Emotional Attention: From Eye Tracking to Computational Modeling

Shaojing Fan, Zhiqi Shen, Ming Jiang, *Member, IEEE*, Bryan L. Koenig, Mohan S. Kankanhalli, *Fellow, IEEE*, and Qi Zhao, *Member, IEEE*

Abstract—Attending selectively to emotion-eliciting stimuli is intrinsic to human vision. In this research, we investigate how emotion-elicitation features of images relate to human selective attention. We create the EMOtional attention dataset (EMOd). It is a set of diverse emotion-eliciting images, each with (1) eye-tracking data from 16 subjects, (2) image context labels at both object- and scene-level. Based on analyses of human perceptions of EMOd, we report an emotion prioritization effect: emotion-eliciting content draws stronger and earlier human attention than neutral content, but this advantage diminishes dramatically after initial fixation. We find that human attention is more focused on awe eliciting and aesthetic vehicle and animal scenes in EMOd. Aiming to model the above human attention behavior computationally, we design a deep neural network (CASNet II), which includes a channel weighting subnetwork that prioritizes emotion-eliciting objects, and an Atrous Spatial Pyramid Pooling (ASPP) structure that learns the relative importance of image regions at multiple scales. Visualizations and quantitative analyses demonstrate the model's ability to simulate human attention behavior, especially on emotion-eliciting content.

Index Terms—Human attention, image sentiment, human psychophysics, convolutional neural network, visual saliency

1 Introduction

Due to the capacity limits of the human brain, not all incoming environmental stimulation can be processed in parallel and evaluated thoroughly [1], [2]. All visual stimuli are in competition to become the focus of the eyes and encoded into visual short-term memory before it is filled up. Such phenomenon is known as selective attention [2], [3], [4], [5], [6]. Selective attention is a hallmark of human visual attention, and it is an important topic among researchers from various domains, ranging from psychology, neuroscience, to computer vision [7], [8], [9], [10], [11], [12].

- *Shaojing Fan, Zhiqi Shen, and Mohan S. Kankanhalli are with the School of Computing, National University of Singapore, Singapore 119613. E-mail: fanshaojing@gmail.com, dcsshenz@nus.edu.sg, mohan@comp.nus.edu.sg.*
- *Ming Jiang and Qi Zhao are with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 USA and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583. E-mail: {mjiang, qzhao}@umn.edu.*
- *Bryan L. Koenig is with the Department of Psychology, Southern Utah University, Cedar City, UT 84720 USA. E-mail: bryankoenig@suu.edu.*

Manuscript received 25 Sept. 2020; revised 16 Mar. 2022; accepted 9 Apr. 2022. Date of publication 21 Apr. 2022; date of current version 6 Jan. 2023.
This work was supported in part by National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, and in part by a University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ).
This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Departmental Ethics Review Committee of NUS under Application No. R-263-000-B25-232, and performed in line with the Declaration of Helsinki.
(Corresponding author: Qi Zhao.)
Recommended for acceptance by R. Cucchiara.
Digital Object Identifier no. 10.1109/TPAMI.2022.3169234

Substantial research finds that the emotional relevance of a stimulus influences selective attention [13], [14], [15], [16], [17], [18]. For example, people preferentially attend to *emotion-eliciting stimuli* (i.e., an object or scene that elicits an emotional response in the observer), such as cute babies or erotic scenes [19], [20]. Although many neuroimaging and behavioral studies have investigated how emotion-eliciting stimuli affect attention [14], [21], [22], few computer vision studies have—due in part to the lack of an eye-tracking dataset that includes emotion-eliciting stimuli. Advances in understanding the relationship between semantics and attention [23], [24], [25], [26], [27], [28] are ahead of those for how sentiment relates with human attention.

In this research, we systematically evaluate how emotion-eliciting features of images relate to human attention. We then model the relations computationally. We first 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 determine how emotion-eliciting content relates to human visual attention. Results indicate that emotion-eliciting content draws human visual attention strongly, quickly, but briefly—which we refer to as the *emotion prioritization effect*. Analyses further find that the emotional tone of scenes as a whole, correlates to human attention. Building on these findings, we propose a deep neural network (DNN) to model human attention computationally. The model (CASNet II) learns the relative importance of salient regions within an image and prioritizes emotion-eliciting content when predicting human attention. Such automatic assessment of visual attention has many applications, such as understanding user behavior, facilitating social advertising, and aiding autonomous driving [29], [30]. Our code,

Fig. 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.

models, and dataset are available online at https://github.com/fanshaojing/emotionalattention/.

We summarize our main contributions as follows:

1) We provide a novel image dataset (EMOd) featuring image sentiment and visual attention. It is the first dataset 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 at both object- and scene-levels. We discover the emotion prioritization effect—for our images, people attend to emotion-eliciting content not only strongly, quickly, but also briefly. We find that the emotional tone of the scene as a whole correlates with different human fixation patterns. For our dataset, awe eliciting and aesthetic animal and vehicle scenes have more focused human attention.
3) We computationally model human attention behavior by designing a deep learning network (CASNet II), and apply it on automated saliency prediction. CASNet II consists of two mechanisms to encode relative importance of regions and objects within an image. First, it employs an Atrous Spatial Pyramid Pooling (ASPP) structure to learn the multi-scale context information. Second, it uses a channel weighting subnetwork to highlight emotion-eliciting objects. Our model, with a much simpler structure but carefully designed to encode emotion prioritization, achieves the top performance on five benchmark datasets when evaluated by the normalized mean of all metrics.

