compared to the optical flow based architecture, the 3D ConvNets based model processes multiple frames in a parallel manner, which significantly speeds up the inference process. To efficiently learn spatio-temporal feature representation from the video volume, our proposed video saliency prediction model will utilize the 3D-ResNet structure as the backbone for spatio-temporal feature extraction.

III. PROPOSED 3D-RESNET BASED VIDEO SALIENCY PREDICTION MODEL

In this section, the concept of our proposed video saliency prediction model is discussed in details. Suppose the input video clip with L consecutive frames and the density ground truth are denoted as X and Y , respectively, the goal of our proposed model is to learn an end-to-end mapping from X to Y . As depicted in Fig. 1, the proposed model consists of 3D-ResNet based spatio-temporal feature extraction module and saliency prediction module. Details of each module are described as follows.

A. 3D-RESNET BASED SPATIO-TEMPORAL FEATURE EXTRACTION

The proposed 3D-ResNet based spatio-temporal feature extraction module consists of an input layer, a 3D convolution

layer and four stacked 3D residual blocks. The input and output features for each layer are 4D tensors with shape $T \times W \times H \times C$, which represents the dimension of temporal, width, height and channel, respectively.

The first 3D convolution layer aims to extract low-level features directly from the input video clip. Each 3D residual block consists of four 3D convolution layers. As suggested in [10], stacked smaller kernels may lead to better performance compared to the setting with larger kernels. Thus, the kernels in these convolution layers are all set as $3 \times 3 \times 3$. Each 3D residual block contains a batch normalization layer [45] and scale layer to accelerate the convergence of training. The input video clip is not down-sampled in the first 3D convolution layer and the first 3D residual block in the temporal dimension, such that the temporal information in the early phase of the network can be preserved. Detailed parameters of each convolution blocks are summarized in Table 1.

B. MULTI-SCALE FEATURE REPRESENTATION BASED SALIENCY PREDICTION

After extracting saliency-related spatio-temporal feature tensor from the input video clip, the saliency prediction module will be executed for inferring the final saliency clip via multiple 3D deconvolution layers. However, the performance could be limited if only the output of the last 3D residual block is utilized, since the local saliency cues across spatial or temporal dimension may get lost due to the scaling operation.

According to the studies in cognitive research [46], human visual system obeys the coarse to fine strategy across spatial scales when detecting salient regions. Traditional hand-crafted features based models usually employed the image pyramids for obtaining multi-scale feature representation, which processed the raw input image for multiple times. However, it is not suitable for real application scenarios due to the increasing of computational complexity in the inference stage. For our proposed 3D-ResNet based feature extraction

TABLE 1. Detailed parameter setting of the proposed architecture.

layer name	parameter	output size
input		$L \times 112 \times 112 \times 3$
3D-Conv1	$3 \times 7 \times 7, 64, \text{stride } 1 \times 2 \times 2$	$L \times 56 \times 56 \times 64$
3D-Res2	$\begin{bmatrix} 3 \times 3 \times 3 \times 64 \\ 3 \times 3 \times 3 \times 64 \end{bmatrix} \times 2$	$L \times 56 \times 56 \times 64$
3D-Res3	$\begin{bmatrix} 3 \times 3 \times 3 \times 128 \\ 3 \times 3 \times 3 \times 128 \end{bmatrix} \times 2$	$\frac{L}{2} \times 28 \times 28 \times 128$
3D-Res4	$\begin{bmatrix} 3 \times 3 \times 3 \times 256 \\ 3 \times 3 \times 3 \times 256 \end{bmatrix} \times 2$	$\frac{L}{4} \times 14 \times 14 \times 256$
3D-Res5	$\begin{bmatrix} 3 \times 3 \times 3 \times 512 \\ 3 \times 3 \times 3 \times 512 \end{bmatrix} \times 2$	$\frac{L}{8} \times 7 \times 7 \times 512$
3D-De-Res4	$4 \times 4 \times 4 \times 256, \text{stride } 2 \times 2 \times 2$	$\frac{L}{2} \times 28 \times 28 \times 256$
3D-De-Res5	$8 \times 8 \times 8 \times 512, \text{stride } 2 \times 4 \times 4$	$\frac{L}{2} \times 28 \times 28 \times 512$
Concat		$\frac{L}{2} \times 28 \times 28 \times 896$
3D-Conv6	$7 \times 3 \times 3 \times 512$	$\frac{L}{4} \times 28 \times 28 \times 512$
3D-DeConv7	$4 \times 4 \times 4 \times 1, \text{stride } 2 \times 2 \times 2$	$\frac{L}{2} \times 56 \times 56 \times 32$
3D-DeConv8	$4 \times 4 \times 4 \times 11, \text{stride } 2 \times 2 \times 2$	$L \times 112 \times 112 \times 1$

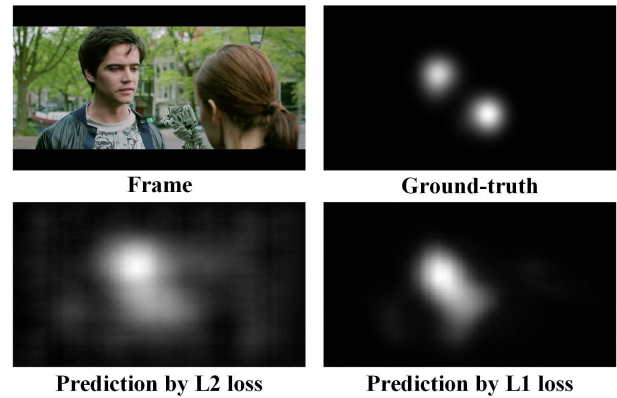
module, the shape of outputs from 3D residual blocks is inherently multi-scale [47]. In this paper, the output feature tensors of the last three 3D residual blocks are utilized for saliency reconstruction. Since the output of the layers “3D-Res4” and “3D-Res5” have coarser spatio-temporal resolution than that of the layer “3D-Res3,” a 3D deconvolution operation is applied to the output of layers “3D-Res4” and “3D-Res5,” respectively.

