Spatio-Temporal Self-Attention Network for Video Saliency Prediction

Ziqiang Wang, Zhi Liu, *Senior Member, IEEE,* Gongyang Li, *Member, IEEE,* Yang Wang, Tianhong Zhang, Lihua Xu, and Jijun Wang

***Abstract*—3D convolutional neural networks have achieved promising results for video tasks in computer vision, including video saliency prediction that is explored in this paper. However, 3D convolution encodes visual representation merely on fixed local spacetime according to its kernel size, while human attention is always attracted by relational visual features at different time. To overcome this limitation, we propose a novel Spatio-Temporal Self-Attention 3D Network (STSANet) for video saliency prediction, in which multiple Spatio-Temporal Self-Attention (STSA) modules are employed at different levels of 3D convolutional backbone to directly capture long-range relations between spatio-temporal features of different time steps. Besides, we propose an Attentional Multi-Scale Fusion (AMSF) module to integrate multi-level features with the perception of context in semantic and spatio-temporal subspaces. Extensive experiments demonstrate the contributions of key components of our method, and the results on DHF1K, Hollywood-2, UCF, and DIEM benchmark datasets clearly prove the superiority of the proposed model compared with all state-of-the-art models.**

***Index Terms*—Video saliency prediction, self-attention, spatio-temporal feature, attention mechanism.**

I. Introduction

HUMANS have a fantastic capability of localizing the most important area in the visual field promptly, named as visual attention mechanism, which facilitates the processing of diverse visual information. In computer vision, modeling the visual attention mechanism is a fundamental research topic, named *saliency prediction* (SP) or *fixation prediction*, which aims to deduce the visual saliency degree of each region in images, expressed in the form of a saliency map (as shown in Fig. 1). SP has been widely applied to various computer vision tasks, such as image captioning [1], photo cropping [2], object segmentation [3]–[6], video compression [7], *etc*.

This work was supported in part by the National Natural Science Foundation of China under Grants 62171269 and 82171544, in part by the Science and Technology Commission of Shanghai Municipality under Grant 21S31903100, and in part by the China Scholarship Council under Grant 202006890079. *(Corresponding author: Zhi Liu.)*

Ziqiang Wang, Zhi Liu, and Gongyang Li are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. Gongyang Li is also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wziqiang@shu.edu.cn; liuzhisjtu@163.com; ligongyang@shu.edu.cn).

Yang Wang is with the Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada (e-mail: ywang@cs.umanitoba.ca).

Tianhong Zhang, Lihua Xu, and Jijun Wang are with Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, Shanghai 200030, China (e-mail: zhang_tianhong@126.com; dr_xulihua@163.com; dr_wangjijun@126.com).

Fig. 1. Visualization of results of the proposed STSANet on sampling frames in two videos.

Traditional solutions for SP leverage hand-crafted features, including low-level cues such as color, orientation, and intensity, as well as high-level contents such as persons and objects [8]–[10]. In the recent years, with the renaissance of neural networks and advance of high-quality datasets and benchmarks [11], [12], deep learning based SP has obtained substantial progress, and the computational models have performed close to human inter-observer level on static datasets. Compared with image SP, video saliency prediction (VSP) is more difficult and develops relatively slowly.

Videos contain spatial information in frames and temporal information between frames. In videos, human attention is not only guided by low-level cues and semantic context in a single frame, but also by relations of features in frames. For example, in a video clip, the same object moving in a scene usually attracts visual attention, as shown in Fig. 1. Consequently, it is crucial for video saliency prediction (VSP) to synchronously exploit spatial and temporal information. Directly using image SP models for VSP triggers poor performance, due to ignoring the temporal information. Recently, deep learning models for VSP have emerged and outperformed the traditional models [13], [14]. Some VSP models [15]–[17] employ both RGB and optical flow backbones to encode appearance and motion information, respectively, and merge them for dynamic saliency inference. However, the motion stream merely considers temporal relations between adjacent frames. This limitation is alleviated by LSTM-based models [18]–[23], which adopt LSTMs to capture temporal long-term relations in a video.

These models utilize convolutional networks and LSTMs to deal with spatial and temporal information separately, therefore, they are unable to synchronously exploit spatial and temporal information, which is instrumental in VSP. To this end, some models [24]–[26] use 3D convolutional networks to process spatio-temporal information jointly, and have shown progressive performance.

In this paper, we further explore VSP based on 3D convolutional networks. Although 3D convolution can encode spatio-temporal information collectively, it processes fixed local spacetime and fails to capture long-range relations between visual features at different time. To remedy this deficiency, we propose a novel Spatio-Temporal Self-Attention (STSA) module to directly learn long-range spatio-temporal dependencies, which is inspired by the self-attention mechanism [27]. In the STSA module, spatio-temporal features are split along the temporal channel, and features at different time steps are separately transformed to an embedding space by embedding convolutional layers. Afterwards, long-range spatio-temporal interactions are built by dot-product attention calculation between features at different time steps.

Based on the STSA module, we propose a Spatio-Temporal Self-Attention Network (STSANet) for VSP. The backbone of our model is a 3D fully convolutional network from S3D [28] pre-trained on Kinetics dataset [29]. We draw four branches from different levels of the backbone, and employ STSA modules on them, respectively, to produce local and global spatio-temporal context at multiple levels. Since the spatio-temporal and semantic gaps exist in the outputs from the four STSA modules, direct fusion, such as summation and concatenation, is not powerful enough to process them well. Instead, we design an Attentional Multi-Scale Fusion (AMSF) module to integrate the multi-level features. The AMSF module consists of an attentional weighting operation and a spatio-temporal multi-scale structure, which are used to alleviate the semantic and spatio-temporal gaps, respectively, during feature fusion.

Overall, our main contributions can be summarized as follows:

- We propose a Spatio-Temporal Self-Attention Network (STSANet) for video saliency prediction, in which self-attention mechanism is integrated into the 3D convolutional network. This integration helps our model achieve superior performance compared with all state-of-the-art models on multiple benchmark datasets.
- We propose a Spatio-Temporal Self-Attention (STSA) module to capture long-range dependencies between time steps of temporal channel. The STSA module is employed at multiple levels of 3D convolutional network to complement the locality of 3D convolution and to produce local and global spatio-temporal representations for video saliency prediction.
- We propose an Attention Multi-Scale Fusion (AMSF) module to fuse spatio-temporal features from STSA modules arranged at different levels of backbone. The AMSF module perceives contextual contents in semantic and spatio-temporal subspaces, and narrows semantic and spatio-temporal gaps during saliency feature fusion.

II. Related Work

In computer vision, saliency models can be divided into two types: *saliency prediction* (SP) and *salient object detection* (SOD). SP aims to predict visual saliency degree of each region in images, while the task objective of SOD is highlighting salient object regions. Numerous computational models have been proposed for image SOD [30]–[36] and video SOD [37]–[40] in recent decades. Also, SP contains image SP and video SP, and this work focuses on video SP (VSP). In this section, we briefly review the SP models and the network design related to our work.

A. Image Saliency Prediction

The earliest work for SP, proposed by Itti *et al.* [8], captured hand-crafted futures in color, intensity, and orientation channels, respectively, and combined them for saliency results. After this work, SP models based on hand-crafted features emerged consecutively [9], [10], [41]. Recently, with the advance of deep learning, data-driven models have outperformed traditional models. *Ensembles of Deep Networks* (eDN) [42] is among the first to apply neural networks to SP task. This model combined features generated from a lot of convolutional layers using a linear classifier. *Deep Gaze I* [43] employed the off-the-shelf features from deep convolutional neural network (CNN) trained on ImageNet [44] for SP, and *Salicon* [45] further fine-tuned pre-trained VGGNet [46] with SP datasets, which verified the effectiveness of transfer learning for SP. After that, a variety of deep SP models [45], [47]–[59] based on VGGNet [46], ResNet [60], DenseNet [61], and NASNet [62] have been proposed successively. At present, SP models have performed close to human inter-observer level on image SP datasets.

B. Video Saliency Prediction

Traditional models related to VSP mainly explored dynamic scene viewing using static and motion information [13], [14], however, hand-crafted spatio-temporal features were not powerful enough for modeling dynamic saliency. A number of VSP models based on deep learning have emerged in recent years.

1) Two-Stream Methods: Bak *et al.* [15] proposed a two-stream network, which employed two convolutional backbones, with RGB images and optical flow maps as inputs respectively, to extract spatio-temporal information and fuse their outputs for saliency inference. Wu *et al.* [16] and Zhang *et al.* [17] further explored the two-stream structure and investigated fusion methods to improve performance.

2) LSTM-based Methods: However, the optical flow network merely considers temporal relations between adjacent frames, hence LSTM is frequently adopted to extend temporal perception. Gorji *et al.* [63] employed multi-stream LSTMs that merged with static SP model for VSP. DeepVS [18] leveraged object and motion networks to extract intra-frame saliency information, and modeled temporal correlation between frames by convLSTMs. Besides, ACLNet [19] designed a neural attention module in a CNN-LSTM structure, which

was supervised with image SP datasets. STRA-Net [20] employed dense residual cross connection to enrich interactions between motion stream and appearance stream during feature extraction, and incorporated an attention mechanism to enhance the spatio-temporal information. SalEMA [64] modified the static SP model by adding a simple exponential moving average for feature fusion in temporal domain, resulting in a low-parametric architecture. Later, Zhang *et al.* [22] utilized spatial and channel attention to select and re-weight spatio-temporal information, and employed an attentive convLSTM to model relations between frames. SalSAC [23] designed a correlation-based convLSTM for VSP, in which adjacent frames were weighted according to similarity between them.

