Video Saliency Prediction Using Spatiotemporal Residual Attentive Networks

Qiuxia Lai, Wenguan Wang, *Member, IEEE*, Hanqiu Sun, *Member, IEEE*, and Jianbing Shen, *Senior Member, IEEE*

***Abstract*— This paper proposes a novel residual attentive learning network architecture for predicting dynamic eye-fixation maps. The proposed model emphasizes two essential issues, *i.e.*, effective spatiotemporal feature integration and multi-scale saliency learning. For the first problem, appearance and motion streams are tightly coupled via dense residual cross connections, which integrate appearance information with multi-layer, comprehensive motion features in a residual and dense way. Beyond traditional two-stream models learning appearance and motion features separately, such design allows early, multi-path information exchange between different domains, leading to a unified and powerful spatiotemporal learning architecture. For the second one, we propose a composite attention mechanism that learns multi-scale local attentions and global attention priors end-to-end. It is used for enhancing the fused spatiotemporal features via emphasizing important features in multi-scales. A lightweight convolutional Gated Recurrent Unit (convGRU), which is flexible for small training data situation, is used for long-term temporal characteristics modeling. Extensive experiments over four benchmark datasets clearly demonstrate the advantage of the proposed video saliency model over other competitors and the effectiveness of each component of our network. Our code and all the results will be available at https://github.com/ashleylqx/STRA-Net.**

***Index Terms*— Dynamic eye-fixation prediction, residual attentive learning, attention mechanism, deep learning, video saliency.**

I. Introduction

HUMANS are able to rapidly orient attention to important areas in visual field and filter out irrelevant information. Such selective process, called as visual attention mechanism, helps humans operate huge amount of visual information in realtime. Visual attention has long been studied in computer vision community dated back to 1990s [1], and shown wide applications such as object segmentation [2], [3], [77], and video summarization [4], to name a few. Integrating attention mechanism into above tasks could allocate limited computation source into the most task-related targets, as well as obtain biologically inspired results.

Recently, inspired by the impressive success of deep learning technique, a lot of deep neural network based saliency models [5]–[12] were proposed for imitating the mechanism of visual attention allocation in static scenes. Those deep static saliency models are able to learn flexible and powerful saliency representations from large-scale eye-tracking data and generally obtain a large margin of performance gain compared with previous non-deep learning static models [13]–[16].

In striking contrast to the significant advance of static visual attention prediction in recent years (back to [5]), there are only very few deep learning based visual attention models [17]–[21] that are specially designed for modeling eye fixation during dynamic free-viewing. Different from static visual attention prediction, in dynamic videos, both spatial and temporal information are essential factors that guide human attention orientation, as suggested by many studies in cognitive science and computer vision. Although rich motion information in the extra temporal domain bring more cues for predicting visual attention, complex motion patterns from the background, inconsistent movements among different foreground patterns, camera motions, and the optical flow computation error, all make video saliency prediction more difficult. Hence, how to effectively integrate saliency features from different domains and scales is one essential but long-time unsolved problem for dynamic visual attention prediction.

Recent deep dynamic visual attention models have shown improved performance, however, none of them well handle the above challenges. Although appearance and motion information are learned by an architecture consisting of convolutional neural network and Long-Short Term Memory (CNN-LSTM) [17], or more informative cues such as optical flow [18] and objectness [19] are explicitly captured by different network streams, those touched few (1) how to fuse saliency features from different domains; and (2) how to learn saliency representations from multi-scales. For two-stream models [18], [19], saliency information from spatial and temporal streams are fused only once (in *early* or *late* fusion strategy), which decreases the representative power and cannot fully leverage rich information encoded in appearance and motion network streams. For CNN-LSTM architecture

Manuscript received February 11, 2019; revised July 2, 2019; accepted August 6, 2019. Date of publication August 23, 2019; date of current version November 4, 2019. This work was supported in part by the Beijing Natural Science Foundation under Grant 4182056, in part by the National Natural Science Foundation of China under Grant 61602183, in part by the CCF-Tencent Open Fund, in part by the Zhijiang Lab's International Talent Fund for Young Professionals, in part by the Fok Ying-Tong Education Foundation for Young Teachers, and in part by the Specialized Fund for the Joint Building Program of the Beijing Municipal Education Commission. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soma Biswas. *(Corresponding author: Jianbing Shen.)*

Q. Lai and H. Sun are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: qxlai@cse.cuhk.edu.hk; hanqiu@cse.cuhk.edu.hk).

W. Wang and J. Shen are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wenguanwang.ai@gmail.com; shenjianbingcg@gmail.com).

Digital Object Identifier 10.1109/TIP.2019.2936112

based model [17], where motion information is captured by a Recurrent Neural Network (RNN), it requires a lot of static and dynamic eye-tracking data for decoupling spatial and temporal saliency information, due to the limited representation power of LSTM. Additionally, they largely neglect the importance of learning saliency information from multi-scales, while only considering saliency information within one single network layer.

Aiming for alleviating the above challenges in video saliency detection and remedying the shortages of previous deep dynamic attention models, we propose a *spatiotemporal residual attentive network*, which emphasizes comprehensive fusion of spatial and temporal saliency features, and multi-scale saliency representation enhancement simultaneously. More specially, inspired by the recent advance of residual network [22] and video classification [23], we design a spatiotemporal residual network with dense residual cross connections between the layers from appearance and motion streams, for encouraging information flow between different stream layers and deeper integration of saliency cues from different domains. For further emphasizing multi-scale saliency representation learning, we introduce a composite attention module that learns attentions from multiple scales, namely a stack of local attentions as well as global attention priors. The local attentions act as a network attention mechanism that suppresses irrelevant saliency features, while the global attentions capture the center bias of visual attention allocation in free-viewing. Considering the relatively small amount of dynamic eye-tracking data, we introduce convolutional Gated Recurrent Unit (convGRU) [24], a lightweight recurrent network, for learning the temporal attention transitions across time. Compared with traditional RNN, it could better capture long-term dependencies and handle gradient vanishing/expoding problem. Its architecture is also relatively simple compared with LSTM, thus it is more applicable for small training data situation.

To summarize, the contributions of this work are three-fold:

- A spatiotemporal residual neural network, which consists of two tightly coupled streams for capturing and fusing saliency features in spatial and temporal domains, is proposed for dynamic visual attention prediction. With the dense residual cross connections among different layers of the network streams, spatial and temporal features are integrated in a more comprehensive way, leading to a powerful spatiotemporal saliency representation.
- A composite attention module is integrated for enhancing spatiotemporal saliency representation with multi-scale information. The composite attention mechanism learns local attentions as well as global attention priors for emphasizing the informative saliency features and filtering out useless information, thus improving the spatiotemporal saliency representation efficiently.
- ConvGRU, as a lightweight recurrent network, is introduced for modeling the attention transitions across video frames, which allows more efficient saliency learning with relative paucity of dynamic eye-tracking data. To the best of our knowledge, this is the first application of convGRU in the task of human visual attention prediction.

Extensive experiments on four large-scale video saliency datasets (*i.e.*, DHF1K [17], Hollywood2 [25], UCF-sports [25] and DIEM [26]) clearly demonstrate that the proposed video saliency model outperforms other competitors.

II. Related Work

A. Computational Models for Visual Attention Prediction

The concept of *visual saliency* in computer vision is typically modeled as two tasks, namely *visual attention prediction* which aims at predicting human eye fixation, and *salient object detection (SOD)* that highlights salient regions in images. The SOD task is driven by object-level applications [27], and covers a wide range of sub-tasks including RGB saliency [28], [29], RGBD saliency [30], video SOD [31]–[36], and co-saliency [37], [38]. We refer interested readers to [39], [40] for more detailed introduction of SOD. In this paper, we focus on the task of predicting visual attention for videos.

The work of Itti *et al.* [1] represents an early attempt that builds a computational model for predicting human eye fixation. After that, a rich set of saliency models are developed in computer vision community. Here, we mainly focus on discussing representative works, especially recent deep learning based visual attention models. According to the input format, these models can be classified as either static or dynamic method.

1) Static Models: Earlier image attention models were inspired by cognitive theories of visual attention, and typically based on bottom-up, stimuli-driven mechanism. They leverage and analyse various biologically inspired features such as color, edge and orientations on different spatial scales [41]–[46] or frequency domain [47]. Then the final attention map is built bottom-up by fusing the saliency estimates from these features, where the locations with distinct features over their neighborhoods usually come out with higher saliency values.

Recently, several deep saliency models have been proposed. Compared with traditional methods, these models achieved significantly better results, benefited from large-scale datasets and the strong learning ability of neural network. Vig *et al*. [5] search for the optimal ensemble of CNNs as deep feature extractor, and train an SVM for predicting the probability of an image region being salient. This is the first work that applies deep learning for visual attention. After that, some models like DeepFix [10], DeepNet [8] and SALICON [6] are built via fine-tuning existing top-performance image classification models (*e.g.*, VGG-16 [48]) on eye-tracking data. Mr-CNN [9] applies a multi-stream network to capture multi-scale saliency. DVA [11] fuses multi-layer features and thus incorporates multi-level saliency predictions within a single network.