After mapping the features to the same scale by the 3D deconvolution operation, the output feature tensors of the third, fourth and fifth 3D residual blocks are concatenated to construct a multi-scale feature representation. Then a 3D convolution layer is employed for dimension reduction across temporal and feature channel scale. Finally, two stacked 3D deconvolution layers are utilized to reconstruct the saliency clip. By altering the kernel stride and kernel number, the resolution of features is altered across the spatio-temporal dimension. The predicted saliency clip S has the same shape as the input density ground truth Y .

C. LOSS FUNCTION FOR VIDEO SALIENCY PREDICTION

For deep learning based architectures, the design of loss function is found to be critical to the final performance. In [48], the authors compared distance-based loss function and probability-distribution-based loss function. They found by utilizing KL-divergence and Bhattacharyya distance, the performance achieves significant improvements compared to Euclidean distance. In [7], the combination of KL-divergence, NSS and CC was introduced as the objective function to train the neural network. These works provide useful evidence that a well-chosen objective function is important for training a specific task deep learning model.

In this paper, the training of our proposed whole framework is in an end-to-end manner, where the density eye fixation map is used as ground truth. The value of output volume of saliency map is between 0 and 1. During the training stage, the parameters of the network are updated via minimizing a loss function. In essence, the optimization of deep learning based saliency prediction is a pixel-wise regression problem, where the squared ℓ_2 norm of error is chosen as the default loss function. As shown in Fig. 2, the structure information of predicted saliency map is not well preserved, since the ℓ_2 norm penalizes large errors but is more tolerant to small errors and easily gets stuck in a local minimum [49].

**FIGURE 2.** Comparisons of the prediction maps based on the L_1 loss and L_2 loss.

Instead of using the ℓ_2 norm, the ℓ_1 norm is employed as loss function, which is defined as

$$\mathcal{L}^{\ell_1}(P) = \frac{1}{N} \sum_{p \in P} |S(p) - Y(p)|, \quad (1)$$

where p is the index of the pixel and P is the 5D tensor with size $B \times L \times W \times H \times C$; $S(p)$ and $Y(p)$ are the values of pixels in the predicted saliency clips and ground truth, respectively; and B is the batch size. During the training stage, all the learnable parameters are updated according to the back-propagated derivative of each pixel p , which is provided as follows.

$$\frac{\partial \mathcal{L}^{\ell_1}(P)}{\partial p} = \text{sign}(S(p) - Y(p)). \quad (2)$$

Compared to the ℓ_2 norm, the ℓ_1 norm does not over-penalize large errors and can reach better minimum. As shown in Fig. 2, ℓ_1 normal based prediction result is visually closer to the ground truth.

D. FLEXIBLE INFERENCE STAGE

During the inference phase, the frames in the video are first decoded and arranged into video clips through a slide window with size L and stride O . Due to the memory limitation, all the frames in the video clip are resized with spatial resolution 112×112 . At the beginning, the first L frames (F_1, \dots, F_L)