3) 3D Convolution-based Methods: Differently, TASED-Net [24] proposed a 3D fully-convolutional encoder-decoder network for VSP and achieved promising performance. Compared with LSTM-based architectures, which process spatial and temporal information separately, a 3D network encodes and decodes spatio-temporal information in a collective way. Moreover, ViNet [26] designed a UNet-like encoder-decoder network based on a 3D backbone, in which features from multiple levels were upsampled with trilinear upsampling and concatenated along the temporal channel. In HD2S [25], multi-level features from a 3D encoder are separately decoded to obtain single-channel conspicuity maps, and integrated them for saliency results. Besides, TSFP-Net [65] employs a feature pyramid structure with top-down feature integration on a 3D convolutional backbone, and combines the multi-level spatio-temporal features to reason the saliency result for a video frame. In our work, we further explore 3D neural network for VSP. The 3D convolution handles a local spatio-temporal space at a time, therefore, 3D convolutional networks lack the ability to directly model long-range spatio-temporal relations. Our model addresses this shortage by employing STSA modules, which build long-range immediate interactions between spatio-temporal features of different time steps, at multiple levels of the 3D backbone.

C. Self-Attention

Self-attention mechanism has been an important issue after [27], in which a lot of self-attention mechanisms were employed to learn text representations for natural language tasks. In self-attention mechanism, input tokens are linearly transformed as queries, keys, and values, respectively, in embedding layers, and then long-range relations between tokens of input sequence are calculated by dot-product attention.

In computer vision, Wang *et al.* [66] proposed the non-local neural network that is a classical implementation of self-attention mechanism in vision tasks. Besides, Oh *et al.* [67] designed memory networks based on self-attention mechanism for video object segmentation. In that model, the memory was updated according to the dependencies between current frame and past frames computed in the form of self-attention mechanism. Moreover, Wang *et al.* [68] adopted the memory and self-attention mechanism for video semantic segmentation. These two models both used 2D deep convolutional networks to encode the frames one by one, and then transformed features

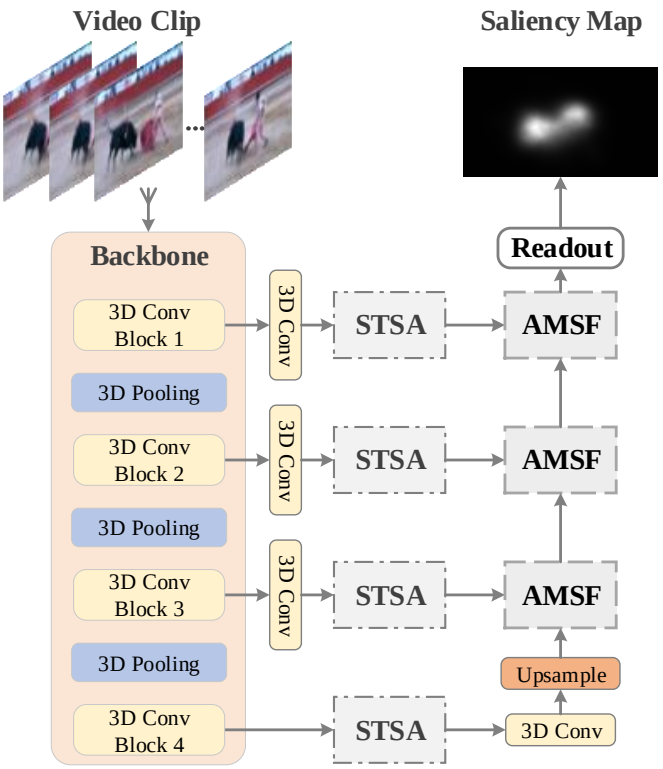


Fig. 2. An overview of the proposed model. Our model contains three main components: 3D convolutional encoder, Spatio-Temporal Self-Attention (STSA) module, and Attentional Multi-Scale Fusion (AMSF) module. A video clip including consecutive frames is used as the input of 3D encoder to generate hierarchical spatio-temporal visual features. Four STSA modules are employed after different 3D convolutional blocks to capture long-range spatio-temporal dependencies between time steps at multiple levels. Afterwards, multi-level features are fused by the AMSF modules to generate video saliency results.

of different frames to an embedding space by embedding layers for attention calculation. Therefore, the amount of computer memory occupied is considerable.

Analogously, COSNet [69] performs a co-attention mechanism [70], [71] for video object segmentation task. This model uses 2D convolutional networks to separately generate embeddings from current frame and a group of reference frames which are uniformly sampled from a video. Subsequently, the co-attention part computes attention summaries from current frame and each reference frame in pairs to capture inter-frame consistency, and then integrates the average of these attention summaries to the embedding of current frame to enhance the capacity of inferring target object.

In our VSP model, we adopt a 3D convolutional network as backbone to directly extract spatio-temporal information from multiple frames, and employ self-attention mechanism to capture long-range dependencies between spatio-temporal features of different time steps. Meanwhile, we further adopt other strategies to compress our model as described in Sec. III-B.

III. The Proposed Model

In this section, we illustrate the proposed Spatio-Temporal Self-Attention Network (STSANet). In Sec. III-A, an overview of our model is given. In Sec. III-B, we describe the proposed

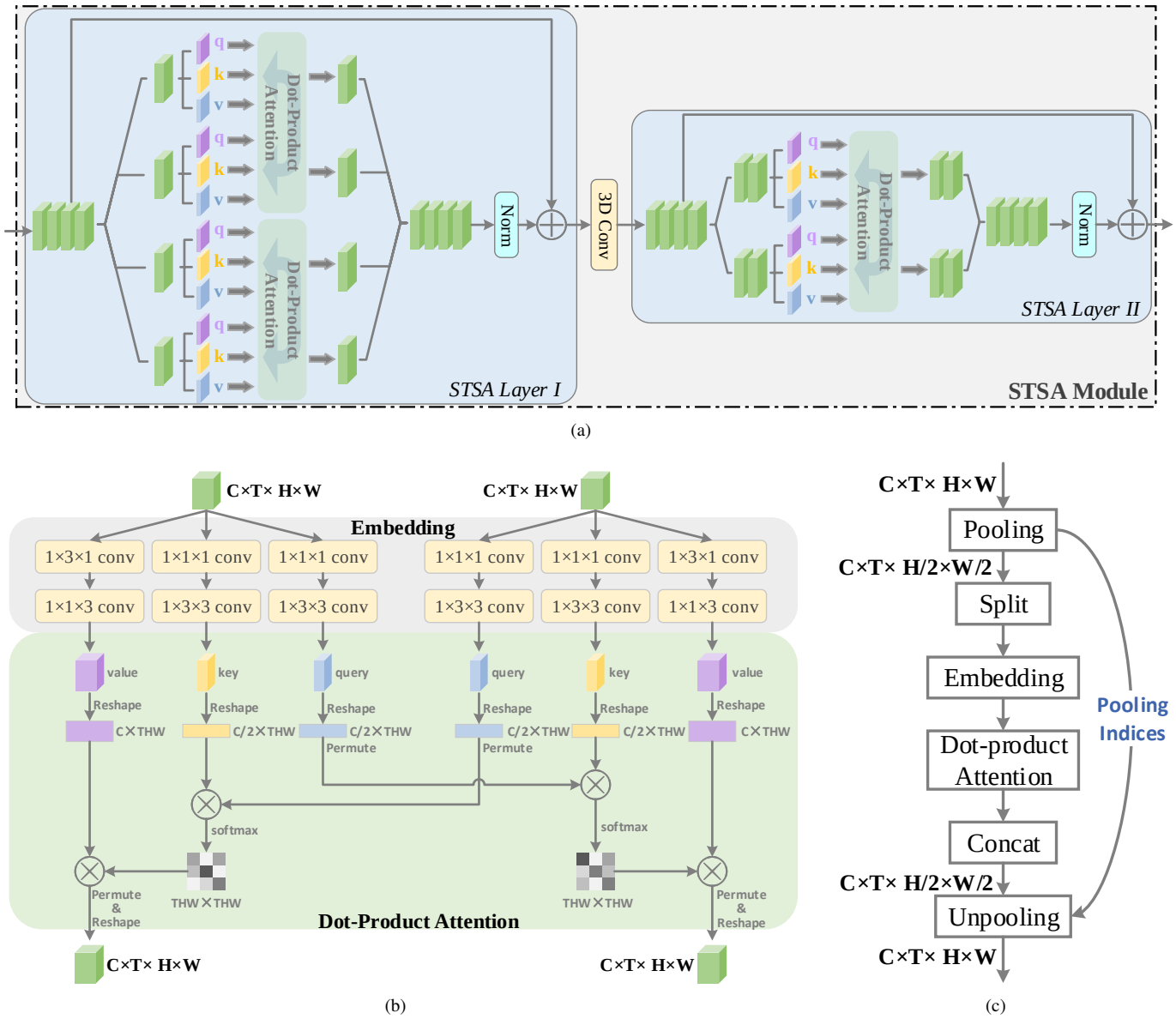


Fig. 3. (a) Spatio-Temporal Self-Attention (STSA) module. (b) An detailed illustration of the embedding and dot-product attention in STSA layers. (c) Spatial bottleneck structure employed on the STSA layers after $Conv\_Block\_1$.

