Emotional Attention: A Study of Image Sentiment and Visual Attention

Shaojing Fan[1], Zhiqi Shen[1], Ming Jiang[2], Bryan L. Koenig[3], Juan Xu[2], Mohan S. Kankanhalli[1], and Qi Zhao[*2]

[1]National University of Singapore [2]University of Minnesota [3]Southern Utah University

Abstract

Image sentiment influences visual perception. Emotion-eliciting stimuli such as happy faces and poisonous snakes are generally prioritized in human attention. However, little research has evaluated the interrelationships of image sentiment and visual saliency. In this paper, we present the first study to focus on the relation between emotional properties of an image and visual attention. We first create the EMOtional attention dataset (EMOd). It is a diverse set of emotion-eliciting images, and each image has (1) eye-tracking data collected from 16 subjects, (2) intensive image context labels including object contour, object sentiment, object semantic category, and high-level perceptual attributes such as image aesthetics and elicited emotions. We perform extensive analyses on EMOd to identify how image sentiment relates to human attention. We discover an emotion prioritization effect: for our images, emotion-eliciting content attracts human attention strongly, but such advantage diminishes dramatically after initial fixation. Aiming to model the human emotion prioritization computationally, we design a deep neural network for saliency prediction, which includes a novel subnetwork that learns the spatial and semantic context of the image scene. The proposed network outperforms the state-of-the-art on three benchmark datasets, by effectively capturing the relative importance of human attention within an image. The code, models, and dataset are available online at https://nus-sesame.top/emotionalattention/.

1. Introduction

People have a remarkable ability to attend selectively to some regions in a scene [50, 9]. Attention selectively follows low-level image properties (*e.g.*, intensity, color) and semantic-level information [26, 15]. Such properties have been incorporated in computational models that predict visual saliency with impressive performance [34, 2, 70]. These models have been used in applications such as automated image annotation and video surveillance [24, 14].

Figure 1: Example images from EMOtional attention dataset (EMOd), along with emotions that observers indicated as strongly elicited by the images and colormaps visualizing human attention.

Substantial research also finds that the emotional relevance of a stimulus influences human attention [17, 13, 67, 39]. For example, people preferentially attend to *emotional stimuli* (*i.e.*, an object or scene that elicits an emotional response in the observer), such as cute babies or erotic scenes [25, 54]. Although many neuroimaging and behavioral studies have investigated how emotional stimuli affect attention [52, 17, 29], few computer vision studies have—due in part to the lack of an eye-tracking dataset that includes emotional stimuli. Advances have been made regarding semantics and attention [57, 27, 70], but much remains unknown about how image sentiment relates with visual saliency.

In this paper, we present the EMOtional attention dataset (EMOd)—a human-annotated dataset focusing on image sentiment and human attention (see Fig. 1). We perform statistical analyses on EMOd to investigate how emotional

*Corresponding author. E-mail: qzhao@umn.edu

Type	Category	Description	Object No.	Image No.
Directly relate to humans	Face (emotional)	Faces with obvious emotional expressions.	899	422
	Face (neutral)	Faces without obvious emotional expressions.	890	443
	Gazed	Objects gazed upon by a human or animal.	111	92
	Touched	Objects touched by a human or animal.	322	244
Relate to other (nonvisual) human senses	Sound	Objects producing sound (*e.g.*, people talking)	995	667
	Smell	Objects with a scent (*e.g.*, a flower, a cup of coffee).	386	309
	Taste	Food, drink, etc.	104	54
	Touch	Notably tactile objects (*e.g.*, a sharp knife).	664	570
To attract attention or to interact with humans	Text	Digits, letters, words, and sentences.	360	169
	Wachability	Objects made to be viewed (*e.g.*, pictures, traffic signs).	186	78
	Operability	Natural or man-made objects held or used with hands.	689	445
Imply motion	Motion	Moving objects, includes gesturing humans/animals.	955	672

Table 1: Descriptions of semantic attributes of objects labeled in EMOd dataset. The fourth and fifth columns indicate the number of objects in each category, and the number of images containing the specific category of objects, respectively.

content relates to human visual attention. Analyses indicate that emotional content attracts human visual attention strongly and briefly—which we refer to as *emotion prioritization effect*. Building on the emotion prioritization effect, we propose a deep neural network (DNN) that learns the relative importance of the salient regions within an image. That is, it accounts for contextual saliency—saliency regarding both spatial and semantic context of the scene. Our main contributions follow.

1. We provide a novel image dataset (EMOd) featuring image sentiment and visual attention, which is the first to include eye-tracking data as well as extensive annotations about image context—emotions, objects, semantics, and scenes—enabling research on these topics together with attention.
2. We evaluate how image sentiment relates to human attention. We observe the emotion prioritization effect—for our images, emotional content not only attracts human visual attention strongly, but also briefly.
3. We propose a novel DNN model with a subnetwork that is able to encode relative importance of regions/objects within an image. The proposed model outperforms state-of-the-art methods on three benchmark datasets.

2. Related work

Predicting human attention: Substantial research has been done on saliency prediction—using computational models to predict human attention [30, 4]. Early saliency prediction models use pixel-level image attributes, such as contrast, color, orientation, and intensity [34, 41, 23]. An earlier advocate for context-aware saliency is [27], which also focuses on low-level image features. Large gain in saliency prediction has resulted from the recent resurgence of deep neural networks [60, 66, 68, 32, 42, 69, 18, 49], such as SALICON [33], DeepGaze [44], and DeepFix [42]. These DNNs endeavor to learn image contexts and achieve considerable performance. However, being trained on the datasets and learning weights as a whole, they enable few insights about how contextual information of multiple objects within an image relate to human attention. Our work takes into consideration contextual information with a new model design that effectively addresses the emotion prioritization effect in attention allocation within an image.