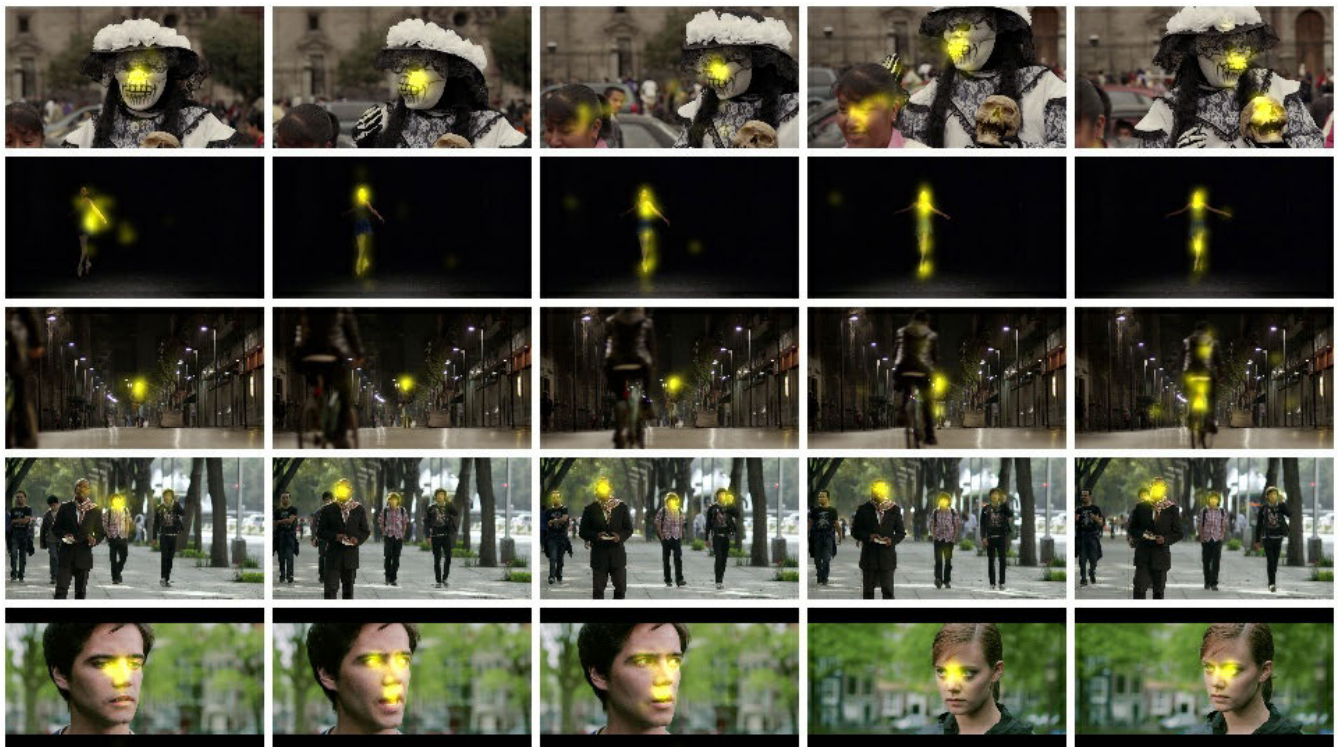


FIGURE 3. Examples selected from the VSG dataset. The red highlighted region denotes the density ground truth. This VSG dataset covers rapid motion scene, such as skateboard and soccer.

are arranged into a 4D tensor and fed into the proposed model, which outputs a volume of saliency map denoted as (S_1, \dots, S_L) . Only the first O saliency maps will be saved. Subsequently, the model will process the next L frames ($F_{O+1} \dots F_{O+L}$). This process will be executed recurrently until all frames have been processed. For real applications, we can adjust the parameter O to make a balance between prediction accuracy and inference speed. The influence of stride O on the model performance will be discussed in Section V.

IV. VIDEO-SET GAZE DATASET

For deep learning based saliency detection framework, large-scale eye fixation databases with diverse video content are fundamental requirements for learning meaningful and representative features. Currently, there are some limitations on the existing datasets. For example, the videos stimuli contained in some datasets are with low-resolution and corrupted by compression artifacts, which may influence the distributions of human gazing. The number of shots in a video clip is also an important factor. If there exist too many shots, the subjects may have to constantly adjust their eyes to the new content and switch their fixation area. In view of these limitations, we have constructed a new challenge dataset termed as the Video-SET Gaze (VGS) dataset, which includes 220 dynamic scenes. For each video clip, the corresponding eye fixation data from subjects are also included in the dataset.

A. EXPERIMENT SETTING

1) STIMULI

The raw video clips used for eye fixation data collection are from VideoSet [50], which is a large-scale just noticeable difference (JND) based compression video quality dataset. The VideoSet contains 220 raw source videos clips with diverse visual contents, such as slow motion scenes, rapid motion scenes, scenes with and without a salient object. The duration of each video clip is 5-seconds. The frame rate of video clips are 24 or 30 fps. In this paper, we only use the uncompressed original videos clips in VideoSet. All videos clips are in spatial resolution of 1920×1080 . Fig. 3 shows some thumbnails images of VideoSet for illustration.

2) EXPERIMENT APPARATUS

The eye tracker used to collect human eye gaze data is Tobii X-120 with sampling rate of 120 Hz. For each frame of a 30 fps video, the eye tracker can capture around four gazing positions from human eyes. The eye tracker is mounted on a 1920×1080 monitor during the experiment.

3) SUBJECTS

37 volunteers are invited to participate in the experiment, aged between 18 and 24. Glasses-wear subjects are required to clean the lens to guarantee the accuracy of gaze data. All of them have no saliency detection or relevant research

background. They are naive to the underlying purposes of the experiment and required to watch the video sequences freely.

4) EXPERIMENT PROCEDURE

The 220 short video sequences are randomized and divided into four groups, and each group contains 55 video clips. Once a video clip is displayed, a 3-second blank-background scenes will be displayed on screen for resting. After the beginning of each group, the volunteers are required to adjust their seat or pose for eye position calibration. Among groups, there is a 3-minute interval for resting. The total time for a volunteer watching videos is around forty minutes, including the resting and calibration. The raw gazing data are collected simultaneously when the videos were playing.

B. DATA COLLECTION AND ANALYSES

The eye tracker records the raw gazing points of the subjects when watching images or video frames on the screen. Compared to the image fixation data, fewer gaze data will be collected for every single frame in a video. The fixation area of each frame in video display mode is relatively smaller compared to the area for static display mode, especially when the video clip has only one salient object. The preprocessing procedure of raw gazing data is discussed as follows.

