Revisiting Video Saliency: A Large-scale Benchmark and a New Model

Wenguan Wang [1], Jianbing Shen* [1], Fang Guo [1], Ming-Ming Cheng [2], Ali Borji [3]

[1]Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China
[2]CCCE, Nankai University, China [3]Department of Computer Science, University of Central Florida, USA
wenguanwang.ai@gmail.com, {shenjianbing, guofang}@bit.edu.cn
cmm@nankai.edu.cn, aborji@crcv.ucf.edu
https://github.com/wenguanwang/DHF1K

Abstract

In this work, we contribute to video saliency research in two ways. First, we introduce a new benchmark for predicting human eye movements during dynamic scene free-viewing, which is long-time urged in this field. Our dataset, named DHF1K (Dynamic Human Fixation), consists of 1K high-quality, elaborately selected video sequences spanning a large range of scenes, motions, object types and background complexity. 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. Second, we propose a novel video saliency model that augments the CNN-LSTM network architecture with an 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 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. We thoroughly examine the performance of our model, with respect to state-of-the-art saliency models, on three large-scale datasets (i.e., DHF1K, Hollywood2, UCF sports). Experimental results over more than 1.2K testing videos containing 400K frames demonstrate that our model outperforms other competitors.

*Corresponding author: *Jianbing Shen*. This work was supported in part by the Beijing Natural Science Foundation under Grant 4182056, the National Basic Research Program of China under Grant 2013CB328805, the Fok Ying Tung Education Foundation under Grant 141067, and the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission.

1. Introduction

Human visual system (HVS) has an astonishing ability to quickly select visually important regions in its visual field. This cognitive process enables humans to easily 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 with the rapid spread of deep learning techniques and the availability of large-scale static gaze datasets (*e.g.*, SALICON [31]). In stark contrast, predicting observers' fixations during dynamic scene free-viewing has less been explored. This task, referred to as *dynamic fixation prediction* or *video saliency detection* is very useful for understanding human attentional behaviors and has several practical real-word applications (*e.g.*, video captioning, compression, question answering, object segmentation, *etc*). It is thus highly desired to have a standard, high-quality dataset composed of diverse and representative video stimuli. Exiting datasets are severely limited in their coverage and scalability, and they only include special scenarios such as limited human activities in constrained situations. None of them includes general, representative, and diverse instances in unconstrained, task-independent scenarios. As a consequence, existing datasets often fail to offer a rich set of fixations for learning video saliency and to assess models. Moreover, the existing datasets did not provide an evaluation server with standalone held out test set to avoid potential dataset overfitting, which hinders further development on this topic.

While saliency benchmarks (*e.g.*, MIT300 [32] and LSUN [68]) have been very instrumental in progressing the static saliency field, such standard widespread benchmarks are missing for video saliency modeling. We believe such benchmarks are highly needed to move the field forward. To this end, we propose a new benchmark "DHF1K (Dynamic Human Fixation 1K)" with a public server for report-

ing evaluation results on a preserved test set. Our benchmark contains a dataset that is unique in terms of generality, diversity and difficulty. It includes 1K videos with more than 600K frames and per-frame fixation annotations from 17 observers. The sequences have been carefully collected to include diverse scenes, motion patterns, object categories, and activities. DHF1K is accompanied with a comprehensive evaluation of several state-of-the-art approaches [16, 52, 50, 23, 13, 20, 37, 30, 2, 28, 18, 24, 57, 47]. 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 would enable a deeper understanding of gaze guidance in free viewing of dynamic scenes.

Further, we propose a novel CNN-LSTM architecture [12, 46] based video saliency model with a supervised attention mechanism. CNN layers are utilized for extracting static features within input frames, while convolutional LSTM (convLSTM) [66] is utilized for sequential fixation prediction over successive frames. An attention module, learned from existing large-scale image saliency datasets, is used to enhance spatially informative features of the CNN. Such a design helps disentangle underlying spatial and temporal factors of dynamic attention and allows convLSTM to learn temporal saliency representations efficiently.