2) Dynamic Models: Although a huge amount of static saliency models have been proposed in the last two decades, there are far less effort in mimicking eye fixation behavior in dynamic scenes. Traditional dynamic models [13]–[15] leverage hand-crafted features in both spatial and temporal domains, and most of them can be viewed as extensions of static models by incorporating extra motion features [49]–[52]. Some other methods resort to features by video compression

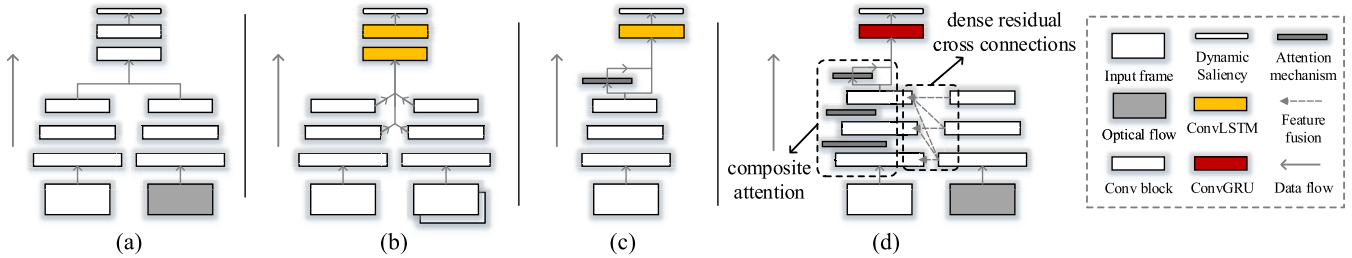


Fig. 1. Architectures of (a) two-stream network [18], (b) OM-CNN + 2C-LSTM network [19], (c) ACLNet model [17], and (d) our proposed spatiotemporal residual attentive network. See Section II-C for more details.

since those features are more related to visual attention and can be highly predictive of human eye fixations [53]–[56].

So far, there are only very few deep dynamic saliency models. Bak *et al*. [18] first propose to utilize a two-stream CNN architecture that learns both appearance and motion saliency information. Later, Bazzani *et al*. [57] assume human attention distribution as Gaussian Mixture Model (GMM) and use LSTM to learn the GMM. Jiang *et al*. [19] propose a multi-stream model that considers appearance, motion, and objectness together. More recently, Wang *et al*. [17] incorporated a supervised static attention mechanism into a CNN-LSTM network and achieved promising results. All above works show the advantages of applying deep learning to this problem, while they largely neglect the issue of effectively fusing multi-domain features and learning saliency representations from multi-scales. In Section II-C, we will offer more detailed discussion for the differences with previous models.

B. Related Network Design

Next we give a brief overview of related network architectures, which inspire the design of our model.

1) Two-Stream Structure: Two-stream network was first proposed by Simonyan and Zisserman [58] for video action recognition, and then used in many spatiotemporal tasks [23], [59]–[61]. It has two parallel streams that take video frames and optical flows as inputs, for capturing appearance and motion information, respectively. Inspired by [62], we propose a spatiotemporal residual network that fuses the spatial and temporal information using dense residual cross connections between the corresponding layers of the appearance and motion streams, which allows more effective interaction during the training process. The motion feature serves as gating signal to the appearance feature, which is able to filter out indeterminant spatial information and inject rich temporal information.

2) Attention Mechanism in Neural Network: It is observed a popular trend to incorporate attention mechanism into network design. Such attention mechanism lets the network focus more on the most important and task-relevant contents. It has shown great successes in computer vision [63]–[65] and natural language processing communities [66], [67]. Such attention can be learned in an implicitly way and enhance network features. Here we introduce a composite attention module that learns a stack of local attentions and global human attention bias, for emphasizing multi-scale saliency representation learning.

3) Gated Recurrent Unit (GRU): The GRU [68] is lighter-weighted over LSTM, while achieving similar performance in capturing long-term dependencies. Also, it can better handle gradient vanishing and exploding problems compared with classical RNN. The convGRU inherits the advantages of fully connected GRU and is further able to preserve spatial information, which is essential for pixel-level saliency prediction task. So far, convGRU was only applied in limited applications such as learning video representation features [24] and video segmentation [69]. In this work, we introduce convGRU for learning attention transitions across video frames.

C. Comparison With Previous Works

In this section, we discuss three representative network architectures of current top-performing deep video saliency models [17]–[19], and detail the differences between these models and our method. This would be better to highlight our contributions. A schematic illustration of these network architectures is shown in Fig. 1.

1) Two-Stream Network [18]: Bak *et al*. propose a two-stream network for video saliency prediction, where two separated branches take video frames and optical flows as inputs, and learn appearance and motion information respectively (see Fig. 1 (a)). However, the appearance and motion features are only fused lately, which is not efficient enough to learn comprehensive spatiotemporal features due to the lack of information exchange between the two streams.

2) OM-CNN + 2C-LSTM Network [19]: Based on the observation that human attention is normally attracted by objects, Jiang *et al*. [19] propose to predict the attention upon both objectness and object motion. Their feature extraction network, *i.e*., OM-CNN, consists of two branches, one is based on YOLO [70], a well-known object detection network, and the other one is modified from FlowNet [71], a CNN-based optical flow estimation network. As shown in Fig. 1 (b), multi-layer features from both two branches are fed into a temporal correlation learning network (2C-LSTM) comprised of two convolutional LSTM (convLSTM) [72] layers.

3) Attentive CNN-convLSTM Network (ACLNet) [17]: ACLNet [17] is based on a typical CNN-LSTM architecture which combines the convolutional and recurrent networks in a single stream to learn the spatial and temporal information, as shown in Fig. 1 (c). It incorporates a supervised attention mechanism to enhance intra-frame salient features with encoded static saliency information learned explicitly from

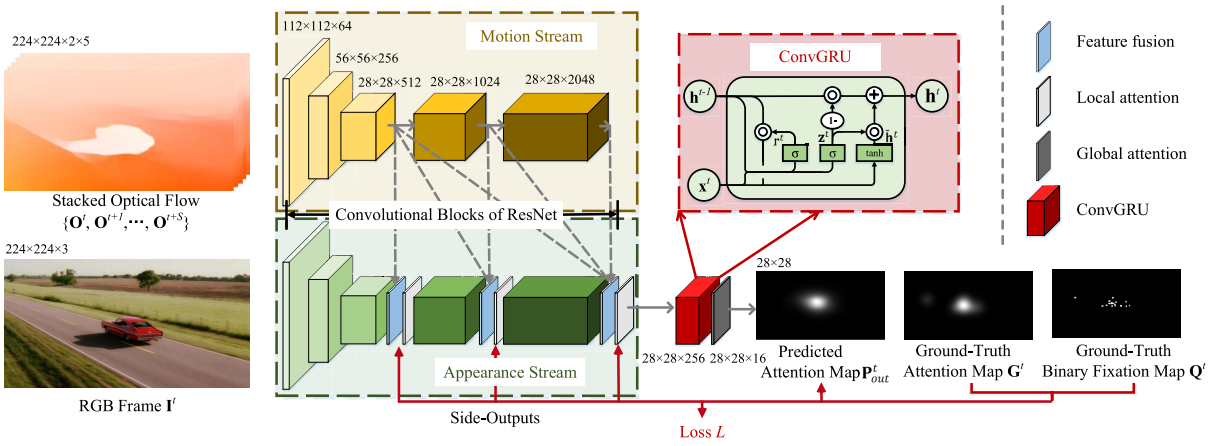


Fig. 2. Architecture overview. Two tightly coupled streams take RGB frames and stacked optical flows as inputs, respectively, and output the comprehensive spatiotemporal saliency representation. The spatiotemporal saliency representations are further enhanced with multi-scale information through a composite attention mechanism. A lightweight recurrent unit, convGRU, is incorporated to facilitate learning the temporal transition of visual attention.

image saliency dataset. However, ACLNet cannot efficiently capture the motion dynamics due to the limited learning power of LSTM and lack of explicit motion representation.

4) The Proposed Model: Our model is significantly differentiated in three aspects. First, by adding multiple residual cross connections between the two stream layers, our model allows multi-domain features fusion in a more efficient and earlier way. Such design learns a more powerful spatiotemporal model, rather than previous methods fusing appearance and motion in a shallow manner. Second, our model is equipped with a composite attention mechanism for enhancing spatiotemporal features in multi-scales. Instead of ACLNet learning attention mechanism from one single layer, the proposed composite attention mechanism emphasizes multi-scales and captures human attention bias as global prior. Third, to model the attention transitions over time, we augment our model with a lightweight recurrent layer, *i.e.*, convGRU, which is more flexible for learning from relatively small dynamic fixation datasets, compared with complex LSTM structure.

III. Our Approach

A. Architecture Overview

Video carries both spatial and temporal information in the form of the individual frame appearance and the motion across frames, respectively. Such nature inspires us to build our model upon a two-stream architecture to better process the appearance and motion information, as shown in Fig. 2.

Two parallel DNN streams are used to handle the spatial and temporal information of an input video, *i.e.* the RGB frames and the optical flows, respectively. To learn the comprehensive spatiotemporal saliency representation, we fuse the saliency features from two streams by incorporating dense residual cross connections among different layers of two DNNs, in Section III-B. To further enhance the spatiotemporal saliency representations with multi-scale information, in Section III-C, we incorporate a composite attention module to learn the local and global attention priors. To learn the temporal attention transitions more efficiently from limited data, we introduce convGRU, a lightweight RNN structure, into our network. More details can be found in Section III-D. In Section III-F, we present the implementation details and training protocols.