The current research extends our previous work [31] in the following ways. (a) We extend the analyses on EMOd from object-level to scene-level, and from spatial to temporal (i.e., eye saccades, attention shift rank [32]). We report three new observations and discuss related insights. (b) We use an improved model (CASNet II) to simulate human attention behavior computationally. CASNet II is built on our previous model (*hereafter*, CASNet I) [31]. Two improvements were made in CASNet II over CASNet I. First, we changed to a customized Atrous Spatial Pyramid Pooling (ASPP) structure [33] to encode multi-scale, contextual saliency. The enlarged receptive fields from ASPP enable CASNet II to learn the relative importance in bigger image regions, extending previous object-based prioritization to a larger image area. Second, we replace the dual-stream architecture in CASNet I with a single stream, and reduce the input image scale. With these changes, CASNet II better models the human emotion prioritization effect, and significantly outperforms CASNet I in saliency prediction on all five benchmark datasets. It also improves the processing speed by almost 300%—CASNet II only takes 0.09 s to process one image whereas CASNet I needs 0.25 s on the same NVIDIA 1080Ti GPU. (c) We test our new model on two additional widely-used benchmark datasets MIT1003 [34] and OSIE [35] to demonstrate its generalizability. (d) We include five latest state-of-the-art methods for a more comprehensive comparison, namely EML-NET [36], DeepGaze II [37], MSI-Net [38], GazeGAN [39], SAM-ResNet [40]. (e) We provide new network visualizations and quantitative analyses to understand how CASNet II outperforms in modeling human attention behavior. (f) We perform new experiments to investigate the models' performance cross emotional and non-emotional datasets. Readers can refer to the supplementary material for a detailed summary of the above improvements, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3169234.

The remainder of the paper is organized as follows. Section 2 describes related research. Section 3 describes the construction of the EMOd dataset. Section 4 presents our analyses and empirical modeling of the psychophysics data. In Section 5, we describe our computational modeling of human attention behavior and test it on five benchmark datasets. In Section 6, we summarize our main findings and potential future applications.

2 Related Research

People have a remarkable ability to selectively attend to some regions in a scene [2], [4], [5], [6]. A plethora of research from multiple disciplines has evaluated selective attention. In this section, we discuss the most relevant research on selective attention, automated human attention prediction, and eye-tracking datasets.

Selective Attention. The preferential processing of high-priority stimuli in the environment is an essential function of selective attention. For example, scientists have reported

a hallmark feature of selective attention to be people's sensitivity to faces [41], [42], [43].

Scientists have also found that human attention generally prioritizes emotion-eliciting content over non-emotion-eliciting content [14], [16], [44], [45]. Emotion-eliciting stimuli–such as smiling faces, babies, and erotic scenes–attract human attention more than neutral stimuli [35], [46]. People tend to focus more on positive parts than negative parts in abstract paintings [47]. Memory of emotional videos modulates eye movements when viewing static scenes from the videos [48]. Salient objects influence the observers' emotional reactions to a whole image [49]. Visual "catchiness" of relevant information in online media can impact observer affect [50]. Also relevant is the non-emotional process referred to as the gaze-cuing effect, in which observers attend to the target of another person's gaze [51]. Our research builds on this prior research from multiple disciplines and extends it in the field of computer science. More specifically, we analyze how emotion-eliciting stimuli relate to human attention allocation on images of general scenes, providing a broad research scope. Furthermore, we computationally model the findings from human participants to show how understanding human attention behavior helps in automated saliency prediction.

Human Attention Prediction. Modeling visual saliency has raised much interest in theory and applications [52], [53], [54], [55], [56], [57]. Early saliency prediction models use pixel-level image attributes, such as contrast, color, orientation, and intensity [58], [59], [60]. An earlier advocate for context-aware saliency is [24], which also focuses on low-level image features. Recently, the resurgence of deep neural networks (DNNs) has resulted in large gains in saliency prediction [61], [62], [63], [64], [65], [66], [67], [68], such as SALICON [69], DeepGaze II [37], EML-NET [36], MSI-Net [38], SAM [40], and DeepFix [70]. While DNN-based models achieve considerable performance improvement, existing models do not explicitly model or offer insights about the relative importance of multiple objects in context. As suggested by Bylinskii and colleagues [71], in order to approach human-level performance, saliency models need to incorporate high-level image concepts, such as text or motion, and reason about the relative importance of image regions.

Building on these suggestions, researchers seek to incorporate increasingly higher-level perceptual properties of images [8], [35], [54], [71]. Their models attempt to encode various high-level concepts. For example, several studies explore how human attention is directed towards faces with emotional expressions [16], [72], [73]. Studies in [74] find for action images that observers have extensive fixation transitions between interacting objects. [75] focuses on human attention on text.

More recently, [76] makes preliminary attempt to incorporate color-based emotion-eliciting information in saliency prediction. [77] and [78] included object sentiments in their saliency prediction networks. However, saliency researchers have not yet attempted to systematically measure or model the relation between emotion-eliciting objects and attention. One major reason could be the lack of a proper dataset with both emotion-eliciting content and eye-tracking data. In our research, we develop a novel eye-tracking dataset focusing on emotional attention. The dataset allows us to comprehensively assess the relation between emotional content and human attention, and inspires a new model design that effectively addresses the emotion prioritization effect in attention allocation within an image.

Eye-Tracking Datasets With Emotion-Eliciting Information. A few datasets feature emotion-eliciting images have been proposed, such as the EMOTIC Dataset [79]; the DeepSent dataset [78], and the Twitter dataset [80]. Without eye tracking data, however, they are unsuitable for our purposes.

Two related datasets that we use as benchmarks for saliency prediction (see Section 5.3) are NUSEF [23] and CAT2000 [81]. NUSEF is 751 emotion-eliciting images that depict mostly faces, nudes, and human actions. CAT2000's training set contains 2,000 images of diverse scenes, such as emotion-eliciting images and cartoons. However, these two datasets have limited emotion-eliciting content and no object labels. Emotion labels are absent in other commonly used eye-tracking datasets (for an overview see [82]). In this research, we present the first eye-tracking dataset to include images of diverse emotion-eliciting scenarios, together with extensive image annotations.

Measuring human attention requires customized eye-tracking equipment, making crowdsourcing difficult. Researchers have been trying to combine crowdsourcing techniques with eye tracking data collection [83], [84]. Some methods for large-scale attention data collection include using webcams and mouse/finger movements [85], [86], [87], [88], [89], [90], but their validity is not completely established for images of diverse scenes. Indeed, [88] reports that their measures of attention are disproportionately influenced by the image semantics, e.g., the number of objects presented in an image. None of these methods to date have been applied specifically to emotion-eliciting images. Thus, how the emotion-eliciting properties of an image impacts attention measurement is unknown. Seeking maximal validity for our dataset, we use the gold-standard: measuring with eye-tracking equipment in controlled laboratory conditions [91].