Our contributions are three-fold. **First**, we introduce a standard benchmark of 1K videos covering a wide range of scenes, motions, activities, *etc*. To the best of our knowledge, the proposed dataset is the largest eye-tracking dataset for *dynamic*, *free-viewing* fixation prediction. **Second**, we present a novel attentive CNN-LSTM architecture for predicting human gaze in dynamic scenes, which explicitly encodes static attention into dynamic saliency representation learning by leveraging both static and dynamic fixation data. **Third**, we present a comprehensive analysis of video saliency models (the first one, to the best of our knowledge) on existing datasets (Hollywood-2, UCF sports), and our new DHF1K dataset. Results show that our model significantly outperforms previous methods.

2. Related Work

2.1. Video Eye-Tracking Datasets

There exist several datasets [43, 44, 25, 17] for dynamic visual saliency prediction, but they are limited and often lack variety, generality and scalability of instances. Some statistics of these datasets are summarized in Table 1.

The **Hollywood-2** dataset [43] comprises all the $1,707$ videos from Hollywood-2 action recognition dataset [42]. The videos are collected from 69 Hollywood movies with 12 action categories, such as eating, kissing and running. The human fixation data were tracked 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 84.5% fixations are located around the faces.

Dataset	Year	Videos	Resolution	Duration(s)	Viewers	Task
CRCNS [25]	2004	50	640×480	6-94	15	task-goal
Hollywood-2 [43]	2012	1,707	720×480	2-120	19	task-goal
UCF sports [43]	2012	150	720×480	2-14	19	task-goal
DIEM [44]	2011	84	1280×720	27-217	$\sim$50	free-view
SFU [17]	2012	12	352×288	3-10	15	free-view
DHF1K(**Ours**)	2017	1,000	640×360	17-42	17	free-view

Table 1. **Statistics of typical dynamic eye-tracking datasets.**

The **UCF sports** fixation dataset [43] contains 150 videos taken from the UCF sports action dataset [49]. The videos cover 9 common sports action classes, such as diving, swinging and walking. Similar to Hollywood-2, the viewers have been biased towards task-aware observation by being instructed to "identify the actions occurring in the video sequence". From the statistics of $1,000$ frames randomly selected from UCF sports, we found 82.3% fixations fall inside the human body area.

The **DIEM** dataset [44] 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 [25, 17], or collected for a special purpose (*e.g.*, salient objects in videos [59]). More importantly, none of the aforementioned datasets includes a preserved test set for avoiding potential data overfitting, which has seriously hampered the research process.

2.2. Computational Models for Fixation Prediction

The study of human gaze patterns in static scenes has received significant interests, which can be dated back to [28, 27]. **Early static saliency models** [36, 69, 15, 7, 18, 22, 33, 63] are mostly based on the *contrast* assumption that conspicuous visual features "pop-out" and involuntarily capture attention (see [5, 6] for review). Computational models compute multiple visual features such as color, edge, and orientation at multiple spatial scales to produce a "saliency map": an image distribution predicting the conspicuity of specific locations and their likelihood in attracting attention [27, 44]. The locations with more distinct feature responses over surroundings usually gain higher saliency values. **Deep learning based static saliency models** [54, 35, 24, 39, 47, 29, 56, 57] have achieved astonishing improvements, relying on the powerful end-to-end learning ability of neural network and the availability of large-scale static saliency datasets [31].

Previous investigations of dynamic human fixation

(a) (b) (c)

Figure 1. **Average annotation maps of three datasets** used in benchmarking: (a) Hollywood-2, (b) UCF sports, (c) DHF1K.

[14, 16, 41, 50, 52, 23, 13, 20, 37] leveraged both static stimulus features and temporal information (*e.g.*, optical flow, difference-over-time, *etc*). Some of those studies [14, 41, 52] can be viewed as extensions of exiting static saliency models with additional motion features. Those models are mainly bound to significant feature engineering and limited representation ability of hand-crafted features. To date, only a few **deep learning based video saliency models** [2, 30] exist in this field. They are mainly based on two-stream network architecture [2] that accounts for color images and motion fields separately, or two-layer LSTM with object information [30]. These works show a better performance and demonstrate the potential advantages in applying neural networks to this problem. However, they do not 1) consider attentive mechanisms; 2) utilize existing large-scale static fixation datasets; and 3) exhaustively assess their performance over large amount of data.