1) RAW GAZE DATA PREPROCESSING

The behaviors of eye movement can be roughly classified into fixation and saccade. The fixation phase refers to the fact that both of eyes will focus on the region of interest (ROI). The saccade characterizes the movement of eyes from one ROI to another. Based on the investigation in [56], the raw gaze point can be classified into fixation point and saccade point by measuring the angular velocity of the current gaze point. If the velocity of a point is larger than a preset threshold, this point will be classified into saccade point and removed from the point sequence. As such, ground truth fixation map is generated based on the raw fixation points. The fixation map is a two-dimension binary mask recording the gazing position, consisting of the fixated point represented with 1 and the non-fixated point represented with 0.

2) GROUND TRUTH DENSITY MAP GENERATION

The fixation map composed of discrete values is difficult for video saliency model training [24]. Thus, the ground-truth density map is generated from the fixation map. The density map is set to be the continuous ground truth for training in the end-to-end model because the distribution of density is similar to a 2D probability distribution, and each value of a pixel represent a saliency confidence value. Due to the micro-tremble of the eye and the limitation of the eye tracker, the fixation point changes within a small range. When generating the density map using Gaussian filtering, the size of the Gaussian kernel is chosen based on the average shaking range of the human eyes.

For each frame, we apply the Gaussian filtering around the gaze location of each participant. Since Gaussian distribution

models the nonuniform distribution of the photoreceptors on the retina (i.e., the eccentricity of the fovea), the size of the Gaussian kernel is set to be 1 degree of visual angle. In our experiment setting, 1-degree visual angle is equivalent to 32 pixels on the 24-inch 1920×1080 monitor.

During the gaze data processing, we also have some interesting observations as follows:

- The distribution of eye fixation points is highly influenced by the high level spatial semantic features such as human face or the temporal information such as rapid motion. For example, the attention of human can be attracted by the moving shuffle-board, as shown in Fig. 3.
- Human eyes use a short time to switch their attention from the center of the screen to the specific saliency objects. The average human reaction time in video free viewing condition is about 0.3s (e.g. 9 frames for a 30 fps video clip). In other words, the minimum temporal length of input video volume for 3D ConvNet should be larger than 9.

V. EXPERIMENT RESULTS

A. EXPERIMENTAL SETUP

1) EYE FIXATION VIDEO DATASETS

For performance evaluation, the experiments are conducted on the most representative video eye-tracking databases, including DIEM [51], HOLLYWOOD2 [52], SAVAM [53], LEDOV [21] and DHF1K [54] and our proposed VSG dataset. The details of each dataset are summarized as follows.

DIEM [51] includes 84 different resolution videos at 30 fps. The duration of each video clip is from 27 to 217 seconds. As such, there are 240,452 frames in total. The scenes are from publicly accessible videos including advertisement, game trailer, movie trailer etc. Most of these videos have frequent cinematic cuts. The free-viewing fixations of around 50 subjects were recorded for each video.

HOLLYWOOD-2 (denoted as HW) [52] is originally a video dataset for action recognition task. The dataset contains 1707 video sequences with 12 action classes, all of which are selected from Hollywood movies. This dataset comprises 48K frames in total. 16 subjects took participation in the data collection.

SAVAM [53] comprises 41 high definition 1080p stereo videos. Fifty subjects aging from 18-56 participated in the data collection experiment. The subjects were asked to freely view the videos. During the experiment, we only use the left eye videos and data for model performance comparison.

LEDOV [21] is an large-scale video saliency dataset published in 2018. It comprises 538 videos and 179,336 frames in total. The videos were selected from different publicly available sources with diverse spatial resolutions and frame rates. 32 subjects aging from 20 to 56 participate in the data collection experiment.

TABLE 2. Descriptions of the existing video saliency datasets.

Dataset	Number of Samples	Total frames	Subjects	Spatial Resolution	Frame Rate	Shots per video	Sample rate (Hz)
DIEM [51]	84	240,452	50	672×544~1280×720	30	55	1000
Hollywood-2 [52]	1707	487,556	16	480×320~720×576	24~30	3.84	500
SAVAM [53]	41	18,360	50	1920×1080	25	8.95	500
LEDOV [21]	538	179,336	32	640×480~3840×2160	15~60	1.21	300
DHF1K [54]	1000	582,605	17	640×360	30	2.74	250
VSG	220	29,850	37	1920×1080	24~30	1.44	120

TABLE 3. Performance comparisons on DIEM [51], Hollywood (HW) [52], SAVAM [53], LEDOV [21], DHF1K [54] and the proposed VSG datasets. The proposed model is compared with seven other saliency models, including static saliency models (SR [26], GBVS [29] and SALICON [8]) and dynamic models (Surprise [55], PQFT [28], UW [17] OM-LSTM [21]). The deep learning based model is in italic. Three metrics (CC, SIM, AUC) are measured to evaluate the performance for these models. The best three results are shown in bold black, blue and green, respectively.