3 EMOtional Attention Dataset

In this section, we provide details on how we constructed EMOtional attention dataset (EMOd), a new dataset of 1,019 emotion-eliciting images, with eye-tracking data and annotations at object and image levels. This dataset is aimed for research on visual saliency and image sentiment.

3.1 Image Collection

The EMOd dataset was constructed from two sources: (1) a subset (321) photos of the International Affective Picture System (IAPS) [100], and (2) a set of 698 photos collected by the authors. From IAPS, we selected 321 photos that were identified as primarily eliciting one emotion in a study by [93]. This subset has also been used in other computer vision research on emotion assessment [98], [101], [102]. The aim of our own collection was to make the dataset more diverse regarding how observers' emotions are evoked. We grouped the 698 images into six types based on how they evoked emotions (parenthetical numbers are how many images were that type): emotion-eliciting objects (29), emotion-eliciting activities (158), emotion-eliciting gist (145), emotion-eliciting spatial layout (105), emotion-eliciting

Fig. 2. User interface of (a) EMOd object-labeling platform, and (b) EMOd image-annotation platform.

color and illumination (121), and emotionally-neutral images (140). Readers can refer to the supplementary material for example images of the six types, available online.

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 $38.94^\circ \times 29.20^\circ$. Subject eye movements were recorded at 1,000 Hz using an Eyelink 1,000 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: Object-Level Annotation

We built an online EMOd object-labeling system based on the LabelMe platform [103] (see Fig. 2a). Three paid undergraduate students from the National University of Singapore labeled the object contour and object name for all objects in each image. Each object was also labeled according to its sentiment category (i.e., negative, neutral, or positive) and semantic category. The design of semantic categories is based on [35], which includes four types: (1) directly relating to humans (i.e., emotional face, neutral face, touched, gazed), (2) relating to other (nonvisual) senses of humans (i.e., sound, smell, taste, touch), (3) designed to attract attention or for interaction with humans (i.e., text, watchability, operability), and (4) objects with implied motion. Table 1 lists all semantic categories. We adopted a similar approach to previous research [104], [105], [106] by keeping the majority votes for object's labels. We had an overall agreement of 82% for all labeled objects, suggesting decent consistency among participants' labels.

3.4 Psychophysics Study III: Scene-Level Annotation

We also built an EMOd scene-annotation platform to collect human perceptions of scene-level attributes (see Fig. 2b). Our attributes list covers both semantic and sentiment aspects of the images, including (1) 10 basic emotions commonly studied in psychology [92], [93]: happiness, surprise, awe, excitement, amusement, contentment, sadness, anger, fear, and disgust; (2) valence, arousal, dominance measured with the Self-Assessment Manikin for non-verbal pictorial assessment [94]; (3) high-level attributes commonly studied in computer vision, such as aesthetics, image quality, photo-realism, depths of field, and symmetry [95], [98], [107]. Table 2 shows the detailed list of the 33 attributes.

For the 698 images we collected, we deployed the EMOd image-annotation platform on AMT and recruited 348 AMT

TABLE 1
Descriptions of Semantic Attributes of Objects Labeled in EMOd Dataset

Type	Category	Description	Object No.	Image No.
Directly related 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
Related 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

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.

TABLE 2
List of 33 Scene-Level Attributes in the EMOd Dataset

Attribute type	Detailed attributes
Emotions [92], [93]	Happiness; Surprise; Awe; Excitement; Amusement; Contentment; Sadness; Anger; Fear; Disgust
Self-Assessment Manikin [94]	Valence; Arousal; Dominance
Semantics [95]	Familiarity; Unusualness; Dynamics; Informativeness; Natural object
Aesthetics [96], [97]	Aesthetics; High quality; Colorfulness; Natural color; Sharpness
Spatial layout [98]	Have objects of focus; Single object focus; Close-up shot; Centered; Symmetry
Naturalness [99]	Photorealism
Related to people [99]	Attractive person; Posing; Eye contact; Positive expression

workers ($> 95\%$ approval rate in Amazon's system) to annotate. For the IAPS data set, due to copyright restrictions, we recruited 10 undergraduate students from the National University of Singapore to annotate them on the platform within the campus intranet. The detailed questionnaire is in the supplementary material, available online. On average, each image was annotated by 10 participants. For each image we computed the score of each attribute by averaging the answers given by the 10 participants, then transformed scores for each attribute to a range of [0, 1] with raw scores of 1 becoming 0 and raw scores of 9 becoming 1. Averaging across the raters for each image, we got an average Cronbach's alpha [108] of. 88 across the 1,019 images in EMOd dataset, indicating a good internal consistency among the annotators [108]. For more details on EMOd construction, human annotations and data reliability, please refer to the supplementary material, available online.

4 Visual Sentiments and Human Attention

We analyzed the data in EMOd to explore how emotion-eliciting properties of images related to human attention, at both object- and scene-levels.

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 to one degree of visual angle and then normalizing the map to maximum 1 (a common method in saliency research [109]). 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 [46].

Our analyses are performed at two levels: (1) object-level, focusing on how the human attention of an individual object is affected by its emotion-eliciting properties; and (2) scene-level, investigating how image sentiment as a whole affects human attention. We use inferential statistical analysis techniques, such as univariate analyses of variance (ANOVA), post-hoc Tukey tests, simple effects analysis, and Spearman's rank correlation. These analyses are standard in behavioral and other sciences. See, for example, [110] for an introduction to these inferential statistics.

4.2 Object-Level Analyses

In this subsection, we report our findings on how observer attention on an object correlates with the object's sentiment category and semantic attributes.

Observation 1 (Emotion prioritization effect). Emotion-eliciting *objects receive more and earlier human attention than* neutral *objects. Furthermore, people attend to emotion-eliciting objects not only strongly, quickly, but also* briefly*—a positive or negative object is more likely to draw human attention at first fixation, but the advantage diminishes quickly during subsequent fixations.*

Observation 1 is based on the following analyses. First, a two-way ANOVA has the 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 ($F(2, 4263) = 22.96$, $p < .001$[1]) and semantic category ($F(12, 4263) = 4.31$, $p < .001$). The larger F score of sentiment over semantics (22.96 *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 (p s[2] $< .001$), and attention scores for negative and positive objects do not significantly differ, $p = .270$ (see Fig. 3a).