Attention and emotion: Psychologist have found that human attention generally prioritizes emotional content over non-emotional content [17, 67, 7]. For example, smiling faces, babies, and erotic scenes attract human attention more than emotionally neutral stimuli [22, 70].

Saliency researchers seek to incorporate increasingly higher-level perceptual properties of images [61, 73, 70, 12], and their models have encoded high-level concepts such as faces [59], interacting objects [56], and text [35]. Saliency researchers have not yet attempted to systematically measure or model the relation between emotion and attention. One major reason could be the lack of a proper dataset with both emotional content and eye-tracking data.

Eye-tracking datasets: Two related datasets that we use as benchmarks (see Sec. 5.2) are NUSEF [57] and CAT2000 [5]. NUSEF is 751 emotion-eliciting images that depict mostly faces, nudes, and human actions. CAT2000's training set contains 2000 images of diverse scenes, such as affective images and cartoons. However, these two datasets have limited emotional content and no object labels. Emotion labels are absent in other commonly used eye-tracking datasets (for an overview see [10]). In this paper, we present the first eye-tracking dataset to include images of diverse emotional scenarios, together with extensive image annotations.

Precise measurement of human attention requires customized eye-tracking equipment, making crowdsourcing difficult. Alternative methods for large-scale attention data collection include using webcams and mouse movements [51, 36, 71, 40, 38], but their validity has not been established for emotional images. Seeking maximal validity for our dataset, we use the gold-standard: measuring with eye-

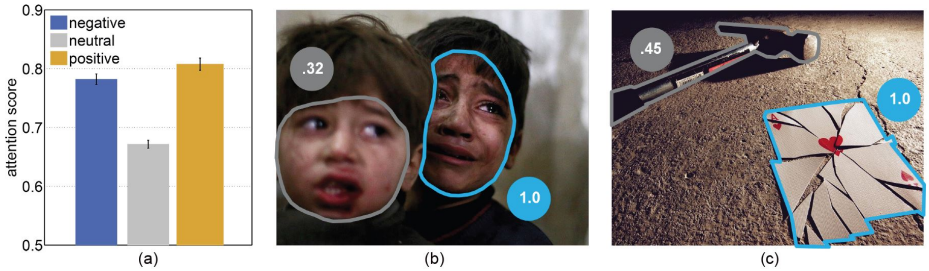

Figure 2: (a) Emotional objects garner more attention than neutral objects. In all figures in this paper, error bars represent standard error. (b, c) Images illustrate how objects in strong emotions (outlined in blue), such as the crying face and broken card, are more salient than neutral/less emotional stimuli (outlined in gray).

tracking equipment in controlled laboratory conditions [21].

3. Construction of EMOd dataset

We constructed EMOd, a new dataset of 1019 emotion-eliciting images, with eye-tracking data and annotations at object and image levels. It is designed for research on visual saliency and image sentiment.

3.1. Image collection

EMOd images were from two sources: (1) 321 emotion-evoking photos selected from the International Affective Picture System (IAPS) [45], and (2) 698 photos collected by the authors using the ontology and attributes in [6, 48] as search terms in an online image search engine (Google Image Search) [1]. We collected the photos to make the dataset more diverse regarding how observers' emotions are evoked, such as the emotion-eliciting objects, activities, and gists.

3.2. Psychophysics study I: eye tracking

Sixteen subjects aged 21 to 35 years old (27.0 ± 4.7) freely observed all EMOd images on a 22-inch LCD monitor. The screen resolution was 1920×1080. The visual angle of the stimuli was about $38.94° \times 29.20°$. Subject eye movements were recorded at 1000Hz using an Eyelink 1000 eye tracker. Each image was presented for 3 seconds, followed by a drift correction that required subjects to fixate in the screen center and press the space bar to continue.

3.3. Psychophysics study II: image annotation

Three paid undergraduate students labeled the following properties of the dominant objects in each image: (1) object contour, (2) object name, (3) sentiment category, selecting from negative, neutral, or positive, and (4) semantic category. We used four types of semantic categories [70]: (i) objects directly relating to humans, (ii) objects relating to nonvisual senses of humans, (iii) objects designed to attract attention or for interaction with humans, and (iv) objects with implied motion. See Table 1 for categories within each type and how many objects and images were coded with each category. Each object could be coded to have one or more categories. For sentiment and semantic labeling, we used only those agreed upon by all three students; objects without unanimous agreement were labeled as "neutral" for sentiment and "other" for semantic category. In total, EMOd has 4302 segmented objects with fine contours, sentiment labels, and semantics labels. The number of positive, neutral, and negative objects are 839, 2429, and 1034, respectively.

We also used online crowd sourcing (Amazon Mechanical Turk (AMT) [8] and campus intranet) to collect perceptions of 33 high-level perceptual attributes such as image aesthetics and elicited emotions (see Fig. 1). For more details on EMOd construction, see the supplementary material.

4. How do sentiments affect human attention?

In this section, we report two findings regarding how emotional properties of images influence human attention. We first explain our analytical methods, then report observations with supporting analyses.