Dataset	Metric	Static Model			Dynamic Model					
		SR	GBVS	<i>SALICON</i>	Surprise	PQFT	UW	<i>OM-LSTM</i>	<i>Proposed</i>	<i>Proposed-S</i>
LEDOV [21]	CC	0.2036	0.4321	0.4758	0.2643	0.2524	0.2559	0.6343	0.5537	0.5574
	SIM	0.3079	0.3940	0.4279	0.3229	0.3257	0.3263	0.5381	0.4784	0.4893
	AUC	0.6915	0.8332	0.8272	0.7304	0.7342	0.7320	0.8817	0.8622	0.8625
DIEM [51]	CC	0.0440	0.1660	0.1110	0.0602	0.0744	0.1168	0.1496	0.1471	0.1442
	SIM	0.1670	0.2191	0.1986	0.1624	0.1809	0.1970	0.2178	0.2119	0.2143
	AUC	0.5768	0.7025	0.6628	0.5687	0.6108	0.6569	0.7081	0.7077	0.7141
HW [52]	CC	0.1055	0.3200	0.4250	0.1915	0.1728	0.1217	0.4480	0.4703	0.4547
	SIM	0.1880	0.2753	0.3210	0.2281	0.2124	0.1954	0.3799	0.3648	0.3717
	AUC	0.6539	0.8298	0.8560	0.7155	0.7193	0.6731	0.8624	0.8737	0.8708
SAVAM [53]	CC	0.1241	0.2977	0.3642	0.1988	0.1767	0.1284	0.3515	0.4291	0.4077
	SIM	0.2107	0.2828	0.3112	0.2474	0.2325	0.2129	0.3406	0.3668	0.3660
	AUC	0.6466	0.7871	0.7864	0.7039	0.6934	0.6600	0.8045	0.8379	0.8320
DHF1K [54]	CC	0.1058	0.2775	0.2498	0.1621	0.1489	0.0966	0.3219	0.3566	0.3278
	SIM	0.1489	0.1830	0.1861	0.1634	0.1434	0.1273	0.2497	0.2539	0.2529
	AUC	0.6569	0.8297	0.7935	0.7134	0.7023	0.6695	0.8412	0.8610	0.8512
VSG	CC	0.1623	0.3392	0.3716	0.1677	0.1910	0.3481	0.3945	0.4013	0.3787
	SIM	0.1926	0.2588	0.2797	0.2096	0.1940	0.2673	0.3178	0.3143	0.3216
	AUC	0.7060	0.8510	0.8233	0.7145	0.7310	0.8489	0.8078	0.8707	0.8630
Inference speed (FPS)		160	2.27	3	6.7	7.14	0.49	16	60	125
Number of parameters		-	-	29M	-	-	-	83M	208M	50M

DHF1K [54] consists of one thousand well-selected high-quality dynamic videos with human fixation ground truth. This dataset totally contains around 600K frames and per-frame fixation annotations from 17 observers.

More detailed information regarding total frames, total subjects, video resolution, frame rate, average shot of the above-mentioned databases are summarized in Table 2.

2) TRAINING/INFERENCE PROTOCOLS

In our experiment, the LEDOV [21] database is used for training. More specifically, all the 538 training videos are segmented into overlapped clips with T consecutive frames. The length of overlap is $\frac{L}{2}$. For data augmentation, horizontal flip operation is employed on the video clips. When applying horizontal flipping, the same operation is performed on corresponding ground truth. Then, the spatial resolution of video clips are resized into 128×128 , with zero-mean and unit-variance normalization. For ground truth generation, the similar operations are implemented on the sequences

of density eye fixation map. In our experiment, we set the length of the video clip as $L = 16$. The proposed model is evaluated on six datasets, including DIEM [51], SAVAM [53], Hollywood-2 [52], DHF1K [54] and the proposed VSG dataset, with in total 2649 video sequences containing more than 900K frames.

3) IMPLEMENTATION DETAILS

In the training phase, the network parameters of the proposed feature extraction module are initialized based on the pre-trained model for video action recognition task [44]. This benefits our network from the perspective of richer and generic features learned on the Sports-1M dataset [30]. The Ada-delta gradient descent algorithm [57] is employed to minimize the L_1 loss between the cube of density eye fixation map and predicted saliency maps. The parameters such as batch size, momentum and weight decay are set as 2, 0.95 and 0.0005, respectively. The adaptively learning rate adjusting policy in Caffe platform [58] is enabled, which can

