Going from Image to Video Saliency: Augmenting Image Salience with Dynamic Attentional Push

Siavash Gorji James J. Clark
Centre for Intelligent Machines, Department of Electrical and Computer Engineering, McGill University
Montreal, Quebec, Canada
siagorji@cim.mcgill.ca clark@cim.mcgill.ca

Abstract

We present a novel method to incorporate the recent advent in static saliency models to predict the saliency in videos. Our model augments the static saliency models with the Attentional Push effect of the photographer and the scene actors in a shared attention setting. We demonstrate that not only it is imperative to use static Attentional Push cues, noticeable performance improvement is achievable by learning the time-varying nature of Attentional Push. We propose a multi-stream Convolutional Long Short-Term Memory network (ConvLSTM) structure which augments state-of-the-art in static saliency models with dynamic Attentional Push. Our network contains four pathways, a saliency pathway and three Attentional Push pathways. The multi-pathway structure is followed by an augmenting convnet that learns to combine the complementary and time-varying outputs of the ConvLSTMs by minimizing the relative entropy between the augmented saliency and viewers fixation patterns on videos. We evaluate our model by comparing the performance of several augmented static saliency models with state-of-the-art in spatiotemporal saliency on three largest dynamic eye tracking datasets, HOLLYWOOD2, UCF-Sport and DIEM. Experimental results illustrates that solid performance gain is achievable using the proposed methodology.

1. Introduction

Visual attention is a temporal selection mechanism in which a subset of available sensory information is chosen for further processing in the human visual system. Visual attention tracking- determining where, and to what, people are paying attention while viewing static photographs or while watching videos and cinematic movies- has attracted much interest recently. Despite the many applications of computational visual attention models for dynamic stimuli such as visual surveillance [3], human-robot interaction [18], video compression [19], and advertising [55], the majority of the existing models focus on static images and spatiotemporal visual attention models are a relatively unexplored problem.

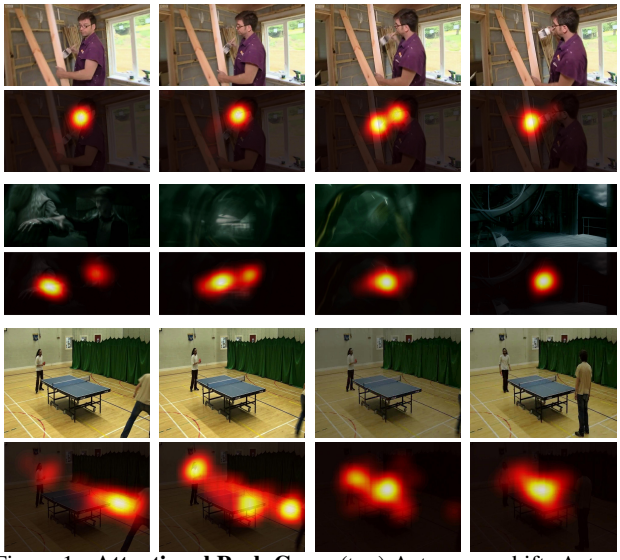

Figure 1: **Attentional Push Cues:** (top) Actor gaze shift: Actors gaze dynamically directs the viewers attention; (middle) Rapid scene changes: The viewers attention is pushed to the center after rapid scene changes; (bottom) Bounce of Attention: An attended actor is moved out of the video frame and pushes the viewers attention to the center.

Almost all computational visual attention models are based on Treisman and Gelades feature integration theory [61] and the Koch and Ullmans [35] feed-forward neural model and are inspired by the pioneering work of Itti et al. [28] where early visual features across multiple scales are linearly combined into a static saliency map. Traditional spatiotemporal saliency models have also benefited from employing early visual features or other static handcrafted features, along with various motion-based features to capture the spatiotemporal nature of dynamic stimuli. The addition of temporal dimension in videos makes dy-

namic visual attention modeling a much more challenging task. Since the added dimension not only requires significantly more data processing, it also needs the computational model to effectively fuse the static and the dynamic features, for even a non-salient (in the static sense) region might have attention allocated to it due to different motion direction. In addition, while the viewing duration of a video frame is limited to a fraction of a second, static images can be viewed leisurely, which complicates moving from image to video saliency.

The recent publication of large-scale eye movement datasets (SALICON [29] and iSUN [66]) has enabled the static visual attention models to benefit from more advanced learning-based techniques, namely deep convolutional neural networks (convnets). The resulting performance gain of the static convnet-based models has been to the extent that newer models achieve only marginal improvements over state-of-the-art (the MIT saliency benchmark [6] and [30]). However, these advancements are yet to be employed by the spatiotemporal saliency models, many of which only consider simple motion cues and are mere straightforward extension of static models (see [32] for a recent review). To the best of our knowledge, the only recent convnet-based spatiotemporal saliency models are: the CMASS method [45], in which shallow neural nets are trained to fuse static hand-crafted features with dense optical flow fields; [10] where a five-layer convnets is trained on RGB color planes and residual motion for each video frame; and the recent work in [39], where RGB color planes, dense optical flow map, depth map and the previous saliency map are fed to a seven-layered encoder-decoder structure. All three models employ very short-term and fixed temporal information, obtained from every two consecutive frames, and do not take into account longer temporal correlations between video frames.