Spatio-Temporal Self-Attention (STSA) module in detail. In Sec. III-C, we introduce the Attentional Multi-Scale Fusion module. In Sec. III-D, we provide the detailed information of the modules. In Sec. III-E, we present the supervision manner and the loss function.

A. Architecture Overview

The architecture of the proposed model is described in Fig. 2. We utilize the fully-convolutional portion of S3D network [28] pre-trained on the Kinetics dataset [29] as the backbone. The backbone is composed of 3D convolutional layers, which have the capability of encoding spatio-temporal information. The input of backbone is a video clip consisting of T consecutive frames, set to 32 in our model.

CNNs are able to encode hierarchical representations, including low-level cues like color contrast, and high-level semantic information like persons or objects, all of which are of value to saliency. Analogously, from shallow to deep layers in 3D CNNs, the outputs correspond to low- and high-level features, respectively. Accordingly, we employ four decoding branches for the output from four 3D convolutional blocks of the backbone, which are separated by three 3D max pooling layers. Each branch has a Spatio-Temporal Self-Attention (STSA) module to directly create global spatio-temporal context at each level. After that, the features from four branches are integrated by Attentional Multi-Scale Fusion (AMSF) modules in a top-down pathway. Lastly, in the readout module, the feature maps are upsampled to the original video resolution, and the features at all time steps are aggregated by 3D convolution to obtain the final saliency map.

B. Spatio-Temporal Self-Attention Module

Convolutional operator updates a position of a feature map by aggregating information in a local window, thereby failing to capture long-range relations between visual features at different time, which play an important role in VSP. To complement the locality of convolutional operator, we propose a Spatio-Temporal Self-Attention (STSA) module that further updates each position at each time step by aggregating global relations with spatio-temporal features at the other time steps.

In Fig. 2, after the backbone, the temporal channel dimensions of multi-level features are uniformly compressed to 4 by 3D convolutional layers. Subsequently, the input features to STSA modules can be denoted as $\mathbf{F} \in \mathbb{R}^{C\times 4\times H\times W}$, where C, 4, H, and W are the dimensions of the semantic channel, temporal channel, height and width, respectively. As described in Fig. 3(a), the STSA module cascades two STSA layers and a 3D convolutional layer in the middle. In the *STSA layer I*, the input feature $\mathbf{F}$ is first split to four sub-features $\{\mathbf{F}^1, \mathbf{F}^2, \mathbf{F}^3, \mathbf{F}^4\} \in \mathbb{R}^{C\times 1\times H\times W}$ along the temporal channel, and each represents spatio-temporal information at different time of a video clip. Then, we use convolutional embedding layers to transform the sub-features to queries, keys, and values, expressed as $\mathbf{F}_q^t \in \mathbb{R}^{C/2\times 1\times H\times W}$, $\mathbf{F}_k^t \in \mathbb{R}^{C/2\times 1\times H\times W}$, and $\mathbf{F}_v^t \in \mathbb{R}^{C\times 1\times H\times W}, t = 1, 2, 3, 4$, respectively. After that, we implement the dot-product attention between sub-features at different time steps to directly capture long-range relations between spatio-temporal features of different time steps. Specifically, this attention mechanism in *STSA layer I* can be formulated as follows:

$$\mathrm{DA}(\mathbf{F}_q^2, \mathbf{F}_k^1, \mathbf{F}_v^1) = Softmax((\mathbf{F}_q^2)^{\mathrm{T}}\mathbf{F}_k^1)(\mathbf{F}_v^1)^{\mathrm{T}}, \quad (1)$$

$$\mathrm{DA}(\mathbf{F}_q^1, \mathbf{F}_k^2, \mathbf{F}_v^2) = Softmax((\mathbf{F}_q^1)^{\mathrm{T}}\mathbf{F}_k^2)(\mathbf{F}_v^2)^{\mathrm{T}}, \quad (2)$$

$$\mathrm{DA}(\mathbf{F}_q^4, \mathbf{F}_k^3, \mathbf{F}_v^3) = Softmax((\mathbf{F}_q^4)^{\mathrm{T}}\mathbf{F}_k^3)(\mathbf{F}_v^3)^{\mathrm{T}}, \quad (3)$$

$$\mathrm{DA}(\mathbf{F}_q^3, \mathbf{F}_k^4, \mathbf{F}_v^4) = Softmax((\mathbf{F}_q^3)^{\mathrm{T}}\mathbf{F}_k^4)(\mathbf{F}_v^4)^{\mathrm{T}}, \quad (4)$$

where $\mathrm{DA}(\cdot)$ is the dot-product attention calculation as depicted in Fig. 3(b), and $Softmax(\cdot)$ indicates the softmax activation function.

The queries, keys and values after embedding layers are first reshaped to $C/2 \times THW$ or $C \times THW$, $T = 1$. Afterwards, the dot-product attention calculation is carried out between $\mathbf{F}^1$ and $\mathbf{F}^2$, $\mathbf{F}^3$ and $\mathbf{F}^4$, respectively. More concretely, as shown in Eq. 1, the similarity matrix is obtained by matrix multiplication of $\mathbf{F}_q^2$ and $\mathbf{F}_k^1$, and normalized by the softmax function, to represent the spatio-temporal relation between $\mathbf{F}^2$ and $\mathbf{F}^1$. After that, the relation is integrated to $\mathbf{F}^1$ by multiplying the similarity matrix and $\mathbf{F}_v^1$. In the same way, the value of $\mathbf{F}^2$ is enhanced by the dot-product attention operation as described in Eq. 2. Meanwhile, the dot-product attention calculation between $\mathbf{F}^3$ and $\mathbf{F}^4$, as formulated in Eq. 3 and Eq. 4, is the same as that between $\mathbf{F}^1$ and $\mathbf{F}^2$. After the two sets of attention calculations, the four outputs are reassembled along the temporal channel after being restored to $C \times 4 \times H \times W$, and an identity shortcut is added as follows:

$$\begin{aligned}\hat{\mathbf{F}} =& \mathbf{F} \oplus \mathrm{Norm}([\mathrm{DA}(\mathbf{F}_q^2, \mathbf{F}_k^1, \mathbf{F}_v^1), \mathrm{DA}(\mathbf{F}_q^1, \mathbf{F}_k^2, \mathbf{F}_v^2),\\ & \mathrm{DA}(\mathbf{F}_q^4, \mathbf{F}_k^3, \mathbf{F}_v^3), \mathrm{DA}(\mathbf{F}_q^3, \mathbf{F}_k^4, \mathbf{F}_v^4)]) \end{aligned} \quad (5)$$

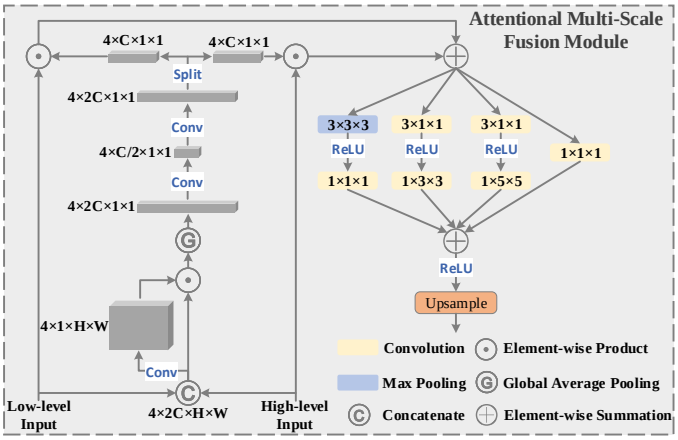


Fig. 4. Attentional Multi-Scale Fusion (AMSF) module.

where $\oplus$ denotes the element-wise summation, $\mathrm{Norm}(\cdot)$ denotes the layer normalization [72], and $[\cdot,\cdot]$ donates the concatenation.

