Revisiting Video Saliency Prediction in the Deep Learning Era

Wenguan Wang, *Member, IEEE*, Jianbing Shen, *Senior Member, IEEE*,
Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji

Abstract—Predicting where people look in static scenes, a.k.a visual saliency, has received significant research interest recently. However, relatively less effort has been spent in understanding and modeling visual attention over dynamic scenes. This work makes three contributions to video saliency research. First, we introduce a new benchmark, called DHF1K (Dynamic Human Fixation 1K), for predicting fixations during dynamic scene free-viewing, which is a long-time need in this field. DHF1K consists of 1K high-quality elaborately-selected video sequences annotated by 17 observers using an eye tracker device. The videos span a wide range of scenes, motions, object types and backgrounds. Second, we propose a novel video saliency model, called ACLNet (Attentive CNN-LSTM Network), that augments the CNN-LSTM architecture with a supervised attention mechanism to enable fast end-to-end saliency learning. The attention mechanism explicitly encodes static saliency information, thus allowing LSTM to focus on learning a more flexible temporal saliency representation across successive frames. Such a design fully leverages existing large-scale static fixation datasets, avoids overfitting, and significantly improves training efficiency and testing performance. Third, we perform an extensive evaluation of the state-of-the-art saliency models on three datasets : DHF1K, Hollywood-2, and UCF sports. An attribute-based analysis of previous saliency models and cross-dataset generalization are also presented. Experimental results over more than 1.2K testing videos containing 400K frames demonstrate that ACLNet outperforms other contenders and has a fast processing speed (40fps using a single GPU). Our code and all the results are available at https://github.com/wenguanwang/DHF1K.

Index Terms—Video saliency, dynamic visual attention, benchmark, deep learning.

1 Introduction

Human visual system (HVS) has an astonishing ability to quickly select and concentrate on important regions in the visual field. This cognitive process allows humans to selectively process a vast amount of visual information and attend to important parts of a crowded scene while ignoring irrelevant information. This selective mechanism, known as visual attention, allows humans to interpret complex scenes in real time.

Over the last few decades, several computational models have been proposed for imitating attentional mechanisms of HVS during static scene viewing. Significant advances have been achieved recently thanks to the rapid spread of deep learning techniques and the availability of large-scale static gaze datasets (*e.g.*, SALICON [2]). In stark contrast, predicting observers' fixations during dynamic scene free-viewing has been under-explored. This task, referred to as *dynamic fixation prediction* or *video saliency detection*, is essential for understanding human attention behaviors and has various practical real-word applications (*e.g.*, video captioning [3], compression [4], question answering [5], object segmentation [6], action recognition [7], *etc.*). It is thus highly desired to have a standard, high-quality benchmark composed of diverse and representative video stimuli. Existing datasets are severely limited in their coverage and scalability, and only include special scenarios such as limited human activities. They lack generic, representative, and diverse instances in unconstrained task-independent scenarios. Consequently, existing datasets fail to offer a rich set of fixations for learning video saliency and to assess models. Moreover, they do not provide an evaluation server with a standalone held out test set to avoid potential dataset over-fitting.

While saliency benchmarks (*e.g.*, MIT300 [8] and SALICON [2]) have been very instrumental in progressing the static saliency field [9], such standard widespread benchmarks are missing for video saliency modeling. We believe such benchmarks are highly desired to drive the field forward. To this end, we propose a new benchmark, named DHF1K (*Dynamic Human Fixation 1K*), with a public server for reporting evaluation results on a preserved test set. DHF1K comes with a dataset that is unique in terms of generality, diversity and difficulty. It has 1K videos with over 600K frames and per-frame fixation annotations from 17 observers. The sequences have been carefully collected to cover diverse scenes, motion patterns, object categories, and activities. DHF1K is accompanied by a comprehensive evaluation of 23 state-of-the-art approaches [10]–[31]. Moreover, each video is annotated with a main category label (*e.g.*, daily activities, animals) and rich attributes (*e.g.*, camera/content movement, scene lighting, presence of humans), which facilitate deeper understanding of gaze guid-

- *W. Wang and J. Shen are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, and also with Inception Institute of Artificial Intelligence, UAE. (Email: wenguanwang.ai@gmail.com, shenjianbing@bit.edu.cn)*
- *J. Xie is with Hikvision Research Institute, USA. (Email: Jianwen.Xie@hikvision.com)*
- *M.-M. Cheng is with College of Computer Science, Nankai University. (Email: cmm@nankai.edu.cn)*
- *H. Ling is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. (Email: hbling@temple.edu)*
- *A. Borji is with MarkableAI. (Email: aliborji@gmail.com)*
- *A preliminary version of this work has appeared in CVPR 2018 [1].*
- *Corresponding author: Jianbing Shen*

ance in free viewing of dynamic scenes.

Further, we propose a novel video saliency model called ACLNet (*Attentive CNN-LSTM Network*), which extends the CNN-LSTM architecture [32] using a supervised attention mechanism. CNN layers are used for extracting static features within input frames [33], while convolutional LSTM (convLSTM) [34] is used for sequential fixation prediction over successive frames. An attention module, learned from existing image saliency datasets, is used to enhance spatially informative features of CNN. Such a design helps disentangle underlying spatial and temporal factors of dynamic attention and allows convLSTM to learn temporal saliency representation efficiently. It also helps avoid overfitting with relatively few video data. With such training protocol leveraging both the dynamic and static fixation data, we build an end-to-end trainable video saliency model and experimentally demonstrate its superior performance and high processing speed.

1.1 Contribution

In summary, our contributions are four-fold:

1) **A large-scale eye-tracking dataset for dynamic free-view fixation prediction.** We introduce a benchmark of 1K videos covering a wide range of scenes, motions, activities, *etc*. Existing video saliency datasets lack variety and generality of common dynamic scenes and fall short in covering challenging situations in unconstrained environments. In contrast, DHF1K makes a significant leap in terms of scalability, diversity and difficulty, and is expected to boost video saliency modeling. To our knowledge, DHF1K is the largest eye-tracking dataset for *dynamic*, *free-viewing* fixation prediction.
2) **A novel attentive CNN-LSTM architecture for dynamic visual attention prediction.** Through a supervised attentive module, the proposed network is able to explicitly encode static attention into dynamic saliency representation learning by leveraging both static and dynamic fixation data. In addition, the attention module is trained in both explicit and implicit manners. As far as we know, we are the first to introduce such neural attention mechanism and training protocol for this task. Results show that our model significantly outperforms previous methods with a much higher processing speed (40 *fps* on a TITAN X GPU).
3) **A comprehensive analysis of fixation prediction models** on existing dynamic eye-tracking datasets (Hollywood-2 [35], UCF sports [35]) and our DHF1K. To the best of our knowledge, such large-scale quantitative study of the performance of visual attention models on dynamic fixation prediction has not been done before in the computer vision community. We present a thorough analysis including a comparison of the performance of static attention models as well as dynamic attention ones, a comparison of performance of deep learning and non-deep learning attention models, a discussion regarding how the performance of visual attention models on dynamic fixation prediction has evolved over the past 20 years and an attribute-level evaluation to provide better insight into the performance of saliency models.
4) **A cross-dataset generalization experiment to quantitatively evaluate dataset bias.** Previous video saliency datasets started out with the goal of being as varied and rich as possible, but fail to answer how well they generalize to real visual world. We perform a cross-dataset generalization study, for the first time in this field, for exploring this essential but largely neglected issue.

This work extends our CVPR 2018 paper [1] in several ways. First, we provide additional details of the proposed DHF1K with respect to previous dynamic eye-tracking datasets and offer a more in-depth discussion of the proposed algorithm. Second, we provide a more comprehensive analysis of saliency models (23 state-of-the-art approaches in total) on three dynamic eye-tracking benchmarks and quantitatively assess their performances, analyze computation time, summarize their features, and review the performance improvement over the past 20 years. Third, more ablation studies are performed for thorough and insightful examination. Forth, we perform an attribute-based study which enables a deeper understanding of the results and points towards promising avenues for future research. Fifth, we perform a cross-dataset generalization analysis to quantitatively measure dataset bias and generalization. Last but not the least, based on our experiments, we draw several important conclusions that are expected to inspire future works in related topics.

1.2 Organization

In Sec. 2, we review previous benchmarks for dynamic visual attention prediction and representative works related to ours. Then, we elaborate our DHF1K dataset in Sec. 3. In Sec. 4, we describe our attentive CNN-LSTM model for dynamic fixation prediction by allowing the use of both static and dynamic eye-tracking data in an explicit and supervised attention module. In Sec. 5, we offer both quantitative and qualitative experimental analyses of our algorithm. Finally, concluding remarks can be found in Sec. 6.

2 Related Work

2.1 Video Eye-Tracking Datasets

There exist several datasets [35]–[38] for dynamic visual saliency prediction, but they are often limited in variety, generality and scalability of instances. Some statistics of these datasets are summarized in Table 1. As seen, these datasets differ in many aspects, such as the number of participants, number of test images, types of stimuli, experimental settings, post processing, *etc*.

Hollywood-2 [35] comprises all the $1,707$ videos from Hollywood-2 action recognition dataset [39]. The videos are collected from 69 Hollywood movies from 12 action categories, such as eating, kissing and running. The human fixation data were collected from 19 observers belonging to 3 groups for free viewing (3 observers), action recognition (12 observers), and context recognition (4 observers). Although this dataset is large, its content is limited to human actions and movie scenes. It mainly focuses on task-driven viewing mode rather than free viewing. With $1,000$ frames randomly sampled from Hollywood-2, we found that 84.5% of fixations are located around on the faces.

UCF sports [35] contains 150 videos taken from the UCF sports action dataset [40]. The videos cover 9 common

TABLE 1
Statistics of typical dynamic eye-tracking datasets.

Dataset	Publication	Year	#Videos	Resolution	Duration(s)	#Viewers	Task	Description
CRCNS [36][1]	TIP	2004	50	640×480	6-94	15	task-goal	Videos typically include synthetic stimuli, outdoors daytime and nighttime scenes, *etc.*
Hollywood-2 [35][2]	TPAMI	2012	1,707	720×480	2-120	19	task-goal	Videos are collected from 69 movies and annotated with 12 action categories, such as eating, kissing and running.
UCF sports [35][2]	TPAMI	2012	150	720×480	2-14	19	task-goal	Videos cover 9 common sports action classes, such as diving, swinging and walking.
DIEM [37][3]	Cognitive Computation	2011	84	1280×720	27-217	∼50	free-view	Videos are collected from publicly accessible video resources, including advertisements, documentaries, *etc.*
SFU [38][4]	TIP	2012	12	352×288	3-10	15	free-view	The eye-tracking data are captured during both the first and second viewings.
DHF1K (**Ours**)[5]	CVPR	2018	1,000	640×360	17-42	17	free-view	Videos were elaborately selected to cover a wide range of scenes, motions, activities, etc. It is the largest eye-tracking dataset for dynamic, free-viewing fixation prediction.