4.1. Definitions and methods

For each image, we compute a *fixation map* by placing at each fixation location a Gaussian distribution with sigma equal one degree of visual angle and then normalizing the map to maximum 1 (a common method in saliency research [46]). Fig. 1 visualizes fixation maps by overlaying colormaps on original images. We define the *attention score* of an object as the maximum fixation-map value that is inside the object's contour. Attention scores thus range between 0 and 1 [22].

The inferential statistical analyses we use—such as univariate analyses of variance (ANOVA), post-hoc Tukey tests, and simple effects analysis—are standard in behavioral and other sciences (for an introduction, see [3]).

4.2. Results

Observation 1 (Emotion prioritization effect): *Emotional* objects attract human attention more than *neutral*

[1] Due to copyright restrictions of the IAPS dataset, all the images shown in this paper are from the author's own collection.

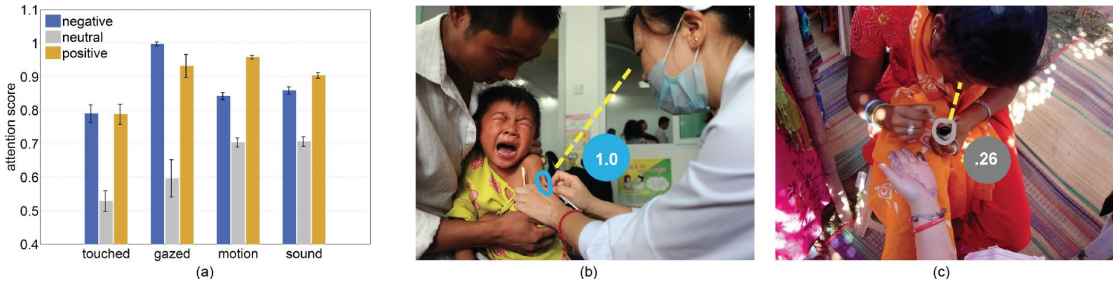

Figure 4: (a) Emotion prioritization is stronger for human-related objects: those being touched, gazed upon, or with motion or sound. (b-c) Examples of gazed-upon objects and their respective attention scores. The emotional gazed-upon object (b) has a higher attention score than the neutral gazed-upon object (c).

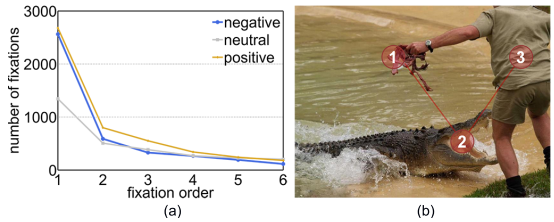

Figure 3: (a) Human observers fixated first on emotional objects more than neutral objects, but the attention prioritization quickly diminishes. (b) Viewers fixated on the emotional objects (*i.e.*, food (1) and crocodile's mouth (2)) before the neutral human body (3).

objects. Furthermore, emotional objects attract attention not only strongly, but also *briefly*—a positive or negative sentiment category strongly increases an object's chance of being attended to at first fixation, but the advantage diminishes quickly during subsequent fixations.

Observation 1 is based on the following analyses. A two-way ANOVA has attention scores of each object as the dependent variable, and sentiment and semantic categories as the independent variables. Attention scores are influenced by both sentiment category ($F(2, 4263) = 21.75, p < .001$[2]) and semantic category ($F(12, 4263) = 4.31, p < .001$). The larger F score of sentiment over semantics (21.75 *v.s.* 4.31) suggests sentiment impacts attention more than semantics. Post hoc Tukey tests indicate that neutral objects have lower attention scores than negative and positive objects (ps[3] $< .001$), and attention scores for negative and positive objects do not significantly differ, $p = .260$ (see Fig. 2).

We also evaluate how the first six fixations are distributed across positive, neutral, and negative objects. We randomly pick an equal number (373) of negative, neutral, and positive objects. We select only from images containing 3 to 6 objects to minimize any effect of image complexity on fixation order. Objects categorized as positive or negative have more fixations than do neutral objects at first fixation, but subsequent fixations show little difference (see Fig. 3). By showing for the first time that attention prioritization diminishes drastically after initial fixation for the EMOd dataset, our findings reveal a more nuanced understanding of the claim that human attention prioritizes emotional stimuli over non-emotional stimuli [17, 67, 7].

Observation 2: The emotion prioritization effect (Observation 1) is stronger for human-related objects than objects unrelated to humans. For example, happy faces are prioritized over neutral faces more than fascinating architecture is over common architecture.

This is indicated by a significant interaction of sentiment category and semantic category, $F(24, 4263) = 3.62, p < .001$, which means that emotion prioritization differs across various combinations of sentiment and semantics. Simple effects analysis shows that the emotion prioritization occurs primarily for semantic categories of "touched", "gazed", "motion", "sound" (see Fig. 4 (a)). Objects being "touched" and "gazed", and objects with "sound" by definition relate to humans. The majority ($\geq 75\%$) of "motion"in EMOd are coded as being on human bodies or human faces, so such objects also relate to people. This suggests that the emotion prioritization effect is stronger on human-related objects. Fig. 4 (b-c) illustrates this interaction using images with gazed-upon objects.

In summary, the emotional properties of images, especially those related to humans, strongly influence visual attention. Building on these findings, we develop a DNN that is adaptive of those emotional properties by using contextual saliency prediction, as described in the next section.