Our elaborate design on these two STSA layers achieves that each of the four features at different time steps ($\mathbf{F}^1$, $\mathbf{F}^2$, $\mathbf{F}^3$ and $\mathbf{F}^4$) is directly updated by long-range spatio-temporal relations with the other three through only three groups of non-overlapping dot-product attention calculations. For instance, $\mathbf{F}^1$ is updated by long-range relation with $\mathbf{F}^2$ in aforementioned *STSA layer I*, and updated by long-range relations with $\mathbf{F}^3$ and $\mathbf{F}^4$ in *STSA layer II*. Meanwhile, $\mathbf{F}^2$, $\mathbf{F}^3$ and $\mathbf{F}^4$ are updated similarly. Specifically, after *STSA layer I* detailed above, *STSA layer II* splits its input feature to two sub-features $\{\mathbf{F}^{12}, \mathbf{F}^{34}\} \in \mathbb{R}^{C\times 2\times H\times W}$ along the temporal channel, and then separately transforms them to queries, keys, and values by embedding layers. Subsequently, the dot-product attention can be expressed in the following formulas:

$$\mathrm{DA}(\mathbf{F}_q^{34}, \mathbf{F}_k^{12}, \mathbf{F}_v^{12}) = Softmax((\mathbf{F}_q^{34})^{\mathrm{T}}\mathbf{F}_k^{12})(\mathbf{F}_v^{12})^{\mathrm{T}}, \quad (6)$$

$$\mathrm{DA}(\mathbf{F}_q^{12}, \mathbf{F}_k^{34}, \mathbf{F}_v^{34}) = Softmax((\mathbf{F}_q^{12})^{\mathrm{T}}\mathbf{F}_k^{34})(\mathbf{F}_v^{34})^{\mathrm{T}}. \quad (7)$$

And the identity shortcut is added as follows:

$$\hat{\mathbf{F}} = \mathbf{F} \oplus \mathrm{Norm}([\mathrm{DA}(\mathbf{F}_q^{34}, \mathbf{F}_k^{12}, \mathbf{F}_v^{12}), \mathrm{DA}(\mathbf{F}_q^{12}, \mathbf{F}_k^{34}, \mathbf{F}_v^{34}]). \quad (8)$$

Notably, we employ a convolutional layer between the two STSA layers to capture more local representations. The convolutional layer halves the dimension of feature in the semantic channel, which is the input feature of the second STSA layer, to suppress memory consumption. Consequently, our STSA module captures local and global features alternately for complementarity.

The STSA modules described above are employed on four branches from the backbone. They enhance the relational visual features between different time steps at multiple levels, which are instrumental in localizing dynamic saliency.

C. Attentional Multi-Scale Fusion

After the STSA modules, the outputs from four branches are fused for saliency inference. The general method is the top-down feature fusion, in which the low-resolution feature is

upsampled by interpolation and integrated with high-resolution feature using element-wise addition or concatenation. In our model, the outputs from four STSA modules represent different contextual information in the temporal, semantic, and spatial subspaces. Direct addition fusion is not powerful enough for this case, since the information gap in different subspaces is not taken into account. To tackle this issue, we design an Attentional Multi-Scale Fusion (AMSF) module. As depicted in Fig. 4, this module can be divided into left and right parts. The attentional weighting operation deals with semantic gap, and the spatio-temporal multi-scale structure alleviates information gap in spatial and temporal subspaces.

We define the output of a pair of adjacent branches as $\{\mathbf{F}_h, \mathbf{F}_l\} \in \mathbb{R}^{C\times 4\times H\times W}$, representing high- and low-level features. Concretely, in the AMSF module, $\mathbf{F}_h$ and $\mathbf{F}_l$ are first concatenated along the semantic channel, and masked by an attention multidimensional matrix $\mathbf{W}_m \in \mathbb{R}^{1\times 4\times H\times W}$ that is generated from $[\mathbf{F}_h, \mathbf{F}_l]$ to enhance the important position and weaken the invalid position, which can be formulated as:

$$\mathbf{F}_M = \mathbf{W}_m \odot [\mathbf{F}_h, \mathbf{F}_l] = \sigma(\mathrm{Conv}([\mathbf{F}_h, \mathbf{F}_l])) \odot [\mathbf{F}_h, \mathbf{F}_l] \quad (9)$$

where σ indicates the sigmoid activation function, $\mathrm{Conv}(\cdot)$ indicates the 3D convolutional layer, and $\odot$ indicates the element-wise product. After that, inspired by the SE Block [73], we employ a global average pooling to squeeze the spatial subspace, and then obtain a semantic weight matrix $[\mathbf{W}_h, \mathbf{W}_l] \in \mathbb{R}^{C\times 4\times 1\times 1}$ with two convolutional layers as follows:

$$[\mathbf{W}_h, \mathbf{W}_l] = \sigma(\mathrm{Conv}(\mathit{Relu}(\mathrm{Norm}(\mathrm{Conv}(\mathrm{GAP}(\mathbf{F}_M)))))), \quad (10)$$

where $\mathit{Relu}(\cdot)$ is the rectified linear unit activation function, and $\mathrm{GAP}(\cdot)$ is the global average pooling. Finally, the weight matrix is split to two parts, and they are used to recalibrate $\mathbf{F}_h$ and $\mathbf{F}_l$ as folows:

$$\mathbf{F}_O = \mathbf{F}_h \odot \mathbf{W}_h \oplus \mathbf{F}_l \odot \mathbf{W}_l, \quad (11)$$

With this attention mechanism, $\mathbf{F}_h$ and $\mathbf{F}_l$ are selected in the semantic subspace with the perception of semantic relationship between the features from adjacent branches.

In addition, as shown in the right part of Fig. 4, for spatial and temporal subspaces, we design a spatio-temporal multi-scale structure, including three convolution branches and a pooling branch, after the attentional weighting operation in the AMSF module. Parallel convolutional layers with different spatial kernel sizes achieve adaptability to different spatial scales of features. Besides, the 3D convolution and the 3D pooling provide the perception of context in the temporal subspace for the fusion module.

To sum up, in the AMSF module, specific calculations with trainable parameters are deployed in three subspaces, to obtain the capability of perceiving and alleviating the spatio-temporal and semantic gaps for accurate saliency results.

D. Implementation Details

1) Spatio-Temporal Self-Attention Module: The output features, from $Conv\_Block\_1$, $Conv\_Block\_2$, $Conv\_Block\_3$, and $Conv\_Block\_4$ of backbone, have the temporal channel dimensions of 16, 16, 8, and 4, respectively. In order to reduce memory occupancy of STSA modules, the temporal channel dimensions of output features from $Conv\_Block\_1$, $Conv\_Block\_2$, and $Conv\_Block\_3$ are uniformly compressed to 4 by 3D convolutional layers, which are set as $4\times 1\times 1$ with temporal stride 4, $4\times 1\times 1$ with temporal stride 4, and $2\times 1\times 1$ with temporal stride 2, respectively. In the STSA module, as given in Fig. 3(b), for query and key embedding layers, we use a $1\times 1\times 1$ convolutional layer to compress the dimension of semantic channel, and a following $1\times 3\times 3$ convolutional layer for spatial information. For the value embedding layers, we adopt the asymmetric convolution (*i.e.,* a $1\times 3\times 1$ convolutional layer followed by a $1\times 1\times 3$ convolutional layer) to obtain spatial information without changing the dimension of semantic channel. These settings make the embedding layers of the STSA module cost-effective and able to capture spatial information.

Specially, in the STSA module on the branch from $Conv\_Block\_1$, we devise a spatial bottleneck structure (*i.e.,* a bottleneck structure established on the spatial subspace) for the STSA layers, because the output from $Conv\_Block\_1$ has a large resolution, which consumes a lot of memory during dot-product attention calculation. As shown in Fig. 3(c), the spatial size of input feature is first reduced by half using a $1\times 2\times 2$ pooling layer, and the pooling indices are temporarily stored. After splitting, embedding, dot-product attention, and concatenation, the feature is restored to the initial resolution using an unpooling layer with the pooling indices. By this means, the dot-product attention calculation is implemented in a bottleneck of small resolution, which greatly reduces the occupancy of memory.

2) Attentional Multi-Scale Fusion: In the attentional weighting operation, the spatial attention part generates the attention matrix $\mathbf{W}_m$ by: $\mathrm{Conv}(1\times 1\times 1) \rightarrow \mathit{Sigmoid}$. After that, the semantic channel attention part gets the weight matrix $[\mathbf{W}_h, \mathbf{W}_l]$ by: $\mathrm{GAP} \rightarrow \mathrm{Conv}(1\times 1\times 1) \rightarrow \mathit{Relu} \rightarrow \mathrm{Norm} \rightarrow \mathrm{Conv}(1\times 1\times 1) \rightarrow \mathit{Sigmoid}$. The spatio-temporal multi-scale structure is inception-like, and the detailed settings can be clearly found in the right part of Fig. 4.

E. Supervision and Loss Function

1) Supervision: The proposed model takes successive 32 frames from a video as input, and produces a saliency map. That is, our model predicts results for videos in a window-sliding manner. Due to the symmetry of our model along the temporal channel, we supervise model training with ground truth of middle frame in the video clip, *i.e.,* the 16-th frame of 32 frames. Therefore, to generate saliency maps for the first 15 frames and the last 16 frames of a video, we repeat the first frame and the last frame, respectively, to construct complete input clips.

$$\mathcal{L}(\mathbf{S}, \mathbf{G}) = \mathrm{KL}(\mathbf{S}, \mathbf{G}) - \mathrm{CC}(\mathbf{S}, \mathbf{G}), \quad (12)$$

where $\mathbf{S}$ and $\mathbf{G}$ are the predicted saliency map and the ground truth, respectively.

TABLE I
COMPARISON RESULTS ON THE TEST SETS OF DHF1K, HOLLYWOOD-2 AND UCF DATASETS. THE BEST TWO ARE MARKED BY RED AND BLUE, RESPECTIVELY.