In addition to the relative lack of research and the short-term temporal span of spatiotemporal visual attention models, recent research has shown that even state-of-the-art in static models, including both traditional hand-crafted features and data-driven convnets, suffer from inability to exploit semantic scene information [57], [5], [8] and [21]. For instance, the effect of the gaze direction of the scene actors on the viewers attention has been studied well, e.g. [53], [37], [9], [60], [4], [8], and has been used as a high-level image semantic in recent static models [47], [50], [21]. Specifically, [21] introduced the idea that by considering the scene actors as active and the viewers as passive participants in a shared attention setting, it becomes possible to augment static saliency models with the gaze direction of the scene actors. The model in [21] formulated the manipulating effect of the actors upon the viewers attention as an *Attentional Push* effect, in which an abstract scene information, i.e. actors gaze here, is used to enhance the saliency of some other image region, i.e. the gazed-at region. The Attentional Push effect is important in the sense that nearly all traditional hand-crafted features and data-driven convnets are restricted to use a local neighborhood of image regions for their power to attract the viewers' attention and employing attentional cues that push the viewers attention can greatly benefit the current models. In addition, while effective in augmenting static visual attention models, the Attentional Push effect becomes stronger in dynamic situations, for the viewer is in a more immersive shared attention setting and is more likely to be affected by Attentional Push.

The model in [21] is limited to a single Attentional Push cue, yet, there are other such cues arising from the literature. One of the most prominent of these is the central bias effect, which have been explicitly infused in many modern static visual attention models [13], [31], [36], [62] and [64]. The fact that even deep convnet-based saliency models such as [36] and [13], which are based on the VGG-16 [58] and the ResNet-50 [24] networks, need to explicitly combine central bias maps with deep features is evidence to the fact that even a seemingly straightforward Attentional Push cue such as central bias cannot be learned with the location invariance feature of convnets. However, the central bias effect can be seamlessly integrated in the shared attention setting, by treating the photographer as an active participant which tries to put semantically salient elements in the center of the frame and thus, pushes the viewers attention. Although the center bias effect decreases with dynamic stimuli [42], its dynamic counterpart, i.e. abrupt scene changes, similarly affect the viewers attention as assessed in [48]. In addition, [59] shows the bounce of the viewers attention back to the center of the screen when tracking an actor which moves off the screen to one side.

In this work we show that not only it is imperative to incorporate Attentional Push in spatiotemporal models, but also noticeable performance improvement is achievable by learning its time-varying effect on the viewers attention in social scenes (everyday scenes depicting human activities). We design a novel spatiotemporal saliency augmentation model which benefits from the recent advent in static saliency to estimate video saliency. Here, we extend the model in [21] by including the photographer in the shared attention setting and augment state-of-the-art in static saliency with dynamic Attentional Push. We propose an end-to-end trainable multi-stream Convolutional Long Short-Term Memory network (ConvLSTM) structure. Our network contains four pathways, a saliency pathway and three Attentional Push pathways, i.e. actors gaze, attentional bounce and abrupt scene changes as shown in 1. The saliency pathway embeds static saliency models and captures the temporal dependencies between consecutive video frames by sequentially analyzing the static saliency maps in the ConvLSTM recurrent mechanism. The first Attentional

Push pathway contains a deep convnet which learns to follow the actors gaze on a 2-D spatial grid. This is then processed by a ConvLSTM to capture the dynamic influence of the actors gaze on viewers attention. In Section 4.3, we report the performance after removing the recurrent structure and show that although static Attentional Push is to some extent effective for dynamic stimuli, solid performance improvement is achievable by employing the dynamic nature of Attentional Push with the recurrent mechanism. The second and the third Attentional Push pathways are responsible to infuse 2-D Gaussian center bias priors upon the detection of attentional bounce and abrupt scene changes. For each case, the Gaussian priors are fed to a ConvLSTM which learns their temporal effect. The multi-pathway structure is followed by an augmenting convnet that combines the outputs of the four ConvLSTM and generates augmented saliency for each video frame. For training, validating and performance evaluation of the proposed model, we use the largest datasets available for video saliency, i.e. HOLLYWOOD2 [43], UCF-Sport [43] and DIEM [44] datasets. Partial annotations for the scene actors head and gaze location are provided and used for training and validation (see Section 4.1).

The contribution of this work is threefold: First, we propose a novel spatiotemporal visual attention model that incorporates state-of-the-art in static saliency and learns long-term temporal dependencies to estimate video saliency. Second, we expand the notion of Attentional Push to dynamic stimuli and show its effectiveness in augmenting static saliency in dynamic scenes. Third, we provide comprehensive experimental evaluation on publicly available video saliency datasets which demonstrates significant improvement in predicting viewers fixation patterns on videos containing human activities. The rest of this paper is organized as follows. Section 2 presents related work. We explain the structure and the training scheme for the proposed model in Section 3. Section 4 outlines the experiments and Section 5 concludes the paper.

2. Related work

We describe closely related work on static and spatiotemporal saliency models and saliency models benefiting from gaze following as a subcomponent.