5. Predicting human attention with contextual information

In this section, we design a DNN guided by our psychophysics findings. Experiments on three benchmark datasets demonstrate the superior performance of the proposed DNN, especially when emotion-eliciting objects stand out in a scene.

[2]We report the results of ANOVAs as, "$F(\mathrm{df}_{condition}, \mathrm{df}_{error}) = F$ value, $p = p$ value". If a p value is smaller than the conventional significance level threshold of .05, we reject the null hypothesis of no difference among the means.

[3]Throughout the paper, ps represents the plural form of p.

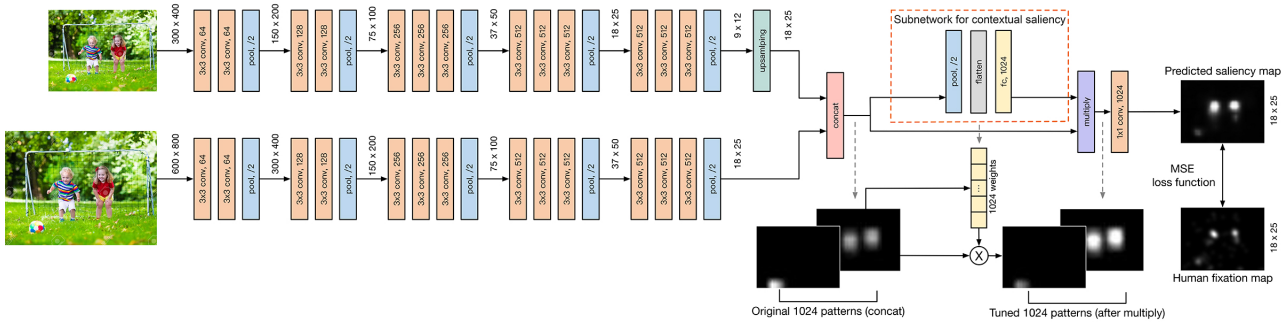

Figure 5: The architecture of the proposed DNN (CASNet). A channel weighting subnetwork (inside the dashed orange rectangle) computes a set of 1024-dimensional feature weights for each image (instead of only one whole image set), to capture the relative importance of the semantic features of a particular image. The gray dashed arrows illustrate how the relative saliency of different regions within an image are modified through the subnetwork.

5.1. Proposed DNN architecture

The proposed DNN architecture is shown in Fig. 5. To address emotion prioritization, we design a channel weighting subnetwork (the orange dashed rectangle) that encodes contextual information, enabling the network to highlight emotion-eliciting objects from the surroundings. Intuitively, by computing a set of 1024-dimensional feature weights for each image (instead of only one whole image set), the subnetwork learns the relative importance of the semantic features of the particular image. Specifically, to compute the weight, we first apply a 2×2 max pooling on the 1024 channels of concatenated feature maps to reduce their dimensionality and spatial variance. We then flatten the output and apply a fully-connected layer to compute a 1024-dimensional vector. Each dimension represents the saliency weight of the corresponding input channel. The fully-connected layer allows the model to learn the relative weights of different objects or regions in a scene based on both their spatial locations and semantic features. Finally, the weights are applied to the input feature by a channel-wise multiplication.

We construct the rest of our network based on a two-stream VGG-16 network architecture. We feed fine-scale images of $800\times600\times3$ pixels to its first stream for extracting relatively high-resolution deep features, while feeding coarser-scale images of $400\times300\times3$ pixels to its second stream for extracting relatively low-resolution deep features. The output of the two network streams are rescaled to the same spatial resolution, and stacked together to form multi-scale deep features of dimension $25\times18\times1024$. Each channel corresponds to an activation map representing a certain visual pattern in the image at different resolutions. We then perform a convolutional layer after the new subnetwork with a 1×1 kernel to reduce the 1024-channel 2D images into a single-channel 2D saliency map of dimension 25×18 pixels. Finally, we resize the saliency map back to the dimension of the original image. The two-stream design is based on SALICON [32], except that we reduce the resolutions of input images from 1600×1200, 800×600 to 800×600, 400×300, and increase the batch size from 1 to 8. We made these changes for better network convergence.

5.2. Experiment settings

Datasets: We test our model on three eye-tracking datasets with emotional content. The first is the EMOd, which includes 1019 emotion-eliciting images. The second is the NUSEF dataset [57], which includes 751 images that depict mostly emotion-eliciting objects and human actions. The third is the training set of CAT2000 [5], which contains 2000 diverse images including emotional, cartoon, social, and so on.

DNN parameters: We initialize the training to the pre-trained parameters for VGG-16 on ImageNet. Mean squared error (MSE) is used as the loss function. The parameters of the DNN are then learned end-to-end on the training images with stochastic gradient descent. The learning rate is 10^{-5} and the batch size is 8. A momentum of 0.9 and a weight decay of 0.0005 are used. We train the model for 30 epochs. Each epoch contains 1250 iterations. We pre-train our network using a mouse contingency based saliency dataset—SALICON [36]. The entire training procedure takes about one day on a single NVIDIA TitanX GPU using Keras with a Tensorflow backend [16, 1].

