SalSAC: A Video Saliency Prediction Model with Shuffled Attentions and Correlation-Based ConvLSTM

Xinyi Wu,[1] Zhenyao Wu,[1] Jinglin Zhang,[3] Lili Ju,[1*] Song Wang[1,2*]
[1]University of South Carolina, USA [2]Tianjin University, China
[3]Nanjing University of Information Science and Technology, China
{xinyiw, zhenyao}@email.sc.edu, jinglin.zhang@nuist.edu.cn, ju@math.sc.edu, songwang@cec.sc.edu

Abstract

The performance of predicting human fixations in videos has been much enhanced with the help of development of the convolutional neural networks (CNN). In this paper, we propose a novel end-to-end neural network “SalSAC” for video saliency prediction, which uses the CNN-LSTM-Attention as the basic architecture and utilizes the information from both static and dynamic aspects. To better represent the static information of each frame, we first extract multi-level features of same size from different layers of the encoder CNN and calculate the corresponding multi-level attentions, then we randomly shuffle these attention maps among levels and multiply them to the extracted multi-level features respectively. Through this way, we leverage the attention consistency across different layers to improve the robustness of the network. On the dynamic aspect, we propose a correlation-based ConvLSTM to appropriately balance the influence of the current and preceding frames to the prediction. Experimental results on the DHF1K, Hollywood2 and UCF-sports datasets show that SalSAC outperforms many existing state-of-the-art methods.

Introduction

Saliency prediction has been introduced for years to learn how people select useful information from complicated visual information they see every day. Thanks to the development of the deep learning techniques and the appearing of many static gaze datasets such as SALICON (Jiang et al. 2015), the performance of the image saliency prediction has been boosted very rapidly.

In recent years, video saliency prediction captures more interest in the AI and vision community due to its huge benefit to other applications, e.g., video captioning, video compression and video object segmentation. Compared to image saliency prediction, it is a much more challenging task. Besides, extracting spatial features accurately as in image saliency prediction, video saliency prediction also needs to balance the memory to be kept (e.g., some moving objects and the newly appearing or disappearing objects) in the current state and the memory to be forgotten (e.g., some static objects) in its preceding states along the temporal dimension; in other words, it needs to consider both spatial and temporal context information to make decision.

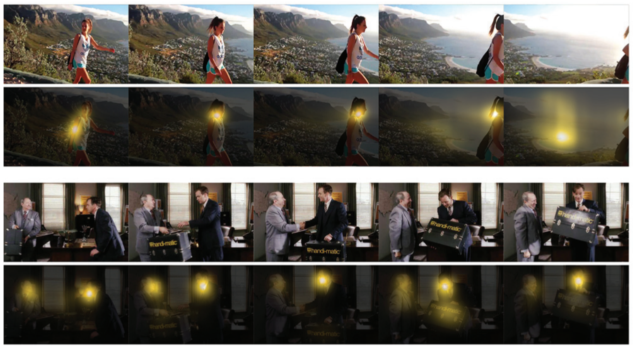

Figure 1: Visualization of the predicted saliency results of two test cases by our network SalSAC.

Many methods have been proposed to tackle these challenges in video saliency prediction mentioned above. To fully extract the spatial features, early works (Gao, Mahadevan, and Vasconcelos 2008; Guo and Zhang 2009) mainly depend on hand-crafted features, such as intensity, color and orientation. In recent years, many approaches (Bak et al. 2017; Wang et al. 2018a) are proposed to use convolutional neural networks (CNN), such as VGG16 and ResNet, as the encoder to extract features from the image sequences. Later, the attention mechanism is also utilized within the networks to further enhance the spatial features (Wang et al. 2018a; Cornia et al. 2018). In this paper, we propose a shuffled attention module, which shuffles multi-level attention maps calculated from the multi-level features to enforce the attention consistency across levels and further enhance the quality of spatial features.

To absorb information from both historic frames and the current one, Convolutional LSTM (ConvLSTM) (Shi et al. 2015) structure is imported into this task in (Gorji and Clark 2018; Jiang et al. 2018; Wang et al. 2018a). Compared to the original LSTM, ConvLSTM can capture useful temporal information in addition to the spatial one. While the ConvLSTM can model certain long-short term memory in the

[*] Co-corresponding authors.

sequence prediction, it is still difficult to accurately capture the newly appearing or disappearing objects. Furthermore, due to the complex nature of human attention, it is always uncertain which one is more important among the historic information and that from the current frame. Inspired by SalEMA (Linardos et al. 2019), which uses the exponential moving average recurrence operation to balance the current state and its preceding state, in this paper we adapt the traditional ConvLSTM to include the relationship between the adjacent frames based on the correlation information. In particular, when the correlation value is large, which means the feature of the current frame is similar to that of its preceding one, the ConvLSTM should consider the feature of the preceding frame more; on the contrary, when the correlation value is small, which means there could be appearing or disappearing of objects, the ConvLSTM should weigh the feature of the current frame more.