Second, we define *attention shift rank* as the descending values indicating the order in which distinct objects are attended by observers, one at a time [32]. Objects with higher attention shift ranks have earlier fixations in a fixation sequence. ANOVA indicates a strong effect of sentiment on attention shift rank for the objects, $F(2, 42993) = 74.16, p < .001$. Post hoc Tukey tests show that both positive and negative objects have higher attention shift rank than neutral objects (ps $< .001$), but negative and positive objects do not significantly differ, $p = .423$, see Fig. 3b. Analyses suggest a strong correlation between objects' attention score and attention shift rank (Spearman's rank correlation $\rho = .80$), indicating that objects that are more salient also draw attention earlier.

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. 4). By showing for the first time that attention prioritization

1. We report the results of ANOVAs as, "F(df $_{condition}$, 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.

2. Throughout the paper, ps represents the plural form of p.

Fig. 3. Emotion-eliciting objects receive more (a) and earlier (b) human attention than neutral objects. In all figures in this paper, error bars represent standard error. Images in (c) and (d) illustrate how emotion-eliciting objects (outlined in blue), such as the crying face and broken card, are more salient and draw attention earlier than neutral/less emotional stimuli (outlined in gray). In each pair of numbers, the first number is the attention score and the second number indicates attention shift rank.

diminishes drastically after initial fixation for the EMOd dataset, our findings reveal a more nuanced understanding of the claim that human attention prioritizes emotion-eliciting stimuli over non-emotion-eliciting stimuli [14], [16], [45].

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 emotion prioritization occurs primarily for semantic categories of "touched," "gazed," "motion," "sound" (see Fig. 5a). Objects being "touched" and "gazed" upon, 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. Figs. 5b and 5c illustrates this interaction using images with gazed-upon objects. To evaluate the gaze-cuing effect, independent samples t-test compares the attention shift rank of faces with gaze cues with faces targeted by gaze cues. Results show that faces with gaze cues have higher attention shift ranks than faces targeted by gaze cues ($t(12) = 3.82, p = .003$), suggesting a potential gaze-cuing effect for faces. Independent samples t-test shows no significant gaze-cuing effect for objects of non-face categories ($t(44) = 0.31, p = .762$), which may be due to the multiple confounding factors on human attention such as object semantics and sentiments.

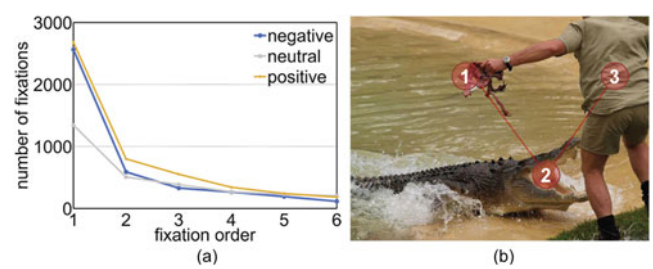


Fig. 4. (a) Human observers fixated first on emotion-eliciting objects more than neutral objects, but the attention prioritization quickly diminishes. (b) Viewers fixated on the emotion-eliciting objects (i.e., food (1) and crocodile's mouth (2)) before the neutral human body (3).

Exploratory analyses evaluated other low- and mid-level factors that might influence attention score and attention shift rank. Results indicate significant effects for object's location, color, and luminance (Spearman's rank correlation, ρs $\geq .32$). Readers can refer to the supplementary materials for details, available online.

Observation 3. *Human attention varies on objects with different semantic attributes. Human faces and human-related objects draw stronger attention.*

Following up on the main effect of object semantic category, post hoc Tukey tests indicate that object categories with highest attention scores are "gazed" upon, "face (emotional)" and "face (neutral)," followed by "sound," "motion," "touch," "smell" and "taste" (see supplementary material for details), available online. This is consistent with previous findings [10], [11], [12], [35], which reports that human faces and human-related objects generally attract more attention.

4.3 Scene-Level Analyses

In addition to the object-level analyses, here we report correlations of human attention with the 33 scene-level attributes (see Table 2 for the detailed list). Previous findings suggest that human attention patterns differ across scene categories [81], and image contexts affect visual attention [24]. Informed by these findings, we compute the Spearman's rank correlation (ρ) between the number of fixations and scene-level attributes separately for each scene category.

Observation 4. *For images of indoor scenes in our dataset, human attention is more focused (i.e., less diffused) with a more focused or a less informative scene.*[3]

Fixation counts positively correlate with the attribute "informativeness" ($\rho = .27, p < .001, n = 59$), and "having focused objects" ($\rho = .25, p < .001, n = 59$) across all images. With a fixed viewing time of 3 seconds, fewer fixations (i.e., longer fixation duration) indicate human attention is more focused.

3. The image "having focused objects" and "informativeness" are among the 33 scene-level attributes rated by our participants. Readers can refer to the supplementary materials for details on attributes annotation, available online.

Fig. 5. (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 emotion-eliciting gazed-upon object—the injection point on the crying child's arm (b) has a higher attention score than the neutral gazed-upon object—the box of dye in the lady's hand (c).

Observation 5. *Human attention is more focused for awe-eliciting and aesthetic images of animal and vehicle scenes.*

Spearman's rank correlation indicate that two attributes indicating positive emotions—"awe" and "aesthetics"—negatively correlate with fixation counts, more so in animal and vehicle scenes ($|\rho|$s $\geq .18$,p $s < .05$, $n = 139, 65$, respectively). In contrast, two attributes indicating negative emotions—"sadness" and "disgust"—positively correlate with fixation counts for animal scenes (ρ s $\geq .17$,p$s < .05$, $n = 139$). Fig. 6 shows example images from vehicle and indoor scenes on the respective attributes. Readers can refer to the supplementary material for a detailed list of all correlated scene attributes, available online. Notably, other factors might influence human fixation allocation, such as the number of objects in a scene and regions toward which people tend to look (e.g., windows). Thus, we need to be cautious in making cause-effect claims regarding scene-level attributes and human attention.