[1] http://ilab.usc.edu/bu/compress/ [2] http://vision.imar.ro/eyetracking/description.php [3] https://thediemproject.wordpress.com
[4] http://www.sfu.ca/~ibajic/ [5] https://github.com/wenguanwang/DHF1K

sports action classes, such as diving, swinging and walking. Similar to Hollywood-2, the viewers were biased towards task-aware observation by being instructed to "identify the actions occurring in the video sequence". Statistics of $1,000$ frames randomly selected from UCF sports suggest that 82.3% of fixations fall inside the human body area.

DIEM [37] is a public video eye-tracking dataset that has 84 videos collected from publicly accessible video resources (*e.g.*, advertisements, documentaries, sport events, and movie trailers, *etc.*). For each video, free-viewing fixations of around 50 observers were collected. This dataset is mainly limited in its coverage and scale.

Other datasets are either limited in terms of variety and scale of video stimuli [36], [38], or collected for special purposes (*e.g.*, salient objects in videos [41]). More importantly, none of the aforementioned datasets includes a preserved test set for avoiding potential data overfitting, which may seriously hamper the research process.

2.2 Computational Models of Fixation Prediction

The study of human gaze pattern in static scenes has received significant interests, and dates back to [26], [42]. Visual attention allocation depends on two types of mechanisms. The *bottom-up* attentional mechanism is driven by external environmental stimuli, involuntarily orienting attention to external, discriminative stimulus features (*exogenous*) - a white spot against a black scene or sudden movement against stable background. Bottom-up attention mainly occurs during pre-attentive vision and free viewing. Alternatively, the *top-down* mechanism is volitional, goal-directed and accompanied by longer-term cognitive factors (*endogenous*). For instance, when inspecting surveillance videos, guards are more likely to allocate their attention to moving people for detecting suspicious behaviors. Only few studies [43]–[45] have been so far specifically designed to model top-down attentional allocation in scenes. Involuntary and exogenous control of attention should be consistent across all human subjects, resulting in a high degree of coordination in multiple viewers' visual attention behaviors given the same stimuli. In contrast, attention across individuals is less coordinated during endogenous control, since the internal cognitive states of the individual and their relation to the current stimuli are less predictable [37].

Early static saliency models [27], [46]–[52] are mostly concerned with the bottom-up visual attention mechanism (see [53], [54] for detailed review). *Contrast* is the most widely used assumption that conspicuous visual features pop out from its surroundings and involuntarily attract human attention. Computational models compute multiple visual features such as color, edge, and orientation at multiple spatial scales to produce a "saliency map": a 2D distribution predicting the conspicuity of specific locations and their likelihood in attracting fixations [37], [42]. The locations with more distinct feature responses over surroundings usually gain higher saliency values.

Recently, **deep learning based static saliency models** [28]–[30], [55]–[59] have achieved great improvements, relying on the powerful end-to-end learning ability of neural networks and the availability of large-scale static saliency datasets [2]. More specially, Vig *et al.* [55] learned deep features from scratch and adopted a linear SVM to classify each local image location to be salient or non-salient (eDN model). This represents an early attempt that applied neural networks to visual attention prediction. Follow-up works mainly focused on exploiting more effective network architectures and leveraging transfer-learning techniques for learning more representative features. For example, DeepFix [56], DeepNet [30] and SALICON net [28] fine-tune VGG-16 [60] pre-trained on image classification task. Mr-CNN [57] was based on multi-streams that learn multiscale saliency information. DVA [29] fused features from multiple layers of VGG-16 for saliency prediction. Pan *et al.* [31] promote the performance of a VGG-16 based saliency predictor with an adversarial training strategy.

The question of how humans distribute their attention while viewing static scenes has drawn a great amount of research effort. However, important dynamic behaviors of HVS in dynamic scenes have not been thoroughly explored. **Previous investigations of dynamic scene viewing** [10]–[16], [61], [62] mainly focus on bottom-up attention orienting, leveraging both static stimulus features and temporal information (*e.g.*, optical flow, difference-over-time, *etc*). Some of these studies [11], [61], [62] can be viewed as extensions of existing static saliency models with additional motion features. Such models are mainly bound to significant feature engineering and limited representation ability

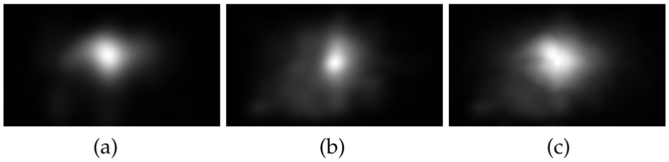


Fig. 1. Average attention maps of three benchmark datasets: (a) Hollywood-2 [35], (b) UCF sports [35], and (c) DHF1K.

of hand-crafted features.

To date, only a few **deep learning based video saliency models** [24], [25], [63], [64] exist in this field. They are mainly based on the two-stream network architecture [65] that accounts for color images and motion fields separately. In [24], an extra stream is added for capturing object information. These works show a better performance and demonstrate the potential advantages in applying neural networks to this problem. In [63], a 7-layer encoder-decoder network is designed to predict visual saliency for RGBD videos. Concurrent with our work, Gorji *et al.* [64] augment static saliency models with multi-stream LSTMs to predict video saliency. However, they do not 1) consider attentive mechanisms, 2) utilize existing large-scale static fixation datasets, and 3) exhaustively assess their performance over a large amount of data.

A related topic is **salient object detection** [66], [67] that uniformly highlights salient object regions in images [68]–[73], videos [74]–[79], image/video groups [80], or RGBD data [81], [82]. However, such algorithms often focus on inferring the main salient object(s), instead of investigating the attention behavior of the HVS during scene free viewing.

2.3 Attention Mechanisms in Neural Networks

Recently, incorporating attention mechanisms into network architectures has shown great successes in several computer vision [83]–[86] and natural language processing tasks [87], [88]. In these studies, the neural attention is differentiable and can be learned in an automatic, top-down and task-specific manner, allowing the network to focus on the most relevant parts in images or sentences. In this paper, we use the trainable neural attention for enhancing intra-frame salient features, thus allowing LSTM to model dynamic representations easily. In contrast to previous models that learn attention implicitly, our attention module encodes strong static saliency information and can be learned from existing static saliency datasets in a supervised manner. This design leads to improved generality and prediction performance. As far as we know, it is the first attempt to incorporate a supervised attention mechanism into the network structure to achieve state-of-art results in dynamic fixation prediction.

3 The DHF1K Dataset

We introduce DHF1K, a large-scale dataset of gaze during free-viewing of videos. DHF1K includes 1K videos with diverse content and length, with eye-tracking annotations from 17 observers. Fig. 1 shows the center biases of DHF1K, compared to Hollywood-2 [35], and UCF sports [35].

TABLE 2
Statistics for video categories in DHF1K dataset.

DHF1K	Human				Animal	Artifact	Scenery
	Daily ac.	Sports	Social ac.	Art			
#sub-classes*	20	29	13	10	36	21	21
#videos	134	185	116	101	192	162	110

*Numbers of sub-classes in each category are reported. For example, *Sports* has sub-classes like *swimming*, *jumping*, *etc*.

3.1 Stimuli

The collection of dynamic stimuli mainly follows the following four principles:

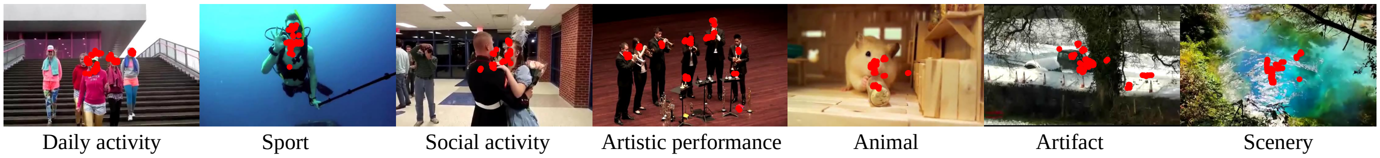
• **Large scale and high quality**. Large scale and high quality are both necessary to ensure the content diversity of a dataset and crucial to guarantee a longer lifespan for a benchmark. To this end, we searched the YouTube engine with about 200 key terms (*e.g.,* dog, walking, car, *etc*). The list of the key terms is mainly built upon the labels of two datasets, MSCOCO [89] and FCVID [90], and is supplemented with about 40 extra keywords proposed by ourselves. The detailed keyword selection process and the full keyword list can be found in the Supplemental Material. From the retrieved results, we carefully selected $1,000$ video sequences. Each video was then converted to a 30 fps Xvid MPEG-4 video file in an AVI format and resized uniformly into 640×360 spatial resolution. Thus, DHF1K comprises a total of $1,000$ video sequences with $582,605$ frames with total duration of $19,420$ seconds.

• **Diverse content**. Stimulus diversity is essential for avoiding overfitting and to delay performance saturation. It offers evenly distributed exogenous control for studying person-external stimulus factors during scene free-viewing. In DHF1K, each video is manually annotated with a category label (totally 150 classes). These labels are further classified into 7 main categories (see Table 2). These semantic annotations enable deeper understanding of high-level stimuli factors guiding human gaze in dynamic scenes and benefit future research. Fig. 2 shows example frames from each category.

• **Varied motion patterns**. Previous investigations [37], [61], [91] suggested that motion is a key factor that directs attention allocation in dynamic viewing. DHF1K is designed to include various motion patterns (*stable-/slow-/fast-motion* of content and camera). Please see Table 3 for the information regarding motion patterns.

• **Various objects**. Previous studies [92]–[94] in cognitive psychology and computer vision have confirmed that objects guide human fixations. Objects in our dataset vary in their categories (*e.g.*, *human*, *animal*, in Table 2) and frequency (Table 4). For each video, five subjects were instructed to count the number of main objects in each image. The majority vote of their counts was considered as the final count.

For completeness, in Tables 5 and 6 we also offer the information regarding scene illumination and the number of humans in the dataset. As demonstrated in [95], luminance is an important exogenous factor for attentive selection. Further, human beings are important high-level stimuli [96], [97] in scene free-viewing.



Daily activity Sport Social activity Artistic performance Animal Artifact Scenery

Fig. 2. Example frames from DHF1K with fixations and corresponding categories. Note that, for better visualization, we use enlarged red dots to represent the human eye fixations. This figure is best viewed in color (zoom in for details).

TABLE 3
Statistics regarding motion patterns.

DHF1K	Content Motion			Camera Motion		
	stable	slow	fast	stable	slow	fast
#videos	126	505	369	343	386	271

TABLE 4
Statistics regarding number of main objects.

DHF1K	#Objects			
	0	1	2	≥3
#videos	56	335	254	355

TABLE 5
Statistics regarding scene illumination.

DHF1K	Scene Illumination		
	day	night	indoor
#videos	577	37	386

TABLE 6
Statistics regarding number of people.

DHF1K	#People			
	0	1	2	≥3
#videos	345	307	236	112

3.2 Apparatus and Technical Specifications

Participants' eye movements were monitored binocularly using a Senso Motoric Instruments (SMI) RED 250 system at a sampling rate of 250 Hz. The dynamic stimuli were displayed on a 19″ display (resolution 1440×900). A headrest was used to stabilize participants' heads at a distance of around 68 cm, as advised by the product manual.