More specifically, in this paper we propose an end-to-end neural network "SalSAC" for video saliency prediction, which takes the CNN-LSTM-Attention (Wang et al. 2018a) as the basic architecture (without using more parameters) and utilizes a shuffled multi-level attention module and a correlation-based ConvLSTM. Static features are first extracted from frames through an encoder CNN. The shuffled attention module is then used to enhance the performance and improve the robustness of the whole network, and the correlation-based ConvLSTM is designed for balancing the weight of the current state and its preceding state depending on the correlation value. Finally, a decoder CNN is fed with the intermediate results to predict the final saliency maps as shown in 1. We conduct various experiments on the DHF1K, Hollywood-2 and UCF-sports datasets and the results on all the datasets clearly demonstrate the effectiveness and accuracy of our method.

The main contributions of this paper are as follows:

- We propose a novel CNN-based neural network trained in the end-to-end manner for video saliency prediction.
- We introduce a random shuffling mechanism for the multi-level attentions calculated from the multi-level features to further enhance the robustness of the network.
- We develop a correlation-based ConvLSTM in our network, which can effectively and adaptively weigh the importance of the current frame and its preceding one to the saliency prediction.
- Our method outperforms many existing state-of-the-art networks on the DHF1K, Hollywood-2 and UCF-sports datasets by using similar and less parameters.

Related work

Saliency Prediction

Saliency prediction for static images has been studied for years and good results have been achieved by both traditional models (Gao and Vasconcelos 2005; Hou and Zhang 2007; Judd et al. 2009) and deep learning based methods (Cornia et al. 2016; 2018; Jiang et al. 2019). This task was later extended into videos, mostly driven by the increasing need of video processing applications. In the early works (Gao, Mahadevan, and Vasconcelos 2008; Guo and Zhang 2009; Rudoy et al. 2013), stimulus modalities and motion features are used to model the saliency in videos. Recently, many CNN based methods (Bak et al. 2017; Leifman et al. 2017; Gorji and Clark 2018; Wang et al. 2018a; Jiang et al. 2018; Cheng et al. 2018; Linardos et al. 2019) are proposed to avoid the limitation of the hand-crafted features and improve the performance of saliency prediction. In (Bak et al. 2017), a two-stream network is adopted to integrate both spatial and temporal information using the pre-computed optical flow. Similarly, in DeepVS (Jiang et al. 2018) two sub-networks are used to obtain the object and motion features and then the fused features are sent into a two-layer ConvLSTM to infer the final saliency maps. In addition to motion features, depth information is used in (Leifman et al. 2017) to enhance saliency prediction. In (Cheng et al. 2018) a spatial-temporal network with weakly-supervised training is proposed to predict saliency for 360° videos.

ConvLSTM

To make use of the temporal information in videos and preserve the spatially informative features, ConvLSTM layers have been widely used in many video tasks (Luo et al. 2017; Tao et al. 2018; Jiang et al. 2018; Wang et al. 2018a). In (Gorji and Clark 2018), a multi-stream ConvLSTM structure is used to learn the dynamic saliency and three attention push cues including gaze following, rapid scene change and attentional bounce. Although the ConvLSTM can hold the temporal information in its cell, it is still hard to accurately balance the influence of the features in the current frame and its preceding one. A simple version of recurrent model by learning a hyper-parameter is used in (Linardos et al. 2019) to weigh the features of the current and preceding frames to replace the ConvLSTM and get a comparable performance. However, this hyper-parameter is learnt during training and then is fixed during the testing process, which may not be suitable for all test cases. It is more reasonable for the network to first consider the relationship between each two consecutive frames and then decide how to appropriately weigh the importance of the current and preceding ones to the saliency prediction. Inspired by this idea, in this work, we utilize the correlation layer proposed in FlowNet (Dosovitskiy et al. 2015) to calculate the correlation of the features between the consecutive frames and then use this information to weigh the current and preceding states. Note that our approach also make the weight to be case-specific even during the testing process.

Attention Mechanism

Attention mechanism has been proven to be very useful for enhancing the performance in many computer vision tasks including image/video saliency prediction. In (Wang et al. 2018a), an attention module is used and trained on an image saliency dataset for improving the quality of the static features. We also make use of the attention mechanism in this work - we calculate multi-level attentions from the corresponding multi-level features and then randomly shuffle them among all levels to leverage the attention consistency