Model	Input Size	FLOPs (G)	Model Size (MB)	Runtime (s)	DHF1K					Hollywood-2				UCF			
					CC ↑	NSS ↑	SIM ↑	AUC ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC ↑	CC ↑	NSS ↑	SIM ↑	AUC ↑
DeepVS [18]	448 × 448	819.5	344	0.050	0.344	1.911	0.256	0.856	0.583	0.446	2.313	0.356	0.887	0.405	2.089	0.321	0.870
ACLNet [19]	224 × 224	164.0	250	0.020	0.434	2.354	0.315	0.890	0.601	0.623	3.086	0.542	0.886	0.510	2.567	0.406	0.897
STRA-Net [20]	224 × 224	540.0	641	0.020	0.458	2.558	0.355	0.895	0.663	0.662	3.478	0.536	0.923	0.593	3.018	0.479	0.910
SalEMA [64]	256 × 192	40.0	364	0.010	0.449	2.574	**0.466**	0.890	0.667	0.613	3.186	0.487	0.919	0.544	2.638	0.431	0.906
TASED-Net [24]	384 × 224	91.5	82	0.060	0.470	2.667	0.361	0.895	0.712	0.646	3.302	0.507	0.918	0.582	2.920	0.469	0.899
Chen *et al.* [21]	320 × 256	371.3	437	0.080	0.476	2.685	0.353	0.900	0.680	0.661	3.804	0.537	0.928	0.599	3.406	0.494	0.917
SalSAC [23]	288 × 160	-	94	0.020	0.479	2.673	0.357	0.896	0.697	0.670	3.356	0.529	0.931	0.671	3.523	0.534	**0.926**
UNISAL [74]	384 × 224	14.4	16	0.009	0.490	2.776	0.390	0.901	0.691	0.673	3.901	0.542	0.934	0.644	3.381	0.523	0.918
HD2S [25]	192 × 128	-	116	0.030	0.503	2.812	**0.406**	0.908	0.700	0.670	3.352	0.551	0.936	0.604	3.114	0.507	0.904
ViNet [26]	384 × 224	115.0	124	0.016	0.511	2.872	0.381	0.908	**0.729**	0.693	3.730	0.550	0.930	0.673	3.620	0.522	0.924
TSFP-Net [65]	352 × 192	-	58	0.011	**0.517**	**2.966**	0.392	**0.912**	**0.723**	**0.711**	**3.910**	**0.571**	**0.936**	**0.685**	**3.698**	**0.561**	0.923
STSANet	384 × 224	193.3	643	0.035	**0.529**	**3.010**	0.383	**0.913**	**0.723**	**0.721**	**3.927**	**0.579**	**0.938**	**0.705**	**3.908**	**0.560**	**0.936**

TABLE II
COMPARISON RESULTS ON DIEM DATASET.

Model	Test Set	CC ↑	NSS ↑	SIM ↑	AUC ↑
DeepVS [18]	i	0.371	2.235	0.238	0.857
ACLNet [19]		0.396	2.368	0.277	0.881
STRA-Net [20]		0.408	2.452	0.306	0.870
Chen *et al.* [21]		0.490	2.346	0.396	0.889
STSANet		**0.625**	**2.618**	**0.505**	**0.901**
ViNet [26]	ii	0.626	2.470	0.483	0.898
TSFP-Net [65]		0.649	2.630	0.529	0.905
STSANet		**0.677**	**2.721**	**0.541**	**0.906**
STSANet	iii	**0.690**	**2.787**	**0.548**	**0.905**

2) Loss Function: In recent years, using the combination of saliency metrics as loss function is common and effective in static and dynamic saliency prediction models based on deep learning [20], [22], [55], [59], [75]. Accordingly, for the proposed model, we investigate the most suitable metric combination experimentally, and finalize the loss function as follows: Kullback-Leibler Divergence (KL) is a common measure of discrepancy between two probability distributions. Here the KL loss is computed as:

$$\mathrm{KL}(\mathbf{S}, \mathbf{G}) = \sum_i \mathbf{G}_i log(\epsilon + \frac{\mathbf{G}_i}{\epsilon + \mathbf{S}_i}), \tag{13}$$

where ϵ is a regularization constant.

Pearson's Correlation Coefficient (CC) loss measures dependencies between two distribution maps, which is formulated as:

$$\mathrm{CC}(\mathbf{S}, \mathbf{G}) = \frac{cov(\mathbf{S}, \mathbf{G})}{sd(\mathbf{S}) \times sd(\mathbf{G})}, \tag{14}$$

where sd is standard deviation, and cov stands for covariance.

IV. EXPERIMENTS AND RESULTS

In this section, we introduce our experiments, and present their results as well as analyses. In Sec. IV-A, several benchmark datasets are described. In Sec. IV-B, we present the training procedure of our model. In Sec. IV-C, we briefly introduce the saliency metrics used in the evaluation. In Sec. IV-D, we compare the proposed model with other state-of-the-art models on different datasets. In Sec. IV-E, we detailedly study the influence of main components of our model. In Sec. IV-F, we compare two different supervision manners on multiple datasets. In Sec. IV-G, we present some failure cases and analyze the limitations of our model as well as the difficulties of VSP task.

A. Datasets

1) DHF1K [19]: It is a large and diverse dataset, including 1K 30 fps videos with 640 × 360 resolution, which are split to 600, 100, and 300 as training, validation, and testing sets, respectively. The corresponding free-viewing data is collected from 17 observers by the eye tracker. The ground truths of testing videos are held out for the evaluation on benchmark website[1].

2) Hollywood-2 [76]: It contains 1,707 videos from Hollywood movies. The annotations come from 19 observers, 3 of which are in a free-viewing mode and the others are driven by action and context recognition tasks. Following the usual split, we use 823 and 884 videos as training and testing sets, respectively.

3) UCF [76]: It consists of 150 videos including kinds of sports action classes, and the annotations are collected in a task-driven manner. In our paper, we adopt the same split as [19] with 103 video for training and 47 videos for testing.

4) DIEM [77]: It has 84 videos based on advertisements, documentaries, game trailers, and movie trailers, *etc*. The annotations of them are collected from about 50 observers in a free-viewing manner. Following [20], [78], we adopt the same 20 videos as testing set, and the rest as training set.

B. Training Procedure

The proposed model is implemented on two NVIDIA TITAN Xp GPUs using *Pytorch* [79]. We initialize the backbone of our model with the weights from S3D [28] pre-trained on the Kinetics dataset [29], and the remaining network is initialized by default settings of *Pytorch*. We train the whole model with the Adam optimizer [80], and the initial learning

[1]https://mmcheng.net/videosal/

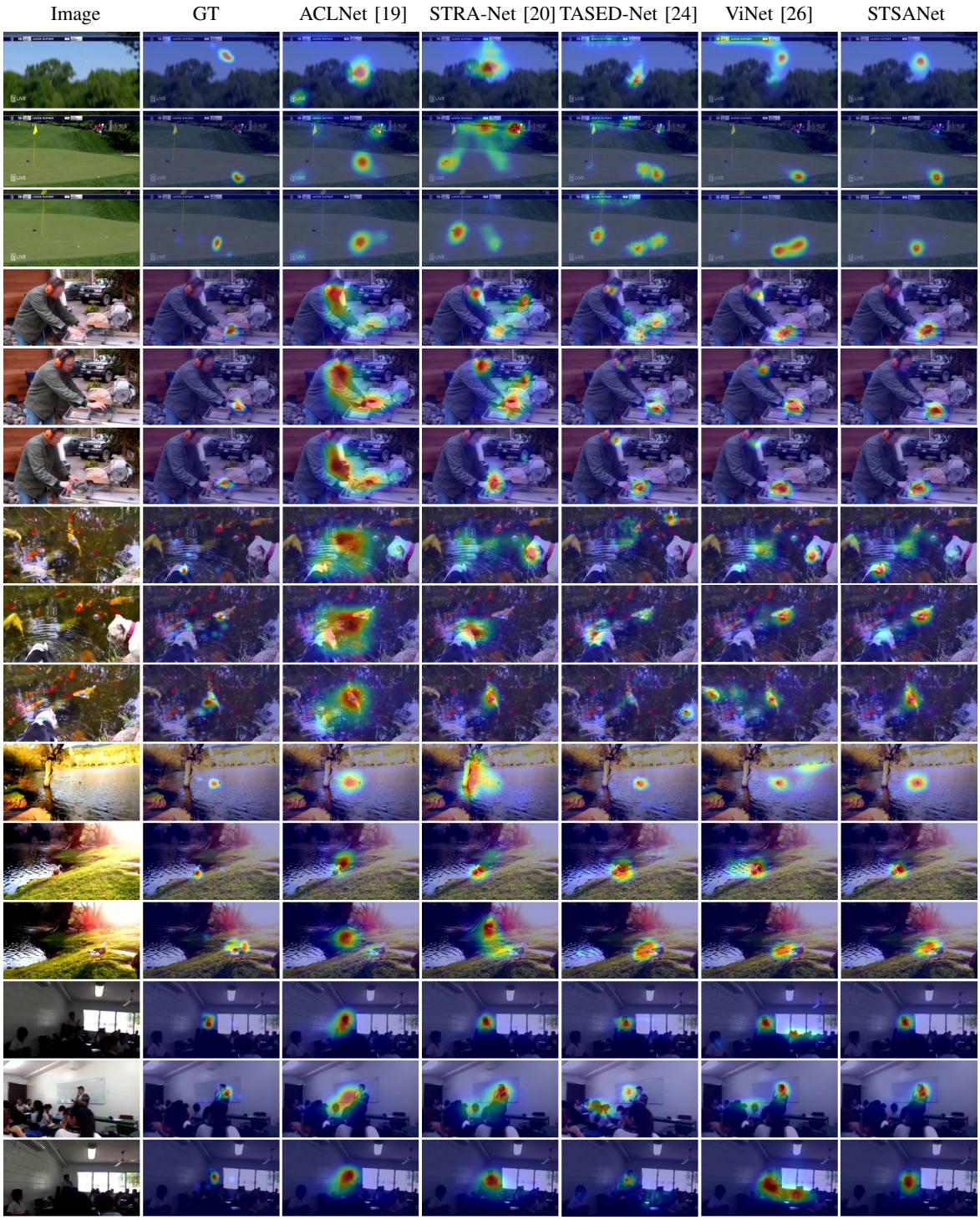


Fig. 5. Qualitative comparisons with start-of-the-art video saliency models on various categories of videos, each of which samples three frames for display.

rate is set to 10^{-4}, which is decreased 10 times when the training loss has been saturated.