There are some **salient object detection models** [40, 1, 11, 61, 58, 60, 4, 62, 21] that attempt to uniformly highlight salient object regions in images or videos. Those models are often task-driven and focus on inferring the main object, instead of investigating the 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 success in several computer vision [67, 9, 55] and natural language processing tasks [51, 48]. In such studies, attention is 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 attention for enhancing intra-frame salient features, thus allowing the LSTM to model dynamic representations more easily. In contrast to previous models learning attentions implicitly, our attention module encodes strong static saliency information and can be learned from existing static saliency dataset in a supervised manner. This design leads to improved generality and prediction performance. 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. DHF1K Dataset

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

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

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

Table 2. **Statistics for video categories** in DHF1K dataset.

DHF1K	Content motion			Camera motion			#Objects			
	stable	slow	fast	stable	slow	fast	0	1	2	≥3
#videos	126	505	369	343	386	271	56	335	254	355

Table 3. **Statistics regarding motion patterns and number of main objects** in DHF1K dataset.

DHF1K	Scene illumination			#People			
	day	night	indoor	0	1	2	≥3
#videos	577	37	386	345	307	236	112

Table 4. **Statistics regarding scene illumination and number of people** in DHF1K dataset.

Stimuli. The collection of dynamic stimuli mainly follows the following 4 principles.

• *Large scale and high quality*. Both scale and quality are necessary to ensure the content diversity of a dataset and is 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*) and carefully selected $1,000$ video sequences from the retrieval results. All videos were converted from their original sources to a 30 fps Xvid MPEG-4 video file in an AVI container and were resized uniformly into 640×360 spatial resolution. Thus, DHF1K comprises a total $1,000$ 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). Those labels are further classified into 7 main categories (see Table 2). Those semantic annotations would enable a deeper understanding of the high-level stimuli factors guiding human gaze in dynamic scenes and be indicative for potential research. In Fig. 2, we show example frames from each category.

• *Varied motion patterns*. Previous investigations [26, 14, 44] suggested that motion is one of the key factors that directs attention allocation in dynamic viewing. For this, DHF1K is designed to span varied motion patterns (*stable-/slow-/fast-motion* of content and camera). Please see Table 3 for the information regarding motion patterns.

• *Various objects*. Previous studies [65, 38, 3] in cognitive and computer vision confirmed that object information is indicative to human fixations. The objects in the dataset vary

Figure 2. **Example frames from DHF1K** with fixations (red dots) and corresponding categories.

in their type (*e.g.*, *human*, *animal*, in Table 2) and frequency (see Table 3). For each video, five subjects were instructed to count the number of the main objects. The majority vote of their counts was considered as the final count.

For completeness, in Table 4, we offer the information of the scene illumination and the amount of humans in the dataset. As demonstrated in [45], luminance is an important exogenous factor for attentive selection. Further, human beings are important high-level stimuli [10, 8] in free-viewing.

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 help participants' heads still at a distance of around 68 cm, as advised by the product manual.

Participants. 17 participants (10 males and 7 females, aging between 20 and 28) who passed the calibration of the eye tracker and had less than 10% fixation dropping rate, were quantified for our eye tracking experiment. All participants had normal or corrected-to-normal vision. All subjects had not seen the stimuli in DHF1K before. All provided informed consent and were naïve to the underlying purposes of the experiment.

Data capturing. The stimuli were equally partitioned into 10 non-overlapping sessions. Participants were required to freeview 10 sessions of videos in random order. In each session, the videos were also displayed at random. Before the experiments, every participant was calibrated using the standard routine in product manual with recommended settings for the best results. 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 can take a rest until she was ready for viewing the next session. Finally, $51,038,600$ fixations were recorded from 17 subjects on $1,000$ videos.

Training/testing split. We split $1,000$ dynamic stimuli into separate 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

Overview. Fig. 3 presents the overall architecture of our video saliency model. It is based on a CNN-LSTM architecture 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 our model is able to produce accurate, spatiotemporal saliency with improved generalization ability. Next, we explain each component of our model in detail.