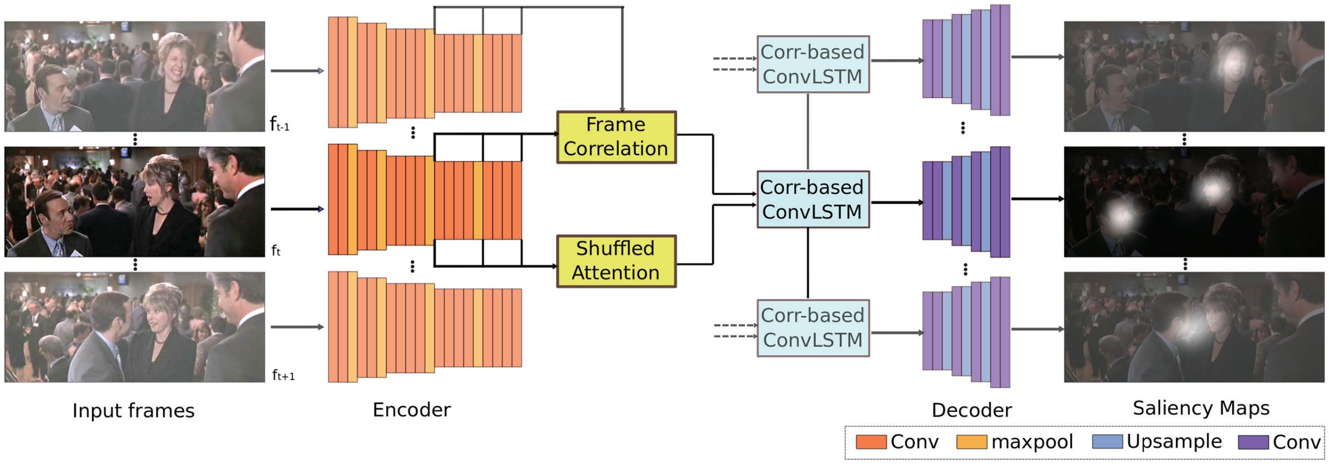

Figure 2: Architecture of the proposed SalSAC network for video saliency prediction, which contains a shuffled attention module and a correlation-based ConvLSTM layer. Here, f_t is the current frame for saliency prediction.

along the layers in the network and finally to enhance the robustness of the whole architecture.

Our Approach

The overall architecture of the proposed SalSAC network for video saliency prediction is shown in Figure 2, which follows the CNN-Attention-ConvLSTM pipeline. The first step is to extract multi-level static image features from the frame sequences using the encoder CNN. Then the shuffled attention module is applied to calculate and shuffle the attentions of the multi-level features. The output of the attention module, together with the calculated correlation between the current frame and its preceding one, are then fed into the correlation-based ConvLSTM layer to model the dynamic changes of the two frames along the temporal dimension. Finally, several convolution and upsampling layers are used as the decoder to process the intermediate results and produce the final saliency prediction.

Encoder-Decoder

Similar to the work of (Cornia et al. 2016) which uses VGG16 as an encoder to extract multi-level features from static images, we take the outputs from three layers of VGG19 (Simonyan and Zisserman 2014) (the third and fourth max-pooling layers and last convolutional layer) as the multi-level spatial feature information. In order to reduce rescaling operations in the decoder, we remove the last max-pooling layer and change the stride of the fourth max-pooling layer to 1 in the original VGG19. Let h and w denote the height and width of the input frame, then the output feature maps at all the three levels are of the same size ($\frac{w}{8} \times \frac{h}{8}$) in the x and y dimensions but have respectively 256, 512 and 512 channels.

For the decoder part, we use six convolutional layers in which the kernel sizes are all 3×3, and three additional bilinear upsampling layers in which the scaling factors are all set to be 2. With this decoder, the final predicted saliency maps are restored back to their original resolution.

Shuffled Attention Module

Previous work (Cornia et al. 2018; Wang et al. 2018a) has shown that the self-attention mechanism plays an important role in further boosting the performance of both static and dynamic saliency detection. Therefore, we add a modified version of the attention module proposed in (Wang et al. 2018b) in our network to help concentrate on more important regions. The multi-level features extracted by the encoder are the input of the attention module. Our modification is inspired by (Guo et al. 2019), which leverages the attention consistency as a constraint for certain image transforms such as scaling, rotation, flipping and translation to enhance the performance of image classification. It is thus expected that the attention maps should preserve certain attention consistency across all levels.

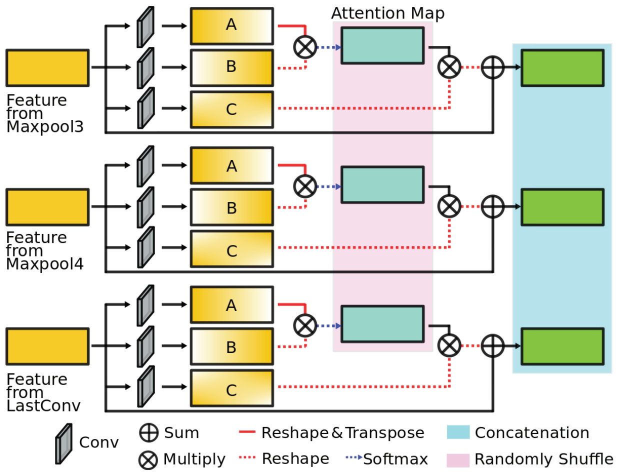

Figure 3: Architecture of the shuffled attention module, which takes the three-level features produced from the encoder as the input.