We further compute the correlation between scene-level attributes and human fixations across time. We use four statistics commonly used in fixation analysis, namely fixation duration, saccade duration, saccade length, and saccade velocity [111]. Below we report the main finding.

Observation 6. *Humans exhibit longer fixation duration and lower saccade velocity on more emotionally positive images, but shorter fixation duration and faster saccade on more informative images and more emotionally negative images.*

We dichotomized each attribute ratings by setting an upper bound threshold as 0.67 (being positive) and a lower bound threshold as. 33 (being negative) for emotional attributes [112]. Fixation patterns differed significantly between the two groups of several high-level attributes, such as "aesthetic," "awe," and "sad" (paired samples t-tests, ps $< .001$). In particular, positive (awe, excitement, happy) images of animals and vehicles had longer fixation duration and lower saccade velocity. This corroborates Observation 5, which shows that human attention is more focused for awe-eliciting and aesthetic animal and vehicle scenes.

Readers can refer to the supplementary material for detailed distribution of these attributes and their correlation with human attention behavior, available online.

In summary, human attention varies according to different emotion-eliciting properties at both object- and scene-levels. A notable caveat for our findings is that we cannot make definitive claims regarding causality. Our methods capture some of the most likely causal variables, but they do not control for other unidentified variables that could correlate with our measured variables and be the true variables influencing human attention. Future research could experimentally isolate the critical variables to increase the internal validity of our correlational findings [113].

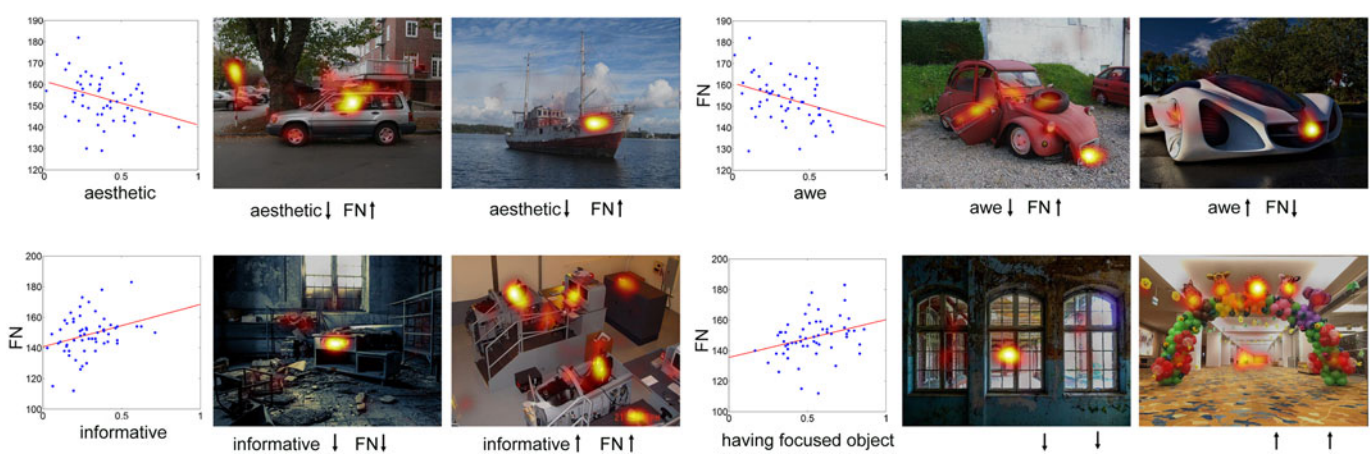


Fig. 6. Correlation of number of fixations with attributes "aesthetics," "awe," "informativeness," and "having focused objects" in scenes of vehicle (top) and indoors (bottom). In the left graph of each set, the x-axis stands for the ratings of respective attribute, the red line is the linear regression line of image points, and each dot represents one image.

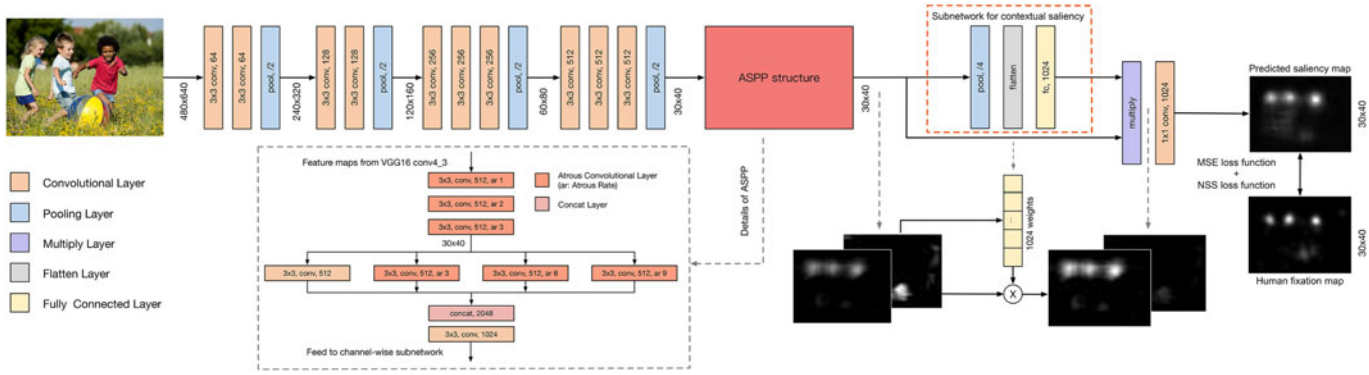


Fig. 7. The architecture of the proposed DNN (CASNet II). An Atrous Spatial Pyramid Pooling (ASPP) structure with four branches (inside the gray dashed rectangle) is used to capture the contextual information for each pixel at multiple resolution scales. A channel weighting subnetwork (inside the dashed orange rectangle) computes a set of 1024-dimensional feature weights for each image (instead of only one for the whole image) 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 Computational Modeling

This section reports how we computationally model human attention. We demonstrate how encoding emotion prioritization can help automated saliency prediction.

5.1 Model Design

The proposed DNN model is illustrated in Fig. 7. In the following paragraphs, we discuss the motivation, mechanisms, and design of our model, which is focused on contextual saliency—saliency regarding both spatial and semantic context of the scene.