B. Spatiotemporal Residual Network

1) Residual Learning: Before going deep into our spatiotemporal residual learning framework, we first give a brief introduction of residual learning [22], which is used as the building block of our model. Let $\mathbf{x}$ and $\mathcal{H}(\mathbf{x})$ denote the input feature and the desired underlying mapping, residual learning is learn the residual of identity mapping $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$, instead of the original, unreferenced mapping $\mathcal{H}(\mathbf{x})$ (see Fig. 3 (a)). Thus $\mathcal{H}(\mathbf{x})$ can be computed as:

$$\mathcal{H}(\mathbf{x}) = \mathbf{x} + \mathcal{F}(\mathbf{x}). \tag{1}$$

Such kind of residual design can ease the training of relatively deep network architecture.

2) Appearance and Motion Streams: Our appearance and motion streams contain several residual blocks (borrowed from the five convolution blocks of ResNet-50 [22]). The *l-th* ($l \in \{1, \ldots, L\}$) residual block is defined as:

$$\mathbf{x}_l = \mathbf{x}_{l-1} + \mathcal{F}\left(\mathbf{x}_{l-1}; \mathbf{W}_l\right), \tag{2}$$

where $\mathbf{x}_{l-1}$ and $\mathbf{x}_l$ are the input and output of the *l-th* block, respectively, and $\mathcal{F}$ is the nonlinear mapping with weights $\mathbf{W}_l$. Given a set of video frames $\{\mathbf{I}^t\}_{t=1}^T$ and the corresponding optical flows $\{\mathbf{O}^t\}_{t=1}^T$. The appearance stream takes one single video frame $\mathbf{I}^t$ as input, and produce the appearance feature. The optical flows $\{\mathbf{O}^t, \mathbf{O}^{t+1}, \ldots, \mathbf{O}^{t+S}\}$ from the neighboring S frames are fed into the motion stream for obtaining corresponding motion features.

3) Dense Residual Connection Across Two Streams: Previous methods largely build the two streams separately; only adopting late fusion 'outside' of the two streams. There is no information exchange between the two streams. Differently, in our model the two streams are more tightly incorporated to learn more comprehensive spatial and temporal features.

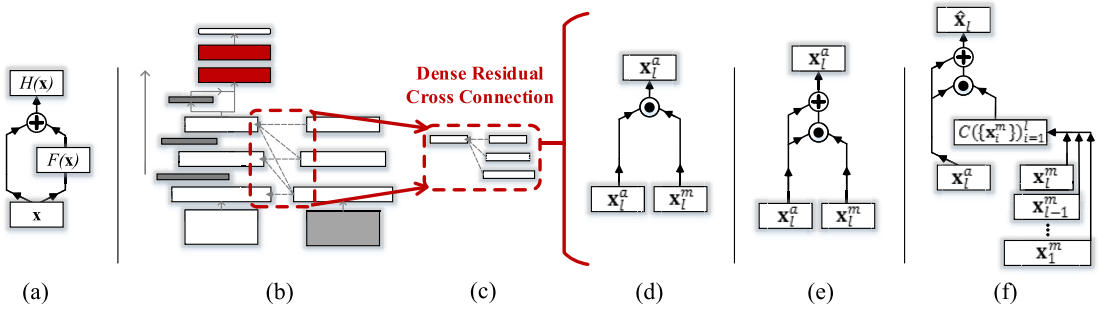


Fig. 3. **(a) Residual learning** is to learn the residual mapping $\mathcal{F}(\mathbf{x})$ instead of the original mapping $\mathcal{H}(\mathbf{x})$. The appearance and motion streams consist of several residual blocks, and are tightly coupled by **(b) dense residual cross connections** between corresponding layers. **(c) Dense residual cross connection** integrates appearance feature with multi-layer motion information for robust spatiotemporal saliency learning. To clarify, the naïve **(d) cross connection** between two streams directly using the motion feature as a gate mechanism to enhance appearance feature, which can be further improved by introducing residual structure, resulting in **(e) residual cross connection** design, and incorporating more comprehensive motion features from previous layers, yielding a **(f) dense residual cross connection** architecture. See Section III-B for more details.

This is achieved via adding dense residual cross connections between the corresponding layers (see Fig. 2 and Fig. 3 (c)).

Considering the fact that spatial stream tends to dominate the motion stream during training [62], we first inject the motion features $\mathbf{x}^m$ into the appearance stream, resulting in a *cross connection* strategy (see Fig. 3 (d)). Intuitively, the appearance feature $\mathbf{x}_l^a$ in *l-th* residual layer of appearance stream can be improved via:

$$\mathbf{x}_l^a \leftarrow \mathbf{x}_l^a \odot \mathbf{x}_l^m, \tag{3}$$

where '$\odot$' denotes the Hadamard product and $\mathbf{x}_l^m$ is the motion signal from the corresponding layer of motion stream, serving as a gate mechanism to enhance the appearance information. However, some tiny values in $\mathbf{x}_l^m$ may greatly suppress the appearance information. To fix ideas, we improve above equation with a residual form:

$$\mathbf{x}_l^a \leftarrow \mathbf{x}_l^a + \mathbf{x}_l^a \odot \mathbf{x}_l^m. \tag{4}$$

With above *residual cross connection* design (Fig. 3 (e)), potentially useful information in the original appearance feature can be preserved even if the motion signal closes to zero.

Further, we densely connect appearance feature in a certain layer l with a stack of motion features from previous lower levels: $\{1, \ldots, l\}$, yielding a *dense residual cross connection* architecture (Fig. 3 (f)), which can be formulated as:

$$\hat{\mathbf{x}}_l = \mathbf{x}_l^a + \mathbf{x}_l^a \odot \mathcal{C}(\{\mathbf{x}_i^m\}_{i=1}^l; \mathbf{W}_l^c), \tag{5}$$

where $\mathcal{C}$ is a nonlinear mapping with weight $\mathbf{W}_l^c$ which concatenates and squeezes the motion signals from different blocks. The $\hat{\mathbf{x}}$ represents the improved feature. The resulted spatiotemporal saliency representation with richer temporal information are more robust w.r.t. complex scenes.

Fig. 4 shows the predictions *w.* and *w/o.* spatiotemporal feature fusions. As seen, the model which effectively fuses saliency information from different domains can focus precisely on where human tend to look at. We further quantitatively demonstrate the effectiveness of the dense residual cross connections in Section IV-C.

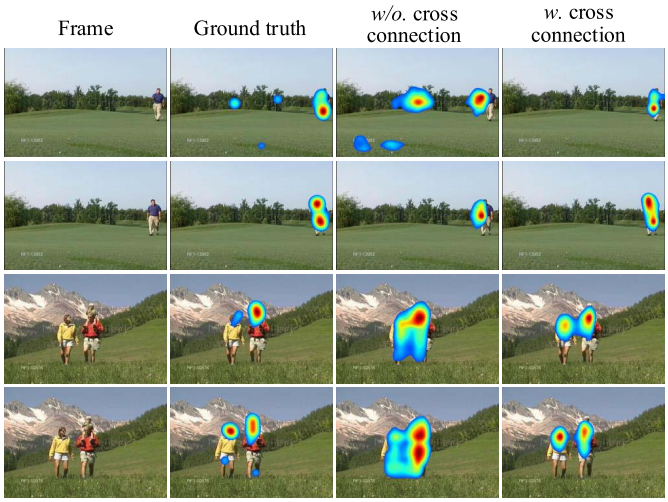


Fig. 4. Model *w.* effective spatiotemporal feature fusion (last column) is able to locate at human attention focuses more precisely than the one *w/o.* such information integration (3*rd* column). See Section III-B for more details.

C. Composite Attention Mechanism

The involvement of attention mechanism in neural network shows effectiveness in many tasks [63]–[67], since attention mechanism imposes the model to attend to the most task-relevant part of the inputs or features. In this section, we propose a composite attention mechanism to emphasize multi-scale saliency representation learning. It has a local attention module for enhancing spatiotemporal features in multi-scales, and a global attention module for capturing prior statistics from dynamic fixation data.

1) Local Attention Module: The fused spatiotemporal features at different levels have diverse receptive fields, which naturally serve as multi-scale saliency representations. To facilitate the learning of more powerful features, the local attention modules are hierarchically embedded to enhance the saliency features of multiple scales.