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 [66] as input. Here, the convLSTM is used for modeling the temporal dynamic nature of this sequential problem, which is achieved by incorporating memory units with gated operations. Additionally, through replacing dot products with convolutional operations, the convLSTM is able to preserve spatial information, which is essential for making spatially-variant pixel-level prediction.

More specifically, the convLSTM utilizes three convolution gates (*input*, *output* and *forget*) to control the flow of signal within the 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:

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

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

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

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

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

i_t, f_t, o_t are the gates. σ and $\tanh$ are the activation functions of logistic sigmoid and hyperbolic tangent, respectively. '$*$' denotes the convolution operator and '$\circ$' represents Hadamard product. 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 [53] are used. For preserving more spatial details, we remove *pool4* and *pool5* layers, which results in $\times 8$ instead of $\times 32$ downsampling. At time step t, with an input frame I_t with 224×224 resolution, we have $\mathcal{X}_t \in \mathbb{R}^{28 \times 28 \times 512}$ and a 28×28 dynamic saliency map from the convLSTM. The kernel size of the conv layer in convLSTM is set as 3.

Attention module. We extend 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:

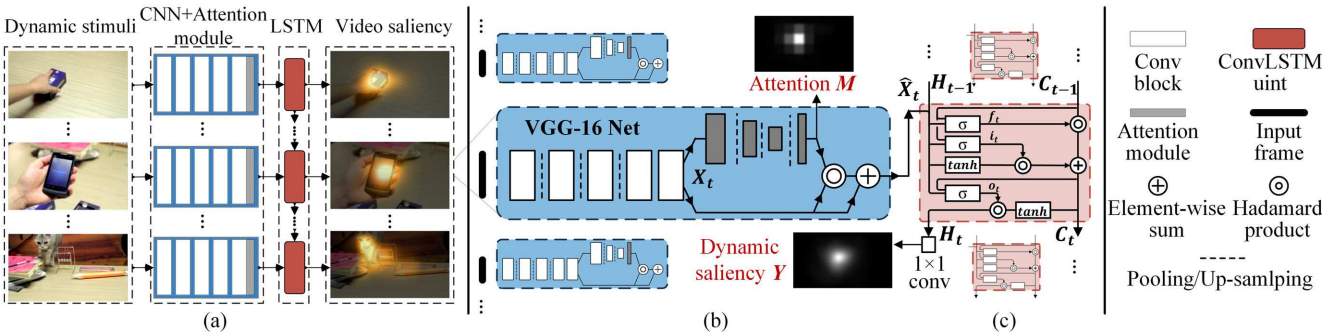

Figure 3. **Network architecture of the proposed video saliency model.** (a) Attentive CNN-LSTM architecture. (b) CNN layers with attention module are used for learning intra-frame static features, where the attention module is learned with the supervision from static saliency data. (c) ConvLSTM used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

• Previous studies [26, 64] shown that human attention is guided by both static and dynamic factors. Through the additional attention module, the CNN is enforced to generate a more explicit spatial saliency representation. This helps disentangle underlying spatial and temporal factors of dynamic attention, and allows convLSTM better capture temporal dynamics.

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

• In VGG-16, we remove the last two pooling layers to obtain a large feature map. This dramatically decreases the receptive field ($212\times212\rightarrow140\times140$), which cannot cover the whole frame (224×224). To remedy this, we insert a set of down- and up-sampling operations into the attention module, which would enhance the intra-frame saliency information with an enlarged receptive field. By this, our model is 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, the attention module generates a downsampled attention map (7×7) with an enlarged receptive field (260×260). Then the small attention map is $\times4$ 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 index of the channel. Here, the attention module work as a feature selector to enhance the 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 [19, 55], we improve Eq. 6 in 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 §5.2 and §5.4, more detailed explorations for the attention module are offered.

Different from previous attention mechanisms that learn task-related attention in an implicit way, our attention module can learn from existing large-scale static fixation data in an explicit and supervised manner (detailed in next part).