We construct the backbone of our model based on the VGG-16 network architecture. The model design is motivated by two aspects. First, our human studies indicate human emotion prioritization is often present in large image regions and not limited to single objects. Second, we notice that the final output saliency map of the computational model depends on the size of the receptive view. A larger receptive view enables the model to capture more information around the targeted pixels in the output saliency map. Ideally, our network should be able to learn contextual information for each pixel at multiple resolution scales. To achieve this, we employ an Atrous Spatial Pyramid Pooling (ASPP) [33] at the last layers of the VGG-16 network (the gray dashed rectangle in Fig. 7). The initial design goal of ASPP is semantic segmentation to detect object boundaries at different scales. We customize it for saliency prediction in two ways. First, we adjust the size of the atrous rates to focus on the holistic context information instead of object boundaries. We then empirically design the pyramid pooling structure with four branches to extract multi-scale context information while ensuring high network efficiency. By doing so, the ASPP structure in CASNet II improves contextual saliency learning with more resolution scales. The enlarged receptive fields allow the model to better learn the relative importance among multiple objects/areas, extending the previous object-based prioritization to prioritization in larger image regions.

In particular, we first replace the three vanilla convolutional layers in block Conv4 to atrous convolutional layers with rates 1, 2, and 3, respectively. Such replacement enlarges the receptive field without increasing the computation overhead. We then apply a four-branch pyramid pooling structure to learn the saliency score for each pixel under different resolution levels of context. The first branch is a 1×1 vanilla convolutional layer representing the same size of feature maps from block Conv4. The other three branches are designed to gain information from larger receptive fields. We set the kernel size to 3×3, and the atrous rates to be 3, 6, and 9, respectively. The larger receptive fields help obtain more holistic contextual information within larger image regions. The ASPP structure produces 1024-dimensional multiple-scale feature maps for later processing.

To address emotion prioritization, we further design a channel weighting subnetwork (the orange dashed rectangle in Fig. 7) that encodes contextual information, enabling the network to highlight emotion-eliciting objects from the surroundings. The model automatically computes a 1024-dimensional feature importance, which corresponds to an image's 1024 feature maps. This allows the subnetwork to learn the relative importance of the image's semantic features. Specifically, to compute the weight, we first apply a 4×4 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 in a channel-wise multiplication.

We feed images of $640\times480\times3$ pixels to the network. The output of backbone network streams are re-scaled to the same spatial resolution, and stacked together to form multi-scale deep features of dimension $40\times30\times1024$. 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

40×30 pixels. Finally, we resize the saliency map back to the dimension of the original image.

5.2 DNN Parameters

We initialize the training to the pre-trained parameters for VGG-16 on ImageNet. A combination of mean squared error (MSE) and Normalized Scanpath Saliency (NSS) is used as the loss function. We set the same weights for NSS and MSE. We use a fixed loss function combination for all experiments. 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 4. 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 [86]. The entire training procedure takes about one day on a single NVIDIA 1080Ti GPU using Tensorflow 2 [114].

5.3 Experiment Datasets

We test our model on five eye-tracking datasets, three of them have image collections focused on emotion-eliciting content. The first is EMOd, with 1019 emotion-eliciting images. The second is the NUSEF dataset [23], which has 751 images that depict mostly emotion-eliciting objects and human actions. The third is the training set of CAT2000 [81], which contains 2000 diverse images including including emotional, cartoon, social, and so on. The other two datasets, MIT1003 [34] and OSIE [35], are widely used in saliency prediction, although they do not focus on emotion-eliciting content. MIT1003 contains 1003 natural indoor and outdoor scenes, and is commonly used on MIT/Tuebingen Saliency Benchmark [120]. OSIE dataset is a collection of 700 aesthetic photographs from Flickr and Google. By testing our algorithms on datasets with different features, we aim to have a comprehensive evaluation of the proposed method.

5.4 Comparison Methods

First, we compare the proposed saliency prediction model (i.e., CASNet II—Context-Adaptive Saliency Network II) with two of our previous versions: i) our model published in CVPR 2018 [31] (i.e., CASNet I, the prior version of CASNet II without the ASPP structure); ii) a model without the weighting subnetwork (i.e., N-CASNet—Not Context-Adaptive Saliency Network). More details are reported in the Ablation Study in Section 5.7.

We further compare our models with 10 others. Eight are state-of-the-art DNN-based models: EML-NET [36], DeepGaze II [37], MSI-Net [38], GazeGAN [39], SAM-ResNet [40], SALICON[4] [64], SalGAN [116], and ML-Net [67]. Two are non-DNN models with top performance in the non-DNN model category: Boolean Map based Saliency (BMS) [117] and Saliency via Sparse Residual & Outlier Detection (SROD) [118]. Two are classic bottom-up approaches: Graph-Based Visual Saliency (GBVS) [119] and Itti-Koch model (IttiKoch) [58]. These models are top performers on the MIT/Tuebingen Saliency Benchmark [120] in their respective categories. To ensure fair comparisons, all DNN-based models are trained on the SALICON dataset to achieve their best possible performance, and all models are directly tested on the five benchmark datasets without training/fine-tuning on them. We disabled the pre-computed center bias in DeepGaze II as we presume all models have no prior knowledge about the test data. For our three versions of CASNet and three comparison models whose codes are publicly available, we run them three times by training on the SALICON dataset, and report the mean and standard deviation.

4. We use the code of OpenSALICON (a publicly available implementation of SALICON) [115].

5.5 Evaluation Metrics

Following the MIT/Tuebingen Saliency Benchmark [120], we use 8 metrics for comprehensive evaluation. The Area Under the ROC Curve (AUC) [121] treats the saliency map as a binary classifier. We use two variants of AUC: AUC-Judd and AUC-Borji [122], and shuffled-AUC (sAUC) [123] 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 [71], [122], [124], as needed in our study. We further use five similarity metrics to measure the similarity between the saliency map and fixation map, namely Normalized Scanpath Saliency (NSS) [125], Linear Correlation Coefficient (CC) [126], histogram intersection (SIM) [127], the Kullback-Leibler divergence (KL) [128], and Information Gain (IG) [122], [129]. See [122] for an introduction of these metrics.