Video Saliency: Most existing spatiotemporal saliency models are based on adding various motion cues to the existing static hand-crafted features in the literature. Among these, some are based on probabilistic modeling while others use various spectral domain transformation for the feature integration stage. An early attempt was proposed by Itti and Baldi [27] where motion energy is used along with orientation, color and intensity contrast as static features. In [22], intensity, color and motion features are combined based on their spectral phase. Similarly, in [16], intensity, color, texture and motion features are extracted and combined based on the discrete cosine transform differences while [14] uses two spatiotemporal Fourier transform to compute video saliency. In [41], a dynamic center-surround model based on the KullbackLeibler (KL) divergence between dynamic patches is proposed. The model in [26] uses incremental coding length to maximize the entropy gain of features on each frame and models the temporal correlation among consecutive frame as a Laplacian distribution. In [15], spatial and temporal dissimilarity (based on motion vectors) are linearly combined. In [49], the difference between the optical flow and accumulated flow map is linearly combined with low-level static features.

There are also spatiotemporal models based on hand-crafted features which use various learning algorithm for the feature integration stage. In [17], static features such as color, intensity and texture are combined with optical flow using uncertainty weighing. In [43] , optical flow-based temporal HoG and MBH descriptor are calculated and their bag of words representation are used to train a multiple kernel learning model. In Rudoy et al. [56], static candidate locations and motion candidates are employed by a random forest regressor. In [33], motion features, based on the number of bits needed to encode a video patch by an optimal encoder, is used to train a Markov random field. Similarly in [65], a video coding feature is used to train a support vector machine for video saliency. The CMASS method [45] uses three-layered fully connected neural nets to fuse static features, ranging from color channels to existing saliency models, with dense optical flow fields. And finally, spatiotemporal models based on feature-learning includes [10], where a five-layer convnets is trained on RGB color planes and residual motion for each video frame and the recent work in [39], which uses RGB color planes, dense optical flow map, depth map and the previous saliency map are fed to a seven-layered encoder-decoder structure.

Static Saliency: The recent advancements in static saliency are mostly benefited from the advent of deep neural networks. The eDN model [62] uses convnet-based feature extractors and linear SVM classifier. Similarly, the model in [40] uses three convnets, each trained for a specific scale, are followed by two fully connected layers. Other models usually benefit from transfer learning in their convnets. More recent models use transfer learning and fine-tune the state-of-the-art models in object recognition. Namely, DeepGaze [38] uses pre-trained AlexNet, SALICON [29] benefits from two pre-trained convnets, DeepFix [36] and ML-Net [11] are based on the pre-trained VGG network. The recent model in [46] contains ten convolutional layers with the first three initialized using the VGG network.

Gaze following: Parks et al. [47] proposed a static attention tracking model which predicts whether the next fixation

Figure 2: **Dynamics of Attentional Push:** Viewers eye fixation pattern after the actor changes his gaze direction. From left to right: time frame 1, time frame +300ms, time frame +600ms and time frame +1500ms.

is gaze related or being saliency driven using a two-state Markov chain. Our model is inspired by the Attentional Push model in [21], which augments static saliency models with the actors gaze for static scenes. While the model in [21] only uses a single Attentional Push cue and is only applicable for static images, in this work, we extend the Attentional Push notion to augment static saliency models to deal with dynamic stimuli. Similarly, Recasens et al. [50] proposed a two-stream convnet to learn the gazed-at object in a scene and recently, they extend the model in [50] to estimate the gaze-at object on videos [51]. Although related, there is a major distinction between saliency prediction and gaze following. While the model in [50] focuses on estimating the gazed-at objects in a scene, from the point of view of the scene actor, our model learns the impact of the actors gaze upon the viewers attention. Although there are cases in which the actors gaze pushes the viewers attention to the gazed-at object(s), this does not hold in all circumstances. Consider situations in which there are multiple objects in the actors gaze direction. From the point of view of [50], the actor is looking at one the objects. From our point of view, the viewers attention is pushed to all of these potential foci of attention with some uncertainty. This is why in our model we limit the input of the gaze following pathway to a cropped region around the actors face and do not feed the whole image content to it. This enables our model to learn and benefit from the manipulating effect of the actors gaze direction on viewers attention. Similarly, when an actor is looking at something that fall outside the current video frame, the model in [51] looks for possible attended objects on separate video frames, while our model learns the manipulating effect of the actors gaze direction both on the same frame, and by using the recurrent structure, consecutive next frames.

3. Network Architecture

While being effective in static scenes, the Attentional Push effect becomes stronger in dynamic situations where the viewer is in a more immersive shared attention setting and is more likely to be affected by the scene actors. Inspired by [21], our attention augmentation model is based on a shared attention setting, in which the viewer, the scene actors and the photographer are all participant in the activity occurring in the scene. While the viewer has no control over what is going on in the scene, the attentional state of the scene actors and the photographer can nonetheless affect the viewer attention. We explicitly model the manipulating effect of the scene actors and the photographer via time-varying Attentional Push maps. In this work, we use three Attentional Push cues, i.e. actors gaze, attentional bounce and abrupt scene changes and combine them with static saliency to estimate the fixation patterns on videos.

Figure 2 shows a time-sampled video frame sequence of a scene where the actor changes his gaze direction. Let us consider the video sequence as separate static images first. Since the contents of the images are similar, we can expect that in a static setting, the viewers fixation patterns would be nearly identical and we can expect the Attentional Push effect of the actors gaze to similarly influence the viewers attention in all three images. However, it is clear that in the actual dynamic setting, the Attentional Push effect varies over time. While being the strongest attentional cue after the actor gaze shift, it becomes less influential and the viewers fixation patterns diverge to other stimuli during the following video frames. This inherent dynamic nature of the Attentional Push effect requires the attention augmentation model to either learn time-varying Attentional Push maps or use a non-static augmentation procedure. Our proposed methodology learns the time-varying manipulating effect of Attentional Push using recurrent mechanisms. As shown in Figure 3, we employ a multi-pathway structure and employ Convolutional LSTM modules to learn the dynamics of each Attentional Push cue. In addition, to benefit from the strong temporal correlation of the fixation patterns in consecutive video frames, a ConvLSTM cell is also used in the saliency pathway. This ensure the propagation of previous information throughout the model. In the following sections, we describe each subsystem and the training procedure of the proposed methodology.