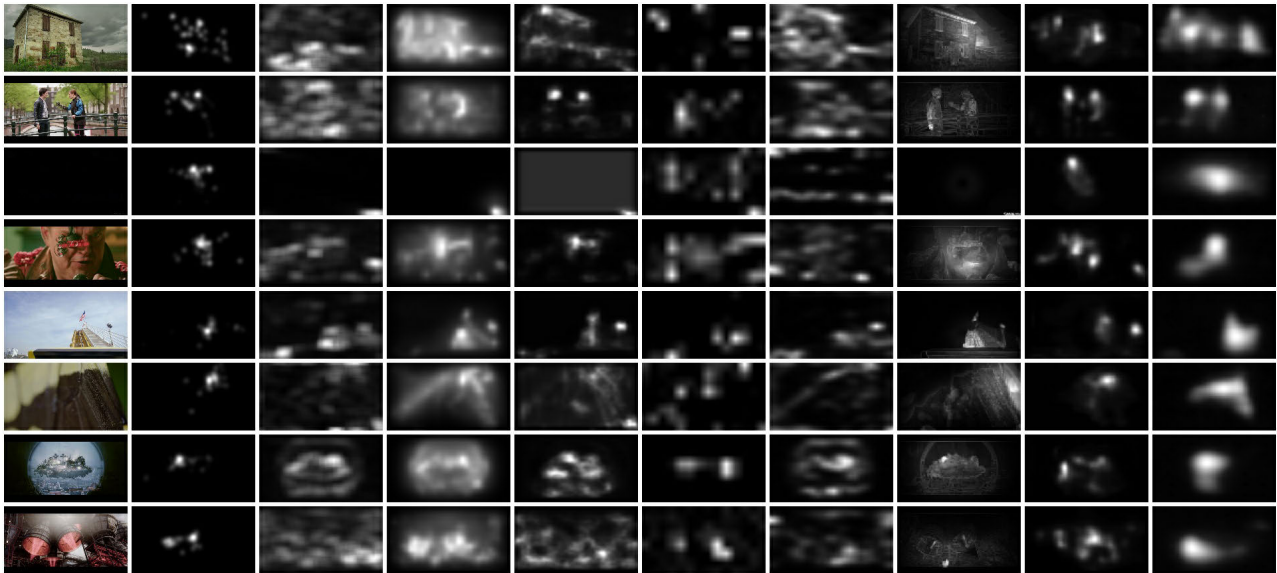


FIGURE 4. Saliency prediction map on VSG dataset. The left two columns are frames and ground truth density map randomly selected from different videos. Eight corresponding prediction from the models (PQFT [28], GBVS [29], SALICON [8], Surprise [55], SR (spectral residual) [26], UW (uncertainty weight) [17], OM-LSTM [21] and proposed) are shown in this figure.

be formulated as

$$R_i = (1 + \alpha i)^{-\gamma} R_0 \quad (3)$$

where R_0 is the initial learning rate and fixed as 10^{-6} . i is the index of iteration. We set $\alpha = 10^{-4}$ and $\gamma = 0.75$ in the experiment. All the training process is completed with a single NVIDIA Titan Xp GPU card. The network was trained over 200K iterations. Detailed information about the training will be provided in our project website.

B. PERFORMANCE COMPARISONS WITH THE STATE-OF-THE-ART METHODS

To demonstrate the effectiveness of the proposed model, we compare the performance with seven competitive saliency detection models, including static models such as Spectral Residual (denoted as SR) [26], GBVS [29] and SALICON [8], as well as dynamic models, including Surprise [55], PQFT [28], Uncertainty Weight (denoted as UW) [17] and OM-LSTM [21]. All the source codes are public available. In particular, SALICON and OM-LSTM are deep learning models. Following the instruction of the source codes provided by authors, these two models are trained on the LEDOV dataset for fair comparison. The settings of the training stage such as data processing and model parameters are set as default.

1) PERFORMANCE COMPARISONS

Three evaluation metrics are employed to measure the performance of saliency detection models on various datasets, including the linear correlation coefficient (CC) [59], the similarity (SIM) and an area under curve variant from Judd *et al.* (denoted as AUC) [60].

For each dataset, the performance is measured upon all the frames. The results are summarized in Table 3, from which it is obvious that the learning based dynamic model outperforms the static model and traditional dynamic model. Specifically, the best performance of non-learning model is UW, which achieves 0.3481, 0.2673 and 0.8489 on CC, SIM and AUC metric respectively on VSG dataset. Comparatively, our proposed model obtain 0.4013, 0.3143 and 0.8707 over these three metrics, which outperforms UW model on VSG dataset. Overall, the performance of proposed model is significantly higher than other methods across various metrics on most of the datasets. For example, on DHF1K dataset, the AUC value of proposed model is 0.8610, which is significantly better than that of the OM-LSTM model (0.8412). Based on the qualitative results provided in Fig. 4, the proposed model can achieve good visualization results on the VSG dataset. Besides, the proposed model is able to detect the fast moving object even if the size of object is relatively small in the frame, as shown in Fig. 5. This qualitative result demonstrates the motion tracking ability of the proposed architecture.

On the other hand, referring to Fig. 6, the proposed model is excellent with the scene with salient content like foreground human, but failed with some outdoor scenes. This may ascribe to the training dataset, since the model is trained to fit the attribute of training dataset. The training dataset LEDOV [21] selected videos only with salient objects, such that the model may fail for these landscape scenarios. Another reason is that there is no salient object in these failed cases, which also make it harder to predict where human tend to watch. We find that human tend to focus on the center of these scenes, but in our proposed model, we didn't add the center prior to the final prediction. Therefore, one practical solution to improve

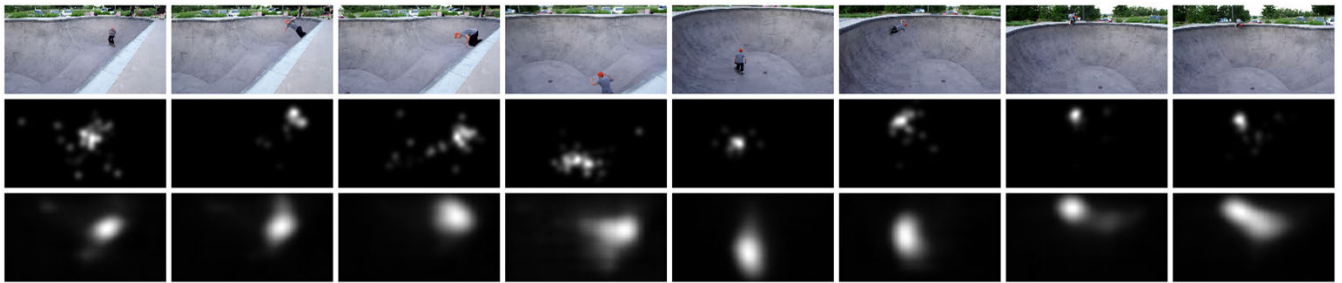


FIGURE 5. Saliency maps predicted by the proposed method selected from a video in the VSG dataset. The first row is the consecutive frame sequence sampled from the same video. The second row is the corresponding ground truth of each frame. The third row is the saliency map predicted by the proposed model. Our model is capable of capturing the fast-moving object in the scene even if the object is tiny, such as the last three frames on the top row.