Loss function. We use the following loss function [24] that considers three different saliency evaluation metrics instead of one. The rationale here is that no single metric can fully capture how satisfactory a saliency map is.

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, that records whether a pixel receives human fixation. The continuous saliency map is obtained via blurring each fixation location with a small Gaussian kernel. Our loss functions 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 to evaluate saliency prediction models. αs 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:

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

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

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

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

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 eye fixations P. Since CC and NSS are similarity metrics, their negatives are adopted for minimization.

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 $\{Q_t^d\}_{t=1}^T$ denote 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 scene.

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 the M, P^s, Q^s indicate the attention map for our static attention module, the ground-truth static fixation map, and the ground-truth static saliency map. In this process, the training of attention module is supervised by the ground-truth static fixation. Note that, in image training batch, we do not train our LSTM module, as it 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 as 20, and the images are randomly sampled from existing static fixation dataset. More implementation details can be found in § 5.1.

5. Experiments

5.1. Experimental Setup

Training/testing protocols. We use the static stimuli (10,000 images) from the training set of SALICON [31] 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 [42], 823 videos for training and 884 videos for testing. For UCF sports, the training and testing sets include 103 and 47 videos, respectively, as suggested by [49]. We randomly sample 10% videos from the training sets of Hollywood-2, and UCF sports as their validation sets. We evaluate our model on the testing sets of DHF1K, Hollywood-2, and UCF sports dataset, in total 1,231 video sequences with more than 400,000 frames.

Implementation details. Our model is implemented in Python on Keras, and trained with the Adam optimizer [34]. During the training phase, 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. We perform early-stopping on the validation set.

Competitors. We compare our model with nine dynamic saliency models: PQFT [16], Seo *et al.* [52], Rudoy *et al.* [50], Hou *et al.* [23], Fang *et al.* [13], OBDL [20], AWS-D [37], OM-CNN [30], and Two-stream [2][1]. For the sake of complementary, we further compare with six state-of-the-art static attention models: ITTI [28], GBVS [18], SALICON [24], DVA [57], Shallow-Net [47], and Deep-Net [47]. OM-CNN, Two-stream, SALICON, DVA, Shallow-Net, and Deep-Net are deep learning models, and others are classical saliency models. Those models are selected due to: 1) their representability of the diversity of the state-of-the-art; or 2) publicly available implementations.

Baselines. We further derive 8 baselines. For each training setting, we derive two baselines: *Our* and *Attention module*, refer to our final dynamic saliency prediction and the intermediate output of our attention module, respectively.

Evaluation metrics. Here, 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). Please refer to [5, 57] for detailed descriptions of these metrics.

Computation load. The whole model is trained in an end-to-end manner. The entire training procedure takes about 30 hours with a single NVIDIA TITAN X GPU and a 4.0GHz Intel processor (in training setting (iv)). Since our model does not need any pre- or post-processing, it takes only about 0.08s to process an frame image of size 224×224.

5.2. Performance comparison

Performance on DHF1K. Table 5 reports the comparative results with the aforementioned saliency models, on the testing set (300 video sequences) of DHF1K dataset. In can be observed that the proposed model consistently and significantly outperforms other competitors, across all the metrics. This can be contributed to our specially designed atten-