3.1. Saliency Pathway

The saliency pathway embeds state-of-the-art static saliency models. Given a video frame $I(t) \in \mathbb{R}^{cols \times rows \times 3}$ at time t, the static saliency $s_{static}(t) \in \mathbb{R}^{cols \times rows \times 1}$ is computed and fed to a ConvLSTMs module. ConvLSTMs are variants of the LSTM [25] where convolutional operations are used instead of the original dot products. This not only significantly reduces the number of parameters, but also exploits the underlying local spatial dependencies between nearby pixels.

Let $x(t)$, $h(t)$ and $c(t)$ denote the input, hidden unit and the memory cell of a ConvLSTM module. The update equations of the ConvLSTM module are:

$$i(t) = \sigma(W_{xi} * x(t) + W_{hi} * h_{t-1} + b_i) \qquad (1)$$

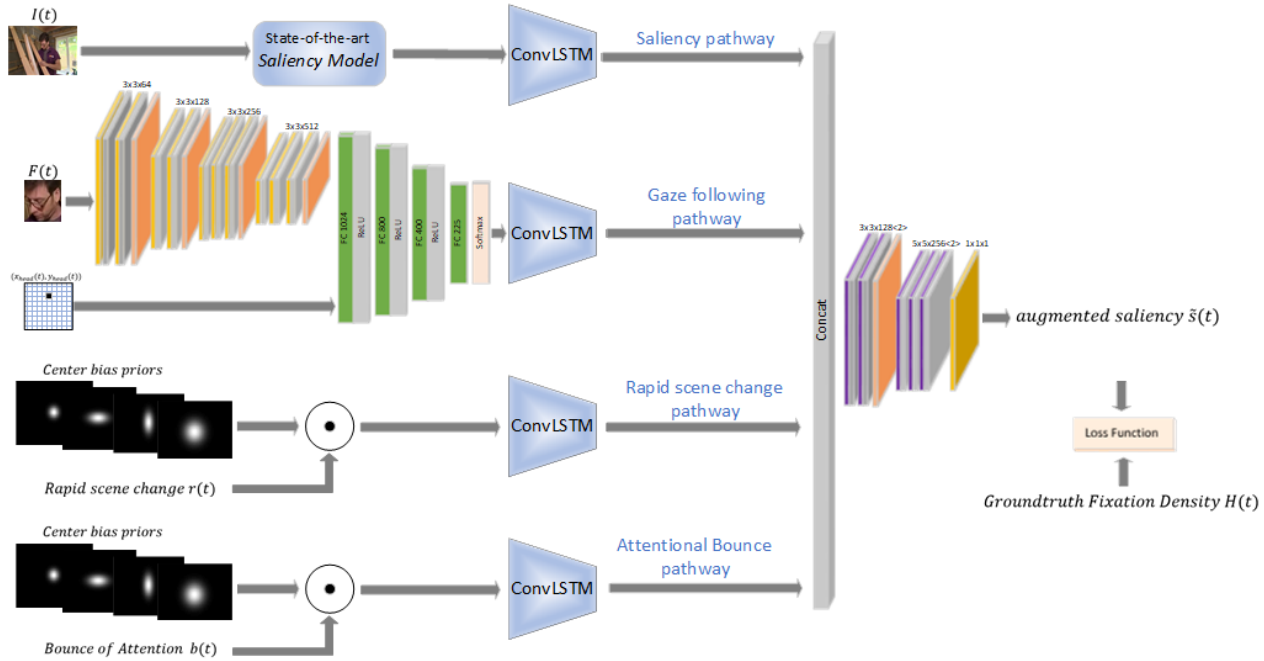

Figure 3: **Network Architecture:** Our network contains four pathways, a saliency pathway and three Attentional Push pathways, gaze following, rapid scene changes and attentional bounce. The network computes the augmented saliency map $\tilde{s}(t)$ for each video frame $I(t)$.

$$f(t) = \sigma(W_{xf} * x(t) + W_{hf} * h_{t-1} + b_f) \quad (2)$$

$$o(t) = \sigma(W_{xo} * x(t) + W_{ho} * h_{t-1} + b_o) \quad (3)$$

$$g(t) = \tanh(W_{xc} * x(t) + W_{hc} * h_{t-1} + b_c) \quad (4)$$

$$c(t) = f(t) \cdot c(t-1) + i(t) \cdot g(t) \quad (5)$$

$$h(t) = o(t) \cdot \tanh(c(t)) \quad (6)$$

where W and b are trainable 2-D convolutional kernels and biases while $i(t)$, $f(t)$ and $o(t)$ denote the input, forget and output gates of the LSTM, respectively. We sequentially pass static saliency maps to the ConvLSTM input by setting $x(t) = s_{static}(t)$, and obtain a refined sequence of time-correlated saliency maps as $s(t) = h(t)$. During training, the saliency ConvLSTM learns to estimate video saliency, by leveraging the temporal correlation between consecutive static saliency maps. This enables our model to benefit from complementary saliency-based and Attentional Push-based information.