5.6 Experiment Results

We report statistical results in Tables 3, 4, 5, 6, and 7. Qualitative results for the EMOd dataset are shown in Fig. 8. Our model (CASNet II), with the channel weighting subnetwork and ASPP structure, is most advantageous on datasets focusing on emotion-eliciting content (i.e., EMOd, OSIE). To have a concise overview of the comparison, we compute an average score over all metrics for each model. Specifically, we use z-score transformation to normalize each column of metrics first. We then compute the mean of all columns of metrics (with a negative weight of KL). No model stands out on every metric (Tables 3, 4, 5, 6, and 7). When evaluated by the normalized mean of all metrics in Table 8, our model achieves the best performance on all datasets, suggesting the efficacy of the model design. Notably, however, different models have advantages on particular metrics, which may be useful for specific applications.

As illustrated in Fig. 8, CASNet II is most advantageous on images showing multiple emotion-eliciting objects (first two rows in Fig. 8) or images without obvious focal objects (last four rows in Fig. 8). This advantage demonstrates the efficacy of the proposed ASPP structure and channel weighting subnetwork.

5.7 Ablation Study

In this subsection, we further evaluate the effectiveness of each component of the model. To do this, we compare the performance of the three versions of our method: i) CASNet II (model with both ASPP structure and channel weighting subnetwork); ii) CASNet I (model with channel weighting subnetwork, but without ASPP structure); iii) N-CASNet

TABLE 3
Results on the EMOd Dataset

	AUC-Judd↑	AUC-Borji↑	sAUC↑	NSS↑	IG↑	CC↑	SIM↑	KL↓
CASNet II (ours)	**0.84** (0.002)	0.81 (0.002)	0.79 (0.000)	1.81 (0.005)	**1.80** (0.004)	**0.68** (0.002)	0.57 (0.006)	**5.55** (0.001)
CASNet I [31]	0.83 (0.001)	0.81 (0.001)	0.79 (0.002)	1.73 (0.003)	1.51 (0.005)	0.66 (0.002)	0.57 (0.002)	5.74 (0.003)
N-CASNet	0.81 (0.008)	0.79 (0.003)	0.77 (0.001)	1.61 (0.007)	1.50 (0.020)	0.60 (0.005)	0.51 (0.035)	5.70 (0.082)
EML-NET [36]	0.83 (0.001)	0.78 (0.003)	0.77 (0.003)	**1.91** (0.003)	0.33 (0.231)	0.70 (0.001)	0.60 (0.001)	6.52 (0.157)
MSI-Net [38]	**0.84** (0.002)	0.81 (0.002)	0.78 (0.004)	1.80 (0.019)	1.28 (0.004)	0.68 (0.006)	0.60 (0.007)	5.89 (0.002)
SALICON [115]	0.83 (0.001)	0.81 (0.001)	0.79 (0.001)	1.64 (0.001)	0.63 (0.001)	0.59 (0.000)	0.52 (0.000)	5.66 (0.000)
DeepGaze II [37]	0.83	**0.82**	**0.80**	1.39	1.26	0.52	0.46	5.93
GazeGAN [39]	0.82	0.80	0.76	1.60	1.21	0.61	0.56	6.62
SAM-ResNet [40]	0.83	0.73	0.72	1.90	0.41	**0.68**	**0.60**	6.46
SalGAN [116]	0.83	0.80	0.78	1.74	1.13	0.66	0.58	5.83
ML-Net [67]	0.82	0.76	0.74	1.74	1.21	0.62	0.56	5.78
BMS [117]	0.77	0.75	0.74	1.12	1.02	0.42	0.45	5.94
SROD [118]	0.74	0.73	0.72	0.98	0.88	0.37	0.42	6.04
GBVS [119]	0.79	0.78	0.75	1.18	1.13	0.47	0.48	5.86
IttiKoch [58]	0.73	0.72	0.70	0.88	0.88	0.35	0.43	6.04

In all subsequent tables in this paper, the best performance in each metric is highlighted in bold. For Tables 3, 4, 5, 6, and 7, the performance of models in the first six rows are the means of three runs. The numbers in the parentheses indicate the standard deviation. "↑" indicates higher values are better. "↓" indicates lower values are better.

TABLE 4
Results on the NUSEF Dataset

	AUC-Judd↑	AUC-Borji↑	sAUC↑	NSS↑	IG↑	CC↑	SIM↑	KL↓
CASNet II (ours)	**0.84** (0.003)	0.79 (0.003)	**0.77** (0.002)	**1.82** (0.001)	**1.36** (0.038)	**0.70** (0.001)	0.57 (0.004)	**5.36** (0.022)
CASNet I [31]	0.83 (0.003)	0.79 (0.003)	0.76 (0.003)	1.75 (0.020)	0.62 (0.063)	0.67 (0.008)	0.58 (0.005)	5.85 (0.043)
N-CASNet	0.81 (0.002)	0.79 (0.003)	0.76 (0.002)	1.67 (0.001)	1.12 (0.011)	0.64 (0.001)	0.49 (0.002)	5.53 (0.008)
EML-NET [36]	0.83 (0.002)	0.75 (0.003)	0.73 (0.003)	1.81 (0.003)	0.31 (0.002)	0.68 (0.002)	0.60 (0.002)	7.15 (0.195)
MSI-Net [38]	**0.84** (0.001)	0.79 (0.001)	0.76 (0.001)	**1.82** (0.003)	0.15 (0.010)	**0.70** (0.001)	**0.61** (0.001)	6.17 (0.007)
SALICON [115]	0.82 (0.001)	**0.80** (0.001)	**0.77** (0.001)	1.68 (0.001)	1.19 (0.003)	0.65 (0.001)	0.53 (0.001)	5.47 (0.002)
DeepGaze II [37]	0.80	0.79	**0.77**	1.33	0.49	0.51	0.46	5.96
GazeGAN [39]	0.82	0.79	0.76	1.64	0.85	0.64	0.57	6.85
SAM-ResNet [40]	0.83	0.70	0.69	1.76	0.46	0.65	0.57	7.25
SalGAN [116]	0.83	0.78	0.75	1.72	0.51	0.66	0.58	5.90
ML-Net [67]	0.82	0.74	0.71	1.66	0.11	0.61	0.55	6.20
BMS [117]	0.77	0.75	0.72	1.08	0.67	0.42	0.44	5.84
SROD [118]	0.74	0.74	0.71	0.95	0.62	0.37	0.42	5.88
GBVS [119]	0.80	0.79	0.74	1.21	0.96	0.49	0.48	5.64
IttiKoch [58]	0.71	0.70	0.67	0.77	0.56	0.31	0.40	5.92

(model without channel weighting subnetwork or ASPP structure). The results are shown in the first three rows in Tables 3, 4, 5, 6, and 7. Fig. 9 gives qualitative examples to show how CASNet II and CASNet I use contextual information to improve saliency prediction by learning the relative importance of emotion-eliciting objects, which more closely matches human emotion prioritization than N-CASNet.