3.3 Participants

17 participants (10 males and 7 females, aging between 20 and 28) who passed the eye tracker calibration and had less than 10% fixation dropping rate, were qualified for our eye tracking experiment. All participants had normal or corrected-to-normal vision. They had not participated in any eye-tracking experiment nor seen the stimuli in DHF1K before. All subjects provided informed consent and were naïve to the underlying purposes of the experiment.

3.4 Data Capturing

The subjects were informed that they would watch a series of unrelated silent video clips[1]. The stimuli were equally partitioned into 10 non-overlapping sessions. Participants were required to freely view 10 sessions of videos in random order. In each session, the videos were also displayed at random. Before the experiments, eye tracker was calibrated using the standard routine in product manual with recommended settings for the best results. The calibration procedure was repeated until an acceptable calibration was obtained as determined by means of validation procedure offered by the product. This procedure expected participants to look at four small circles near the middle of the screen. The calibration was considered to be acceptable if a fixation was shown for each circle and no fixation appeared in an obvious outlier position. To avoid eye fatigue, each video presentation was followed by a 5-second waiting interval with black screen. After undergoing a session of videos, the participant took a rest until she was ready for viewing the next session. In this way, the video stimuli were shown to each subject in a different random order, and each of the video stimuli was viewed by all the 17 subjects. The raw data recorded by the eye tracker consisted of time and position values for each frame. We filter out the fixations which are outside of frames. Finally, $51,038,600$ fixations were recorded from 17 subjects on $1,000$ videos.

To convert the discrete fixation map into a continuous saliency map, we convolve each fixation location (of all subjects) with a small Gaussian filter. Following [8], [98], the size of the Gaussian is set to about one degree of visual angle ($\sim$30 image pixels in our case). The finally stored continuous saliency map is normalized to a range of 0-1.0.

3.5 Training/Testing Split

We split $1,000$ dynamic stimuli into training, validation and test sets. Following random selection, we arrive at a unique split consisting of 600 training and 100 validation videos with publicly available fixation records, as well as 300 test videos with annotations held-out for benchmarking purpose.

4 Our Approach

Fig. 3 presents the overall architecture of our ACLNet. It is based on a CNN-LSTM structure that combines convolutional network and recurrent model to exploit both spatial and temporal information for predicting video saliency. The CNN-LSTM network is extended with a supervised attention mechanism, which explicitly captures static saliency information and allows the LSTM to focus on learning dynamic information. The attention module is trained from rich static eye-tracking data. Thus, ACLNet is able to produce accurate video saliency with improved generalization ability. Next, we elaborate each component of ACLNet.

4.1 The CNN-LSTM Architecture

Formally, given an input video $\{I_t\}_t$, we first obtain a sequence of convolutional features $\{\mathcal{X}_t\}_t$ from CNN. Then, the features $\{\mathcal{X}_t\}_t$ are fed into a convLSTM [34] as input. Here, the convLSTM is used for modeling the temporal nature of this sequential problem, which is achieved by incorporating memory units with gated operations. Additionally,

[1] Note that the collected dynamic stimuli are accompanied with audio, but we use silent videos during data capturing. That is because in this work, we specifically focus on exploring the influence of visual stimuli in human attention behavior during dynamic scene viewing.

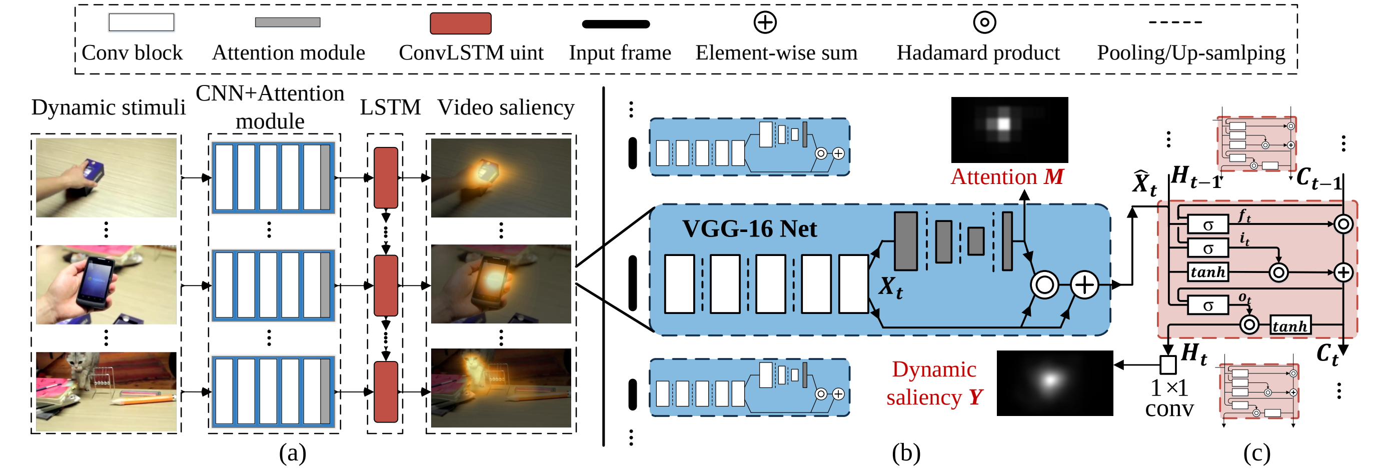


Fig. 3. Network architecture of the proposed video saliency model ACLNet. (a) Attentive CNN-LSTM architecture, (b) CNN layers with attention module are used for learning intra-frame static features, where the attention module is learned with the supervision from static saliency data, and (c) ConvLSTM used for learning sequential saliency representations.

by replacing dot products with convolution operations, the convLSTM is able to preserve spatial information, which is essential for making spatially-variant pixel-wise prediction.

More precisely, the convLSTM utilizes three convolution gates (*input*, *output* and *forget*) to control the flow of signal within a cell. With the input feature $\mathcal{X}_t$ at time step t, the convLSTM outputs a hidden state $\mathcal{H}_t$ and maintains a memory cell $\mathcal{C}_t$ for controlling state update and output:

$$i_t = \sigma(W_i^{\mathcal{X}} * \mathcal{X}_t + W_i^{\mathcal{H}} * \mathcal{H}_{t-1} + W_i^{\mathcal{C}} \circ \mathcal{C}_{t-1} + b_i), \quad (1)$$

$$f_t = \sigma(W_f^{\mathcal{X}} * \mathcal{X}_t + W_f^{\mathcal{H}} * \mathcal{H}_{t-1} + W_f^{\mathcal{C}} \circ \mathcal{C}_{t-1} + b_f), \quad (2)$$

$$o_t = \sigma(W_o^{\mathcal{X}} * \mathcal{X}_t + W_o^{\mathcal{H}} * \mathcal{H}_{t-1} + W_o^{\mathcal{C}} \circ \mathcal{C}_t + b_o), \quad (3)$$

$$\mathcal{C}_t = f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(W_c^{\mathcal{X}} * \mathcal{X}_t + W_c^{\mathcal{H}} * \mathcal{H}_{t-1} + b_c), \quad (4)$$

$$\mathcal{H}_t = o_t \circ \tanh(\mathcal{C}_t), \quad (5)$$

where i_t, f_t, o_t are the gates. σ and $\tanh$ are respectively the activation functions of logistic sigmoid and hyperbolic tangent, '$*$' denotes the convolution operator and '$\circ$' represents Hadamard product. All the inputs $\mathcal{X}$, cell memory $\mathcal{C}$, hidden states $\mathcal{H}$ and gates i, f, c are 3D tensors of the same dimension. Ws and bs are weights and biases which can be learned with back-propagation. The dynamic fixation map can be obtained via convolving the hidden states $\mathcal{H}$ with a 1×1 kernel (see Fig. 3 (c)).

In our implementation, the first five *conv* blocks of VGG-16 [60] are used. For preserving more spatial details, we remove *pool4* and *pool5* layers, which results in ×8 instead of ×32 downsampling. At time step t, with an input frame I_t of resolution 224×224, we have $\mathcal{X}_t \in \mathbb{R}^{28\times28\times512}$ and a 28×28 dynamic saliency map from the convLSTM. The kernel size of the conv layer in convLSTM is set to 3.

4.2 Neural Attention Module

We extend the above CNN-LSTM architecture with an attention mechanism, which is learned from existing static fixation data in a supervised manner. Such design is mainly driven by the following three motivations:

- Previous studies [91], [99] have shown that human attention is guided by both static and dynamic factors. Through the additional attention module, CNN is enforced to generate a more explicit spatial saliency representation. This helps disentangle underlying spatial and temporal factors of dynamic attention, and allows the convLSTM to better capture temporal dynamics.

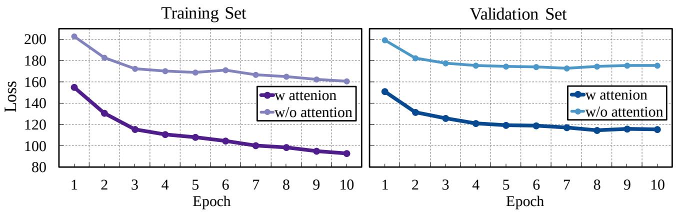


Fig. 4. Performance of ACLNet with or without the attention module on the training and validation sets of DHF1K. The attention module significantly improves training efficiency and performance.

- The CNN-LSTM architecture introduces a large number of parameters for modeling spatial and temporal patterns. However, for sequential data such as videos, obtaining labeled data is costly. Even with large-scale datasets like DHF1K with 1K videos, the amount of training data is still insufficient, considering the high correlation among those frames from the same video. The supervised attentive module is able to leverage existing rich static fixation data to improve the generalization power of ACLNet.

- In VGG-16, we remove the last two pooling layers to obtain a large feature map. This dramatically decreases the receptive field (212×212→140×140), which can not cover the whole frame (224×224). To remedy this, we insert a set of down- and up-sampling operations into the attention module, which enhance the intra-frame saliency information with an enlarged receptive field. ACLNet is thus able to make more accurate predictions from a global view.

As demonstrated in Fig. 3 (b), our attentive module is built upon the *conv5-3* layer, as an additional branch of several conv layers interleaved with pooling and upsampling operations. Given the input feature $\mathcal{X}$, with pooling layers (detailed in Sec. 5.1), the attention module generates a downsampled attention map (7×7) with an enlarged re-

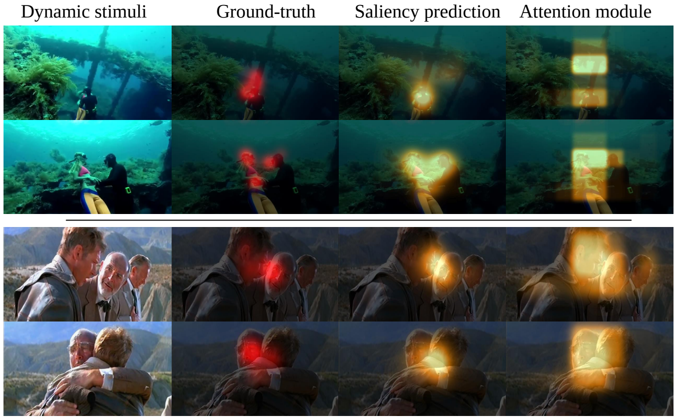


Fig. 5. Illustration of the attention maps predicted by our ACLNet and the attention module on two dynamic stimuli. Best viewed in color.

ceptive field (260×260). Note that our attention module is equipped with a *sigmoid* function, which relaxes the sum-to-one constraint in the *soft-max* based neural attention [83]. Then, the small attention map is $\times 4$ upsampled as the same spatial dimensions of $\mathcal{X}$. Let $M \in [0,1]^{28\times28}$ be the upsampled attention map, the feature $\mathcal{X} \in \mathbb{R}^{28\times28\times512}$ from *conv5-3* layer can be further enhanced by:

$$\hat{\mathcal{X}}^c = M \circ \mathcal{X}^c, \tag{6}$$

where $c \in \{1,\ldots,512\}$ is the channel index, and '$\circ$' indicates Hadamard product defined in Sec. 4.1. Here, the attention module works as a feature selector to enhance feature representation.