3.2. Gaze Following

We formulate the problem of estimating the actors gaze as classifying the gazed-at location to one a pre-defined set of possible locations on an $M \times M$ spatial grid. We used a similar structure as [21] for the static gaze estimation network. The network is based on the VGG-16 model and consists of fourteen weight layers, four of which are fully connected layers, and four max-pool layers, three of which having strides of two. We provide a cropped image region around the actor's head and the location of the head within the $M \times M$ spatial grid. Given the head location of the actor as $(x_{head}(t), y_{head}(t))$, we extract a close-up head region as $F(t)$ and resize it to 224×224 pixels. We provide partial head and gaze location annotations for the training set (see Section 4.1). For testing, a YOLO9000-based face detector [52] is used to locate the actors head. This makes the final feature maps to be of the size of $(28 \times 28 \times 512)$. The first fully connected layer is responsible to project the above into a compressed representation, which is then concatenated with the flattened head location and is fed through the remaining weight layers. A softmax layer is applied to the output of the last layer to obtain a 2-D probability distribution of the actor's gaze over the $M \times M$ spatial grid. The above static Attentional Push map is then fed to a ConvLSTM module to deal with the dynamic aspect of Attentional Push effect. When a new gaze shift occurs, the LSTM learns to use the forget gate $f(t)$ to erase the previous memory and to transfer the current Attentional Push input to the memory cell (Eqn. 5) and therefore, to the output $AP_1(t)$, given by Eqn. 6. On the other hand, during the subsequent video frames for which the input Attentional Push map remains mostly the same, the LSTM learns to apply temporal inhibition of the current input.

3.3. Attentional Bounce and Rapid Scene Change

Bounce of attention occurs when an attended scene actor moves off the screen to one side. As shown in Figure 1, this pushes the viewers attention to the center of

the screen. To incorporate the bounce of attention and rapid scene change, we use a set of 2-D Gaussian functions with diagonal covariance matrices. Similar to [36], we use 16 Gaussian blobs with fixed horizontal and vertical variance as static Attentional Push maps. For each video frame, a binary map is generated based on the detection of bounce of attention $b(t)$ and rapid scene change $r(t)$, where $r(t), b(t) \in \mathbb{R}^{cols \times rows}$, and $\{r_{ij}\}, \{b_{ij}\} \in \{0, 1\}$. Using element-wise multiplication, these signal are used as gates to control the corresponding ConvLSTM modules. Upon detection of the bounce of attention or rapid scene change, the corresponding signal is set which allows the Gaussian priors to be fed to the ConvLSTM module. The LSTM learns to forget the previous memory and to transfer the input Attentional Push maps to its hidden state and therefore, its output ($AP_2(t)$ and $AP_3(t)$ in Figure 3). When the detection signals go back to zero during the subsequent video frames, zero-filled maps are fed to the LSTM instead, which learns to apply temporal inhibition on the subsequent output frames. To detect rapid scene changes we adopt the method in [34] which is based on comparing the edge strength and orientation of consecutive video frames. To detect bounce of attention, we adopted the tracking method in [63] on the head location data.

3.4. Augmented Saliency

To fuse the saliency and the Attentional Push pathways, we use a set of trainable dilated convolutional layers. Having strides of larger than one, effectively increases the receptive field of each convolutional kernel without increasing the network parameters. The last convolutional layer has a (1×1) kernel which effectively maps the deep features of the previous layer into the augmented saliency map, $\tilde{s}(t)$. The augmenting convnet is trained to learn an optimal combination strategy to fuse the complementary information given by the saliency and Attentional Push ConvLSTMs.

4. Evaluation and Comparison

4.1. Datasets

We use the three largest video eye tracking datasets to train, validate and test the performance of proposed methodology which are summarized in Table 1.

DIEM is a widely used dataset, containing 84 videos and free-viewing fixation data from 50 subjects. The dataset contains videos from various categories and a wide range of duration (20 to more than 200 seconds). We use 40 videos (more than 104k frames) containing human activities from the DIEM dataset, ranging from movie trailers, news segments, advertisements and sport scenes. We use 30 videos for training, 5 for validation and 5 for performance evaluation. We provide partial head location and gaze annotations for 8 training videos (12k frames) which are used during the fine-tuning of the whole model.

HOLLYWOOD2 is the largest dynamic eye tracking dataset containing 823 training and 884 validation sequences, with free-viewing fixation data for 3 subjects (we only used the data under the free-viewing condition). The videos in this dataset are short video sequences from a set of 69 Hollywood movies, containing 12 different human action classes, ranging from answering phone, eating, driving, running and etc. We use all the training sequences and split the validation set into a 442 validation and 442 test sequences. We also provide partial head location and gaze annotations on 35 training videos (11k frames) which are used during the fine-tuning of the whole model.

UCF-Sports dataset contains 150 videos on 9 sports action classes with an average duration of 6.39 seconds. We divide the videos of this dataset onto a training set containing 100 videos, a validation set with 10 videos and a test set with 40 videos. We provide partial head location and gaze annotations for 40 training videos (2500 frames) which are used during the fine-tuning of the whole model.

In addition, we use the large-scale static gaze following dataset, GazeFollow [50], for pre-training the gaze-follow convnet, as suggested in [21].