Let $\hat{\mathbf{x}}_l \in \mathbb{R}^{W_l \times H_l \times C_l}$ be the spatiotemporal saliency feature at the *l-th* layer of the appearance branch. The goal of the local attention module is to learn an attention mask $\mathbf{M}_l \in [0, 1]^{W_l \times H_l}$ to softly weight $\hat{\mathbf{x}}_l$ based on the saliency

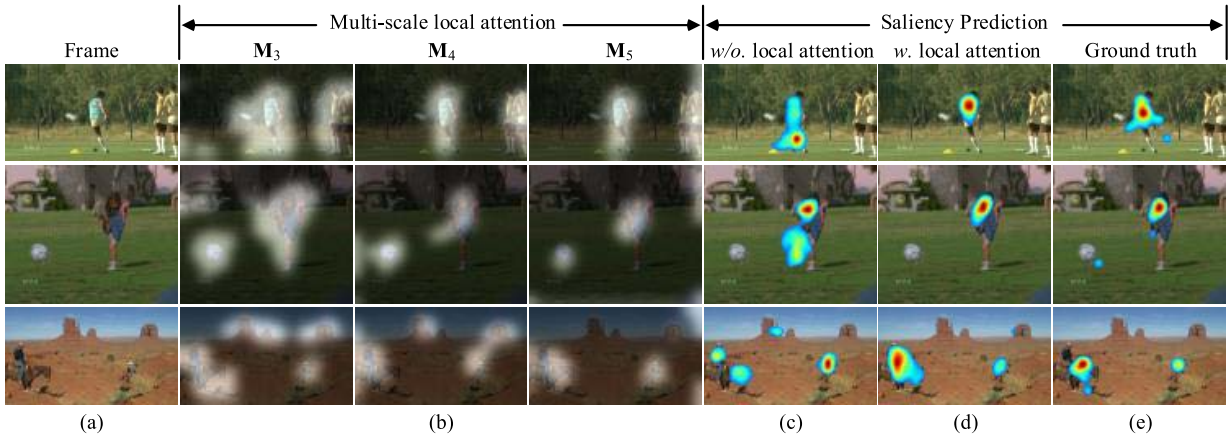


Fig. 5. Illustration of multi-scale local attention module. (a) Input video frame. (b) Multi-scale local attentions. (c) Saliency predictions *w/o.* local attention. (d) Saliency predictions *w.* local attention. (e) Ground-truth saliency maps.

information it carries. Specially, to reflect the importance of different regions with the attention mask, we apply a *softmax* operation to produce a normalized importance weight $\mathbf{M}_l$:

$$(\mathbf{M}_l)_{i,j} = \frac{exp\left(\mathcal{M}\left(\hat{\mathbf{x}}_l; \mathbf{W}_l^{att}\right)_{i,j}\right)}{\sum_{w=1}^{W_l}\sum_{h=1}^{H_l} exp\left(\mathcal{M}\left(\hat{\mathbf{x}}_l; \mathbf{W}_l^{att}\right)_{w,h}\right)}, \quad (6)$$

where $\mathbf{W}_l^{att}$ is the weight of non-linear mapping $\mathcal{M}$ that maps the spatiotemporal feature $\hat{\mathbf{x}}_l$ to a single channel mask, and the footnote i, j denotes the 2D spatial coordinate ($i \in \{1, \ldots, W_l\}$, $j \in \{1, \ldots, H_l\}$). In this way, the local attention module learns the normalized probability of each region of current spatiotemporal feature being important. The attention masks from different levels highlight multi-scale important areas and guide the network to gradually focus on the salient part. The enhanced feature can be given as:

$$(\hat{\mathbf{x}}_l)^c \leftarrow \mathbf{M}_l \odot (\hat{\mathbf{x}}_l)^c, \quad (7)$$

where $c \in \{1, \ldots, C_l\}$ is the channel index. However, directly multiplying the mask with feature may mistakenly discard important information at earlier training stage when the local attention module has not been well trained with selectiveness. In view of this, we introduce an identity mapping again, and re-formulate the local attention module in Eq. (7) as:

$$(\hat{\mathbf{x}}_l)^c \leftarrow (\hat{\mathbf{x}}_l)^c + \mathbf{M}_l \odot (\hat{\mathbf{x}}_l)^c. \quad (8)$$

The local attention module with residual structure avoids introducing drastic changes into the original features, thus allowing more effective learning. To further guarantee the discriminative ability of the enhanced spatiotemporal representations, we deeply supervise [73] the side-outputs of the multi-scale saliency features with ground truth fixation data, as illustrated in Fig. 2. Fig. 5 (b) visualizes the learned attentions at different scales, where the higher level attention gradually focus on the area where human tends to look at. With the enhancement of multi-scale attention, our model is able to make more precise predictions (see Fig. 5 (c) and (d)).

The enhanced saliency feature $\hat{\mathbf{x}}_l$ is then fed to the appearance stream and contributes to the appearance feature of next residual block. The calculation of the appearance features in Eq. (2) can be re-written as:

$$\mathbf{x}_l^a = \hat{\mathbf{x}}_{l-1} + \mathcal{F}\left(\hat{\mathbf{x}}_{l-1}; \mathbf{W}_l^a\right). \quad (9)$$

2) Global Attention Module: Photographers tend to position the important objects around the center when taking photos or recording videos, which encourages people to look for objects of interest at center when they are used to watching such kind of photos or videos [74]. Previous studies also showed that people tend to watch the center when there are no highly attractive contents in the images or videos [75]. Here we propose a global attention module to directly learn the nature bias of human fixation from the dynamic visual attention data.

The prior statistics of visual attention can be approximately modeled as a mixture of Gaussians [57]. Each Gaussian can be modeled by a 2D Gaussian function:

$$f(x, y) = \frac{1}{2\pi\rho_x\rho_y} exp\left(-\left(\frac{(x-\mu_x)^2}{2\rho_x^2} + \frac{(y-\mu_y)^2}{2\rho_y^2}\right)\right), \quad (10)$$

where μ_x, μ_y and ρ_x, ρ_y are means and standard deviations along two axes, respectively.

To learn the parameters of the mixture of Gaussians, a prior learning module [76] is incorporated which can be trained end-to-end with the dynamic fixation data. The corresponding prior maps are generated using the learned parameters, and then concatenated to the spatiotemporal features to further enhance the features with the global attention that are directly learned from training data. The visualized prior maps under different training settings are shown in Fig. 6, where different training data may result in different global priors.

D. ConvGRU

RNN is suitable for handling sequential data and is able to model the temporal relation within the sequence. Each recurrent cell has a hidden unit to maintain a dynamic memory according to the current hidden state and the new input:

$$\begin{aligned} \mathbf{h}^t &= \mathbf{W}_h^h \phi(\mathbf{h}^{t-1}) + \mathbf{W}_x^h \mathbf{x}^t, \\ \mathbf{y}^t &= \mathbf{W}_h^y \phi(\mathbf{h}^t), \end{aligned} \quad (11)$$

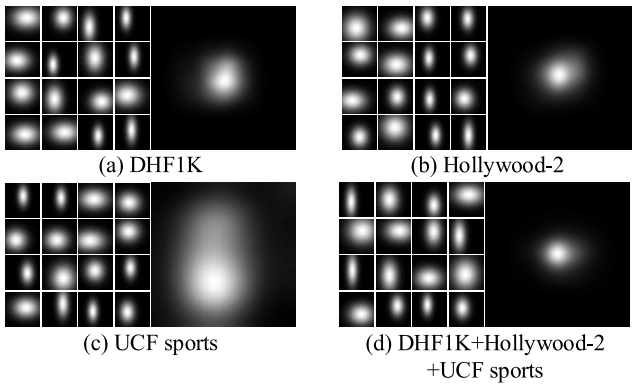


Fig. 6. Visualized prior maps using different training sets from (a) DHF1K, (b) Hollywood-2, (c) UCF sports, and (d) DHF1K+Hollywood-2+UCF sports. See Section IV-A for details about the training settings. For each sub-figure, the left shows the learned gaussian function maps, and the right shows the integrated prior map under certain network parameters.

where $\mathbf{x}, \mathbf{y}, \mathbf{h}$ represent the input, output, and hidden states, respectively, $\mathbf{W}$s are the learnable weights and ϕ is the activation function. For video related tasks, a natural thought is to resort to RNN for modeling the temporal transitions. However, such *vanilla* RNN easily suffers from gradient vanishing/exploding. One promising solution is a newly proposed recurrent unit, GRU [68], which alleviates the above problem by introducing gating units to control the information flow across time steps:

$$\begin{aligned}
\mathbf{z}^t &= \sigma(\mathbf{W}_h^z \mathbf{h}^{t-1} + \mathbf{W}_x^z \mathbf{x}^t),\\
\mathbf{r}^t &= \sigma(\mathbf{W}_h^r \mathbf{h}^{t-1} + \mathbf{W}_x^r \mathbf{x}^t),\\
\tilde{\mathbf{h}}^t &= tanh(\mathbf{W}_h^h (\mathbf{r}^t \odot \mathbf{h}^{t-1}) + \mathbf{W}_x^h \mathbf{x}^t),\\
\mathbf{h}^t &= \mathbf{z}^t \tilde{\mathbf{h}}^t + (\mathbf{1} - \mathbf{z}^t)\mathbf{h}^{t-1},
\end{aligned} \tag{12}$$

where the update gate $\mathbf{z}^t$ controls how much previous hidden state $\mathbf{h}^{t-1}$ to be kept, and the reset gate $\mathbf{r}^t$ defines how to combine current input with the previous hidden state. Compared with another popular candidate, *i.e.*, LSTM, GRU can model the long-term relationships with simpler architecture, thus is more applicable when the training datasets are relatively small.