The above attention module may lose useful information for learning a dynamic saliency representation, as the attention module only considers static saliency information in still video frames. For this, inspired by the recent advances of attention mechanism and residual connection [86], [100], we improve Equ. 6 in the residual form:

$$\hat{\mathcal{X}}^c = (1+M) \circ \mathcal{X}^c. \tag{7}$$

With the residual connection, both the original CNN features and the enhanced features are combined and fed to the LSTM model. In Fig. 4, we observe that the model with the supervised attention module gains better training efficiency on the training set and improved performance on the validation set. Fig. 5 visualizes the attention maps predicted by the proposed ACLNet and the attention module on two videos, showing that with the differentiable attention module, ACLNet is able to capture the visually important regions during dynamic scene viewing. In Sec. 5.3 and 5.7, more detailed quantitative explorations of the attention module are offered.

Different from previous attention mechanisms that learn task-related attention in an implicit manner, our attention module can learn from existing large-scale static fixation data in an explicit and supervised manner, as described in the following subsections.

4.3 Loss Functions

We use the loss function as in [28] that considers three different saliency evaluation metrics instead of one. The rationale is that no single metric can fully capture how satisfactory a saliency map is. We use different metrics to capture several quality factors.

We denote the predicted saliency map as $Y \in [0,1]^{28\times28}$, the map of fixation locations as $P \in \{0,1\}^{28\times28}$ and the continuous saliency map (distribution) as $Q \in [0,1]^{28\times28}$. Here the fixation map P is discrete, and records whether a pixel receives human fixation. The continuous saliency map is obtained via blurring each fixation location with a small Gaussian kernel (see Sec. 3.4). Our loss function is defined as follows:

$$\mathcal{L}(Y,P,Q)=\mathcal{L}_{KL}(Y,Q)+\alpha_1\mathcal{L}_{CC}(Y,Q)+\alpha_2\mathcal{L}_{NSS}(Y,P), \tag{8}$$

where $\mathcal{L}_{KL}$, $\mathcal{L}_{CC}$ and $\mathcal{L}_{NSS}$ are the *Kullback-Leibler (KL) divergence*, the *Linear Correlation Coefficient (CC)*, and the *Normalized Scanpath Saliency (NSS)*, respectively, which are derived from commonly used metrics [53] to evaluate saliency prediction models. α_1 and α_2 are balance parameters and are empirically set to $\alpha_1 = \alpha_2 = 0.1$.

$\mathcal{L}_{KL}$ is widely adopted for training saliency models and is chosen as the primary loss in our work:

$$\mathcal{L}_{KL}(Y,Q) = \sum\nolimits_x Q(x)\log\left(\frac{Q(x)}{Y(x)}\right). \tag{9}$$

$\mathcal{L}_{CC}$ measures the linear relationship between Y and Q:

$$\mathcal{L}_{CC}(Y,Q) = -\frac{\mathrm{cov}(Y,Q)}{\rho(Y)\rho(Q)}, \tag{10}$$

where $cov(Y,Q)$ is the covariance of Y and Q, and $\rho(\cdot)$ stands for standard deviation.

$\mathcal{L}_{NSS}$ is derived from the NSS metric:

$$\mathcal{L}_{NSS}(Y,P) = -\frac{1}{N}\sum\nolimits_x \overline{Y}(x)P(x), \tag{11}$$

where $\overline{Y} = \frac{Y-\mu(Y)}{\rho(Y)}$ and $N = \sum_x P(x)$. It is calculated by taking the mean of scores from the normalized saliency map $\overline{Y}$ (with zero mean and unit standard deviation) at human fixations P. Since CC and NSS are similarity metrics, their negatives are adopted for minimization.

4.4 Training Protocol

Our model is iteratively trained with sequential fixation and image data. In training, a video training batch is cascaded with an image training batch. More specifically, in a video training batch, we apply a loss defined over the final dynamic saliency prediction from LSTM. Let $\{Y_t^d\}_{t=1}^T$, $\{P_t^d\}_{t=1}^T$, and $\{Q_t^d\}_{t=1}^T$ be the dynamic saliency predictions, the dynamic fixation sequence and the continuous ground-truth saliency maps, we minimize the following loss:

$$\mathcal{L}^d = \sum\nolimits_{t=1}^T \mathcal{L}(Y_t^d, P_t^d, Q_t^d). \tag{12}$$

In this process, the attention module is trained in an implicit way, since we do not have the groundtruth fixation of each frame in static scenes.

In an image training batch, we only train our attention module via minimizing

$$\mathcal{L}^s = \mathcal{L}(M, P^s, Q^s), \tag{13}$$

where M, P^s, Q^s indicate the attention map for our static attention module, the ground-truth static fixation map, and

TABLE 7
Statistics and features of saliency prediction algorithms used in our evaluation.

Model	Ref.	Year	Pub.	Input Size	Deep Learning	Run-time(s)	Implementation	Network Architecture	Size (MB)
*ITTI	[26]	1998	TPAMI	full size		0.9	Matlab	Classic	
*GBVS	[27]	2007	NIPS	full size		2.7	Matlab+C	Classic	
*SALICON	[28]	2015	ICCV	$\max\{w,h\}$=800	✓	0.3†	Caffe	VGG-16 [60]	117
*Shallow-Net	[30]	2016	CVPR	320×240	✓	0.08†	Python+Theano	self-designed network	244
*Deep-Net	[30]	2016	CVPR	320×240	✓	0.1†	Python+Caffe	VGG-16 [60]	103
* SalGAN	[31]	2017	CVPR-workshop	256×192	✓	0.02†	Python+Theano	VGG-16 [60]	130
*DVA	[29]	2018	TIP	$\max\{w,h\}$=256	✓	0.1†	Python+Caffe	VGG-16 [60]	100
PQFT	[10]	2010	TIP	64×64		1.2	Matlab	Classic	
Seo *et al.*	[11]	2009	JoV	full size		2.3	Matlab	Classic	
Rudoy *et al.*	[12]	2013	CVPR	h =144		180	Matlab	Classic	
Hou *et al.*	[13]	2009	NIPS	120×80		0.7	Matlab	Classic	
Fang *et al.*	[14]	2014	TIP	full size		147	Matlab	Classic	
OBDL	[15]	2015	CVPR	h = 288		0.8	Matlab	Classic	
AWS-D	[16]	2017	TPAMI	full size		9	Matlab	Classic	
PMES	[17]	2001	ICIP	full size		579	Matlab	Classic	
MAM	[19]	2002	ICIP	full size		778	Matlab	Classic	
PIM-ZEN	[18]	2003	ICME	full size		43	Matlab	Classic	
PIM-MCS	[20]	2004	ICASSP	full size		10	Matlab	Classic	
MCSDM	[21]	2009	ICIS	full size		15	Matlab	Classic	
MSM-SM	[22]	2013	SPL	full size		8	Matlab	Classic	
PNSP-CS	[23]	2014	TCSVT	full size		895	Matlab	Classic	
OM-CNN	[24]	2018	ECCV	448×448	✓	0.05†	Python+Tensorflow	VGG-16 [60]+YOLO [101]+ FlowNet [102]+2×LSTM	344
Two-stream	[25]	2018	TMM	640×480	✓	20†	Python+Caffe	2×Deep-Net [30] (optical flow as extra input)	315
ACLNet	-	2018	CVPR	224×224	✓	0.02†	Python+Tensorflow	VGG-16 [60]+convLSTM	250

*Static attention model. †Runtime with GPU.

the ground-truth static saliency map, respectively. In this process, the training of attention module is supervised by the ground-truth static fixation. Note that, in the image training batch, we do not train our LSTM module, which is used for learning the dynamic representation.

For each video training batch, 20 consecutive frames from the same video are used. Both the video and the start frame are randomly selected. For each image training batch, we set the batch size to 20, and the images are randomly sampled from existing static fixation dataset. More implementation details can be found in Sec. 5.1.

5 Experiments

First, Sec. 5.1 details our experimental settings. Analyses of model size and runtime can be found in Sec. 5.2. In Sec. 5.3, quantitative experiments on three eye-tracking benchmarks (Hollywood-2 [35], UCF sports [35], and DHF1K) in comparison with 23 popular visual attention models demonstrate the robustness, effectiveness, and efficiency of our algorithm. Further, Sec. 5.4 provides more insights into the experimental results and gives suggestions for further work. Qualitative results and attribute-level evaluation are reported in Sec. 5.5 and 5.6, respectively. To better understand the contributions of different ingredients of ACLNet, in Sec. 5.7, we implement several variants of our method to conduct ablative studies. In Sec. 5.8, we perform a cross-dataset generalization experiment to study the generalization of current video saliency datasets.

5.1 Experimental Setup

5.1.1 Training and Testing Protocols

We use the static stimuli ($10,000$ images) from the training set of the SALICON [2] dataset for training our attention module. For dynamic stimuli, we consider 4 settings: using the training set(s) from **(i)** DHF1K, **(ii)** Hollywood-2, **(iii)** UCF sports, and **(iv)** DHF1K+Hollywood-2+UCF sports. For DHF1K, we use the original training/validation/testing splitting (600/100/300). For Hollywood-2, following [39], we use 823 videos for training and 884 videos for testing. Note that the videos are further divided into short clips during training and testing. For UCF sports, the training and testing sets include 103 and 47 videos, respectively, as suggested by [40]. We randomly sample 10% videos from the training sets of Hollywood-2, and UCF sports as their validation sets. We evaluate ACLNet on the testing sets of DHF1K, Hollywood-2, and UCF sports dataset, in total $1,231$ video sequences with more than 400K frames.

5.1.2 Implementation Details

ACLNet is implemented in Python on Keras, and trained with the Adam optimizer [103]. Our attention module is implemented as: downsampling(×2) → conv(1×1, 64) → conv(3×3, 128) → downsampling(×2) → conv(1×1, 64) → conv(3×3, 128) → conv(1×1, 1) → upsampling(×4). The conv layer is represented as (kernel, channel). The implementation of our model can be found at https://github.com/wenguanwang/DHF1K. During training, the learning rate was set to 0.0001 and was decreased by a factor of 10 every 2 epochs. The network was trained for 10 epochs. The whole model is trained in an end-to-end manner. The entire training procedure takes about 30 hours using a single NVIDIA TITAN X GPU (in training setting (iv)).