4.2. Evaluation protocol

Static saliency models: We evaluate the performance of the proposed model with several state-of-the-art in spatiotemporal saliency models. To illustrate the effectiveness of dynamic Attentional Push in augmenting static saliency models, we use several neural network-based and traditional static saliency models and train and test the performance of the network in Figure 3. We use four neural network-based, i.e. eDN [62], ML-Net [12], SalNet [46] and SAM-ResNet [13], and two best-performing traditional static saliency models, BMS [67] and RARE [54]. For evaluation, we report the performance of the models using three popular evaluation metrics: the Area Under the ROC Curve (AUC), the Normalized Scan-path Saliency (NSS), and the Correlation Coefficient (CC) to ensure that the main qualitative conclusions are independent of the choice of metric. We use the implementation of the evaluation scores from [7].

Training: To the best of our knowledge, an eye tracking database for video sequences containing the actors gaze information is yet to be developed. As noted in 4.1, we provide head and gaze annotations for a subset of the training sets (25k frames), which constitutes a small portion of the available data. Furthermore, if not pre-trained, the attentional bounce and the rapid scene change ConvLSTM modules are likely to diverge during training, given the sparse nature of the corresponding detections. Therefore, we do not proceed by training the whole model in Figure 3, and

Table 1: Summaries of the used datasets. The last three columns indicate the number of videos for each case.

Dataset	Annotations	Viewers	Added annotations	Training	Validation	Test
DIEM	Eye movement	50	Partial Head & gaze location	30	5	5
HOLLYWOOD2	Eye movement	3	Partial Head & gaze location	823	442	442
UCF-Sports	Eye movement	3	Partial Head location	100	10	40
GazeFollow	Head and gaze location	Crowd	-	119125	3018	-

instead pre-train each of the four pathways separately, and then fine tune the model using the annotated portion of the training sets. To pre-train each pathway, we use stochastic gradient descent to minimize the KL divergence between the corresponding ConvLSTM output and the ground truth fixation density map. Given two probability distribution maps $P, Q \in \mathbb{R}^2$, the KL-divergence loss measures the loss of information when P is used to estimate Q and is given by $KL(P,Q) = \Sigma_i Q_i \log(\frac{Q_i}{P_i})$ where i varies over all pixel locations. The kernel parameters of all ConvLSTMs are initialize by the Xavier method [20], and their hidden states and memory cells are initialized to zero.

We pre-train the saliency ConvLSTM using all the training samples listed in Table 1 with a learning rate of 1×10^{-4} and a weight decay of 5×10^{-5}. This way, the saliency ConvLSTM is trained to estimate video saliency, by leveraging the temporal correlation between consecutive video frames. This later enables the augmenting layers to benefit from complementary saliency-based and Attentional Push-based information. We use temporal segments containing 16 consecutive video frames from the training sets. Although the training segments mostly contain more than 100 frames, we use training on shorter video clips as a method of data augmentation. The training stops if the performance saturates on the validation set, to prevent over-fitting.

The gaze-following pathway is pre-trained in two steps. Following [21], the static gaze-following layers are first trained on the GazeFollow dataset. We initialize the convolutional layers with the VGG-16 network while the fully connected layers are randomly initialized by the Xavier method. For this phase of training, stochastic gradient descent is used to minimize the multinomial logistic regression loss between the soft-max output and the ground truth gaze location, with a learning rate of 1×10^{-5} for the fully connected layers and a learning rate of 1×10^{-7} for the convolutional layers. Drop-out and batch normalization are used after each of the fully connected layers to speedup convergence. Then, we train the gaze-following layers and the corresponding ConvLSTM by minimizing the error on the annotated subset of the training set. Here, we set the learning rate of the ConvLSTM to 1×10^{-5} while the learning rate of the static gaze-following layers are set to 1×10^{-7}. This enables the gaze-following ConvLSTM to learn the temporal dynamics of the Attentional Push map in estimating dynamic fixation patterns. To pre-train the attentional bounce and the rapid scene change ConvLSTMs, we first generate the corresponding detection signals $r(t)$ and $b(t)$ for the entire training set in Table 1 and select temporal segments containing 16 consecutive video frames around each positive detection. Given the smaller number of training instances, we also use an overlap of 10 frames in cutting the video clips and train the ConvLSTMs with learning rate of 1×10^{-6}, a weight decay of 5×10^{-5} and a dropout rate of 0.25.

After pre-training, we then use the annotated portion of the training set to fine-tune the whole model. The augmenting convolutional layers are randomly initialized with the Xavier method and are trained by back propagating the KL divergence loss between the augmented saliency maps and the ground truth fixation densities with a learning rate 1×10^{-5}. During the fine-tuning stage, we set the learning rate of the pre-trained modules to 1×10^{-7}. We use the validation performance to stop the training. A YOLO9000-based face detector [52] is used during the validation and for performance evaluation.

4.3. Results

In this section, we compare the accuracy of the augmented saliency models in predicting video saliency with three state-of-the-art in spatiotemporal saliency models, OBDL [33] , Rudoy [56] and PQFT [23]. Table 2 compares the prediction performance. The results clearly show that the augmented saliency consistently improves upon the static saliency models and achieve considerable performance gain over spatiotemporal saliency models on all three test sets. The results indicated that the augmented eDN and augmented SAM-ResNet outperform all other models with a significant margin. Interestingly, although the Rudoy model outperform four of the static saliency models, including the convnet-based ML-Net and SalNet, all the augmented saliency models achieve considerable gain over the Rudoy model showing that not only it is possible to benefit from the recent static saliency models in dynamic scenes, augmenting them with the dynamic Attentional Push maps results in solid performance improvement over the spatiotemporal models.