Our model is first trained with the training set of DHF1K dataset, and the validation set is used to monitor the convergence. Then we test the results on the DHF1K benchmark. For the Hollywood-2, UCF, and DIEM datasets, we fine-tune the proposed model from the weights trained on DHF1K, and use the testing sets to monitor the convergence. All input frames are resized to 384×224 and the batch size is set to 3.

C. Metrics

Saliency metrics involved in our experimental results and analyses include Normalized Scanpath Saliency (NSS), Pearson's Correlation Coefficient (CC), Similarity (SIM), Kullback–Leibler Divergence (KL), and variants of Area Under ROC Curve (AUC (AUC-Judd) and sAUC (shuffled AUC)). Specifically, the distribution-based metrics, including SIM, CC, and KL, are obtained by comparing results with the fixation continuous maps. The location-based metrics, con-

TABLE III
ABLATION STUDY ON MAIN COMPONENTS OF THE PROPOSED MODEL.

Model	CC ↑	NSS ↑	SIM ↑	AUC ↑	KL ↓
Single-stream	0.506	2.839	0.387	0.910	1.462
UNet-like	0.523	2.983	0.396	0.917	1.388
UNet-like + STSA	0.534	3.049	0.403	0.919	1.362
UNet-like + AMSF	0.531	3.033	0.406	0.919	1.359
Ours	**0.539**	**3.082**	**0.411**	**0.920**	**1.344**

taining NSS, AUC-Judd, and sAUC, are calculated with the binary maps of fixation points. More specific characteristics of saliency metrics can be found in [81].

D. Comparison with the State-of-the-Art Models

We compare our model with state-of-the-art VSP models based on deep learning in recent years, including DeepVS [18], ACLNet [19], STRA-Net [20], SalEMA [64], Chen *et al.* [21], TASED-Net [24], SalSAC [23], UNISAL [74], HD2S [25], ViNet [26], and TSFP-Net [65], on the DHF1K benchmark and the testing sets of Hollywood-2, UCF, and DIEM datasets.

Table I reports a range of attributes of models, where input size, model size, and runtime mostly come from corresponding papers and the DHF1K benchmark website[1], and FLOPs is measured using the publicly available codes of the models. In terms of these attributes, our STSANet is at an intermediate level of computational efficiency. In fact, the STSA module is a major contributor to computational consumption, but it also results in considerable performance gains for the STSANet.

On the DHF1K benchmark[1], our model outperforms all other models on CC, NSS, and AUC metrics, and ranks second on sAUC. More concretely, on CC and NSS, the performance of our model exceeds that of other models by 2.3% and 1.5%, respectively. The SalEMA model shows the highest score on SIM metric, however, it gets low scores on other four metrics compared with state-of-the-art models.

The results on the test sets of Hollywood-2 and UCF datasets are shown in the right of Table I, where results of other models come from their corresponding papers and the DHF1K benchmark website[1]. On the Hollywood-2 dataset, our method achieves the best performance in terms of all metrics. On the UCF dataset, our model outperforms others by a large margin on CC and NSS metrics, and ranks first on AUC metric. Specifically, compared with the second best model, our model has improved by 2.9% and 5.7% on CC and NSS, respectively. On SIM metric, the score of our model ranks second but very close to the first one.

The results on the DIEM datasets are reported in Table II. The LSTM-based models, including DeepVS [18], ACLNet [19], STRA-Net [20], and Chen *et al.* [21], all use the first 300 frames of 20 test videos as test set (Test set i) in the corresponding papers, while the two 3D convolution-based models, ViNet [26] and TSFP-Net [65], use 17 of 20 test videos as test set (Test set ii). Accordingly, we evaluate our model on different test sets for fair comparisons with other models. On the Test set i, our model outperforms others by a large margin(*e.g.* $\frac{0.625-0.490}{0.490} = 27.6\%$ on CC). Besides,

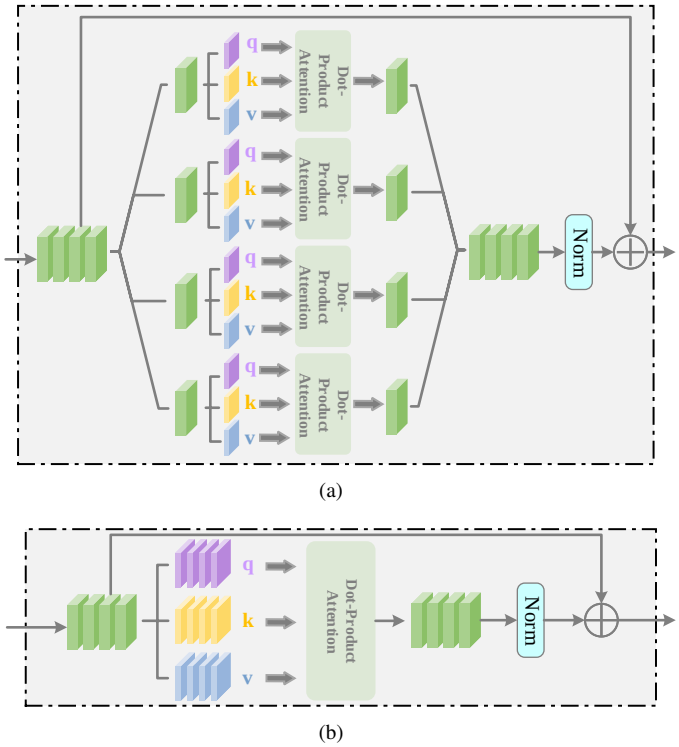


Fig. 6. Structures of two variant of STSA module: (a) *w/o* Temporal Relations. (b) Single Similarity Matrix.

our model achieves a substantial improvement compared with other 3D convolution-based models on the Test set ii, such as 4.3% (0.649 to 0.677) on CC and 3.5% (2.630 to 2.721) on NSS. Finally, we present the results of our model on all 20 test videos of DIEM dataset (Test set iii).

In Fig. 5, we further qualitatively compare our model with some representative state-of-the-art VSP models, including LSTM-based ACLNet [19] and STRA-Net [20], as well as 3D convolution-based TASED-Net [24] and ViNet [26]. It can be clearly observed that our model achieves more accurate results than others on different indoor and outdoor videos.

E. Ablation Analysis

In this section, we conduct comprehensive ablation experiments for STSANet on the validation set of DHF1K dataset, which contains around 60K frames. We first investigate the contribution of main components in our model, and then further provide detailed ablation studies on our STSA and AMSF modules, respectively.

1) The Contributions of Main Components: We quantitatively evaluate the effectiveness of main components in the proposed model, and the results are reported in Table III. We construct a single-stream network based on the 3D backbone as a baseline, in which feature maps are gradually restored to initial image size by 3D convolution and trilinear interpolation layers after backbone. Next, the UNet-like structure is added over the baseline for exploiting the multi-level information of encoder, which brings obvious improvements in terms of all metrics. Based on the UNet-like encoder-decoder architecture, the scores of CC and NSS increase 2.1% (0.523 to 0.534)

TABLE IV
THE UPPER PART IS ABLATION STUDY ON THE STRUCTURE OF STSA MODULE, WHERE *w/o* TEMPORAL RELATIONS MEANS IMPLEMENTING SELF-ATTENTION MECHANISM SEPARATELY ON EACH TEMPORAL CHANNEL, AND SINGLE SIMILARITY MATRIX MEANS IMPLEMENTING SELF-ATTENTION MECHANISM DIRECTLY ON THE ENTIRE FEATURE. THE LOWER PART IS ABLATION STUDY ON THE STSA MODULES AT DIFFERENT LEVELS, WHERE *rm* STSA_N MEANS REMOVING THE STSA MODULE AFTER $Conv\_Block\_n$.