5.1.3 Compared Computational Saliency Models

We compare our model with sixteen dynamic saliency models including: PQFT [10], Seo *et al.* [11], Rudoy *et al.* [12], Hou *et al.* [13], Fang *et al.* [14], OBDL [15], AWS-D [16], PMES [17], PIM-ZEN [18], MAM [19], PIM-MCS [20],

TABLE 8
Quantitative results on **DHF1K**. The best scores are marked in **bold**. Training settings (Sec. 5.1) for video saliency datasets: (i) DHF1K, (ii) Hollywood-2, (iii) UCF sports, and (iv) DHF1K+Hollywood-2+UCF sports. Symbol ∗ indicates non-deep learning models. See Sec. 5.3 for details. These notes are the same for Table 9 and Table 10.

	Dataset Method	DHF1K				
		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Baseline	*center prior	0.854	0.238	0.503	0.302	0.167
	*chance	0.500	0.330	0.500	0.000	0.000
Dynamic models	*PQFT [10]	0.699	0.139	0.562	0.137	0.749
	*Seo *et al.* [11]	0.635	0.142	0.499	0.070	0.334
	*Rudoy *et al.* [12]	0.769	0.214	0.501	0.285	1.498
	*Hou *et al.* [13]	0.726	0.167	0.545	0.150	0.847
	*Fang *et al.* [14]	0.819	0.198	0.537	0.273	1.539
	*OBDL [15]	0.638	0.171	0.500	0.117	0.495
	*AWS-D [16]	0.703	0.157	0.513	0.174	0.940
	*PMES [17]	0.545	0.093	0.502	0.055	0.237
	*MAM [19]	0.551	0.108	0.500	0.041	0.214
	*PIM-ZEN [18]	0.552	0.095	0.498	0.062	0.280
	*PIM-MCS [20]	0.551	0.094	0.499	0.053	0.242
	*MCSDM [21]	0.591	0.110	0.500	0.047	0.247
	*MSM-SM [22]	0.582	0.143	0.500	0.058	0.245
	*PNSP-CS [23]	0.543	0.085	0.499	0.028	0.121
	OM-CNN [24]	0.856	0.256	0.583	0.344	1.911
	Two-stream [25]	0.834	0.197	0.581	0.325	1.632
Static models	*ITTI [26]	0.774	0.162	0.553	0.233	1.207
	*GBVS [27]	0.828	0.186	0.554	0.283	1.474
	SALICON [28]	0.857	0.232	0.590	0.327	1.901
	Shallow-Net [30]	0.833	0.182	0.529	0.295	1.509
	Deep-Net [30]	0.855	0.201	0.592	0.331	1.775
	DVA [29]	0.860	0.262	0.595	0.358	2.013
	SalGAN [31]	0.866	0.262	**0.709**	0.370	2.043
Training setting (i)	ACLNet	0.885	0.311	0.553	0.415	2.259
	Attention module	0.854	0.251	0.545	0.332	1.755
Training setting (ii)	ACLNet	0.878	0.297	0.543	0.388	2.125
	Attention module	0.855	0.250	0.541	0.318	1.703
Training setting (iii)	ACLNet	0.866	0.277	0.596	0.362	1.951
	Attention module	0.852	0.260	0.582	0.350	1.945
Training setting (iv)	ACLNet	**0.890**	**0.315**	0.601	**0.434**	**2.354**
	Attention module	0.870	0.273	0.577	0.380	2.077

MCSDM [21], MSM-SM [22], PNSP-CS [23], OM-CNN [24], and Two-stream [25]. For the sake of completeness, we further include seven state-of-the-art static attention models: ITTI [26], GBVS [27], SALICON [28], SalGAN [31], DVA [29], Shallow-Net [30], and Deep-Net [30]. Among all these models, OM-CNN, Two-stream, SALICON, SalGAN, DVA, Shallow-Net, and Deep-Net are deep learning models, and others are classical saliency one. These models are selected due to: 1) representing the diversity of the state-of-the-art; or 2) publicly available implementations. We re-implemented [25] since the official code does not run properly. For SALICON [28], we use the open source implementation in https://github.com/CLT29/OpenSALICON. For other methods with publicly available implementations, we use the parameters provided by authors and keep them fixed for all the experiments. In Table 7, detailed statistics and features of above saliency models are summarized.

5.1.4 Baseline Models

We derive 8 baselines from the proposed ACLNet. For each training setting, we derive two baselines: ACLNet and *Attention module*, referring to our final dynamic saliency prediction and the intermediate output of our attention module, respectively. We also offer another two baselines: *center prior* and *chance*. Baseline *center prior* is obtained as the averaged saliency map over the training set of Hollywood-2, UCF sports, or DHF1K dataset. Baseline *center prior* is a weak baseline that randomly selects pixels as salient.

TABLE 9
Quantitative results on **Hollywood-2** [35].

	Dataset Method	Hollywood-2				
		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Baseline	*center prior	0.869	0.331	0.615	0.421	1.808
	*chance	0.500	0.330	0.500	0.000	0.000
Dynamic models	*PQFT [10]	0.723	0.201	0.621	0.153	0.755
	*Seo *et al.* [11]	0.652	0.155	0.530	0.076	0.346
	*Rudoy *et al.* [12]	0.783	0.315	0.536	0.302	1.570
	*Hou *et al.* [13]	0.731	0.202	0.580	0.146	0.684
	*Fang *et al.* [14]	0.859	0.272	0.659	0.358	1.667
	*OBDL [15]	0.640	0.170	0.541	0.106	0.462
	*AWS-D [16]	0.694	0.175	0.637	0.146	0.742
	*PMES [17]	0.696	0.180	0.620	0.177	0.867
	*MAM [19]	0.630	0.153	0.562	0.099	0.494
	*PIM-ZEN [18]	0.670	0.167	0.598	0.134	0.667
	*PIM-MCS [20]	0.663	0.163	0.570	0.118	0.584
	*MCSDM [21]	0.618	0.147	0.524	0.067	0.288
	*MSM-SM [22]	0.683	0.180	0.561	0.132	0.682
	*PNSP-CS [23]	0.647	0.146	0.548	0.077	0.370
	OM-CNN [24]	0.887	0.356	0.693	0.446	2.313
	Two-stream [25]	0.863	0.276	0.710	0.382	1.748
	SalGAN [31]	0.901	0.393	**0.789**	0.535	2.542
Static models	*ITTI [26]	0.788	0.221	0.607	0.257	1.076
	*GBVS [27]	0.837	0.257	0.633	0.308	1.336
	SALICON [28]	0.856	0.321	0.711	0.425	2.013
	Shallow-Net [30]	0.851	0.276	0.694	0.423	1.680
	Deep-Net [30]	0.884	0.300	0.736	0.451	2.066
	DVA [29]	0.886	0.372	0.727	0.482	2.459
Training setting (i)	ACLNet	0.905	0.471	0.757	0.577	2.517
	Attention module	0.880	0.415	0.748	0.529	2.283
Training setting (ii)	ACLNet	0.912	0.519	0.754	0.609	3.049
	Attention module	0.885	0.416	0.690	0.490	2.113
Training setting (iii)	ACLNet	0.884	0.449	0.749	0.534	2.647
	Attention module	0.898	0.429	0.763	0.543	2.409
Training setting (iv)	ACLNet	**0.913**	**0.542**	0.757	**0.623**	**3.086**
	Attention module	0.878	0.479	0.686	0.478	2.060

5.1.5 Evaluation Metrics

There are several ways to measure the agreement between model predictions and human eye movements [29], [53]. In our experiments, we employ five classic metrics, namely Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd (AUC-J), and shuffled AUC (s-AUC).

5.2 Runtime Analysis

In Table 7, we report the speed of our model and other saliency models. For all the methods, we include their computation time of optical flow (if used) and exclude the I/O time. For the non-deep learning methods, ITTI [26] is the fastest method (0.9s per frame on CPU) among static models and Hou *et al.* [13] is the fastest dynamic saliency model (0.7s per frame on CPU). Since our model does not need any pre- or post-processing, it takes only about 0.024s to process a frame of size 224 × 224, which is faster than previous deep dynamic attention models: OM-CNN (0.05s) and Two-stream (20s). We also observe ACLNet is the fastest

TABLE 10
Quantitative results on **UCF sports** [35].

	Dataset / Method	UCF sports AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Baseline	*center prior	0.834	0.299	0.566	0.350	1.585
	*chance	0.500	0.330	0.500	0.000	0.000
Dynamic models	*PQFT [10]	0.825	0.250	0.722	0.338	1.780
	*Seo *et al.* [11]	0.831	0.308	0.666	0.336	1.690
	*Rudoy *et al.* [12]	0.763	0.271	0.637	0.344	1.619
	*Hou *et al.* [13]	0.819	0.276	0.674	0.292	1.399
	*Fang *et al.* [14]	0.845	0.307	0.674	0.395	1.787
	*OBDL [15]	0.759	0.193	0.634	0.234	1.382
	*AWS-D [16]	0.823	0.228	0.750	0.306	1.631
	*PMES [17]	0.756	0.263	0.714	0.349	1.788
	*MAM [19]	0.669	0.213	0.624	0.218	1.130
	*PIM-ZEN [18]	0.760	0.234	0.702	0.306	1.657
	*PIM-MCS [20]	0.777	0.238	0.695	0.303	1.596
	*MCSDM [21]	0.756	0.228	0.626	0.230	1.091
	*MSM-SM [22]	0.752	0.262	0.634	0.280	1.584
	*PNSP-CS [23]	0.755	0.210	0.628	0.218	1.091
	OM-CNN [24]	0.870	0.321	0.691	0.405	2.089
	Two-stream [25]	0.832	0.264	0.685	0.343	1.753
Static models	*ITTI [26]	0.847	0.251	0.725	0.356	1.640
	*GBVS [27]	0.859	0.274	0.697	0.396	1.818
	SALICON [28]	0.848	0.304	0.738	0.375	1.838
	Shallow-Net [30]	0.846	0.276	0.691	0.382	1.789
	Deep-Net [30]	0.861	0.282	0.719	0.414	1.903
	DVA [29]	0.872	0.339	0.725	0.439	2.311
	SalGAN [31]	0.876	0.332	0.762	0.470	2.238
Training setting (i)	ACLNet	0.894	0.403	0.742	0.517	2.559
	Attention module	0.853	0.333	0.719	0.435	1.946
Training setting (ii)	ACLNet	0.874	0.364	0.727	0.452	2.186
	Attention module	0.860	0.322	0.656	0.367	1.667
Training setting (iii)	ACLNet	**0.905**	**0.496**	**0.767**	**0.603**	**3.200**
	Attention module	0.884	0.354	0.743	0.500	2.339
Training setting (iv)	ACLNet	0.897	0.406	0.744	0.510	2.567
	Attention module	0.877	0.379	0.685	0.411	1.899

one among all the deep-learning models and our real-time processing speed brings high applicability. In addition, our model (250 MB) is smaller than deep dynamic attention models: OM-CNN (344 MB) and Two-stream (315 MB).