We perform ablation analysis to assess the relative impact of each component in the augmented saliency. For this, we use the eDN model and train an augmented eDN model, with one or more components of the model in Fig-

Table 2: Average evaluation scores for the augmented saliency vs. static and spatiotemporal saliency models on the DIEM, HOLLYWOOD2 and UCF-Sports test sets.

	DIEM			HOLLYWOOD2			UCF-Sports		
	AUC	NSS	CC	AUC	NSS	CC	AUC	NSS	CC
ML-Net [12]	0.67	0.46	0.13	0.73	0.72	0.26	0.69	0.7	0.22
augmented ML-Net	0.82	1.84	0.41	0.84	1.85	0.41	0.83	1.88	0.49
SalNet [46]	0.72	1.31	0.26	0.73	1.16	0.31	0.70	0.87	0.21
augmented SalNet	0.85	1.94	0.54	0.84	1.79	0.43	0.84	1.72	0.42
SAM-ResNet [13]	0.88	1.98	0.43	0.87	1.96	0.46	0.89	2.01	0.49
augmented SAM-ResNet	0.91	2.34	0.54	0.91	2.29	0.55	0.92	2.31	0.58
eDN [62]	0.88	1.43	0.32	0.87	1.53	0.33	0.88	1.44	0.33
augmented eDN	0.90	2.21	0.42	0.90	2.11	0.49	0.90	2.15	0.49
RARE [54]	0.75	0.54	0.08	0.76	0.68	0.14	0.78	0.69	0.16
augmented RARE	0.84	1.16	0.26	0.83	1.32	0.27	0.85	1.19	0.35
BMS [67]	0.77	1.28	0.28	0.76	1.08	0.26	0.77	1.15	0.17
augmented BMS	0.85	1.66	0.35	0.85	1.68	0.36	0.84	1.55	0.35
Spatiotemporal Models									
OBDL [33]	0.74	1.16	0.26	0.79	1.45	0.32	0.78	1.08	0.30
Rudoy [56]	0.78	1.31	0.36	0.79	1.37	0.33	0.78	1.34	0.34
PQFT [23]	0.70	0.8	0.19	0.70	0.7	0.14	0.7	0.75	0.16
STS [1]	0.88	2.18	0.48				0.82	2.13	0.48
RMDN [2]				0.90	2.64	0.61			

Table 3: Ablation analysis of the proposed methodology. The results are based on eDN saliency and the DIEM test set.

	NSS
Augmented saliency	2.21
No dynamic Attentional Push	1.53
Saliency and gaze following pathways	1.98
No gaze following	1.63
No attentional bounce	2.06
No rapid scene change	2.01
No saliency	1.89
Saliency pathway only	1.57
Static Saliency	1.43

ure 3 disabled at a time. We only report the NSS score for comparison. The first and the last entries in Table 3 are the baseline performance of the eDN and augmented eDN as reported in Table 2. The second row reports the performance by removing all the ConvLSTMs, which reduces the model into the model in [21]. The results indicates that if the dynamic nature of Attentional Push is not employed, the augmented model would perform marginally better compared to the static model. The third entry is the result of augmenting the static saliency model with dynamic gaze following, which achieves considerable performance. Overall, the results indicate that while dynamic gaze following has the strongest effect, other Attentional Push push cues also contribute to the performance of the augmented saliency.

5. Conclusion

We presented a framework which benefits from the recent development in static saliency models in predicting the fixation patterns on videos. Our model extends the notion of Attentional Push and learns the dynamic influence of it upon the viewers attention. Our multi-stream structure could be readily extended to incorporate other abstract attentional cues which cannot be learned either as the results of model restrictions or the limited amount of available training data. We performed extensive experimental tests and found the augmented saliency models to outperform both the static and spatiotemporal saliency models.

References

[1] Ç. Bak, A. Erdem, and E. Erdem. Two-stream convolutional networks for dynamic saliency prediction. *CoRR*, abs/1607.04730, 2016. 8

[2] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent mixture density network for spatiotemporal visual attention. In *International Conference on Learning Representations (ICLR)*, 2017. 8

[3] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, pages 1–11, 2009. 1

[4] A. Borji, D. Parks, and L. Itti. Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision*, 14(13):1–32, 2014. 2

[5] A. Borji, D. N. Sihite, and L. Itti. What stands out in a scene? a study of human explicit saliency judgment. *Vision Research*, 91:62 – 77, 2013. 2

[6] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu. 2

[7] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint*, arXiv:1604.03605, 2016. 6

[8] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *Computer Vision – ECCV 2016: 14th European Conference*, 2016. 2

[9] M. S. Castelhano, M. Wieth, and J. M. Henderson. I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pages 251–262. Springer Berlin Heidelberg, 2007. 2

[10] S. Chaabouni, J. Benois-Pineau, and C. B. Amar. Transfer learning with deep networks for saliency prediction in natural video. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1604–1608, Sept 2016. 2, 3

[11] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493, Dec 2016. 3

[12] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *23rd International Conference on Pattern Recognition (ICPR)*, 2016. 6, 8

[13] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *CoRR*, abs/1611.09571, 2016. 2, 6, 8