5.3. Evaluation metrics

We use 9 metrics for comprehensive evaluation. The Area Under the ROC Curve (AUC) [28] treats the saliency map as a binary classifier. We use two variants of AUC: AUC-Judd and AUC-Borji [11], and shuffled-AUC (sAUC) [64] which alleviates the effects of center bias. Although comprehensive and commonly used in the community, AUC by nature is not able to distinguish between cases where models predict different relative importance values for different regions of an image [11, 12, 20], as needed in our study. We further use six similarity metrics to measure the similarity between the

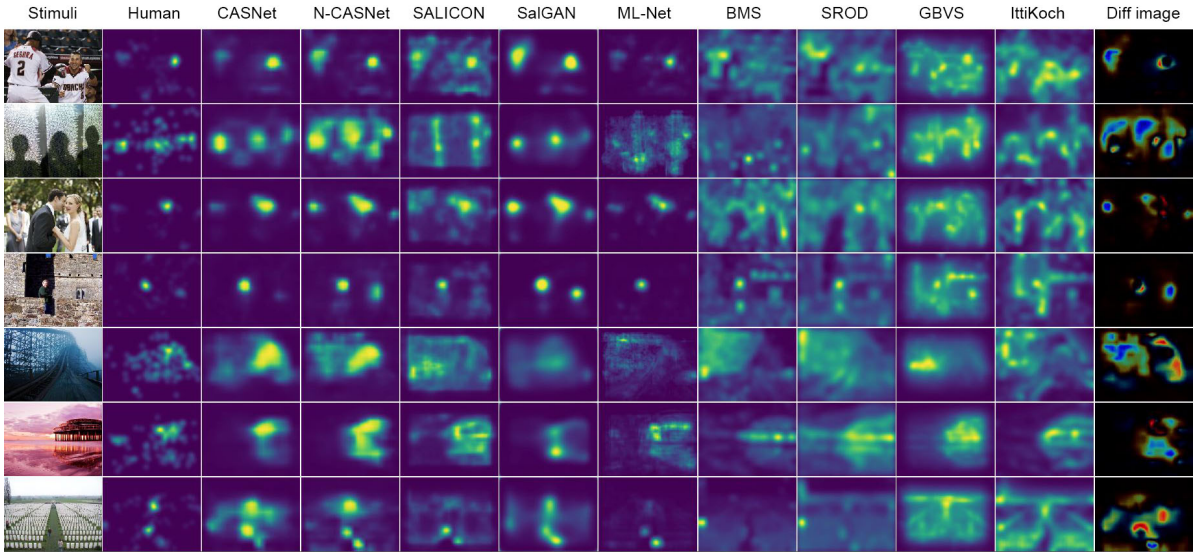

Figure 6: Qualitative results generated by our saliency model in comparison with state-of-the-art methods. Our model (CASNet) outperforms other models in both location and order, by taking into consideration contextual information (*e.g.*, encoding relative importance of objects in the first four rows and highlighting areas of interest in scene images in the last three rows). The last column (Diff image) visualizes the difference between predictions from CASNet and N-CASNet: colors close to orange/red indicate increased saliency after applying the subnetwork for contextual saliency, whereas colors close to blue/green indicate decreased saliency.

Metric	CASNet	N-CASNet	SALICON	SalGAN	ML-Net	BMS	SROD	GBVS	IttiKoch
AUC-Judd	**0.83**	0.82	0.82	**0.83**	0.82	0.77	0.74	0.79	0.73
AUC-Borji	**0.80**	0.79	**0.80**	**0.80**	0.76	0.75	0.73	0.78	0.72
sAUC	**0.78**	0.77	**0.78**	**0.78**	0.74	0.74	0.72	0.75	0.70
NSS	**1.75**	1.61	1.59	1.74	1.74	1.12	0.98	1.18	0.88
IG	**1.58**	1.48	1.45	1.13	1.21	1.02	0.88	1.13	0.88
CC	**0.66**	0.61	0.59	**0.66**	0.62	0.42	0.37	0.47	0.35
SIM	**0.58**	0.55	0.53	**0.58**	0.56	0.45	0.42	0.48	0.43
EMD	**2.66**	3.04	3.02	2.76	2.84	4.06	4.43	3.42	4.20
KL	**5.54**	5.61	5.67	5.83	5.78	5.94	6.04	5.86	6.04

Table 2: Results on the EMOd dataset. The best performance in each metric is highlighted in bold. For all metrics larger values indicate higher performance, except smaller is better for EMD and KL.

saliency map and fixation map, namely Normalized Scanpath Saliency (NSS) [55], Linear Correlation Coefficient (CC) [47], histogram intersection (SIM) [62], the Earth Movers Distance (EMD) [58], the Kullback-Leibler divergence (KL) [37], and Information Gain (IG) [43, 11]. See [11] for an introduction of these metrics.

5.4. Results

Statistical results are reported in Tables 2 – 4. Qualitative results are shown in Figures 6 – 7 .

Comparison with state-of-the-arts models: We report results for our model both with the subnetwork for contextual saliency prediction (*i.e.*, CASNet—Context-Adaptive Saliency Network) and without the subnetwork (*i.e.*, N-CASNet—Not Context-Adaptive Saliency Network). We compare our saliency prediction models with 7 others. Three are state-of-the-art DNN-based models: SALICON[4] [32], SalGAN [53], and ML-Net [18]. Two are non-DNN models with top performance in the non-DNN model category: Boolean Map based Saliency (BMS) [72] and Saliency via Sparse Residual & Outlier Detection (SROD) [63]. Two are classic bottom-up approaches: Graph-Based Visual Saliency (GBVS) [31] and Itti-Koch model (IttiKoch) [34]. These models are top performers on MIT benchmark [10] in their respective categories[5]. To ensure fair comparisons, all DNN-based

[4]We use the code of OpenSALICON (a publicly available implementation of SALICON) [65].