Model	CC ↑	NSS ↑	SIM ↑	AUC ↑	KL ↓
w/o Temporal Relations	0.532	3.040	0.402	0.919	1.358
Single Similarity Matrix	0.532	3.021	0.395	0.919	1.357
w/o STSA Layer I	0.530	3.026	0.397	0.918	1.370
w/o STSA Layer II	0.533	3.026	0.404	0.919	1.357
rm STSA_1	0.534	3.051	0.404	0.919	1.353
rm STSA_2	0.532	3.022	0.401	0.919	1.356
rm STSA_3	0.535	3.056	0.407	0.919	1.350
rm STSA_4	0.536	3.047	0.405	0.919	1.352
Ours	**0.539**	**3.082**	**0.411**	**0.920**	**1.344**

and 2.2% (2.983 to 3.049), respectively, after adding the proposed STSA modules. As for the AMSF module, replacing the traditional top-down feature fusion with AMSF improves the CC score by 1.5%. Besides, adding the combination of STSA and AMSF over the UNet-like encoder-decoder model increases the scores of CC, NSS, and SIM by 3.1% (0.523 to 0.539), 3.3% (2.983 to 3.082), and 3.8% (0.396 to 0.411), respectively, and optimizes the score of KL by 3.2% (1.388 to 1.344). Overall, continuous performance improvements are shown when adding main components into the baseline in turn. From the baseline to the full settings, the scores of CC, NSS, and KL are totally optimized by 6.5% (0.506 to 0.539), 8.6% (2.839 to 3.082), and 8.1% (1.462 to 1.344), respectively.

2) Ablation Study on STSA Module: As reported in Table IV, we conduct an in-depth ablation study on our STSA module. As mentioned in Sec. III-B, the STSA module achieves that each feature at different time steps is directly updated by long-range spatio-temporal relations with the others through the cooperation of two STSA layers, for enhancing the relational features between time steps. To study the importance of temporal relations contained in two STSA layers, we construct a variant of STSA module without capturing temporal relations between time steps, namely *w/o* Temporal Relations, as depicted in Fig. 6(a). In this variant, self-attention mechanism is implemented separately at different time steps. Specifically, for feature at each time step, its own query and key are multiplied for similarity matrix that is then integrated to value by matrix multiplication. Consequently, this variant captures long-distance spatial relations separately at each time step, without temporal relations between different time steps. We replace the STSA module with this variant in our model and the evaluation results indicate the contribution of capturing long-range relations between spatio-temporal features of different time steps in our STSA module.

In the second variant named as Single Similarity Matrix, as depicted in Fig. 6(b), self-attention mechanism is applied to the whole feature $\mathbf{F} = [\mathbf{F}^1, \mathbf{F}^2, \mathbf{F}^3, \mathbf{F}^4] \in \mathbb{R}^{C\times 4\times H\times W}$. Concretely, the input feature $\mathbf{F}$ is transformed to query $\mathbf{F}_q^t \in \mathbb{R}^{C/2\times 4\times H\times W}$, key $\mathbf{F}_k^t \in \mathbb{R}^{C/2\times 4\times H\times W}$, and value $\mathbf{F}_v^t \in \mathbb{R}^{C\times 4\times H\times W}$ by embedding layers. In the dot-product attention, a single similarity matrix with size $4HW \times 4HW$ is generated from query and key, to capture spatio-temporal relationship. The experimental results of the Single Similarity Matrix show worse performance than the proposed model as reported in Table IV. In this variant, the elements in the similarity matrix are selected in temporal subspace during the softmax function. For example, if the elements representing the relation between $\mathbf{F}^1$ and itself get high values, the elements representing the relations between $\mathbf{F}^1$ and sub-features at other time steps ($\mathbf{F}^2$, $\mathbf{F}^3$, and $\mathbf{F}^4$) will get low values, which amounts to a disguised weakening of these elements, and hence leads to an inability to adequately capture the relations between $\mathbf{F}^1$ and sub-features at other time steps. Our STSA module avoids this drawback, because it directly captures long-range relations between each sub-feature and the other three in two STSA layers.

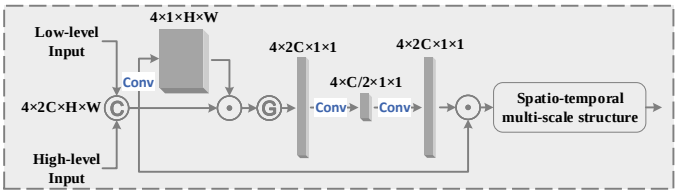


Fig. 7. Structure of the variant of AMSF module: replacing addition fusion with concatenation fusion in AMSF module.

TABLE V
ABLATION STUDY ON THE AMSF MODULE. CONCAT-BASED MEANS REPLACING ADDITION FUSION WITH CONCATENATION FUSION. AM AND STMS INDICATE THE ATTENTION WEIGHTING OPERATION AND SPATIO-TEMPORAL MULTI-SCALE STRUCTURE, RESPECTIVELY.

Model	CC ↑	NSS ↑	SIM ↑	AUC ↑	KL ↓
Concat-Based	0.532	3.015	0.397	0.919	1.355
w/o AW	0.533	3.033	0.401	0.919	1.352
w/o STMS	0.528	3.019	0.403	0.918	1.372
Ours	**0.539**	**3.082**	**0.411**	**0.920**	**1.344**

Furthermore, in order to inspect the necessity of two STSA layers in the STSA module, the STSA layer I and STSA layer II are removed separately. The comparison results demonstrate that removing either STSA layer I or STSA layer II results in degradation of performance compared with the full settings, in which each of features at different time steps can be updated by long-range relations with spatio-temporal features of all the other time steps through the cooperation of two STSA layers.

Moreover, to validate the effectiveness of STSA modules employed on different levels, we separately remove one of four STSA modules from the proposed STSANet for evaluation. As shown in the lower part of Table IV, the performance deteriorates in terms of all metrics after removing STSA module at any level, which suggests that all STSA modules at different levels have contribution to saliency results.

3) Ablation Study on AMSF Module: To further verify the contribution of the design of AMSF module, we change the AMSF module by removing the attentional weighting (AW) operation and spatio-temporal multi-scale (STMS) structure,

TABLE VI
COMPARISON RESULTS OF DIFFERENT SUPERVISION MANNERS ON DHF1K, HOLLYWOOD-2, AND UCF DATASETS.

Dataset	Supervision	CC ↑	NSS ↑	SIM ↑	AUC ↑	KL ↓
DHF1K	middle-supervised	**0.539**	**3.082**	**0.411**	**0.920**	**1.344**
	last-supervised	0.537	3.057	0.405	**0.920**	**1.344**
Hollywood-2	middle-supervised	0.721	**3.927**	**0.579**	0.938	0.769
	last-supervised	**0.722**	3.908	0.576	**0.939**	**0.764**
UCF	middle-supervised	**0.705**	**3.908**	**0.560**	**0.936**	**0.851**
	last-supervised	0.700	3.789	0.541	0.935	0.873

respectively. As shown in Table V, the two variants with incomplete structures result in performance degradation in terms of all metrics. Our complete AMSF modules integrates multi-level features with the perception of context in all three subspaces for better saliency results.

Besides, as shown in Fig. 7, we construct a variant of the AMSF module about feature fusion, namely Concat-Based, which replaces addition fusion with concatenation fusion in the AMSF module. The experimental comparison shows that the AMSF module helps the proposed model go to a better convergence compared with the Concat-Based variant. In our AMSF module, addition fusion allows highlighting the intersecting saliency regions in high- and low-level feature maps, which helps to collectively consider the multi-level saliency information during feature fusion. Moreover, addition fusion brings fewer parameters than concatenation fusion.

F. Discussion on Supervision Manner

Previous work [24] used the ground truths corresponding to last frames of input video clips for supervise learning. In our work, as mentioned in Sec. III-E, the ground truth of middle frame in the video clip is used to supervise model training. We refer to the two supervision manners as middle-supervised and last-supervised for short, and compare them experimentally. As show in Table VI, we adopt two different supervision manners to train our model on DHF1K, Hollywood-2, and UCF datasets, respectively. The evaluation results show that the middle-supervised and last-supervised have close performance on the DHF1K and Hollywood-2 datasets. On the UCF dataset, the middle-supervised manner presents a relative advantage in terms of all metrics. As a result, other experiments are implemented in the middle-supervised manner.

G. Failure Cases and Analyses

Here we present some failure cases and analyze the limitations of our model as well as the difficulties of VSP task. In Fig. 8, for the first video (the first two rows), human fixations are mainly on the wood being processed. However, the results of computational models are scattered on other objects. Varied scene and all sorts of objects make it difficult for models to accurately localize the wood, *i.e.*, the main character of the video. On the other hand, our STSANet infers saliency results only from a video clip, so the global context from an entire video is not taken into account, which causes difficulty in highlighting the main character in such a complex video. In the second video (the last two rows), human fixations are distributed on the interaction of the water snake with other objects, *i.e.*, the water snake and water plants (the third row), and the water snake and fish (the last row). Although models are able to capture the motions of the snake, it is hard for them to perceive the interaction, and thus they fail to precisely locate the part of the snake's body.

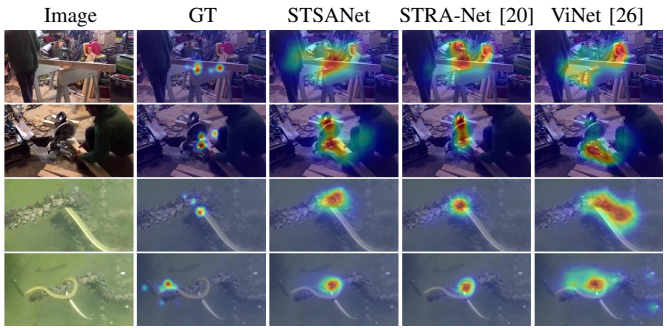


Fig. 8. Some failure cases on DHF1K dataset.