Different from the approach taken in (Guo et al. 2019) which imposes an extra term in the loss function to restrain the attention maps of the multi-level features to be close, we instead randomly shuffle the attention maps among the three levels. The structure of the shuffled attention module is illustrated in Figure 3, where the shuffle operation is embedded into the non-local block (Wang et al. 2018b). First, each of the three-level features is fed as input into three convolutional layers to get three feature maps (A, B and C), and the first two feature maps (A and B) takes a multiplication (i.e., A^{T}B after reshaping the spatial dimension into a vector form), then the result is passed into a softmax layer to obtain the attention map. Second, the three attention maps are randomly shuffled for each iteration during training, e.g., the original three attention maps for Maxpool3, Maxpool4 and LastConv levels are applied to Maxpool4, LastConv and Maxpool3 levels respectively, and then multiplied respectively to the third feature map (C) at each level, which are next added to the input multi-level features in a residual connection manner. The three levels of outputs are finally concatenated together as the final static image information of the current frame.

Correlation Operation

We use the correlation layer firstly introduced in (Dosovitskiy et al. 2015) to calculate the relationship between features extracted from two consecutive frames. The structure of the correlation operation is presented in Figure 4.

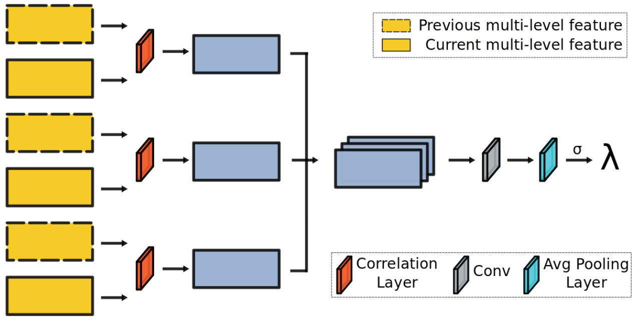

Figure 4: The correlation operation based on the three-level features between each pair of two consecutive frames.

The correlation operation is done for each of the three-level features separately. For the level k ($k \in \{1, 2, 3\}$), the correlation map C_i is calculated using the following formula:

$$C_k = \mathcal{C}(f_{t-1}^k, f_t^k), \tag{1}$$

where f_{t-1}^k and f_t^k are respectively the features of the current frame and its preceding frame, and $\mathcal{C}$ represents the correlation function for measuring the relationship between the two given features, which is a dot-product operator along the channel dimension here. The correlation results of all levels are then concatenated and passed into a convolutional layer and a global average pooling layer. Finally, a factor λ is obtained to measure the dynamic changes between the two consecutive frames as follows:

$$\lambda = \sigma\left(\frac{64}{wh}\sum_{(i,j)}(WC + b)\right), \tag{2}$$

where W and b denote the weights and bias for the 3×3 convolutional layer respectively, C is the concatenation of C_1, C_2 and C_3, and σ the sigmoid function to restrict the range of the value to $[0, 1]$.

The factor λ will be used to measure the similarity between two frames. If λ is large, it means that the dynamic change is small, thus the historic information is important for predicting the current frame. When there are objects showing up or disappearing in the current frame, difference between the two consecutive frames is large , resulting in a small λ, and in this case the historic information is less important.

Correlation-Based ConvLSTM

In (Jiang et al. 2018; Wang et al. 2018a), ConvLSTM has been used to capture the temporal information of videos, which consists of three gated operations: the input, forget and output gates. The ConvLSTM layer sometimes cannot capture the newly appearing objects very well, which however is important on the video saliency prediction task. To address this issue, we propose a correlation-based ConvLSTM by integrating the correlation information between the current and preceding frames into ConvLSTM, Figure 5 illustrates the difference between the traditional ConvLSTM and the proposed correlation-based ConvLSTM.

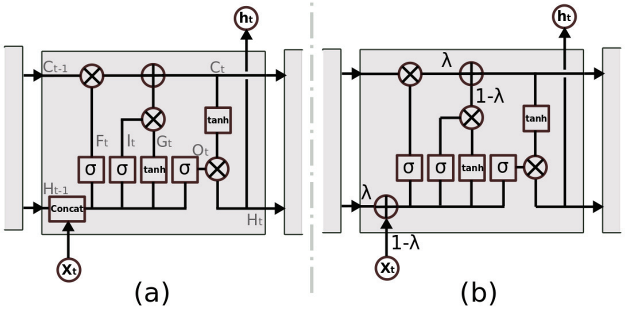

Figure 5: An illustration of (a) the traditional ConvLSTM and (b) the proposed correlation-based ConvLSTM.