[5]To be fair, we exclude DNN models that use or learn center bias (*e.g.*,, SAM-ResNet [19]). We include as many top performing models as possible, but models/code of some are not publicly available, such as Deep Gaze 2 [44] and DeepFix [42].

Metric	CASNet	N-CASNet	SALICON	SalGAN	ML-Net	BMS	SROD	GBVS	IttiKoch
AUC-Judd	**0.83**	**0.83**	0.82	**0.83**	0.82	0.77	0.74	0.80	0.71
AUC-Borji	0.77	0.77	**0.79**	0.78	0.74	0.75	0.74	0.79	0.70
sAUC	0.75	0.74	**0.76**	0.75	0.71	0.72	0.71	0.74	0.67
NSS	**1.75**	1.67	1.59	1.72	1.66	1.08	0.95	1.21	0.77
IG	**1.35**	1.29	1.24	0.51	0.11	0.67	0.62	0.96	0.56
CC	**0.67**	0.64	0.61	0.66	0.61	0.42	0.37	0.49	0.31
SIM	**0.58**	0.56	0.54	**0.58**	0.55	0.44	0.42	0.48	0.40
EMD	2.75	2.93	3.08	**2.72**	2.91	4.31	4.72	3.68	4.75
KL	**5.37**	5.41	5.56	5.90	6.20	5.84	5.88	5.64	5.92

Table 3: Results on NUSEF dataset. The best performance in each metric is highlighted in bold.

Metric	CASNet	N-CASNet	SALICON	SalGAN	ML-Net	BMS	SROD	GBVS	IttiKoch
AUC-Judd	**0.82**	0.81	0.80	0.81	0.79	0.78	0.77	0.80	0.71
AUC-Borji	0.79	0.77	0.78	**0.80**	0.73	0.77	0.76	0.79	0.70
sAUC	0.76	0.74	0.75	**0.77**	0.70	0.73	0.72	0.75	0.66
NSS	**1.50**	1.36	1.35	1.45	1.31	1.15	1.07	1.24	0.76
IG	**0.46**	0.30	0.27	0.08	0.04	-0.13	-0.11	0.18	-0.25
CC	**0.58**	0.52	0.52	0.56	0.49	0.44	0.41	0.49	0.30
SIM	**0.57**	0.53	0.52	0.53	0.51	0.49	0.48	0.50	0.42
EMD	**2.42**	2.89	2.86	3.21	3.08	3.12	3.31	3.12	3.97
KL	**5.82**	5.93	6.03	6.08	6.08	6.21	6.06	5.99	6.29

Table 4: Results on CAT2000 dataset. The best performance in each metric is highlighted in bold.

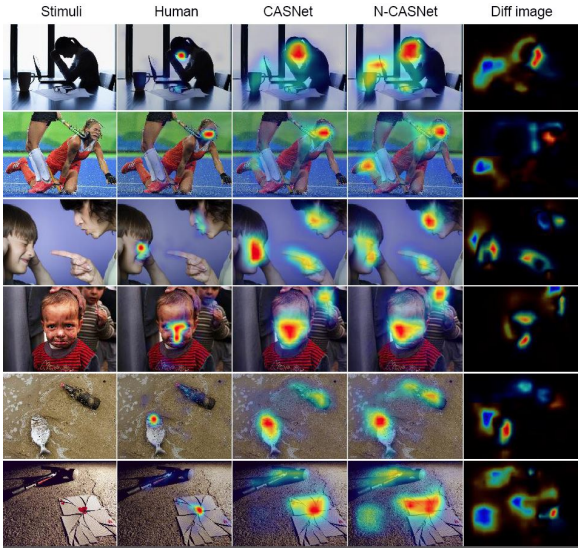

Figure 7: CASNet outperforms N-CASNet for co-occurrence of face (touched) with non-face object (first two rows), e-motional face with neutral face (thrid and fourth rows), and emotional object with neutral object (last two rows). The last column (Diff image) visualizes the difference between predictions from CASNet and N-CASNet.

models are trained on the SALICON dataset to achieve their best possible performance, and all models (including ours) are directly tested on the three benchmark datasets without training/fine-tuning on them.

As shown in Tables 2 – 4, our model with the contextual saliency subnetwork (CASNet) has the best overall performance across datasets, without additional center bias mechanism. CASNet's advantage is greatest on EMOd. This is perhaps because EMOd focuses more than the other datasets on emotional content, and CASNet is most advantageous on emotional images. CASNet consistently outperforms on AUC-Judd, NSS, IG, CC, SIM, and KL. For other metrics, CASNet is not always the best but it is close to the best.

Performance on predicting contextual saliency: As suggested in [43, 11], NSS and IG take into account the relative importance of the salient regions, thus are the best evaluation measures for contextual saliency. CASNet beats the other methods on these two metrics across all three datasets, demonstrating its advantage on contextual saliency. Notably, CASNet consistently outperforms N-CASNet on all datasets (Table 2 – 4), and its advantage is largest on NSS and IG. This suggests the effectiveness of learning the relative weights of salient regions inside an image through the proposed subnetwork. Fig. 7 illustrates how CASNet uses contextual information to improve saliency prediction by learning the relative importance of emotional objects, which more closely matches human emotion prioritization than N-CASNet.

5.5. Analysis

To better understand the models, we further explore their performance on EMOd (as it has intensive object labels).