5.3 Quantitative Evaluation and Model Comparison

The section presents quantitative evaluation results on DHF1K, Hollywood-2 and UCF sports datasets.

• **Performance on DHF1K.** Table 8 reports the comparative results with the aforementioned saliency models on the test set (300 video sequences) of DHF1K. It can be observed that our model consistently and significantly outperforms other competitors in all metrics. This can be attributed to our specially designed attention module which allows our model to explicitly learn static and dynamic saliency representations in CNN and LSTM separately. Notice that our model does not even use any optical flow algorithm. This significantly improves the applicability of our model and demonstrates the effectiveness of our training protocol in leveraging both static and dynamic stimuli.

• **Performance on Hollywood-2.** We further test our model on Hollywood-2 where the testing set comprises 884 video sequences. The results are summarized in Table 9. Again, our model performs significantly higher than other methods across various metrics. Besides, when we go insight into the performance with training settings, the performance improves by increasing the amount of training data. This suggests that the large-scale training data volume is important for the performance of neural networks.

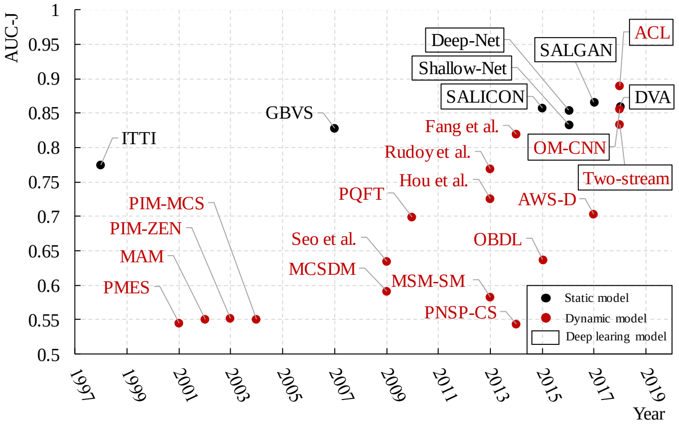


Fig. 6. Dynamic saliency prediction performance over time, evaluated on the DHF1K test set. The static (dynamic) saliency models are plotted as black (red) dots, and the deep learning based models are represented by black boxes. It can be observed a performance improvement starting in 2015, corresponding to the application of deep learning techniques to visual saliency detection. See Sec. 5.4 for details.

• **Performance on UCF sports.** On the test set (47 video sequences) of UCF sports, ACLNet again generates consistently better results than other state-of-the-art solutions (see Table 10). Interestingly, we find that with small amount of training data (training setting (iii), 103 video stimuli from UCF sports), ACLNet achieves a very high performance, even better than the model (*ACLNet, training setting (iv)*) trained with large-scale data (1.5K video stimuli). This can be explained by the lack of diversity in the training data, as the videos in UCF sports are highly related (with similar scenes and actors) and due to small scale. This is also consistent with our observation on UCF sports videos where 82.3% fixations are located on the human body (see Sec. 2.1).

5.4 Further Analyses

Now we provide detailed analyses to gain deeper insights on previous studies and suggest hints for future research.

• **Dynamic saliency models: deep *vs* non-deep learning.** In dynamic scenes, previous deep learning based dynamic saliency models (*i.e.*, OM-CNN, Two-stream) show significant improvements over classic dynamic models (*e.g.*, PQFT, Seo *et al.*, Rudoy *et al.*, Hou *et al.*, and Fang *et al.*). This demonstrates the strong learning capacity and premise of neural networks for modeling dynamic saliency.

• **Non-deep learning models: static *vs* dynamic.** An interesting finding is that classic dynamic methods (*i.e.*, PQFT, Seo *et al.*, Rudoy *et al.*, Hou *et al.*, and Fang *et al.*) do not perform as well as their static counterparts: ITTI and GBVS. This is probably due to two reasons. First, the perceptual cues and underlying mechanisms of visual attention allocation during dynamic viewing are more complex and still not clear. Second, previous studies are more focused on computational models of static saliency, while less efforts were paid for modeling dynamic saliency.

• **Deep learning models: static *vs* dynamic.** Compared with state-of-the-art deep learning based static models (*i.e.*,

TABLE 11
Attribute-based study w.r.t. content motion, camera motion, number of objects, scene illumination and number of people in DHF1K dataset. *ND-avg* indicates the average score of three top-performing heuristic models: GBVS [27], Fang *et al*. [14], and ITTI [26]. *D-avg* refers to the average score of three top-performing deep learning models: ACLNet, SalGAN [31] and DVA [29], according to Table 8. Symbol $*$ indicates non-deep learning models. See Sec. 5.6 for details. These notes are the same for Table 12.

Metric	Method	Content Motion			Camera Motion			#Objects				Scene Illumination			#People			
		stable	slow	fast	stable	slow	fast	0	1	2	≥3	day	night	indoor	0	1	2	≥3
AUC-J ↑	*ITTI [26]	0.768	0.798	0.799	0.779	0.801	0.803	0.799	0.828	0.794	0.767	0.807	0.751	0.780	0.798	0.801	0.788	0.783
	*PQFT [10]	0.685	0.692	0.709	0.715	0.692	0.683	0.710	0.710	0.718	0.671	0.690	0.665	0.710	0.671	0.726	0.700	0.693
	DVA [29]	0.880	0.862	0.855	0.862	0.866	0.855	0.883	0.879	0.870	0.839	0.861	0.822	0.867	0.857	0.876	0.860	0.850
	OM-CNN [24]	0.867	0.852	0.857	0.844	0.860	0.862	0.889	0.874	0.862	0.831	0.858	0.788	0.858	0.847	0.866	0.863	0.849
	ACLNet	0.896	0.886	0.886	0.884	0.887	0.891	0.919	0.902	0.892	0.866	0.888	0.812	0.891	0.876	0.898	0.895	0.880
	**ND-avg*	0.814	0.822	0.829	0.815	0.826	0.830	0.835	0.849	0.829	0.797	0.831	0.769	0.817	0.826	0.836	0.816	0.809
	D-avg	0.883	0.867	0.868	0.864	0.872	0.873	0.897	0.887	0.876	0.846	0.872	0.803	0.871	0.862	0.880	0.877	0.860

DVA, Deep-Net), previous deep learning based dynamic models (*i.e.*, OM-CNN, Two-stream) only obtain slightly better performance (or are on par). Although strong motion information (*i.e.*, optical flow, motion network) have been encoded into OM-CNN and Two-stream, their performance are still limited. We attribute this to the inherent difficulties of video saliency prediction and previous models' neglect of utilizing existing rich static saliency data.

- **Performance change over the past 20 years.** Fig. 6 plots the s-AUC over time, evaluated on the DHF1K test set. The first observation is that the performance gradually improved over time, which demonstrates the progress of visual saliency computation models. We also find a relatively rapid performance improvement starting in 2015, with the application of deep learning techniques to visual saliency modeling. A closer look reveals surprisingly that the ITTI model, as an early proposed saliency model, achieves far better performance than most non-deep learning dynamic saliency models. This indicates that previous heuristic video saliency models may be over-fitted over small datasets.

5.5 Qualitative Evaluation and Model Comparison

Fig. 7 gives visual results of ACLNet and four representative saliency models: ITTI [26], DVA [29], PQFT [10], and OM-CNN [24] on UCF sports [35] (a, b), Hollywood-2 [35] (c, d) and DHF1K (e, f). ITTI and PQFT are popular heuristic models which focus on static and dynamic saliency prediction, respectively. The other two, DVA and OM-CNN, are deep learning methods, showing promising performance among previous static and dynamic saliency models respectively, according to our prior quantitative study.

In Fig. 7 (a), most saliency models successfully detect semantically-meaningful parts (which typically attract human attention), such as human and text. However, previous methods fail to discriminate the correct relative importance among different parts. They assign high saliency to the diver while wrongly highlighting the importance of advertising text. ITTI performs worse as parts of the background are detected as salient. PQFT, yet another heuristic method, improves the results significantly. This highlights the importance of dynamic information in video saliency prediction. DVA also performs well, showing the advantage of applying neural networks in this field. But it is still worse than OM-CNN and ACLNet, which explicitly utilize motion information or model temporal dynamics using LSTMs. Fig. 7 (b) shows a crowded scene. In this case, ITTI fails to find the salient regions, due to the noise in the crowded background. PQFT is more favored, as the noise responses from parts of the background are successfully removed. In some frames, its performance is even better than DVA, showing again the importance of modeling temporal dynamics in this problem.

From Fig. 7 (c) we observe that, although OM-CNN accurately focuses on human faces, it fails to discriminate the most important one. This suggests a high-level understanding of the video content is needed. PQFT seems to be less effective, perhaps because the motion information is not important in this case, and introduces noise. This demonstrates how to fuse appearance, motion, and semantic information is essential in designing a heuristic dynamic saliency model. The difficulty of fusing motion and appearance features may be the main reason that PQFT gains lower overall performance than ITTI, though it makes better predictions in some cases. As depicted in Fig. 7 (d), deep learning methods such as ACLNet, DAV, and OM-CNN, show advantage over heuristic methods, as they can detect semantically-meaningful parts effectively. Besides, the third and forth columns show two adjacent frames which are almost the same. However, interestingly OM-CNN yields different results for these two very similar frames indicating the potential instability of deep learning models. Thus, exploring more stable and interpretable deep saliency models may be a promising and essential direction. Fig. 7 (e) shows a challenging scene with a highly-cluttered background and similar appearance distributions of the foreground and background. Traditional methods like ITTI and PQFT face difficulties while deep models perform more favorably. Among deep models, OM-CNN performs the worst as it fails to find the objects. Fig. 7 (d) gives an example that challenges all the methods. Clearly, ITTI and PQFT, dominated by the low-level handcrafted features, fail to interpret such a difficult scene. Though deep methods implicitly leverage semantically-rich features, they fall short to reason about the high-level knowledge, *i.e.*, the most important player, and tactical awareness behind their actions and movements.