V. CONCLUSION

This paper proposes a novel Spatio-Temporal Self-Attention 3D neural network (STSANet) for video saliency prediction. We employ Spatio-Temporal Self-Attention (STSA) modules at different levels of the 3D convolutional backbone to overcome the locality of 3D convolution. In each STSA module, spatio-temporal features are split along the temporal channel and features at different time steps are transformed by embedding layers, then dot-product attention calculation is implemented between features of different time steps. By this means, at multiple levels, features at each time step can be updated by long-range dependencies with features of other time steps. The integration of 3D networks and spatio-temporal self-attention mechanism brings performance gains as shown in ablation experiments. Accordingly, this method has the potential to be applied to other video tasks. Furthermore, we design an Attentional Multi-Scale Fusion (AMSF) module for the integration of multi-level spatio-temporal features. The AMSF module contains an attentional weighting operation and a spatio-temporal multi-scale structure, which separately aim to alleviate the semantic and spatio-temporal gaps between features of different levels. Extensive experiments demonstrate outstanding performance of the proposed model compared with all state-of-the-art video saliency prediction methods.

REFERENCES

[1] L. Zhou, Y. Zhang, Y.-G. Jiang, T. Zhang, and W. Fan, "Re-caption: Saliency-enhanced image captioning through two-phase learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 694–709, 2020.
[2] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2019.
[3] G. Li, Z. Liu, R. Shi, and W. Wei, "Constrained fixation point based segmentation via deep neural network," *Neucomputing*, vol. 368, pp. 180–187, 2019.
[4] G. Li, Z. Liu, R. Shi, Z. Hu, W. Wei, Y. Wu, M. Huang, and H. Ling, "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1461–1475, 2021.
[5] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[6] W. Wang, J. Shen, X. Lu, S. C. H. Hoi, and H. Ling, "Paying attention to video object pattern understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2413–2428, 2021.

[7] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2014.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[10] M. Cerf, J. Harel, W. Einhaeuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 241–248.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2106–2113.

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

[13] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 171–177, 2010.

[14] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[15] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1688–1698, 2018.

[16] Z. Wu, L. Su, and Q. Huang, "Learning coupled convolutional networks fusion for video saliency prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2960–2971, 2019.

[17] K. Zhang and Z. Chen, "Video saliency prediction based on spatial-temporal two-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3544–3557, 2019.

[18] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 625–642.

[19] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 220–237, 2021.

[20] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2020.

[21] J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, "Video saliency prediction using enhanced spatiotemporal alignment network," *Pattern Recognition*, vol. 109, p. 107615, 2021.

[22] K. Zhang, Z. Chen, and S. Liu, "A spatial-temporal recurrent neural network for video saliency prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 572–587, 2021.

[23] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, "SalSAC: A video saliency prediction model with shuffled attentions and correlation-based convlstm," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 12 410–12 417.

[24] K. Min and J. J. Corso, "TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2394–2403.

[25] G. Bellitto, F. P. Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Hierarchical domain-adapted feature learning for video saliency prediction," *arXiv preprint arXiv:2010.01220*, 2020.

[26] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "ViNet: Pushing the limits of visual modality for audio-visual saliency prediction," *arXiv preprint arXiv:2012.06170*, 2020.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and Ł. Kaiser, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 6000–6010.

[28] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murph, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[29] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[30] W. Zhang, Q. M. J. Wu, G. Wang, and H. Yin, "An adaptive computational model for salient object detection," *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 300–316, 2010.

[31] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5961–5970.

[32] Y. Zhou, A. Mao, S. Huo, J. Lei, and S.-Y. Kung, "Salient object detection via fuzzy theory and object-level enhancement," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 74–85, 2019.

[33] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1913–1927, 2020.

[34] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 324–336, 2020.

[35] Q. Ren, S. Lu, J. Zhang, and R. Hu, "Salient object detection by fusing local and global contexts," *IEEE Transactions on Multimedia*, vol. 23, pp. 1442–1453, 2021.

[36] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access)*, 2021.

[37] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2527–2542, 2017.

[38] X. Zhou, Z. Liu, C. Gong, and W. Liu, "Improving video saliency detection via localized estimation and spatiotemporal refinement," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2993–3007, 2018.

[39] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2018.

[40] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8546–8556.

[41] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 545–552.

[42] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2798–2805.

[43] M. Kümmerer, L. Theis, and M. Bethge, "Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet," in *International Conference on Learning Representations (ICLR)*, 2015.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[45] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 262–270.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[47] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 598–606.

[48] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5753–5761.

[49] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *International Conference on Pattern Recognition (ICPR)*, 2016, pp. 3488–3493.

[50] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[51] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4799–4808.

[52] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.

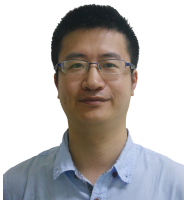
[53] S. F. Dodge and L. J. Karam, "Visual saliency prediction using a mixture of deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4080–4090, 2018.
[54] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.
[55] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.
[56] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2163–2176, 2020.
[57] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? Dataset and model," *IEEE Transactions on Image Processing*, vol. 29, pp. 2287–2300, 2020.
[58] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder-decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.
[59] N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi, "Tidying deep saliency prediction architectures," in *International Conference on Intelligent Robots and Systems (IROS)*, 2020.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.
[62] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8697–8710.
[63] S. Gorji and J. J. Clark, "Going from image to video saliency: Augmenting image salience with dynamic attentional push," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7501–7511.
[64] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. Giro-I-Nieto, and K. Mcguinness, "Simple vs complex temporal recurrences for video saliency prediction," in *British Machine Vision Conference (BMVC)*, 2019, pp. 1–12.
[65] Q. Chang, S. Zhu, and L. Zhu, "Temporal-spatial feature pyramid for video saliency detection," *arXiv preprint arXiv:2105.04213*, 2021.
[66] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.
[67] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9225–9234.
[68] H. Wang, W. Wang, and J. Liu, "Temporal memory attention for video semantic segmentation," *arXiv preprint arXiv:2102.08643*, 2021.
[69] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3618–3627.
[70] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 289–297.
[71] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," in *International Conference on Learning Representations (ICLR)*, 2017.
[72] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2021.
[73] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.
[74] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 419–435.
[75] Z. Wang, Z. Liu, W. Wei, and H. Duan, "SalED: Saliency prediction with a pithy encoder-decoder architecture sensing local and global information," *Image and Vision Computing*, vol. 109, no. 104149, 2021.
[76] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, 2015.
[77] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, 2011.
[78] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.
[79] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.
[80] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[81] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

Ziqiang Wang received the B.E. degree from Shanghai University, Shanghai, China, in 2019. He is currently pursuing the M.E. degree at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His current research interests include computer vision and deep learning.

Zhi Liu (M'07-SM'15) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002 and 2005, respectively. He is currently a Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC member/session chair in ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, WIAMIS 2013, *etc*. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*.

Gongyang Li received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include image/video object segmentation and saliency detection.

Yang Wang received the B.Sc. degree from the Harbin Institute of Technology, Harbin, China, the M.Sc. degree from the University of Alberta, Edmonton, AB, Canada, and the Ph.D. degree from Simon Fraser University, Burnaby, BC, Canada, all in computer science. He was previously a NSERC Postdoc Fellow with the University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently an Associate Professor of computer science with the University of Manitoba, Winnipeg, MB, Canada. His research interests include computer vision and machine learning.



Tianhong Zhang MD. PhD. Senior Psychiatrist, Director of Early Identification and Intervention for Clinical High Risk of Psychosis, Clinical PI for SHARP (ShangHai At Risk for Psychosis) program, Shanghai Mental Health Centre, Master's tutor in Shanghai Jiaotong University School of Medicine. Adjunct Professor of Medical College, University of Ottawa, Secretary and Member of CSNP Schizophrenia Research Alliance, Member of Schizophrenia Collaborative Group of Psychiatric Society of Chinese Medical Association, BMC psychiatry, Psychiatry Research Editor, Frontier in Psychiatry Guest Editor.

Lihua Xu MD. PhD. Psychiatrist, Attending Doctor. Her research direction is the biomarker research of high risk syndrome of psychosis. To be responsible for the clinical evaluation, follow-up and data management of high risk syndrome of psychosis.

Jijun Wang MD. PhD. Chief Physician, Doctoral Supervisor of Shanghai Mental Health Center (mental health center affiliated to Shanghai Jiaotong University School of Medicine), director of brain film image eye movement research office, PI of Shanghai heavy mental disease laboratory, member of psychiatry basic and clinical branch of Chinese Neuroscience Society, member of EEG and EMG branch of Shanghai Medical Association, director of Youth Committee of Shanghai Overseas Chinese joint committee. As editor of the Journal of psychiatry, BMC psychiatry and reviewer of various international journals (Cochrane Database Syst. Rev., Biological Psychology, International Journal of Psychiatry, etc.).