Emotion prioritization: Do the models exhibit emotion

prioritization like humans do? To see, we perform the same analyses as in Sec. 4.2, except calculating an object's attention score as the highest value of the normalized (predicted) saliency map in the object's contour. We compute the average predicted saliency scores of negative, neutral, and positive objects in EMOd by CASNet. The result (Fig. 8) is similar to Fig. 2. This suggests that the proposed model has a considerable ability to model human emotion prioritization. An ANOVA (object saliency scores as the dependent variable, object emotion types as the independent variable) for each model further shows that CASNet has the largest F value (CASNet: 92.17, N-CASNet: 87.95, SalGAN: 81.22, SALICON: 82.44, ML-Net: 69.99), indicating that CASNet prioritizes emotional objects more than comparing methods.

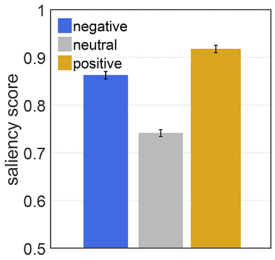

Figure 8: Emotional objects are predicted as being more salient than neutral objects by CASNet, which is consistent with the emotion prioritization effect of human observers.

CNN visualization: We perform a direct comparison before and after adding the channel weighting subnetwork for (a) CASNet with local weights frozen to 1 during training ("before" version, equivalently N-CASNet), and (b) regular CASNet ("after" version). We select 6 highly emotional images for 4 emotions and extract their highest-response patches (192×192 pixels in size) on their strongest weighted channel. The responses in "after" version show stronger emotions, suggesting that the subnetwork directs model's attention to more emotional content (Fig. 9).

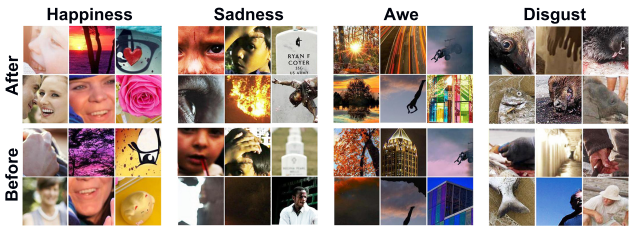

Figure 9: Examples of highest-response patches from before (bottom block) and after (top block) channel weighting. Patches of the same image are ordered in the same sequence.

Relative saliency of co-occurring objects: The capability of the proposed channel weighting subnetwork is not limited to emotion prioritization, but more broadly, it is able to predict the relative importance of co-occurring objects in general. To show this, we identify all images with co-occurring category objects (see Table 1 for the 12 categories). For each image, we calculate the difference in attention score for those two objects for human ground truth data. We also calculate the same difference score as predicted by CASNet and N-CASNet. By correlating the differences of each model with the human ground truth across images, we evaluate the degree to which models predict the relative saliency of the co-occurring objects. We calculate separate Spearman's rank correlations for all types of object co-occurrences (*e.g.*, faces with gazed-upon objects, gazed-upon objects with touched-objects). A larger correlation indicates that the model does a better job at predicting the relative saliency of co-occurring objects in the ground truth data. A paired t-test shows that CASNet has a higher correlation with human ground truth than N-CASNet (.74 *v.s.* .71, $p < .00001$) across all types of object co-occurrences. See Fig. 10 for examples.

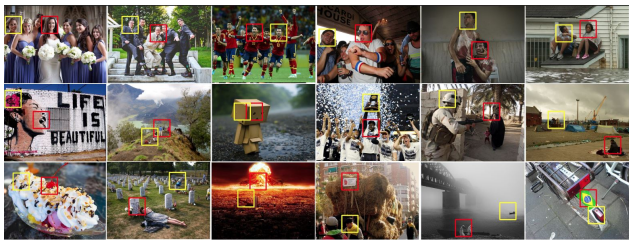

Figure 10: The most salient patches predicted by N-CASNet (yellow sqaure) and CASNet (red square). CASNet correctly prioritizes the most salient faces within an image (top row), people/body parts over other objects (middle row), and the most salient non-human objects.

6. Conclusion

In this paper, we present EMOd—a new emotional attention dataset for research on visual saliency and emotion-eliciting stimuli. Analyses on EMOd show that eye fixations correlate with human affective responses to the visual content of the images. We report the emotion prioritization effect, the strong and rapid, but brief, attentional bias towards emotional objects. To computationally address the emotion prioritization effect, we develop a novel DNN (CASNet) that encodes the relative importance of multiple salient regions and accounts for contextual importance for human attention. To our knowledge, this is the first attempt to quantify the relationships among human affective responses and visual attention on complex scenes, with a new DNN model that effectively mimics human attention in this context.

Acknowledgements

We thank Dr. Tian-Tsong Ng for helpful discussions. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centre in Singapore Funding Initiative, and a University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ).