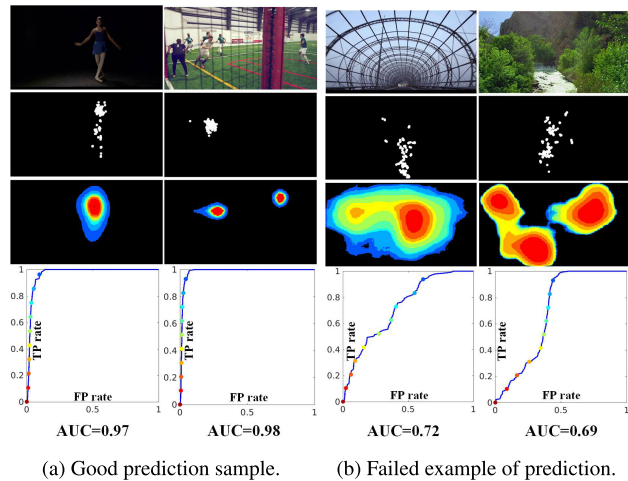


FIGURE 6. Examples of prediction on single frames and corresponding receiver operating characteristic curve.

the predict accuracy for these failed scenes is adding a center prior weighting to the final prediction.

2) COMPUTATIONAL COMPLEXITY

For practical application, computation load is significantly important, since the video saliency prediction is a fundamental part of video analysis system. The computation load in terms of inference speed for all the models are summarized in Table 3. The inference speed of proposed model with default setting is 60 fps, which around four times faster than that of the OM-LSTM model. Different from the optical flow based models, the proposed model can process consecutive frames in parallel, which significantly reduces the computational cost.

C. ABLATION STUDY

Benefiting from the architecture design, the proposed model is able to use a skip frame inference method to further improve the inference speed. We use eight frames overlapping between the previous and current inference clips by default. To explore the contribution of overlapping frames to the final prediction accuracy, we designed another experiment

TABLE 4. Comparison of different stride setting for inference sliding window on the datasets.

Dataset	Metric	Stride		
		8	12	16
LEDOV [21]	CC	0.4291	0.4262	0.4234
	SIM	0.3668	0.3656	0.3637
	AUC	0.8379	0.8365	0.8354
DIEM [51]	CC	0.1471	0.1456	0.1460
	SIM	0.2119	0.2112	0.2112
	AUC	0.7077	0.7082	0.7071
HW [52]	CC	0.4703	0.4674	0.4639
	SIM	0.3648	0.3644	0.3617
	AUC	0.8737	0.8730	0.8716
SAVAM [53]	CC	0.4291	0.4262	0.4234
	SIM	0.3668	0.3656	0.3637
	AUC	0.8379	0.8365	0.8354
DHF1K [54]	CC	0.3564	0.3555	0.3536
	SIM	0.2538	0.2545	0.2526
	AUC	0.8610	0.8606	0.8597
VSG	CC	0.4013	0.4027	0.4034
	SIM	0.3143	0.3159	0.3152
	AUC	0.8707	0.8713	0.8714
Inference speed (FPS)		60	81	107

which change the processing stride during the inference phrase. As shown in Table 4, the inference speed of the proposed model is highly dependent on the setting of stride parameter of the sliding window, whereas performance loss can be ignored. With the longest inference stride (16 frames), the **Proposed** model can achieve a inference speed of 107 frames per second, nearly double its speed of default setting (60 fps). The result indicates that although the relationship of the inference clip sequence is not explicitly considered in our design, the effect of the dependency of the clips cannot be ignored. The reason could be attributed to the human attention mechanism which relies more on short time dependency instead of long-time memory.

To further improve the inference speed, a simplified version of the original architecture (denoted as **Proposed-S**) is proposed. Compared with the original

proposed version, **Proposed-S** doesn't apply a feature pyramid concatenation operation. As shown in Fig. 1, the trimmed version removes the feature pyramid module in the original version. The direct result is lesser parameter is needed in **Proposed-S**. As it shows in the last line of Table 3, the **Proposed** method has more than 200 million parameters, whereas the **Proposed-S** shrink the number to about 50 million without great loss of accuracy performance. Based on the results in Table 3, the inference speed of **Proposed-S** is able to achieve two times faster than the original version, with the processing speed of 125 fps. Without dramatical loss of accuracy, it is possible for this model to perform a real-time video analysis task.

VI. CONCLUSION

This paper designed an end-to-end low complexity 3D-ResNet architecture for video saliency prediction. The proposed framework captures saliency-relevant spatio-temporal features in an integrated manner, such that the video saliency can be feasibly inferred. Moreover, a large-scale eye fixation dataset is built to better reflect the video saliency performance. With high inference speed, the experimental results show that proposed architecture is competitive with state-of-the-art video saliency model and image saliency model, and superior saliency prediction performance in terms of both accuracy and computational speed is achieved.