Our correlation-based ConvLSTM still uses the classic three-gated operations, however, the inputs to all gates are modified to:

$$(1 - \lambda)X_t + \lambda H_{t-1}, \tag{3}$$

where X_t is the output feature from the shuffled attention module, H_{t-1} represents the hidden state and λ is the factor calculated from the correlation operation. With the new input described in Eq. (3), the three gated operations are then redefined as follows:

$$I_t = \sigma(W_i((1 - \lambda)X_t + \lambda H_{t-1}) + b_i), \tag{4}$$

$$F_t = \sigma(W_f((1 - \lambda)X_t + \lambda H_{t-1}) + b_f), \tag{5}$$

$$O_t = \sigma(W_o((1-\lambda)X_t + \lambda H_{t-1}) + b_o), \qquad (6)$$

where I_t, F_t, O_t are the input, forget and output gates, respectively, $\{W_i, W_f, W_o\}$ denote the weights, and $\{b_i, b_f, b_o\}$ denote the bias. Then the new cell state C_t, i.e., the output of the correlation-based ConvLSTM, is also balanced by the factor λ and becomes:

$$C_t = \lambda F_t \odot C_{t-1} + (1-\lambda) I_t \odot G_t \qquad (7)$$

with

$$G_t = \tanh(W_c((1-\lambda)X_t + \lambda H_{t-1}) + b_c), \qquad (8)$$

where $\odot$ denotes the element-wise product of the vectors. The hidden state H_t is defined as:

$$H_t = O_t \odot \tanh(C_t). \qquad (9)$$

By using the correlation-based ConvLSTM, we could adaptively balance the importance of the information from the current frame and its preceding one since λ changes along the sequence of frames.

Loss Function

Following the work of (Cornia et al. 2016; 2018; Wang et al. 2018a), we choose the popular Kullback-Leibler divergence as one part of the loss function, which is often used to measure the distance between two distributions. Let us first calculate the probabilistic representation T and F for the ground-truth saliency map GT and the predicted saliency map P, respectively:

$$T(i,j) = \frac{GT(i,j)}{SGT + \epsilon}, \qquad F(i,j) = \frac{P(i,j)}{SP + \epsilon}, \qquad (10)$$

where $SGT = \sum_{(i,j)} GT(i,j)$, $SP = \sum_{(i,j)} P(i,j)$, and ϵ is set to be 10^{-20} to avoid being divided by zero. Then the Kullback-Leibler divergence L_{KL} is defined as follows:

$$\mathcal{L}_{KL} = \sum_{(i,j)} T(i,j) log\left(\frac{T(i,j)}{F(i,j)} + \epsilon\right). \qquad (11)$$

Meanwhile, the Mean Squared Error (L_{MSE}) is used as the other part of the loss function to measure the L^2 distance between the ground truth and the prediction:

$$\mathcal{L}_{MSE} = \frac{1}{hw} \sum_{(i,j)} (GT(i,j) - P(i,j))^2. \qquad (12)$$

The final loss function is denoted as follows:

$$\mathcal{L}_{MSE} = \mathcal{L}_{KL} + \beta \mathcal{L}_{MSE}, \qquad (13)$$

where β is a weighting parameter, which is set to be 100 in our experiments.

Experimental results

Datasets

We carry out tests and comparisons on three datasets: DHF1K (Wang et al. 2018a), Hollywood-2 (Marszałek, Laptev, and Schmid 2009) and UCF-sports (Mathe and Sminchisescu 2014).

DHF1K is a newly collected dataset for free viewing video saliency prediction. It contains 1,000 videos, in which the first 700 videos have high quality annotations published and the remaining 300 videos are held as a benchmark for testing. Following (Wang et al. 2018a), we use the first 700 videos for training (including 100 for validation) and the last 300 for evaluation.

Hollywood-2 is the largest dynamic eye-tracking dataset which contains a total of 1,707 videos. Different from DHF1K, Hollywood-2 is collected under task-driven, given that 16 of 19 observers are aware of the action and context while annotating. In this paper, following (Marszałek, Laptev, and Schmid 2009), we use 823 videos as the training set and 884 videos as the testing set.

UCF-sports contains 150 videos which cover 9 common sports action categories. Similar to Hollywood-2, its collection is also driven by task purpose. Following (Wang et al. 2018a), we use 103 videos for training and 47 videos for testing.

In addition, we use SALICON (Jiang et al. 2015), an image saliency dataset containing a training set of 10,000 static images, for pre-training part of the network in order to improve its ability of capturing the static image information.

Metrics

For evaluation of the performance of video saliency prediction, we use five common visual saliency metrics as in (Borji and Itti 2012; Bylinskii et al. 2018): AUC-Judd (AUC-J), shuffled AUC (s-AUC), Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC). For all of these metrics, larger value means better performance.