To better formulate the visual attention transitions across video frames, we adapt GRU for handling multi-dimensional inputs by replacing the matrix multiplications with convolution operations, which is formulated as:

$$\begin{aligned}
\mathbf{z}^t &= \sigma(\mathbf{W}_h^z * \mathbf{h}^{t-1} + \mathbf{W}_x^z * \mathbf{x}^t),\\
\mathbf{r}^t &= \sigma(\mathbf{W}_h^r * \mathbf{h}^{t-1} + \mathbf{W}_x^r * \mathbf{x}^t),\\
\tilde{\mathbf{h}}^t &= tanh(\mathbf{W}_h^h * (\mathbf{r}^t \odot \mathbf{h}^{t-1}) + \mathbf{W}_x^h * \mathbf{x}^t),\\
\mathbf{h}^t &= \mathbf{z}^t \tilde{\mathbf{h}}^t + (\mathbf{1} - \mathbf{z}^t)\mathbf{h}^{t-1}.
\end{aligned} \tag{13}$$

where '$*$' is the convolution operator and all the parameters, *i.e.*, $\mathbf{x}$, $\mathbf{z}$, $\mathbf{r}$, $\mathbf{h}$, $\mathbf{W}$s, are 3D tensors. An illustration of convGRU architecture can be found in Fig. 2. The convGRU module in our network takes the sequence of the top-level spatiotemporal features $\{\hat{\mathbf{x}}_L^t\}_{t=1}^T$ as inputs to learn the dynamic fixation transitions. The output sequence of features accord with the spatiotemporal pattern of dynamic visual attention and are temporally smoother in modeling the attention transitions.

E. Loss Function

Similar to [6], our loss function is a linear combination of four loss terms adapted from saliency evaluation metrics, which can comprehensively measure the quality of the predictions. The proposed loss function is defined as:

$$\begin{aligned}
\mathcal{L}(\mathbf{P}, \mathbf{Q}, \mathbf{G}) = \mathcal{L}_{kl}(\mathbf{P}, \mathbf{G}) + \omega_1 \mathcal{L}_{cc}(\mathbf{P}, \mathbf{G})\\
+\omega_2 \mathcal{L}_{nss}(\mathbf{P}, \mathbf{Q}) + \omega_3 \mathcal{L}_{sim}(\mathbf{P}, \mathbf{G}),
\end{aligned} \tag{14}$$

where $\mathbf{P}$ is the predicted attention map, $\mathbf{Q}$ is the ground-truth binary fixation map, and $\mathbf{G}$ is the continuous ground-truth attention map. ωs are scalars to balance the four loss terms.

$\mathcal{L}_{kl}$ is derived from the *Kullback-Leibler (KL) divergence* metric. It views $\mathbf{P}, \mathbf{Q}$ as probability distributions and measures the information loss of using $\mathbf{P}$ to approximate $\mathbf{G}$:

$$\mathcal{L}_{kl}(\mathbf{P}, \mathbf{G}) = \sum\nolimits_i G_i log\left(\frac{G_i}{P_i}\right). \tag{15}$$

$\mathcal{L}_{cc}$ is adopted from the *Linear Correlation Coefficient (CC)* metric which measures the linear relationship between two random variables. The calculation is defined as follows:

$$\mathcal{L}_{cc}(\mathbf{P}, \mathbf{G}) = -\frac{cov(\mathbf{P}, \mathbf{G})}{\rho(\mathbf{P})\rho(\mathbf{G})}, \tag{16}$$

where $cov(\mathbf{P}, \mathbf{G})$ is the covariance of $\mathbf{P}$ and $\mathbf{G}$, and $\rho(\cdot)$ represents the standard deviation.

$\mathcal{L}_{nss}$ is from the *Normalized Scanpath Saliency (NSS)* metric calculating the average saliency value at positive positions:

$$\mathcal{L}_{nss}(\mathbf{P}, \mathbf{Q}) = -\frac{1}{N}\sum\nolimits_i \frac{P_i - \mu(\mathbf{P})}{\rho(\mathbf{P})} \cdot Q_i, \tag{17}$$

where $\mu(\cdot)$ and $\rho(\cdot)$ are the mean and the standard deviation, respectively. N is the number of positive pixels in $\mathbf{Q}$.

$\mathcal{L}_{sim}$ is adopted from the *Similarity (SIM)* metric concerning the intersection between two probability distributions. Larger interaction indicates stronger similarity:

$$\mathcal{L}_{sim}(\mathbf{P}, \mathbf{G}) = -\sum\nolimits_i min(P_i', G_i'). \tag{18}$$

In the calculation of $\mathcal{L}_{sim}$, $\mathbf{P}$ and $\mathbf{G}$ are normalized to be probability distributions P_i' and G_i' satisfying $\sum_i P_i' = 1$ and $\sum_i G_i' = 1$. Note that *CC, NSS* and *SIM* are similarity metrics, hence their negatives are adopted as the loss terms.

Considering the supervisions applied at multi-scale spatiotemporal features and the final output, the overall loss is:

$$\mathcal{L}_{total} = \alpha_{out}\mathcal{L}(\mathbf{P}_{out}, \mathbf{Q}, \mathbf{G}) + \sum\nolimits_{l=3}^{L} \alpha_l \mathcal{L}(\mathbf{P}_l, \mathbf{Q}, \mathbf{G}), \tag{19}$$

where $\mathbf{P}_{out}$ is the final prediction, $\mathbf{P}_l$ is the side output of the enhanced spatiotemporal feature $\hat{\mathbf{x}}_l$. Note that the incorporation of spatiotemporal feature starts from the output of the $3rd$ block and $L = 5$. The α_{out} and $\{\alpha_l\}_{l=3}^L$ are the weights of final-output loss and side-output losses, respectively.

F. Implementation Details

1) Network Details: We build our spatiotemporal residual network based on two ResNet-50 [22] for handling appearance and motion information, respectively. To obtain feature maps with finer resolution, for each stream we remove the fully-connected layer and change the strides of the last two blocks

to 1. To preserve the receptive field, we further modify the dilation rates of the last two blocks to 2 following [76].

The composite attention mechanism includes local attentions and global attention priors. The local attention module is built as: $Conv(3 \times 3, \lfloor \frac{C}{2} \rfloor) \rightarrow ReLU \rightarrow Batch\ Normalization$ (BN) $\rightarrow Conv(1 \times 1, 1) \rightarrow Softmax$, where C is the channel number of the input feature. Multiple local attention modules are hierarchically embedded to enhance multi-scale spatiotemporal features from different levels. The side output of each enhanced spatiotemporal feature is obtained through a $Conv(1 \times 1, 1)$ layer with *Sigmoid* activation. To capture the global attention priors, we incorporate a prior module [76] for learning a set of Gaussian parameters $\{\mu_x^k, \mu_y^k, \rho_x^k, \rho_y^k\}_{k=1}^K$ from the training data, and set $K = 16$ in our experiments.

The convGRU layer contains 256 filters of kernel size 3. As shown in Fig. 2, given an input video sequence $\{\mathbf{I}^t \in [0, 255]^{224\times224\times3}\}_{t=1}^T$ with typical 224×224 spatial resolution, the convGRU takes the top-level enhanced spatiotemporal feature sequence $\{\hat{\mathbf{x}}_L^t \in \mathbb{R}^{28\times28\times2048}\}_{t=1}^T$ as inputs for learning dynamic attention transitions. The output feature maps are concatenated with K global attention maps generated with the parameters learned by global attention module. The concatenated features are fed to a $Conv(1 \times 1, 1)$ layer with *Sigmoid* activation to generate the final attention predictions $\{\mathbf{P}_{out}^t \in [0, 1]^{28\times28}\}_{t=1}^T$. The outputs are bilinearly up-sampled to the size of the ground truths for quantitative evaluation.

2) Training Details: We train the spatiotemporal residual module and our convGRU in tandem. We first initialized the spatiotemporal feature learning part (the two streams) with the weights of ResNet-50 pre-trained on ImageNet, and trained it under certain settings, with deep supervision on the three side-outputs. Then we randomly initialized the temporal transition learning part and trained the whole network, where the spatiotemporal feature extraction part has been well trained.

The video batch size is 1, and the frame batch size is 5, *i.e*. we choose 5 frames from a single video as a training data batch. The optical flows are generated before training using Flownet 2.0 [78], and the absolute values of the magnitudes are confined between 0.1 and 20 to avoid noise. The number of stacked optical flows, S, is set to be 5, *i.e*. we feed 5 consecutive optical flows into the motion stream each time a single frame is fed into the appearance stream. Our model is implemented using *Keras* with *Tensorflow* backend. We choose Adam [79] with initial learning rate 10^{-4} to minimize the total loss in Eq. (19). We set $\alpha_{out} = \alpha_l = 1$ to equally weight the final-output loss and side-output losses. We set $\omega_1 = 0.2$, $\omega_2 = \omega_3 = 0.1$ in Eq. (14) to emphasize the importance of $\mathcal{L}_{kl}$.

IV. Experiments

In Section IV-A, we review the datasets on which we train our models and make comparisons with other models. The qualitative and quantitative comparisons can be found in Section IV-B, and the ablation study is shown in Section IV-C.

A. Datasets

There are four widely used datasets for dynamic visual attention, which differ in number of video samples, number of participants, experimental settings, viewing purpose *etc*.