5.6 Attribute-based Study

As stated in Sec. 3.1, to enable a deeper analysis and understanding of the performance of saliency models, we annotate the video sequences in DHF1K with a set of seven main categories (*i.e.*, *daily activity*, *sport*, *artistic performance*, *social activity*, *animal*, *artifact*, and *scenery*), and five attributes

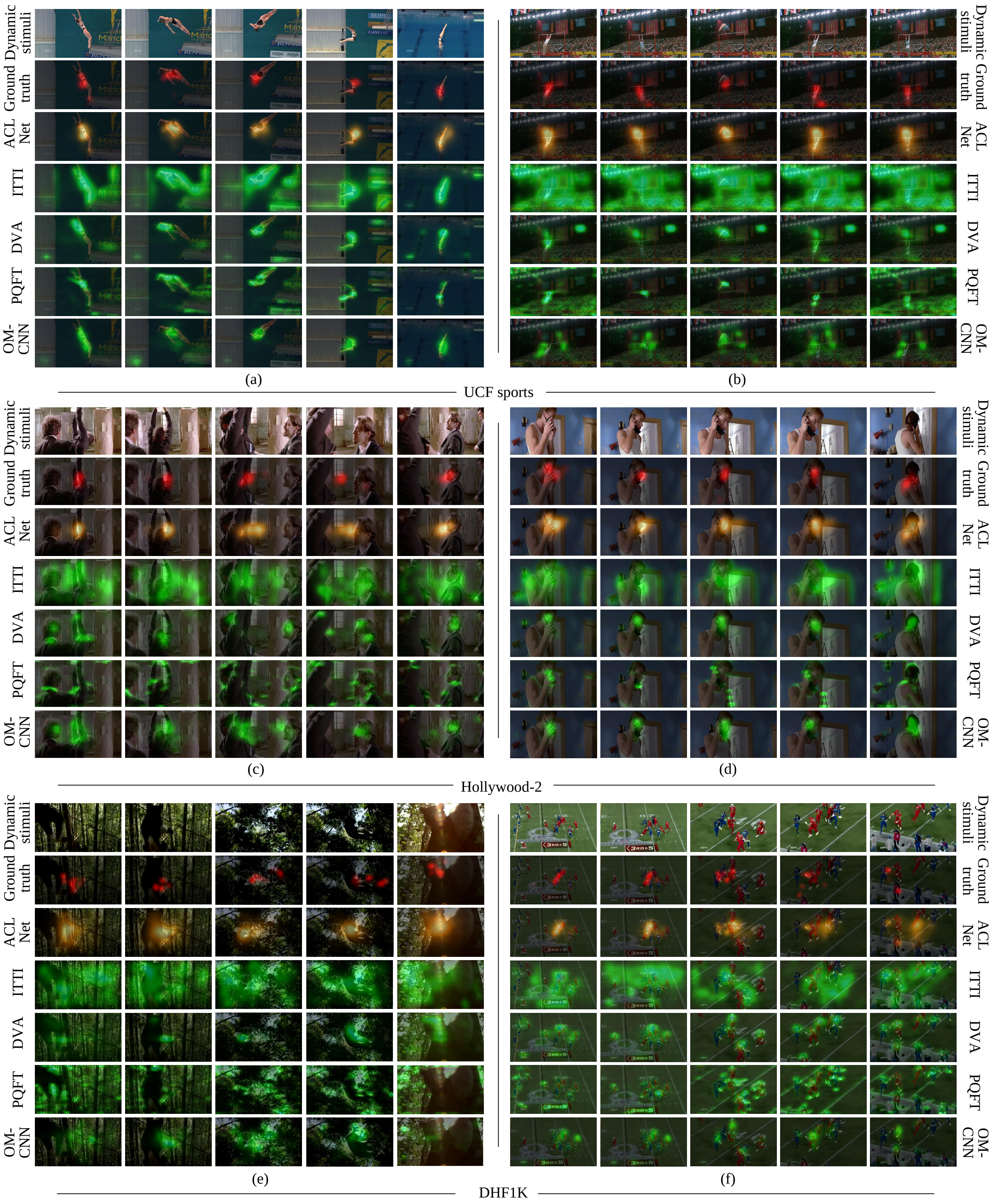


Fig. 7. Qualitative results of our ACLNet and four representative saliency models: ITTI [26] (non-deep static saliency model), DVA [29] (deep static saliency model), PQFT [10] (non-deep dynamic saliency model), and OM-CNN [24] (deep dynamic saliency model) on three video saliency datasets: UCF sports [35] (a, b), Hollywood-2 [35] (c, d) and DHF1K (e, f). Best viewed in color. It can be observed that the proposed ACLNet is able to handle various challenging scenes well and produces more accurate video saliency results than other competitors. See Sec. 5.5 for details.

TABLE 12
Attribute-based study w.r.t. video category on DHF1K dataset.

Metric	Method	Human				Animal	Artifact	Scenery
		Daily Activ.	Sports	Social Activ.	Art			
AUC-J↑	*ITTI [26]	0.777	0.821	0.767	0.786	0.822	0.794	0.760
	*PQFT [10]	0.716	0.713	0.651	0.757	0.678	0.726	0.619
	DVA [29]	0.865	0.868	0.838	0.887	0.888	0.855	0.803
	OM-CNN [24]	0.838	0.882	0.836	0.888	0.880	0.831	0.805
	ACLNet	0.883	0.907	0.870	0.909	0.905	0.865	0.844
	**ND-avg*	0.812	0.846	0.794	0.827	0.852	0.815	0.788
	D-avg	0.861	0.888	0.848	0.897	0.893	0.851	0.818

TABLE 13
Ablation study on DHF1K dataset. See Sec. 5.7 for details.

Aspects	Variants	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
ACLNet	training setting (iv) (*1.5K videos+10K images*)	**0.890**	**0.315**	**0.601**	**0.434**	**2.354**
Attention module	attention module (*1.5K videos+10K images*)	0.870	0.273	0.577	0.380	2.077
	w/o attention (*1.5K videos*)	0.847	0.236	0.579	0.306	1.685
	w/ implicit attention (*1.5K videos*)	0.854	0.238	0.586	0.343	1.762
	implicit attention module (*1.5K videos*)	0.831	0.217	0.554	0.307	1.573
	w/ center bias (*1.5K videos*)	0.851	0.230	0.582	0.324	1.731
	w/o residual connection (*1.5K videos+10K images*)	0.874	0.303	0.594	0.401	2.174
	w/o downsampling (*1.5K videos+10K images*)	0.870	0.298	0.583	0.389	2.085
Training	reduced training samples (*1.5K videos+5K images*)	0.877	0.297	0.588	0.372	2.098
convLSTM	w/o convLSTM (*1.5K videos+10K images*)	0.867	0.269	0.573	0.382	2.034
	chance	0.500	0.330	0.500	0.000	0.000

regarding content motion, camera motion, number of objects, scene illumination and number of people. By using these annotations, in this section, we construct subsets of the testset of DHF1K with different dominant features and analyze the performance of saliency models (*i.e.*, ACLNet , ITTI [26], DVA [29], PQFT [10], and OM-CNN [24]) for each video attributes/categories.

Table 11 and 12 report the performance on subsets of the testset of DHF1K (characterized by a particular attribute/category), measured by AUC-J. Due to limited space, we provides the results of ACLNet and the four representative saliency models: ITTI [26], DVA [29], PQFT [10], and OM-CNN [24] used in our qualitative study in Sec. 5.5. In addition, two extra baselines: *ND-avg* and *D-avg*, are included. *ND-avg* represents the average results of three top-performing non-deep learning models: GBVS [27], Fang *et al.* [14] and ITTI [26]; *D-avg* indicates average results of three top-performing deep learning models: ACLNet, SalGAN [31] and DVA [29], according to Table 8. Overall, from Table 11 and 12, it can be observed that ACLNet consistently outperforms other competitors in all settings. This verifies again the effectiveness of ACLNet. Next, we provide more detailed attribute-based analyses.

• **What is the most challenging situation for deep/non-deep learning based saliency models?** As demonstrated in Table 11, nighttime setting poses the greatest challenge to both non-deep learning (*ND-avg*: 0.769) and deep learning saliency models (*D-avg*: 0.803). This is sensible since the visually important regions are not easily discriminated from the background in dim environments. The scenes with multiple objects (≥3) also represent a major difficulty to current state-of-the-art saliency models (*ND-avg*: 0.797, *D-avg*: 0.846). In such cases, the relative importance among several main objects is needed to be accurately assigned. Unfortunately, such high-level scene understanding/reasoning is a hard task even for current top-performing heuristic and deep saliency models. This is consistent with the observation in [97]. Bylinskii *et al.* [97] found that humans tend to fixate people that are central to an event, or stand out from the crowd (discriminated by high-level factors such as facial expression, age, accessories, *etc.*). Interestingly, saliency models, either non-deep learning (*ND-avg*: 0.809) or deep learning models (*D-avg*: 0.860), also perform worse over scenes with multiple people (≥3). This hints again that the assignment of relative importance to objects (people) is one of the main challenges in this field.

• **What is the most challenging scene for deep/non-deep learning based saliency models?** As shown in Table 12, among different video categories, scenery scenes are very challenging to saliency models (*ND-avg*: 0.788, *D-avg*: 0.818). The main reason because it is hard to determine obvious salient areas on these cases, thus saliency models do not predict fixations well. To solve this issue, a deeper exploration of pure stimuli-driven human visual attention behavior is needed. Another difficult subset is social activity videos (*ND-avg*: 0.794, *D-avg*: 0.848). In this case, humans typically interact with each other (*e.g.*, hug, conversation, cooperation) or manipulate objects (*e.g.*, instrument). Thus, commonsense regarding human social behavior may be an essential factor that should be considered when creating an effective saliency model.

• **Do deep saliency models bring additional benefits other than improving performance?** The results in Table 11 and 12 demonstrate deep saliency models consistently improve performance over all the attributes and categories, especially when compared with heuristic methods. However, it is interesting to see the most difficult subsets (*i.e.*, nighttime setting, multiple objects, multiple people, scenery videos, and social activity videos) for the heuristic methods are all exactly the hardest ones for deep models (even with the same rank of difficulties). These observations imply that, although deep learning techniques greatly advance the state-of-the-art, they do not bring much insight into this problem. The performance improvement is mainly driven by the availability of large-scale data and the strong learning ability of neural networks. Efforts towards exploring the underlying mechanisms of human attention allocation behavior are still highly-needed to move this field forward.

5.7 Ablation Study

Now we perform detailed analysis of our proposed approach in several aspects on DHF1K. We verify the effectiveness of the proposed mechanism, and examine the influence of different training protocols, as summarized in Table 13.

• **Effect of attention mechanism.** By disabling the attention module and training only with video stimuli (baseline:

TABLE 14
Results for cross-dataset generalization experiment. Performance (left: AUC-J, right: SIM) for dynamic saliency prediction when training on one dataset (rows) and testing on another (columns), *i.e.*, each row is: training on one dataset and testing on all the datasets. “Self” refers to training and testing on the same dataset (same as diagonal). “Mean Others” indicates average performance on all except self. See Sec. 5.8 for details.

Metric	AUC-J ↑					
Test on: / Train on:	DHF1K	Hollywood-2	UCF sports	Self	Mean others	Percent drop ↓
DHF1K	**0.833**	0.852	0.842	0.833	0.847	**-2%**
Hollywood-2	0.818	**0.859**	0.822	0.859	0.820	5%
UCF sports	0.820	0.828	**0.851**	0.824	0.875	3%
Mean others	0.819	0.840	0.832	-	-	-

Metric	SIM ↑					
Test on: / Train on:	DHF1K	Hollywood-2	UCF sports	Self	Mean others	Percent drop ↓
DHF1K	**0.219**	0.330	0.302	0.219	0.316	**-44%**
Hollywood-2	0.214	**0.365**	0.262	0.365	0.238	35%
UCF sports	0.204	0.327	**0.348**	0.348	0.266	25%
Mean others	0.209	0.328	0.282	-	-	-

w/o attention), we observe clear performance drop (*e.g.* AUC-J: 0.890→0.847), showing the effectiveness of the attention module and showing that leveraging static stimuli indeed improves the predication accuracy in dynamic scenes. Our attention module is trained over existing static eye-tracking datasets in an explicit manner. We replace our attention module with a traditional implicit attention mechanism, which is achieved via replacing the last sigmoid activation with a spatial softmax operation and only using dynamic eye-tracking data. We find that the implicit attention mechanism boosts the performance, compared to the model without attention. But it is worse than the proposed explicit attention module, which can be directly trained from data.