Model Specification

Training setting We first take the training set of the SALICON dataset to pre-train the network without the correlation-based ConvLSTM layer for 15 epochs. Then the training videos from DHF1K, Hollywood-2 or UCF-sports are used to train the respective network for another 30 epochs with the batch size set to be 1, respectively. We take Adam as the optimizer with weight decay to be 0.0001. The learning rate is set 0.0001 at first, then for every 10 epochs, it drops by a factor of 10. Our network SalSAC is implemented using Pytorch and trained with one Nvidia 1080Ti GPU.

More details The length of the clip is randomly picked from each video which ranges from 1 to 60 frames. This is for keeping the training and testing consistency due to different lengths of the videos. During the training phase, the three multi-level attention maps are randomly shuffled at each iteration, while for the testing part they are always fixed to their original order. We use the data augmentation strategies of cropping, randomly horizontal flipping and randomly adjusting brightness during both the pre-training and training processes. More specifically, the frames of the DHF1K are resized to 160×288 and for the Hollywood-2 and UCF-sports datasets, the frames are resized to 224×224.

Table 1: Ablation studies on the validation set of DHF1K.

Factors	Model Variants of SalSAC	AUC-J	NSS	CC	SIM	s-AUC
Baseline	w/o Atten. & ConvLSTM, w/ single level feature & pre-train	0.881	2.408	0.446	0.335	0.676
Full Settings	-	**0.898**	**2.624**	**0.480**	**0.364**	**0.729**
Attention	w/o attention	0.894	2.580	0.472	0.357	0.698
	w/ a single-scale attention	0.896	2.600	0.475	0.362	0.700
	w/ unshuffled multi-scale attention	0.896	2.613	0.477	0.361	0.700
ConvLSTM	w/o ConvLSTM	0.892	2.461	0.455	0.345	0.682
	w/ traditional ConvLSTM	0.894	2.547	0.466	0.346	0.702
	w/ correlation-based ConvLSTM v1	0.896	2.558	0.469	0.351	0.709
	w/ correlation-based ConvLSTM v2	0.897	2.588	0.473	0.357	0.696
Training	w/o pre-train	0.890	2.594	0.468	0.352	0.688

Ablation Studies

To validate the design of our network SalSAC, we explore some model variants on the validation set of the DHF1K. The model variants and their performance are reported in Table 1. The results clearly verify the effectiveness of all important design features in SalSAC.

Impact of attention The attention mechanism obviously does improve the saliency prediction performance. Note that, we calculate the attention map of the feature extracted from the last convolutional layer as the single-scale attention. With the help of multi-scale attention and shuffling operation, the "full-settings" SalSAC achieves the best performance over all other settings for attention prediction.

Impact of ConvLSTM We can see that there is a sudden drop on NSS (2.624 to 2.461) when the ConvLSTM layer is removed from SalSAC, which shows that the ConvLSTM can make use of the temporal information to help the saliency prediction.

In addition, to find a best way to balance the influence of the current frame and its preceding one, we test three variants of the ConvLSTM layer, namely v1, v2 and the proposed correlation-base ConvLSTM used by SalSAC. For v1, we use the traditional ConvLSTM but change the final output to:

$$C_t = \lambda C_t^{trad} + (1-\lambda)X_t, \tag{14}$$

where C_t^{trad} denotes the output of traditional ConvLSTM, which means we only balance the output of the ConvLSTM layer. For v2, we change Eq. (7) for the proposed correlation-based ConvLSTM to:

$$C_t = F_t \odot C_{t-1} + I_t \odot G_t, \tag{15}$$

which means we only balance the input of the ConvLSTM layer.

From Table 1, we observe that all kinds of balancing strategies can help enhance the performance on all the metrics, except for a little drop on the s-AUC of v2 (0.702 to 0.696) compared with the traditional ConvLSTM. By balancing both the input and the output of the ConvLSTM, the "full-settings" SalSAC achieves the best performance in terms of all five evaluation metrics.

Impact of pre-training It is observed that without pre-training on the static image dataset SALICON, SalSAC gets the lowest AUC-J score (0.890) across all the model variants and has worse performance than SalSAC with pre-training ("full settings") on all metrics, thus the pre-training on the image saliency dataset does help enhance the quality of extracting static spatial information.

Table 2: Sizes (MB) of the parameters of all the models.

Name	Size	Name	Size	Name	Size
STSConvNet	315	DeepVS	344	ACLNet	250
SalEMA	364	STRA-Net	641	SalSAC	93.5

Comparison Tests

We compare our network SalSAC with many existing state-of-the-art competitors including STRA-Net (Lai et al. 2019), ACLNet (Wang et al. 2018a), SalEMA (Linardos et al. 2019), DeepVS (Jiang et al. 2018) and STSConvNet (Bak et al. 2017), which are all dynamic saliency prediction models. We specially note that SalSAC uses smaller number of parameters than most of the competitor models, as shown in Table 2. STRA-Net uses the most parameters which is about 6.8 times of our SalSAC. Performance results of all these methods on the DHF1K test dataset, Hollywood-2 dataset and UCF-sports datasets are reported in Table 3.