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

	Dataset / Method	DHF1K					Hollywood-2					UCF sports				
		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Dynamic models	*PQFT [16]	0.699	0.139	0.562	0.137	0.749	0.723	0.201	0.621	0.153	0.755	0.825	0.250	0.722	0.338	1.780
	*Seo *et al.* [52]	0.635	0.142	0.499	0.070	0.334	0.652	0.155	0.530	0.076	0.346	0.831	0.308	0.666	0.336	1.690
	*Rudoy *et al.* [50]	0.769	0.214	0.501	0.285	1.498	0.783	0.315	0.536	0.302	1.570	0.763	0.271	0.637	0.344	1.619
	*Hou *et al.* [23]	0.726	0.167	0.545	0.150	0.847	0.731	0.202	0.580	0.146	0.684	0.819	0.276	0.674	0.292	1.399
	*Fang *et al.* [13]	0.819	0.198	0.537	0.273	1.539	0.859	0.272	0.659	0.358	1.667	0.845	0.307	0.674	0.395	1.787
	*OBDL [20]	0.638	0.171	0.500	0.117	0.495	0.640	0.170	0.541	0.106	0.462	0.759	0.193	0.634	0.234	1.382
	*AWS-D [37]	0.703	0.157	0.513	0.174	0.940	0.694	0.175	0.637	0.146	0.742	0.823	0.228	0.750	0.306	1.631
	OM-CNN [30]	0.856	0.256	0.583	0.344	1.911	0.887	0.356	0.693	0.446	2.313	0.870	0.321	0.691	0.405	2.089
	Two-stream [2]	0.834	0.197	0.581	0.325	1.632	0.863	0.276	0.710	0.382	1.748	0.832	0.264	0.685	0.343	1.753
Static models	*ITTI [28]	0.774	0.162	0.553	0.233	1.207	0.788	0.221	0.607	0.257	1.076	0.847	0.251	0.725	0.356	1.640
	*GBVS [18]	0.828	0.186	0.554	0.283	1.474	0.837	0.257	0.633	0.308	1.336	0.859	0.274	0.697	0.396	1.818
	SALICON [24]	0.857	0.232	0.590	0.327	1.901	0.856	0.321	0.711	0.425	2.013	0.848	0.304	0.738	0.375	1.838
	Shallow-Net [47]	0.833	0.182	0.529	0.295	1.509	0.851	0.276	0.694	0.423	1.680	0.846	0.276	0.691	0.382	1.789
	Deep-Net [47]	0.855	0.201	0.592	0.331	1.775	0.884	0.300	0.736	0.451	2.066	0.861	0.282	0.719	0.414	1.903
	DVA [57]	0.860	0.262	0.595	0.358	2.013	0.886	0.372	0.727	0.482	2.459	0.872	0.339	0.725	0.439	2.311
Training setting (i)	Ours	0.885	0.311	0.553	0.415	2.259	0.905	0.471	0.757	0.577	2.517	0.894	0.403	0.742	0.517	2.559
	Attention module	0.854	0.251	0.545	0.332	1.755	0.880	0.415	0.748	0.529	2.283	0.853	0.333	0.719	0.435	1.946
Training setting (ii)	Ours	0.878	0.297	0.543	0.388	2.125	0.912	0.519	0.754	0.609	3.049	0.874	0.364	0.727	0.452	2.186
	Attention module	0.855	0.250	0.541	0.318	1.703	0.885	0.416	0.690	0.490	2.113	0.860	0.322	0.656	0.367	1.667
Training setting (iii)	Ours	0.866	0.277	0.596	0.362	1.951	0.884	0.449	0.749	0.534	2.647	**0.936**	**0.599**	**0.816**	**0.742**	**4.122**
	Attention module	0.852	0.260	0.582	0.350	1.945	0.898	0.429	0.763	0.543	2.409	0.910	0.399	0.777	0.562	2.650
Training setting (iv)	Ours	**0.890**	**0.315**	**0.601**	**0.434**	**2.354**	**0.913**	**0.542**	**0.757**	**0.623**	**3.086**	0.897	0.406	0.744	0.510	2.567
	Attention module	0.870	0.273	0.577	0.380	2.077	0.878	0.479	0.686	0.478	2.060	0.877	0.379	0.685	0.411	1.899

* Non-deep learning model.

Table 5. **Quantitative results on DHF1K, Hollywood2, and UCF sports** datasets. The best scores are marked in **bold**. Training settings (§5.1) for video saliency datasets: (i) DHF1K, (ii) Hollywood-2, (iii) UCF sports, and (iv) DHF1K+Hollywood-2+UCF sports.

tion module, which makes our model explicitly learn static and dynamic saliency representations in CNN and LSTM separately. Our model even does not use any optical flow algorithm (different with Fang *et al.* [13], Two-stream [2]). This significantly improves the applicability of our model and demonstrates the effectiveness of our training protocol that leveraging both static and dynamic stimuli.