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 5

[2] R. J. Baddeley and B. W. Tatler. High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision research*, 46(18):2824–2833, 2006. 1

[3] R. A. Bailey. *Design of comparative experiments*, volume 25. Cambridge University Press, 2008. 3

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015. 2

[5] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future Datasets"*, 2015. 2, 5

[6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia*, pages 223–232, 2013. 3

[7] T. Brosch, G. Pourtois, and D. Sander. The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion*, 24(3):377–400, 2010. 2, 4

[8] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011. 3

[9] C. Bundesen, S. Vangkilde, and A. Petersen. Recent developments in a computational theory of visual attention (tva). *Vision research*, 116:210–218, 2015. 1

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

[11] 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. 5, 6, 7

[12] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *ECCV*, pages 809–824. Springer, 2016. 2, 5

[13] M. G. Calvo and P. J. Lang. Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motivation and Emotion*, 28(3):221–243, 2004. 1

[14] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian. Saliency-aware nonparametric foreground annotation based on weakly labeled data. *IEEE transactions on neural networks and learning systems*, 27(6):1253–1265, 2016. 1

[15] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–11, 2009. 1

[16] F. Chollet. Keras. *GitHub repository*, 2015. 5

[17] R. J. Compton. The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and cognitive neuroscience reviews*, 2(2):115–129, 2003. 1, 2, 4

[18] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. *arXiv preprint arXiv:1609.01064*, 2016. 2, 6

[19] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*, 2016. 6

[20] A. Das, R. K. Kumar, D. R. Kisku, and G. Sanyal. Attention identification via relative saliency of localized crowd faces. In *Proceedings of the 10th International Conference on Informatics and Systems*, pages 101–106. ACM, 2016. 5

[21] A. Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer Science & Business Media, 2007. 3

[22] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 2008. 2, 3

[23] S. Engmann, B. M. Hart, T. Sieren, S. Onat, P. König, and W. Einhäuser. Saliency on a natural scene background: Effects of color and luminance contrast add linearly. *Attention, Perception, & Psychophysics*, 71(6):1337–1352, 2009. 2

[24] J. Fan, Y. Gao, H. Luo, and G. Xu. Automatic image annotation by using concept-sensitive salient objects for image content representation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 361–368. ACM, 2004. 1

[25] J. M. Fawcett, E. J. Russell, K. A. Peace, and J. Christie. Of guns and geese: A meta-analytic review of the weapon focus literature. *Psychology, Crime & Law*, 19(1):35–66, 2013. 1

[26] C. C. Fowlkes, D. R. Martin, and J. Malik. Local figure–ground cues are valid for natural images. *Journal of Vision*, 7(8):2–3, 2007. 1

[27] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012. 1, 2

[28] J. GREEN Dand SWETS. Signal detection theory and psychophysics, 1988. 5

[29] R. Gupta. Commentary: Neural control of vascular reactions: Impact of emotion and attention. *Frontiers in Psychology*, 7, 2016. 1

[30] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2010. 2

[31] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 6

[32] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, 2015. 2, 5, 6

[33] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, December 2015. 2

[34] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998. 1, 2, 6

[35] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *ECCV*, pages 512–528. Springer, 2014. 2

[36] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 2, 5

[37] J. M. Joyce. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*, pages 720–722. Springer, 2011. 6

[38] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *TOCHI (2017). DOI: http://dx. doi. org/10.1145/3131275*, 2017. 2

[39] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[40] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *CVPR*, pages 2176–2184, 2016. 2

[41] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche. Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial vision*, 13(2):201–214, 2000. 2

[42] S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *arXiv preprint arXiv:1510.02927*, 2015. 2, 6

[43] M. Kümmerer, T. S. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015. 6, 7

[44] M. Kümmerer, T. S. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 2, 6

[45] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. *Technical report A-8*, 2008. 3

[46] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013. 3

[47] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision research*, 47(19):2483–2498, 2007. 6

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[49] N. Liu and J. Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *arXiv preprint arXiv:1610.01708*, 2016. 2

[50] R. Marois and J. Ivanoff. Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6):296–305, 2005. 1

[51] B. Ni, M. Xu, T. V. Nguyen, M. Wang, C. Lang, Z. Huang, and S. Yan. Touch saliency: Characteristics and prediction. *IEEE Transactions on Multimedia*, 16(6):1779–1791, 2014. 2

[52] A. Öhman, A. Flykt, and F. Esteves. Emotion drives attention: detecting the snake in the grass. *Journal of experimental psychology: general*, 130(3):466–474, 2001. 1

[53] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 6

[54] H. Pashler. *Attention*. Psychology Press, 2016. 1

[55] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005. 6

[56] S. Ramanathan, H. Katti, R. Huang, T.-S. Chua, and M. Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 729–732. ACM, 2009. 2

[57] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *ECCV*, pages 30–43. Springer, 2010. 1, 2, 5

[58] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 6

[59] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, pages 1147–1154, 2013. 2

[60] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[61] M. Spain and P. Perona. Some objects are more equal than others: Measuring and predicting importance. In *ECCV*, pages 523–536. Springer, 2008. 2

[62] M. J. Swain and D. H. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991. 6

[63] H. Tang, C. Chen, and X. Pei. Visual saliency detection via sparse residual and outlier detection. *IEEE Signal Processing Letters*, 23(12):1736–1740, 2016. 6

[64] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision research*, 45(5):643–659, 2005. 5

[65] C. Thomas. Opensalicon: An open source implementation of the salicon saliency model. *arXiv preprint arXiv:1606.00110*, 2016. 6

[66] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Pattern Recognition*, pages 2798–2805, 2014. 2

[67] P. Vuilleumier. How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences*, 9(12):585–594, 2005. 1, 2, 4

[68] L. Wang, H. Lu, X. Ruan, and M.-H. Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, pages 3183–3192, 2015. 2

[69] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841. Springer, 2016. 2

[70] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 1, 2, 3

[71] 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

[72] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013. 6

[73] Q. Zhao and C. Koch. Learning saliency-based visual attention: A review. *Signal Processing*, 93(6):1401–1407, 2013. 2