DHF1K For this dataset it is observed that our SalSAC achieves the best scores in four (NSS, CC and s-AUC, and having the same best score on AUC-J as STRA-NetAUC-J) out of the total five metrics. Figure 6 shows the visualization of a sample from the DHF1K validation set and its attention and saliency prediction results by SalSAC, where we can see that the attention maps are very consistent with the final predictions.

Hollywood-2 SalSAC achieves similar performance as STRA-Net and performs better than other models on the Hollywood-2 dataset (the best AUC-J and CC scores are obtained by SalSAC, NSS and s-AUC by STRA-Net and SIM

Table 3: Comparison results on the DHF1K, Hollywood-2 and UCF-sports datasets. The best scores are shown in bold.

Dataset / Model	DHF1K					Hollywood-2					UCF-sports				
	AUC-J	NSS	CC	SIM	s-AUC	AUC-J	NSS	CC	SIM	s-AUC	AUC-J	NSS	CC	SIM	s-AUC
STSConvNet	0.834	1.632	0.325	0.197	0.581	0.863	1.748	0.382	0.276	0.710	0.832	1.753	0.343	0.264	0.685
DeepVS	0.856	1.911	0.344	0.256	0.583	0.887	2.313	0.446	0.356	0.693	0.870	2.089	0.405	0.321	0.691
ACLNet	0.890	2.354	0.434	0.315	0.601	0.913	3.086	0.623	**0.542**	0.757	0.897	2.567	0.510	0.406	0.744
SalEMA	0.890	2.574	0.449	**0.466**	0.667	0.919	3.186	0.613	0.487	0.708	0.906	2.638	0.544	0.431	0.740
STRA-Net	0.895	2.558	0.458	0.355	0.663	0.923	**3.478**	0.662	0.536	**0.774**	0.910	3.018	0.593	0.479	0.751
SalSAC	**0.896**	**2.673**	**0.479**	0.357	**0.697**	**0.931**	3.356	**0.670**	0.529	0.712	**0.926**	**3.523**	**0.671**	**0.534**	**0.806**

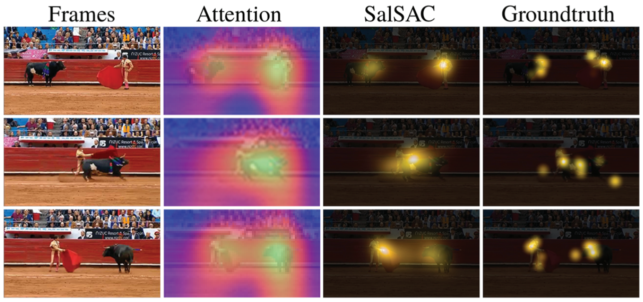

Figure 6: Visualization of the results of a sample from the DHF1K validation set by SalSAC.

by ACLNet). Figure 7 visualizes the testing results of a complicated case in the test set of Hollywood-2 by SalSAC and STRA-Net, which indicates that SalSAC can better handle dynamic changes.

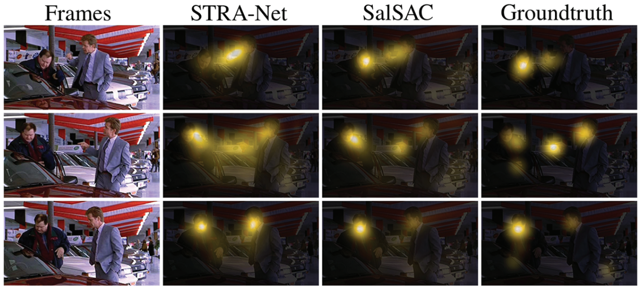

Figure 7: Visualization of the results of a testing sample from the Hollywood-2 testing set, in which SalSAC captures the man's hand better than STRA-Net.

UCF-sports For the UCF-sports dataset, SalSAC clearly achieves the best performance in all five metrics. The visualization result of a sample from the UCF-sports testing set is shown in Figure 8.

More discussions Compared with the Hollywood-2 and UCF-sports datasets, the DHF1K dataset is indeed more complicated due to the free-viewing design. On the other hand, the closer we get to the real human fixation, more helpful the free-viewing datasets will be in the future. It is also observed that the optical flow is not necessary for predicting the video saliency, i.e., the motion information can be seen as part of the temporal information. Moreover, not all the moving objects will arouse interests of human due to the limitation and concentration of human attention. Thus the optical flow could bring noise and more computations for the prediction.