To gain more insight about the attention module, *i.e.*, what does the attention module learn, we offer two baselines *attention module* and *implicit attention module*. The two baselines represent the attention maps predicted by the proposed attention module (trained with both implicit and explicit manners) and the implicit attention module (trained only in an implicit way), respectively. From their results we draw two conclusions. First, both attention modules can capture visual importance, as the performance is significantly above chance. Second, the attention module trained in both implicit and explicit manners performs better than the implicit attention module, hence resulting in better final dynamic fixation prediction results.

For baseline *w/ center bias*, we replace our attention module with a pre-computed center prior, which is calculated by averaging all the saliency maps over the training set in DHF1K dataset. The model equipped with the center prior gains higher performance over all the metrics, except the s-AUC which is specially designed for alleviating the bias borrowed by the center prior.

To explore the effect of the residual connection in attention module (Equ. 8), we train the model based on Equ. 5 (without residual connection). We observe a minor decrease showing that employing residual connection could avoid distorting spatial features in frames.

In our attention module, we apply down-sampling for enlarging the receptive field. We also study the influence of such design. We find that the attention module with enlarged receptive field leads to better performance, since the model could make prediction in global view.

• **Training.** We assess different training protocols. By reducing the amount of static training stimuli from 10K to 5K, we observe a performance drop (*e.g.*, AUC-J: 0.890→0.877). The baseline (*w/o attention*) can also be viewed as the model without any static training stimuli, which gains worse performance (*e.g.*, AUC-J: 0.890→0.847).

• **Effect of convLSTM.** To study the influence of convLSTM, we re-train our model without convLSTM (using training setting (iv)) and obtain a baseline: *w/o convLSTM*. We observe a drop in performance showing that the dynamic information learned in convLSTM boosts the performance.

5.8 Cross-Dataset Generalization

Datasets play an important role in advancing visual saliency prediction, not just as source for training models, but also as a means for measuring and comparing performance. Datasets are collected with the goal of representing the visual world, summarizing the algorithm as a single benchmark performance number. A concern thus comes into view: it is necessary to evaluate how well a particular dataset represent the real visual world. Or more specially, quantitatively measuring the dataset’s generalization ability. Here, we follow [104] to assess how general video saliency dataset are. We study cross-dataset generalization, *e.g.*, training on DHF1K and testing on Hollywood-2.

Following [104], for each dataset, we re-train our ACLNet (w/o attention module and static training data) with 103 videos and test it on 47 videos. Both the numbers of training and test videos are the maximum ones possible due to the limited size of the UCF sports dataset. Results are summarized in Table 14. Each column corresponds to the performance when training on all the datasets respectively and testing on one dataset. Each row corresponds to training on one dataset and testing on all the datasets. Note that since our training/testing protocol is different from the one used in benchmarks mentioned in previous sections, the actual performance numbers are not meaningful. Rather, it is the relative performance difference that matters. Not surprisingly, we observe that the best results are achieved when training and testing on the same dataset. By looking at the numbers across one row, we can determine how good a dataset is at generalizing to the others. By looking at the numbers across each column, we can determine how easy a dataset is for the other datasets. We find that DHF1K is the most difficult dataset (lowest column averages across two metrics; AUC-J: 0.819, SIM: 0.209) and generalizes the best (highest row averages on *Mean others* and lowest row averages on *Percent drop*).

Overall, this analysis demonstrates that the proposed DHF1K dataset has made significant improvement in terms of generalization and hardness, compared with previous eye-tracking datasets: Hollywood-2 and UCF sports.

6 Discussion and Conclusion

In this paper, we presented the Dynamic Human Fixation 1K (DHF1K) dataset, which is a large-scale carefully designed and systematically collected benchmark dataset for video saliency analysis. It contains 1K videos capturing representative instances, diverse contents and various motions, with human eye-tracking records and attribute annotations. We further proposed a novel deep learning based video saliency model ACLNet, which leverages a supervised attention mechanism to explicitly capture static saliency information and help LSTM better capture dynamic saliency representations over successive frames. Then, we performed extensive experiments on DHF1K, Hollywood-2, and UCF-sports datasets. To the best of our knowledge, our experiments form the largest scale performance evaluation of dynamic saliency models. We compared our model with previous visual saliency models and showed that it outperforms other contenders and runs very efficiently. We also performed attribute-level evaluation, and assessed the generalization ability of video saliency datasets. Our analyses and benchmark are expected to motivate future interests in this field.

References

[1] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1072–1080.

[3] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim, "Supervising neural attention models for video captioning by human gaze data," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, vol. 29, no. 1, pp. 1–14, 2011.

[5] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 932–937.

[6] 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.

[7] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Dynamically encoded actions based on spacetime saliency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2755–2764.

[8] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

[9] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *arXiv preprint arXiv:1904.09146*, 2019.

[10] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[11] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 15–15, 2009.

[12] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1147–1154.

[13] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2008.

[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] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5501–5510.

[16] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 893–907, 2017.

[17] Y.-F. Ma and H.-J. Zhang, "A new perceived motion based shot content representation," in *Proceedings of the International Conference on Image Processing*, 2001, pp. 426–429.

[18] G. Agarwal, A. Anbu, and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed mpeg domain," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2003.

[19] Y.-F. Ma and H.-J. Zhang, "A model of motion attention for video skimming," in *Proceedings of the International Conference on Image Processing*, 2002.

[20] A. Sinha, G. Agarwal, and A. Anbu, "Region-of-interest based compressed domain video transcoding scheme," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[21] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang, "A motion attention model based rate control algorithm for h.264/avc," in *Proceedings of the IEEE/ACIS International Conference on Computer and Information Science*, 2009, pp. 568–573.

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

[23] Y. Fang, W. Lin, Z. Chen, C. M. Tsai, and C. W. Lin, "A video saliency detection model in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27–38, 2014.

[24] 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*, 2018.

[25] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Transactions on Multimedia*, 2018.

[26] 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.

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

[28] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[29] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

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

[31] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition–workshop*, 2017.

[32] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[33] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 353–369.

[34] 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 *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[35] 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.

[36] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[37] 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.

[38] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic, "Eye-tracking database for a set of standard video sequences," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 898–903, 2012.

[39] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2929–2936.

[40] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

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

[42] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[43] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[44] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top-down saliency using natural statistics." *Visual Cognition*, vol. 17, no. 6-7, p. 979, 2009.

[45] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 470–477.

[46] 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.

[47] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32–32, 2008.

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

[49] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2006, pp. 155–162.

[50] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[51] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[52] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 8, pp. 2014–2027, 2017.

[53] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[54] 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.

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

[56] 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.

[57] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 392–404, 2018.

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

[59] 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*, 2018.

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

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

[62] 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.

[63] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar, "Learning gaze transitions from depth to improve video saliency estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1698–1707.

[64] S. Gorji and J. J. Clark, "Going from image to video saliency: Augmenting image salience with dynamic attentional push," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[65] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[66] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[67] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[68] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[69] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5025–5034, 2016.

[70] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.

[71] 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*, 2019.

[72] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[73] W. Wang, S. Zhao, J. S. Shen, S. C. H. Steven Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[74] 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.

[75] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[76] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 715–731.

[77] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Transactions on Image Processing*, 2019.

[78] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[79] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[80] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[81] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for RGB-D action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1651–1664, 2016.

[82] 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 Transactions on Image Processing*, vol. 27, no. 2, pp. 568–579, 2018.

[83] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015.

[84] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[85] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2956–2964.

[86] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[87] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2015, pp. 379–389.

[88] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2016, pp. 2249–2255.

[89] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[90] Y. Jiang, Z. Wu, J. Wang, X. Xue, and S. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 352–364, 2018.

[91] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[92] J. M. Wolfe, "Guided search 4.0," *Integrated Models of Cognitive Systems*, pp. 99–119, 2007.

[93] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[94] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 742–756, 2015.

[95] B. C. Motter, "Neural correlates of attentive selection for color or luminance in extrastriate area V4," *Journal of Neuroscience*, vol. 14, no. 4, pp. 2178–2189, 1994.

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

[97] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *European Conference on Computer Vision*, 2016.

[98] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, pp. 18–18, 2008.

[99] M. Wischnewski, A. Belardinelli, W. X. Schneider, and J. J. Steil, "Where to look next? Combining static and dynamic proto-objects in a tva-based model of visual attention," *Cognitive Computation*, vol. 2, no. 4, pp. 326–343, 2010.

[100] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[101] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[102] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.

[103] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[104] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.

Wenguan Wang received his PhD degree from Beijing Institute of Technology in 2018. He is currently a senior scientist at Inception Institute of Artificial Intelligence (IIAI), UAE. From 2016 to 2018, he was a joint Ph.D. candidate in Department of Statistics, University of California, directed by Prof. Song-Chun Zhu. He received the Baidu Scholarship in 2016 and ACM China Doctoral Dissertation Award in 2018. His current research interests include computer vision, image processing and deep learning.

Jianbing Shen (M'11-SM'12) is a Professor at Beijing Institute of Technology. He has published about 100 journal and conference papers such as *IEEE TPAMI*, *IEEE CVPR*, and *IEEE ICCV*. He has also obtained many flagship honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. His research interests include computer vision and deep learning. He is an Associate Editor of IEEE TIP, IEEE TNNLS and Neurocomputing.

Jianwen Xie received his Ph.D degree in statistics from University of California, Los Angeles (UCLA) in 2016. He is currently a senior research scientist at Hikvision Research Institute, USA. Before joining Hikvision, he was a staff research associate and postdoctoral researcher in the Center for Vision, Cognition, Learning, and Autonomy (VCLA) at UCLA from 2016 to 2017. His research focuses on generative modeling and learning with applications in computer vision.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program, *etc*. He is an Associate Editor of IEEE TIP and serves as Area Chairs for CVPR 2019 and ICCV 2019.

Haibin Ling received the PhD degree from University of Maryland in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. After that, he joined Siemens Corporate Research as a research scientist. He then joined Temple University as an Assistnat Professor in 2008 and was later promoted to Associate Professor. In fall 2019, he will join SUNY Stony Brook University as an Empire Innovation Professor. He is an Associate Editor of *IEEE Trans. on PAMI*, *Pattern Recognition*, and *CVIU*, and served as Area Chairs for *CVPR* 2014, 2016 and 2019.

Ali Borji received the PhD degree in cognitive neurosciences from the Institute for Studies in Fundamental Sciences (IPM), Tehran, IRAN in 2009. He hid a postdoc at the University of Southern California from 2010 to 2014. He is currently a senior research scientist at MarkableAI, NYC. His research interests include visual attention, visual search, object and scene recognition, machine learning, neurosciences, and biologically plausible vision models.