REFERENCES

- [1] X. Ding and Z. Chen, "Improving saliency detection based on modeling photographer's intention," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 124–134, Jan. 2019.
- [2] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.
- [3] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, Aug. 2017.
- [4] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–15.
- [5] C. Ackerman and L. Itti, "Robot steering with spectral image information," *IEEE Trans. Robot.*, vol. 21, no. 2, pp. 247–251, Apr. 2005.
- [6] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 155–170, Jan. 2018.
- [7] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2016.
- [8] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 262–270.
- [9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [12] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *Proc. ECCV*, 2016, pp. 809–824.
- [13] Y. Yang, B. Li, P. Li, and Q. Liu, "A two-stage clustering based 3D visual saliency model for dynamic scenarios," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 809–820, Apr. 2019.
- [14] M. Xu, Y. Ren, Z. Wang, J. Liu, and X. Tao, "Saliency detection in face videos: A data-driven approach," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1335–1349, Jun. 2018.
- [15] K. Fu, I. Y.-H. Gu, and J. Yang, "Saliency detection by fully learning a continuous conditional random field," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1531–1544, Jul. 2017.
- [16] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Appl. Sci. Neural Netw., Fuzzy Syst., Evol. Comput.*, vol. 5200, pp. 64–79, Dec. 2003.
- [17] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.
- [18] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Proc. 2nd ACM ICMR*, 2012, pp. 1–7.
- [19] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1688–1698, Jul. 2018.
- [20] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1147–1154.
- [21] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *Proc. ECCV*, 2018, pp. 602–617.
- [22] N. Imamoglu, W. Lin, and Y. Fang, "A saliency detection model using low-level features based on wavelet transform," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 96–105, Jan. 2013.
- [23] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187–198, Feb. 2012.
- [24] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan, "Static saliency vs. dynamic saliency: A comparative study," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 987–996.
- [25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
- [26] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
- [27] X. Cui, Q. Liu, and D. Metaxas, "Temporal spectral residual," in *Proc. 17th ACM Int. Conf. Multimedia*, vol. 86, 2009, pp. 617–620.
- [28] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
- [29] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.
- [30] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
- [31] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.
- [32] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with hevc features," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 369–385, Jan. 2017.
- [33] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1072–1080.
- [34] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.
- [35] 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. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2758–2766.
- [36] W. Wang, J. Shen, J. Xie, and F. Porikli, "Super-trajectory for video segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1671–1679.

- [37] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos and spherical images," *Int. J. Comput. Vis.*, vol. 126, no. 11, pp. 1199–1219, 2018.
- [38] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
- [39] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4141–4150.
- [40] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
- [42] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.
- [43] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5534–5542.
- [44] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
- [45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [46] R. J. Watt, "Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 10, pp. 2006–2021, 1987.
- [47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
- [48] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5753–5761.
- [49] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
- [50] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang, Y. Zhang, J. Huang, S. Kwong, and C.-C. J. Kuo, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 292–302, Jul. 2017.
- [51] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cogn. Comput.*, vol. 3, no. 1, pp. 5–24, 2011.
- [52] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.
- [53] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, and F. Alexey, "Semi-automatic visual-attention modeling and its application to video compression," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 1105–1109.
- [54] 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. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4894–4903.
- [55] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 547–554.
- [56] M. Nyström and K. Holmqvist, "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data," *Behav. Res. Methods*, vol. 42, no. 1, pp. 188–204, 2010.
- [57] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
- [58] 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 *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
- [59] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.
- [60] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis. Sep./Oct. 2009*, pp. 2106–2113.



ZHENHAO SUN received the M.S. degree from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2019, where he is currently a Research Assistant with the Research Institute for Future Media Computing. His research interests include image/video saliency detection and deep learning.



XU WANG (M'15) received the B.S. degree from South China Normal University, Guangzhou, China, in 2007, the M.S. degree from Ningbo University, Ningbo, China, in 2010, and the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2014. In 2015, he joined the College of Computer Science and Software Engineering, Shenzhen University, as an Assistant Professor. His research interests include video coding and processing, visual quality assessment, and visual saliency detection.



QIUDAN ZHANG received the B.E. and M.S. degrees from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong. Her research interests include computer vision, visual attention, and deep learning.



JIANMIN JIANG received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 1994. From 1997 to 2001, he was a Full Professor of computing with the University of Glamorgan, Pontypridd, U.K. In 2002, he joined the University of Bradford, Bradford, U.K., as a Chair Professor of digital media and the Director of the Digital Media and Systems Research Institute. He was a Full Professor with the University of Surrey, Guildford, U.K., from 2010 to 2014, and a Distinguished Chair Professor (1000-Plan) with Tianjin University, Tianjin, China, from 2010 to 2013. He is currently a Distinguished Chair Professor and the Director of the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He has published around 400 refereed research articles. His current research interests include, image/video processing in compressed domain, digital video coding, medical imaging, computer graphics, machine learning and AI applications in digital media processing, and retrieval and analysis. He was a Chartered Engineer and a Fellow of RSA. He was a member of EPSRC College, U.K., and an EU FP-6/7 Evaluator.

...