Performance on Hollywood-2. We further test our model on Hollywood-2 dataset, where the testing sets comprises 884 video sequences. The results are summarized in Table 5. Again, our model consistently significantly higher than other methods across various metrics. Besides, when we go insight into the performance with training settings, the performance would increase with increasing amount of training data. This suggests that the large-scale training data volume is important for the performance of neural network.

Performance on UCF sports. With the test set (47 video sequences) of UCF sports dataset, we again observe the proposed model provides consistently good results, compared to related state-of-the-art (see Table 5). Interestingly, we find that, with small amount of training data (training setting (iii), 103 video stimuli from UCF sports dataset), the proposed model achieves a very high performance, even better than the model (*Our, training setting (iv)*) trained with large-scale data (1.5K video stimuli). This could be explained by lack of diversity in the video training data, as the videos in UCF sports dataset are highly related (with similar scenes and actors) and small scale. This is also consistent with our research for UCF sports which shows that 82.3% fixations are located on the human body area (see § 2.1).

5.3. Analysis

Based on our extensive experiments, we provide more detailed analyses, which would give deeper insights of previous studies and suggest some 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. et al.*, Rudoy *et al.*, Hou *et al.*, Fang *et al.*). This demonstrates the strong learning ability of neural network and the promise of developing neural network in this challenging area.

Non-deep learning models: static *vs* dynamic. An interesting finding is classic dynamic methods (*i.e.*, PQFT, Seo *et al.*, Rudoy *et al.*, Hou *et al.*, Fang *et al.*) did not perform better than their static counterparts: ITTI, GBVS. This is probably due to two reasons. First, the perceptual cues and

Figure 4. **Qualitative results** of our video saliency model on three datasets. Best viewed in color.

Aspects	Variants	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Baseline	training setting (iv) *(1.5K videos+10K images)*	**0.890**	**0.315**	**0.601**	**0.434**	**2.354**
Attention module	w/o attention *(1.5K videos)*	0.847	0.236	0.579	0.306	1.685
	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

Table 6. **Ablation study on DHF1K**. See §5.4 for details.

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.*, DVA, Deep-Net), previous deep learning based dynamic models (*i.e.*, OM-CNN, Two-stream) only obtain slightly better performance (or only competitive). 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 into the inherent difficulties of video saliency prediction and previous models' neglect of utilizing existing rich static saliency data.

5.4. Ablation study

In this section, we offer a more detailed exploration of our proposed approach in several aspects with DHF1K dataset. We verify the effectiveness of the proposed mechanism, and examine the influence of different training protocols. The results are summarized in Table 6.

Effect of attention mechanism. By disabling the attention module, and only training with video stimuli we observe a performance drop (*e.g.*, AUC-J: 0.890→0.847), verifying the effectiveness of attention module and showing that the leverage of static stimuli indeed improves the predication accuracy in dynamic scenes. For exploring the effect of the residual connection in attention module (Eq. 8), we train the model based on Eq. 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 would gain 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 get a baseline: *w/o convLSTM*. We observe a drop of performance; showing that the dynamic information learnt in convLSTM could boost the performance.

6. Discussion and Conclusion

In this work, we presented "Dynamic Human Fixation (DHF1K)", a large-scale carefully designed and systematically collected benchmark dataset to facilitate research in video saliency modeling. To the best of our knowledge, our work is the most comprehensive performance evaluation of video saliency models. DHF1K contains 1K videos, which capture representative instances, diverse contents and various motions, with human eye-tracking annotations.

Further, we proposed a novel deep learning based video saliency model, which encodes a supervised attention mechanism to explicitly capture static saliency information and help LSTM better capture dynamic saliency representations over successive frames. We performed extensive experiments on DHF1K, Hollywood-2, and UCF-sports datasets, and analyzed the performance of our model compared to previous attention models in dynamic scenes. Our experimental results demonstrate that our proposed model outperforms other competitors and is quite efficient.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 3

[2] C. Bak, A. Kocak, E. Erdem, and A. Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE TMM*, 2017. 2, 3, 6, 7