1) DHF1K [17]: includes 1K videos with diverse contents of 7 main categories, such as human activities, arts, animals *etc*. It also covers varied motion patterns and various foreground objects for studying human attention allocation in dynamic viewing. The eye-tracking data come from 17 observers performing free-viewing. These videos are randomly split into 600/100/300 training/validation/test samples, respectively.

2) Hollywood-2 [25]: consists of 1,707 videos from the Hollywood-2 action recognition dataset [91], containing 12 actions such as eating, running, kissing, *etc*. The groundtruth fixations are the eye-tracking data from 19 observers. The training and testing sets include 823 and 884 videos, respectively, and can be further divided into multiple shots following the ground-truth shot boundaries.

3) UCF Sports [25]: contains 150 videos from the UCF sports action dataset [92] with 9 sports actions such as diving, golfing *etc*. The participants were instructed to identify the actions during experiments, which inevitably bias the fixation data with special viewing purpose. Statistics reveals that most of the fixations are located at human body regions [17]. It contains 103 videos for training and 47 videos for testing.

4) DIEM [26]: has 84 videos with varying styles, in which the gaze data is collected from 50 participants per video. The videos are from professional movies which are intentionally edited to attract human attention to certain objects in the scene. Following [53], the testing set contains 20 selected videos, and the remaining videos belong to the training set.

We use the standard experiment protocol in [17] for fare comparison, *i.e*., considering 4 training settings with the training set(s) from (i) DHF1K, (ii) Hollywood-2, (iii) UCF sports, and (iv) DHF1K + Hollywood-2 + UCF sports.[1] The performance is evaluated on the testing sets of DHF1K, Hollywood-2 and UCF sports. To further evaluate the generalization ability of the proposed method, we also conduct experiments on the DIEM, whose videos are not used for training.

B. Comparison Results

We quantitatively compare our model with 22 saliency models, including 17 dynamic models (PQFT [80], Seo and Milanfar [81], Rudoy *et al*. [82], Hou and Zhang [83], Fang *et al*. [55], OBDL [53], AWS-D [84], PMES [85], PIM-ZEN [87], MAM [86], PIM-MCS [88], MCSDM [89], MSM-SM [54], PNSP-CS [90], OM-CNN [19], Two-stream [18],[2] and ACLNet [17]) and 6 static models (ITTI [1], GBVS [49], SALICON [6], DVA [11], Shallow-Net [8], and Deep-Net [8]), among which, OM-CNN, Two-stream, SALICON, DVA, Shallow-Net, and Deep-Net are deep learning models, and others are classical ones. We consider a variety of different metrics, including AUC-Judd, shuffled AUC (s-AUC), Linear Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), and Similarity (SIM). All the

[1] We would recommend later researchers to compare with our model under training setting (iv).

[2] We re-implemented [18] since the official codes cannot run correctly.

Fig. 7. Qualitative comparisons with other methods on: (a) DHF1K; (b) Hollywood-2; (c) UCF sports; (d) DIEM. For each dataset, we present one example video with three representative frames for demonstration.

metrics are widely used and accepted for evaluating a visual attention model. See [39] for detailed introductions.

1) Performance on DHF1K: We evaluate our model on the testing set of DHF1K containing 300 videos. The qualitative comparison over 3 frames, which are selected to show the temporal transitions of visual attention in Fig. 7 (a). There exist inconsistent movements of foreground objects at different time points, which raises challenges for attention prediction. Our model can precisely predict the visual attention locations and comprehensively model dynamic attention transitions across frames. This can be attributed to the efficient spatiotemporal feature fusion, the composite attention mechanism and the con-vGRU which models long-term dependencies. The quantitative evaluation results are shown in Table I. Our model consistently outperforms other methods in all the metrics.

2) Performance on Hollywood-2: Hollywood-2 is a large dataset with 884 testing video sequences, all of which are movie scenes. Fig. 7 (b) shows an outdoor scene where a woman is hurrying on a bustling street. Despite various shooting environments and complex action patterns, our model consistently provides better predictions w.r.t. all other methods. Table II shows the quantitative evaluation results. Again, our model achieves better results over most of the metrics.

3) Performance on UCF Sports: The testing set of UCF sports contains 47 sports videos. The visual comparisons are shown in Fig. 7 (c), where an athlete swinging on the high bar. As seen, our attention model can capture dynamic attentions of different motions and provide better prediction results. The evaluation results are shown in Table III. Our model achieves the best performance compared with all other models.

4) Performance on DIEM: Following [53], we evaluate our model under four training settings on the testing set of DIEM, where the first 300 frames of each video are condidered. The qualitative comparison over 3 frames are shown in Fig. 7 (d). Table IV shows the quantitative results. Our model achieves

TABLE I

QUANTITATIVE RESULTS ON DHF1K. (THE BEST SCORES ARE MARKED IN **BOLD**. THIS NOTE IS THE SAME FOR OTHER TABLES.)

	Methods	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
	*center prior	0.854	0.238	0.503	0.302	0.167
Dynamic Models	*PQFT [80]	0.699	0.139	0.562	0.137	0.749
	*Seo *et al.* [81]	0.635	0.142	0.499	0.070	0.334
	*Rudoy *et al.* [82]	0.769	0.214	0.501	0.285	1.498
	*Hou *et al.* [83]	0.726	0.167	0.545	0.150	0.847
	*Fang *et al.* [55]	0.819	0.198	0.537	0.273	1.539
	*OBDL [53]	0.638	0.171	0.500	0.117	0.495
	*AWS-D [84]	0.703	0.157	0.513	0.174	0.940
	*PMES [85]	0.545	0.093	0.502	0.055	0.237
	*MAM [86]	0.551	0.108	0.500	0.041	0.214
	*PIM-ZEN [87]	0.552	0.095	0.498	0.062	0.280
	*PIM-MCS [88]	0.551	0.094	0.499	0.053	0.242
	*MCSDM [89]	0.591	0.110	0.500	0.047	0.247
	*MSM-SM [54]	0.582	0.143	0.500	0.058	0.245
	*PNSP-CS [90]	0.543	0.085	0.499	0.028	0.121
	OM-CNN [19]	0.856	0.256	0.583	0.344	1.911
	Two-stream [18]	0.834	0.197	0.581	0.325	1.632
	ACLNet [17]	0.890	0.315	0.601	0.434	2.354
Static Models	*ITTI [1]	0.774	0.162	0.553	0.233	1.207
	*GBVS [49]	0.828	0.186	0.554	0.283	1.474
	SALICON [6]	0.857	0.232	0.590	0.327	1.901
	Shallow-Net [8]	0.833	0.182	0.529	0.295	1.509
	Deep-Net [8]	0.855	0.201	0.592	0.331	1.775
	DVA [11]	0.860	0.262	0.595	0.358	2.013
Ours	Training setting (i)	0.890	0.323	0.634	0.427	2.305
	Training setting (ii)	0.885	0.332	0.635	0.433	2.400
	Training setting (iii)	0.866	0.307	0.655	0.378	2.108
	Training setting (iv)	**0.895**	**0.355**	**0.663**	**0.458**	**2.558**

* Non-deep learning model.

TABLE II

QUANTITATIVE RESULTS ON HOLLYWOOD-2 DATASET

	Methods	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
	*center prior	0.869	0.331	0.615	0.421	1.808
Dynamic Models	*PQFT [80]	0.723	0.201	0.621	0.153	0.755
	*Seo *et al.* [81]	0.652	0.155	0.530	0.076	0.346
	*Rudoy *et al.* [82]	0.783	0.315	0.536	0.302	1.570
	*Hou *et al.* [83]	0.731	0.202	0.580	0.146	0.684
	*Fang *et al.* [55]	0.859	0.272	0.659	0.358	1.667
	*OBDL [53]	0.640	0.170	0.541	0.106	0.462
	*AWS-D [84]	0.694	0.175	0.637	0.146	0.742
	*PMES [85]	0.696	0.180	0.620	0.177	0.867
	*MAM [86]	0.630	0.153	0.562	0.099	0.494
	*PIM-ZEN [87]	0.670	0.167	0.598	0.134	0.667
	*PIM-MCS [88]	0.663	0.163	0.570	0.118	0.584
	*MCSDM [89]	0.618	0.147	0.524	0.067	0.288
	*MSM-SM [54]	0.683	0.180	0.561	0.132	0.682
	*PNSP-CS [90]	0.647	0.146	0.548	0.077	0.370
	OM-CNN [19]	0.887	0.356	0.693	0.446	2.313
	Two-stream [18]	0.863	0.276	0.710	0.382	1.748
	ACLNet [17]	0.913	**0.542**	0.757	0.623	3.086
Static Models	*ITTI [1]	0.788	0.221	0.607	0.257	1.076
	*GBVS [49]	0.837	0.257	0.633	0.308	1.336
	SALICON [6]	0.856	0.321	0.711	0.425	2.013
	Shallow-Net [8]	0.851	0.276	0.694	0.423	1.680
	Deep-Net [8]	0.884	0.300	0.736	0.451	2.066
	DVA [11]	0.886	0.372	0.727	0.482	2.459
Ours	Training setting (i)	0.909	0.446	0.736	0.543	2.581
	Training setting (ii)	0.922	0.536	**0.779**	0.650	3.475
	Training setting (iii)	0.892	0.428	0.742	0.509	2.497
	Training setting (iv)	**0.923**	0.536	0.774	**0.662**	**3.478**

* Non-deep learning model.

competitive results without training or finetuning on the training set of DIEM.