[14] X. Cui, Q. Liu, and D. Metaxas. Temporal spectral residual: Fast motion saliency detection. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 617–620, New York, NY, USA, 2009. ACM. 3

[15] L. Duan, T. Xi, S. Cui, H. Qi, and A. C. Bovik. A spatiotemporal weighted dissimilarity-based method for video saliency detection. *Signal Processing: Image Communication*, 38(Supplement C):45 – 56, 2015. Recent Advances in Saliency Models, Applications and Evaluations. 3

[16] 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*, 24(1):27–38, Jan 2014. 3

[17] Y. Fang, Z. Wang, W. Lin, and Z. Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE Transactions on Image Processing*, 23(9):3910–3921, Sept 2014. 3

[18] J. Ferreira and J. Dias. Attentional mechanisms for socially interactive robots- a survey. *Autonomous Mental Development, IEEE Transactions on*, 6(2):110–125, June 2014. 1

[19] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, and F. Alexey. Semiautomatic visual-attention modeling and its application to video compression. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, page 11051109, 2014. 1

[20] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*, 2010. 7

[21] S. Gorji and J. J. Clark. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 2, 4, 5, 6, 7, 8

[22] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. 3

[23] 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*, 19(1):185–198, Jan 2010. 7, 8

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016. 2

[25] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 4

[26] X. Hou and L. Zhang. Dynamic visual attention: searching for coding length increments. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 681–688. Curran Associates, Inc., 2009. 3

[27] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 631–637 vol. 1, June 2005. 3

[28] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998. 1

[29] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 3

[30] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 2

[31] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2106–2113, 2009. 2

[32] Y. Kavak, E. Erdem, and A. Erdem. A comparative study for feature integration strategies in dynamic saliency estimation. *Signal Processing: Image Communication*, 51(Supplement C):13 – 25, 2017. 2

[33] S. H. Khatoonabadi, N. Vasconcelos, I. V. Baji, and Y. Shan. How many bits does it take for a stimulus to be salient? In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, June 2015. 3, 7, 8

[34] Y.-M. Kim, S. W. Choi, and S.-W. Lee. Fast scene change detection using direct feature extraction from mpeg compressed videos. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 3, pages 174–177 vol.3, 2000. 6

[35] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985. 1

[36] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, Sept 2017. 2, 3, 6

[37] G. Kuhn and A. Kingstone. Look away! eyes and arrows engage oculomotor responses automatically. *Attention, Perception and Psychophysics*, 71:314–327, 2009. 2

[38] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint*, arXiv/1411.1045, 2014. 3

[39] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar. Learning gaze transitions from depth to improve video saliency estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[40] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[41] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):171–177, Jan 2010. 3

[42] M. Mancas, V. P. Ferrera, N. Riche, and J. G. Taylor. *From Human Attention to Computational Attention*. Springer, 2016. 2

[43] 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*, 37, 2015. 3

[44] P. K. Mital, T. J. Smith, R. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011. 3

[45] T. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan. Static saliency vs. dynamic saliency: A comparative study. In *Proceedings of the 21st ACM International Conference on Multimedia*, page 987996, 2013. 2, 3

[46] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3, 6, 8

[47] D. Parks, A. Borji, and L. Itti. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research*, 116, Part B:113 – 126, 2015. 2, 3

[48] T. Po-He, C. Ran, C. I. G. M., M. D. P., and I. Laurent. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7):4, 2009. 2

[49] P. Polatsek, W. Benesova, L. Paletta, and R. Perko. Novelty-based spatiotemporal saliency detection for prediction of gaze in egocentric video. *IEEE Signal Processing Letters*, 23(3):394–398, March 2016. 3

[50] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2, 4, 6

[51] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba. Following gaze in video. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4

[52] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. 5, 7

[53] P. Ricciardelli, E. Bricolo, S. M. Aglioti, and L. Chelazzi. My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individuals gaze. *Neuroreport*, 13(17):2259–2264, 2002. 2

[54] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642–658, 2013. 6, 8

[55] R. Rosenholtz, A. Dorai, and R. Freeman. Do predictions of visual perception aid design? *ACM Trans. Appl. Percept.*, 8(2):12:1–12:20, Feb. 2011. 1

[56] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1147–1154, June 2013. 3, 7, 8

[57] J. Shen and L. Itti. Top-down influences on visual attention during listening are modulated by observer sex. *Vision Research*, 65(Supplement C):62 – 76, 2012. 2

[58] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556, 2014. 2

[59] T. J. Smith. The attentional theory of cinematic continuity. *Projections*, 6(1):1–27, 2012. 2

[60] R. Subramanian, V. Yanulevskaya, and N. Sebe. Can computers learn from humans to see better?: Inferring scene semantics from viewers' eye movements. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 33–42. ACM, 2011. 2

[61] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. 1

[62] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2798–2805, June 2014. 2, 3, 6, 8

[63] L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3119–3127, Dec 2015. 6

[64] Z. Wu, L. Su, Q. Huang, B. Wu, J. Li, and G. Li. Video saliency prediction with optimized optical flow and gravity center bias. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2016. 2

[65] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang. Learning to detect video saliency with hevc features. *IEEE Transactions on Image Processing*, 26(1):369–385, Jan 2017. 3

[66] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint*, arXiv:1504.06755, 2015. 2

[67] J. Zhang and S. Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015. 6, 8