[3] A. Borji. What is a salient object? A dataset and a baseline model for salient object detection. *IEEE TIP*, 24(2):742–756, 2015. 3

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015. 3

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2013. 2, 6

[6] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE TIP*, 22(1):55–69, 2013. 2

[7] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2006. 2

[8] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *ECCV*, 2016. 4

[9] 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 *ICCV*, 2015. 3

[10] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *NIPS*, 2008. 4

[11] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 3

[12] 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 *CVPR*, 2015. 2

[13] Y. Fang, Z. Wang, W. Lin, and Z. Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE TIP*, 23(9):3910–3921, 2014. 2, 6, 7

[14] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*, 2008. 2, 3

[15] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *NIPS*, 2005. 2

[16] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1):185–198, 2010. 2, 6, 7

[17] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic. Eye-tracking database for a set of standard video sequences. *IEEE TIP*, 21(2):898–903, 2012. 2

[18] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2007. 2, 6, 7

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[20] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan. How many bits does it take for a stimulus to be salient? In *CVPR*, 2015. 2, 6, 7

[21] Q. Hou, M.-M. Cheng, X. Hu, Z. Tu, and A. Borji. Deeply supervised salient object detection with short connections. In *CVPR*, 2017. 3

[22] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 2

[23] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 2008. 2, 6, 7

[24] X. Huang, C. Shen, X. Boix, and Q. Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015. 2, 5, 6, 7

[25] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE TIP*, 13(10):1304–1318, 2004. 2

[26] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005. 3, 5

[27] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 2

[28] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 2, 6, 7

[29] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. In *CVPR*, 2016. 2

[30] L. Jiang, M. Xu, and Z. Wang. Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM. *arXiv preprint arXiv:1709.06316*, 2017. 2, 3, 6, 7

[31] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *CVPR*, 2015. 1, 2, 6

[32] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 1

[33] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009. 2

[34] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[35] S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE TIP*, 2017. 2

[36] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE TPAMI*, 28(5):802–817, 2006. 2

[37] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo. Dynamic whitening saliency. *IEEE TPAMI*, 39(5):893–907, 2017. 2, 6, 7

[38] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 3

[39] N. Liu, J. Han, T. Liu, and X. Li. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE TNNLS*, 2016. 2

[40] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR*, 2007. 3

[41] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE TPAMI*, 32(1):171–177, 2010. 2, 3

[42] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2, 6

[43] S. Mathe and C. Sminchisescu. Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE TPAMI*, 37(7):1408–1424, 2015. 2

[44] 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*, 3(1):5–24, 2011. 2, 3

[45] B. C. Motter. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience*, 14(4):2178–2189, 1994. 4

[46] H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016. 2

[47] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016. 2, 6, 7

[48] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP*, 2016. 3

[49] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2, 6

[50] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, 2013. 2, 6, 7

[51] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, 2015. 3

[52] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009. 2, 3, 6, 7

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[54] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014. 2

[55] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, 2017. 3, 5

[56] W. Wang and J. Shen. Deep cropping via attention box prediction and aesthetics assessment. In *ICCV*, 2017. 2

[57] W. Wang and J. Shen. Deep visual attention prediction. *IEEE TIP*, 27(5):2368–2378, 2018. 2, 6, 7

[58] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015. 3

[59] W. Wang, J. Shen, and L. Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE TIP*, 24(11):4185–4196, 2015. 2

[60] W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 2018. 3

[61] W. Wang, J. Shen, L. Shao, and F. Porikli. Correspondence driven saliency transfer. *IEEE TIP*, 25(11):5025–5034, 2016. 3

[62] W. Wang, J. Shen, R. Yang, and F. Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2018. 3

[63] W. Wang, J. Shen, Y. Yu, and K.-L. Ma. Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE TVCG*, 23(8):2014–2027, 2017. 2

[64] 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*, 2(4):326–343, 2010. 5

[65] J. M. Wolfe. Guided search 4.0. *Integrated models of cognitive systems*, pages 99–119, 2007. 3

[66] S. Xingjian, 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 *NIPS*, 2015. 2, 4

[67] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 3

[68] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[69] 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*, 8(7):32–32, 2008. 2