C. Ablation Study

In this section, we show the effectiveness of each main component described in Section III by experimenting on several model variants under training setting (iii), and evaluate these models on the testing set of UCF sports.

1) Spatiotemporal Feature Fusion: We incorporate dense residual cross connections between two streams to fuse the spatial and temporal features (Section III-B). To validate the effectiveness of the feature fusion for learning better spatiotemporal representations, we experiment on four variants, namely *w/o. cross connection*, *w/o. dense connection* and *w/o. residual mapping*, meanwhile preserving the composite attention mechanism, the convGRU and the deep supervision at the features of 3*rd*-5*th* blocks unless otherwise specified.

The *w/o. cross connection* refers to the model without any early interactions between two streams. The feature fusion only happens at the last block where the outputs from two streams are concatenated and squeezed. To adapt the composite attention mechanism for this situation, we double the number of local attention modules, and embed these modules into both streams to separately enhance the spatial and temporal features, *i.e.* $\{\mathbf{x}_l^a\}_{l=3}^5$ and $\{\mathbf{x}_l^m\}_{l=3}^5$. The model *w/o. dense connection* indicates that the residual cross connections are only added between the equivalent blocks of the two streams in the model. For *w/o. residual mapping*, the fusion is performed by directly multiplying a condensed stack of motion features from several previous blocks to the appearance feature of current block. In the last two cases, all the fusions start from the output of the 3*rd* block, as described in Section III-F. Table V shows that gating appearance feature using motion information can result in better spatiotemporal saliency representations. The dense and the residual mapping structures further boost the prediction precision.

2) Composite Attention Mechanism: Our composite attention mechanism consists of the local and global attention module, in Section III-C. We experiment on four variants w.r.t. the attention mechanism, without modifying the spatiotemporal feature fusion strategy and the convGRU. The first case *w/o. composite attention* is obtained by disabling both the local and global attention modules in the complete model. The second case *w/o. global attention* only disables the global attention module, while keeping embedded local attention modules functioning. The third case *w/o. local attention* disables all the local attention modules in the complete model, and only learns attention using the global attention module. To further verify the residual structure in the local attention module, the fourth case *w/o. local residual structure* removes the identity mappings of the local attention modules from the complete model, which corresponds to the case in Eq. (7).

TABLE III
QUANTITATIVE RESULTS ON UCF SPORTS DATASET

	Methods	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
	*center prior	0.834	0.299	0.566	0.350	1.585
Dynamic Models	*PQFT [80]	0.825	0.250	0.722	0.338	1.780
	*Seo *et al.* [81]	0.831	0.308	0.666	0.336	1.690
	*Rudoy *et al.* [82]	0.763	0.271	0.637	0.344	1.619
	*Hou *et al.* [83]	0.819	0.276	0.674	0.292	1.399
	*Fang *et al.* [55]	0.845	0.307	0.674	0.395	1.787
	*OBDL [53]	0.759	0.193	0.634	0.234	1.382
	*AWS-D [84]	0.823	0.228	0.750	0.306	1.631
	*PMES [85]	0.756	0.263	0.714	0.349	1.788
	*MAM [86]	0.669	0.213	0.624	0.218	1.130
	*PIM-ZEN [87]	0.760	0.234	0.702	0.306	1.657
	*PIM-MCS [88]	0.777	0.238	0.695	0.303	1.596
	*MCSDM [89]	0.756	0.228	0.626	0.230	1.091
	*MSM-SM [54]	0.752	0.262	0.634	0.280	1.584
	*PNSP-CS [90]	0.755	0.210	0.628	0.218	1.091
	OM-CNN [19]	0.870	0.321	0.691	0.405	2.089
	Two-stream [18]	0.832	0.264	0.685	0.343	1.753
	ACLNet [17]	0.905	0.496	0.767	0.603	3.200
Static Models	*ITTI [1]	0.847	0.251	0.725	0.356	1.640
	*GBVS [49]	0.859	0.274	0.697	0.396	1.818
	SALICON [6]	0.848	0.304	0.738	0.375	1.838
	Shallow-Net [8]	0.846	0.276	0.691	0.382	1.789
	Deep-Net [8]	0.861	0.282	0.719	0.414	1.903
	DVA [11]	0.872	0.339	0.725	0.439	2.311
Ours	Training setting (i)	0.902	0.426	0.728	0.525	2.582
	Training setting (ii)	0.887	0.434	0.701	0.518	2.581
	Training setting (iii)	**0.914**	**0.535**	**0.790**	**0.645**	**3.472**
	Training setting (iv)	0.910	0.479	0.751	0.593	3.018

* Non-deep learning model.

TABLE IV
QUANTITATIVE RESULTS ON DIEM DATASET

	Methods	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
	*center prior	0.714	0.112	0.526	0.125	0.783
Dynamic Models	*PQFT [80]	0.724	0.126	0.649	0.144	0.856
	*Seo *et al.* [81]	0.723	0.130	0.568	0.116	0.665
	*Rudoy *et al.* [82]	0.775	0.150	0.618	0.260	1.390
	*Hou *et al.* [83]	0.735	0.142	0.589	0.128	0.735
	*Fang *et al.* [55]	0.823	0.167	0.636	0.251	1.423
	*OBDL [53]	0.762	0.165	0.694	0.221	1.289
	*AWS-D [84]	0.774	0.150	0.695	0.216	1.252
	*PMES [85]	0.657	0.122	0.607	0.142	0.817
	*MAM [86]	0.579	0.089	0.552	0.072	0.408
	*PIM-ZEN [87]	0.660	0.114	0.615	0.132	0.757
	*PIM-MCS [88]	0.662	0.110	0.607	0.124	0.709
	*MCSDM [89]	0.663	0.105	0.558	0.084	0.466
	*MSM-SM [54]	0.619	0.092	0.571	0.107	0.624
	*PNSP-CS [90]	0.637	0.091	0.559	0.074	0.417
	OM-CNN [19]	0.857	0.238	0.693	0.371	2.235
	Two-stream [18]	0.859	0.256	0.682	0.366	2.171
	ACLNet [17]	**0.881**	0.277	0.693	0.396	2.368
Static Models	*ITTI [1]	0.791	0.132	0.653	0.196	1.103
	*GBVS [49]	0.813	0.156	0.633	0.214	1.198
	SALICON [6]	0.793	0.171	0.674	0.270	1.650
	Shallow-Net [8]	0.838	0.188	0.620	0.297	1.646
	Deep-Net [8]	0.849	0.164	0.697	0.291	1.650
	DVA [11]	0.868	0.237	**0.721**	0.386	2.347
Ours	Training setting (i)	0.874	0.270	0.678	0.382	2.210
	Training setting (ii)	0.855	0.295	0.651	0.376	2.265
	Training setting (iii)	0.831	0.235	0.654	0.230	1.798
	Training setting (iv)	0.870	**0.306**	0.678	**0.408**	**2.452**

* Non-deep learning model.

The evaluation results are shown in Table V. By comparing *w/o. composite attention* and *w/o. global attention*, or *w/o. local attention* and *Ours*, we may conclude that the involvement of local attention module can remarkably improve almost all the five metrics except AUC-J. The AUC-J can be further enhanced after incorporating the global attention module, as seen from *w/o. composite attention* and *w/o. local attention*, or *w/o. global attention* and *Ours*. From *w/o. local residual structure* and *Ours*, we see that the identity mapping in the local attention module is essential for precise prediction.

3) ConvGRU: The convGRU is a lightweight RNN which is able to model the temporal transitions of dynamic visual attention, and can be well trained with relatively small datasets due to its simple structure. The ability of modeling temporal visual attention trainsition is significant for preidicting the saliency shifts when human is faced with multiple saliency targets [93] or is observing dynamic scenes [36].

To varify the effectiveness of the convGRU module in the complete model, we first replace the convGRU with a convolutional layer which has the same number of filters and kernel size (*w. convolution*). To compare convGRU with convLSTM, we replace the convGRU with a convLSTM of the same number of filters and kernel size (*w. convLSTM*).

As shown in Table V, convGRU outperforms the other two variants in almost all the metrics, suggesting that convGRU would be the better choice for video saliency prediction compared with feed-forward convolution and convLSTM. The example in Fig. 8 demonstrates the ability of the convGRU module to model the saliency shifts, *e.g.* from the ball to the head. The *w. convolution* keeps focusing on the head, while the *w. convLSTM* focus on the head instead of the ball in the 1*st* frame and it pays almost equal attention on the body when it is expected to look at the head in the 4*th* frame.