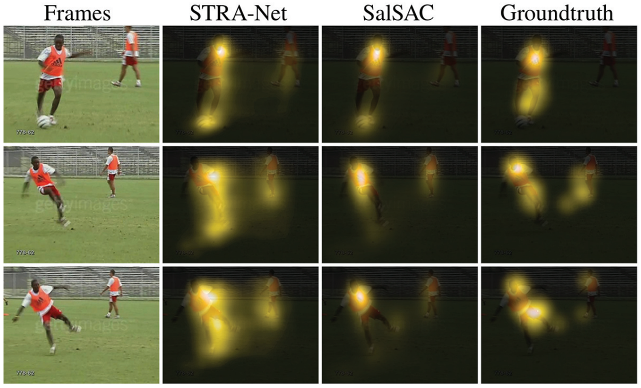

Figure 8: Visualization of the testing results of a sample from the UCF-sports testing set, in which SalSAC produces more concentrated saliency compared with STRA-Net.

Conclusion

In this paper we have proposed and investigated a new CNN-LSTM-Attention architecture "SalSAC" for video saliency prediction, which includes two novel components: a shuffled attention module and a correlation-based ConvLSTM layer. The shuffled attention module preserves the attention consistency as the network goes deeper, while the correlation-based ConvLSTM is used for effectively and adaptively balancing the importance of the current frame and its preceding one. Experiments performed on the DHF1K, Hollywood-2 and UCF-sports datasets also have demonstrated that our SalSAC can outperform many existing state-of-the-art methods by using similar number of or less parameters.

Acknowledgments. This work was supported, in part, by NSFC under grants U1803264, 61672376, 41775008, and 61702275.

References

Bak, C.; Kocak, A.; Erdem, E.; and Erdem, A. 2017. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia* 20(7):1688–1698.

Borji, A., and Itti, L. 2012. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1):185–207.

Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; and Durand, F. 2018. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(3):740–757.

Cheng, H.-T.; Chao, C.-H.; Dong, J.-D.; Wen, H.-K.; Liu, T.-L.; and Sun, M. 2018. Cube padding for weakly-supervised saliency prediction in 360 videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1420–1429.

Cornia, M.; Baraldi, L.; Serra, G.; and Cucchiara, R. 2016. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 3488–3493. IEEE.

Cornia, M.; Baraldi, L.; Serra, G.; and Cucchiara, R. 2018. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing* 27(10):5142–5154.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2758–2766.

Gao, D., and Vasconcelos, N. 2005. Discriminant saliency for visual recognition from cluttered scenes. In *Advances in Neural Information Processing Systems*, 481–488.

Gao, D.; Mahadevan, V.; and Vasconcelos, N. 2008. The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in Neural Information Processing Systems*, 497–504.

Gorji, S., and Clark, J. J. 2018. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7501–7511.

Guo, C., and Zhang, L. 2009. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing* 19(1):185–198.

Guo, H.; Zheng, K.; Fan, X.; Yu, H.; and Wang, S. 2019. Visual attention consistency under image transforms for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 729–739.

Hou, X., and Zhang, L. 2007. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8. Ieee.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1072–1080.

Jiang, L.; Xu, M.; Liu, T.; Qiao, M.; and Wang, Z. 2018. Deepvs: A deep learning based video saliency prediction approach. In *European Conference on Computer Vision (ECCV)*, 602–617.

Jiang, L.; Wang, Z.; Xu, M.; and Wang, Z. 2019. Image saliency prediction in transformed domain: A deep complex neural network method. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2106–2113. IEEE.

Lai, Q.; Wang, W.; Sun, H.; and Shen, J. 2019. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*.

Leifman, G.; Rudoy, D.; Swedish, T.; Bayro-Corrochano, E.; and Raskar, R. 2017. Learning gaze transitions from depth to improve video saliency estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 1698–1707.

Linardos, P.; Mohedano, E.; Nieto, J. J.; O'Connor, N. E.; Giro-i Nieto, X.; and McGuinness, K. 2019. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*.

Luo, Z.; Peng, B.; Huang, D.-A.; Alahi, A.; and Fei-Fei, L. 2017. Unsupervised learning of long-term motion dynamics for videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2203–2212.

Marszałek, M.; Laptev, I.; and Schmid, C. 2009. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2929–2936. IEEE Computer Society.

Mathe, S., and Sminchisescu, C. 2014. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(7):1408–1424.

Rudoy, D.; Goldman, D. B.; Shechtman, E.; and Zelnik-Manor, L. 2013. Learning video saliency from human gaze using candidate selection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1147–1154.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 802–810.

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

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8174–8182.

Wang, W.; Shen, J.; Guo, F.; Cheng, M.-M.; and Borji, A. 2018a. Revisiting video saliency: A large-scale benchmark and a new model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4894–4903.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7794–7803.