TABLE V
ABLATION STUDY ON UCF SPORTS DATASET. SEE SECTION IV-C FOR DETAILS

Aspects	Variants	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
	Ours Training setting (iii)	**0.914**	**0.535**	**0.790**	**0.645**	**3.472**
Spatiotemporal Feature Fusion	*w/o.* cross connection	0.896	0.512	0.775	0.613	3.110
	w/o. dense connection	0.909	0.518	0.772	0.626	3.384
	w/o. residual mapping	0.906	0.496	0.764	0.591	3.151
Composite Attention Mechanism	*w/o.* composite attention	0.910	0.515	0.772	0.620	3.321
	w/o. global attention	0.908	0.521	0.773	0.627	3.382
	w/o. local attention	0.912	0.514	0.766	0.619	3.355
	w/o. local residual structure	0.898	0.501	0.754	0.603	3.296
ConvGRU	*w.* convolution	0.907	0.524	0.774	0.639	3.403
	w. convLSTM	0.911	0.522	0.782	0.629	3.376
Side Outputs	$\hat{\mathbf{x}}_3$ side output	0.893	0.461	0.752	0.558	2.963
	$\hat{\mathbf{x}}_4$ side output	0.906	0.516	0.775	0.621	3.340
	$\hat{\mathbf{x}}_5$ side output	0.912	0.529	0.788	0.640	3.354
Loss function	$\mathcal{L}_{sim}$	0.908	0.522	0.778	0.631	3.355

4) Side Outputs: We study the discriminative property of the enhanced spatiotemporal features by evaluating the side outputs. Three additional baselines are introduced, namely

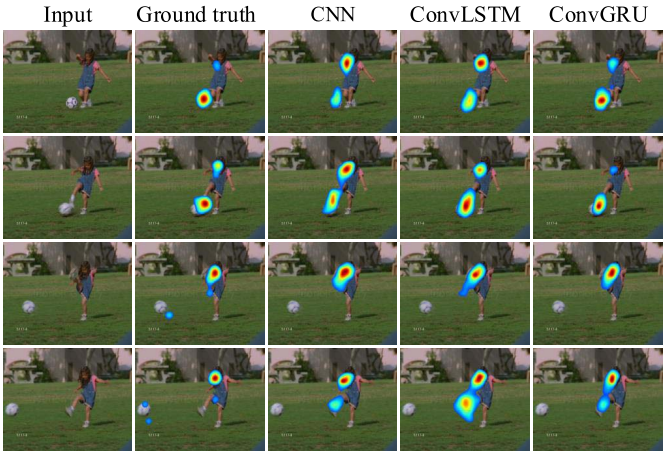


Fig. 8. Visual comparisons for temporal transition learning using CNN, convLSTM and convGRU.

$\hat{\mathbf{x}}_3$ *side output*, $\hat{\mathbf{x}}_4$ *side output* and $\hat{\mathbf{x}}_5$ *side output*, which correspond to the side outputs from three enhanced spatiotemporal representations. As shown in Table V, the side outputs gradually becomes better since higher layers can make finer predictions. The final prediction achieves the best performance.

V. Discussion and Conclusion

This paper proposes a spatiotemporal residual attentive network for dynamic attention prediction. It extends two-stream architecture with dense residual cross connections, encouraging information exchange between two streams and resulting in more powerful spatiotemporal saliency representations. To further enhance the spatiotemporal features with multi-scale information, we propose a composite attention mechanism which learns a stack of local attentions as well as global attention priors to filter out unrelated information. We further incorporate the lightweight convGRU to model the temporal characteristics of dynamic fixation, allowing more efficient learning with small training dataset. Extensive results demonstrate the superiority of the proposed model to precisely locate dynamic human fixations as well as capture the temporal attention transitions. The ablation experiments show the effectiveness of each component in our model.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[2] J. Shen, J. Peng, X. Dong, L. Shao, and F. Porikli, "Higher order energies for image segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4911–4922, Oct. 2017.

[3] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[4] G. Evangelopoulos *et al.*, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[5] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[6] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 262–270.

[7] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 362–370.

[8] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 598–606.

[9] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 392–404, Feb. 2016.

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

[11] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

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

[13] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Appl. Sci. Neural Netw., Fuzzy Syst., Evol. Comput. VI*, vol. 5200, pp. 64–79, Dec. 2003.

[14] L. Zhang, M. H. Tong, and G. W. Cottrell, "SUNDAY: Saliency using natural statistics for dynamic analysis of scenes," in *Proc. AAAI Annu. Cognit. Sci. Conf.*, 2009, pp. 2944–2949.

[15] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan, "Regularized feature reconstruction for spatio-temporal saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3120–3132, Aug. 2013.

[16] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.

[17] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4894–4903.

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

[19] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 602–617.

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

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

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[23] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3468–3476.

[24] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[25] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.

[26] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cogn. Comput.*, vol. 3, no. 1, pp. 5–24, 2011.

[27] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. H. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3064–3074.

[28] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 186–202.

[29] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1448–1457.

[30] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1–10.

[31] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[32] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[33] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4819–4831, Oct. 2019.

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

[35] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper ConvLSTM for video salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 715–731.

[36] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8554–8564.

[37] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[38] W. Wang, J. Shen, and H. Lin, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, 2019.

[39] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*. [Online]. Available: https://arxiv.org/abs/1904.09146

[40] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[41] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[42] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[43] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 481–488.

[44] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 155–162.

[45] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[46] M. Xu, Y. Ren, and Z. Wang, "Learning to predict saliency on face images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3907–3915.

[47] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[49] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 545–552.

[50] S.-H. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1063–1069.

[51] F. Zhou, S. B. Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3358–3365.

[52] J. Shen, J. Peng, and L. Shao, "Submodular trajectories for better motion segmentation in videos," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2688–2700, Jun. 2018.

[53] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5501–5510.

[54] K. Muthuswamy and D. Rajan, "Salient motion detection in compressed domain," *IEEE Signal Process. Lett.*, vol. 20, no. 10, pp. 996–999, Oct. 2013.

[55] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sep. 2014.

[56] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with hevc features," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 369–385, Jan. 2017.

[57] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[58] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[59] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4694–4702.

[60] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 461–470.

[61] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1933–1941.

[62] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7445–7454.

[63] C. Cao *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2956–2964.

[64] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 21–29.

[65] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.

[66] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2015, pp. 379–389.

[67] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2016, pp. 2249–2255.

[68] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*. [Online]. Available: https://arxiv.org/abs/1409.1259

[69] M. Siam, S. Valipour, M. Jagersand, and N. Ray, "Convolutional gated recurrent networks for video segmentation," in *Proc. Int. Conf. Image Process.*, Sep. 2017, pp. 3090–3094.

[70] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[71] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2758–2766.

[72] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[73] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Statist.*, 2015, pp. 562–570.

[74] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, p. 4, 2009.

[75] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 541–554, Mar. 2013.

[76] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2016.

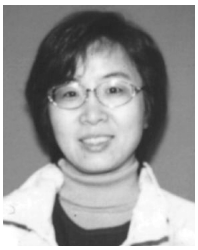
[77] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, "Real-time superpixel segmentation by DBSCAN clustering algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5933–5942, Dec. 2016.

[78] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jul. 2017, pp. 2462–2470.

[79] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[80] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.
[81] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.
[82] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1147–1154.
[83] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 681–688.
[84] V. Leborán, A. García-Díaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 893–907, May 2017.
[85] Y.-F. Ma and H.-J. Zhang, "A new perceived motion based shot content representation," in *Proc. Int. Conf. Image Process.*, Oct. 2001, pp. 426–429.
[86] Y.-F. Ma and H.-J. Zhang, "A model of motion attention for video skimming," in *Proc. Int. Conf. Image Process.*, Sep. 2002, p. 1.
[87] G. Agarwal, A. Anbu, and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed MPEG domain," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2003, p. II-133.
[88] A. Sinha, G. Agarwal, and A. Anbu, "Region-of-interest based compressed domain video transcoding scheme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. III-161.
[89] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang, "A motion attention model based rate control algorithm for H.264/AVC," in *Proc. IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, Jun. 2009, pp. 568–573.
[90] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.
[91] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2929–2936.
[92] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[93] H.-C. Nothdurft, "Attention shifts to salient targets," *Vis. Res.*, vol. 42, no. 10, pp. 1287–1306, May 2002.

Qiuxia Lai received the B.E. and M.S. degrees from the School of Automation, Huazhong University of Science and Technology, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. Her research interests include image/video processing and deep learning.

Wenguan Wang received the Ph.D. degree from the Beijing Institute of Technology in 2018. He was a joint Ph.D. student at the Department of Statistics, University of California, directed by Prof. S.-C. Zhu during 2016–2018. He is currently a Senior Scientist with the Inception Institute of Artificial Intelligence (IIAI), United Arab Emirates. His current research interests include visual relation understanding and graph neural networks.



Hanqiu Sun received the M.S. degree in electrical engineering from The University of British Columbia and the Ph.D. degree in computer science from the University of Alberta, Canada. She is currently an Associate Professor with The Chinese University of Hong Kong. Her research interests include virtual reality, interactive graphics/animation, real-time hypermedia, virtual surgery, mobile image/video synopsis and navigation, and touch-enhanced and dynamics simulations.

Jianbing Shen (M'11–SM'12) is currently the Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), United Arab Emirates, and an Adjunct Honorary Professor with the Beijing Institute of Technology. He has published about 100 journal and conference articles, among them eight papers are selected as the ESI Highly Cited Papers. His research interests include computer vision, deep learning, autonomous driving, and medical image analysis. He is also an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and other journals.