Contribution of Channel Weighting Subnetwork. To analyze the contribution of the channel weighting subnetwork, we compare the performance of CASNet I and N-CASNet. On all five datasets, CASNet I consistently outperforms N-CASNet. The results demonstrate the efficacy of our contextual saliency mechanism. Furthermore, as suggested in [122], [129], NSS and IG take into account the relative importance of the salient regions, thus are the best evaluation measures for contextual saliency. CASNet II beats the other methods on these two metrics across all three datasets, demonstrating its advantage on contextual saliency. Notably, CASNet I consistently outperforms N-CASNet on all datasets (Tables 3, 4, 5, 6, and 7), 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.

Contribution of ASPP Structure. The channel weighting subnetwork discussed above aims to highlight emotion-eliciting objects. However, this is insufficient to encode the holistic contextual information, which we found was important for observers of EMOd. The ASPP structure is used to model contextual saliency at multiple scales. With the ASPP structure, the largest receptive field of CASNet II covers almost 90% of the whole image area (receptive view: 580*580, input image size: 480*640), whereas without ASPP, the largest receptive view of CASNet I only makes up for 32% of the image region (receptive view: 196*196, input image size: 300*400). Readers can refer to Table S5 and Figs. S7-S8 in the supplementary material for more details, available online. As shown in Tables 3, 4, 5, 6, and 7, CASNet II (with

TABLE 5
Results on the CAT2000 Dataset

	AUC-Judd↑	AUC-Borji↑	sAUC↑	NSS↑	IG↑	CC↑	SIM↑	KL↓
CASNet II (ours)	0.83 (0.003)	**0.81** (0.004)	**0.78** (0.004)	1.55 (0.008)	0.42 (0.016)	0.60 (0.004)	0.56 (0.001)	**5.82** (0.012)
CASNet I [31]	0.81 (0.006)	0.79 (0.004)	0.76 (0.003)	1.48 (0.013)	0.23 (0.194)	0.57 (0.005)	0.54 (0.021)	5.96 (0.123)
N-CASNet	0.77 (0.032)	0.75 (0.014)	0.73 (0.007)	1.24 (0.101)	0.06 (0.210)	0.48 (0.039)	0.47 (0.048)	6.08 (0.133)
EML-NET [36]	0.83 (0.001)	0.78 (0.003)	0.75 (0.003)	1.62 (0.005)	**0.61** (0.003)	0.61 (0.003)	0.58 (0.003)	6.79 (0.139)
MSI-Net [38]	0.82 (0.001)	0.80 (0.001)	0.77 (0.001)	1.48 (0.014)	0.17 (0.024)	0.57 (0.004)	0.57 (0.001)	6.00 (0.017)
SALICON [115]	0.81 (0.001)	0.80 (0.002)	0.76 (0.001)	1.43 (0.002)	0.28 (0.008)	0.55 (0.001)	0.52 (0.001)	5.91 (0.006)
DeepGaze II [37]	0.80	0.79	0.76	1.24	-0.15	0.48	0.48	6.22
GazeGAN [39]	0.83	**0.81**	**0.78**	1.52	-0.58	0.59	0.57	6.52
SAM-ResNet [40]	**0.84**	0.76	0.74	**1.77**	-0.21	**0.65**	**0.59**	6.25
SalGAN [116]	0.81	0.80	0.77	1.45	0.08	0.56	0.53	6.08
ML-Net [67]	0.79	0.73	0.70	1.31	0.04	0.49	0.51	6.08
BMS [117]	0.78	0.77	0.73	1.15	-0.13	0.44	0.49	6.21
SROD [118]	0.77	0.76	0.72	1.07	-0.11	0.41	0.48	6.06
GBVS [119]	0.80	0.79	0.75	1.24	0.18	0.49	0.50	5.99
IttiKoch [58]	0.71	0.70	0.66	0.76	-0.25	0.30	0.42	6.29

TABLE 6
Results on the MIT1003 Dataset

	AUC-Judd↑	AUC-Borji↑	sAUC↑	NSS↑	IG↑	CC↑	SIM↑	KL↓
CASNet II (ours)	**0.88** (0.001)	**0.86** (0.001)	**0.83** (0.002)	2.25 (0.007)	**2.08** (0.014)	0.65 (0.000)	0.47 (0.004)	5.39 (0.013)
CASNet I [31]	0.87 (0.002)	**0.86** (0.002)	0.82 (0.003)	2.09 (0.027)	1.96 (0.013)	0.61 (0.007)	0.47 (0.003)	5.46 (0.010)
N-CASNet	0.85 (0.002)	0.83 (0.002)	0.80 (0.002)	1.99 (0.001)	1.65 (0.015)	0.57 (0.001)	0.38 (0.003)	5.69 (0.010)
EML-NET [36]	**0.88** (0.001)	0.83 (0.003)	0.80 (0.002)	**2.40** (0.008)	1.47 (0.118)	**0.67** (0.002)	**0.55** (0.002)	5.78 (0.079)
MSI-Net [38]	**0.88** (0.001)	**0.86** (0.002)	0.82 (0.003)	2.20 (0.010)	2.03 (0.036)	0.64 (0.005)	0.50 (0.005)	5.41 (0.023)
SALICON [115]	0.86 (0.001)	0.85 (0.001)	0.82 (0.001)	1.97 (0.014)	1.81 (0.001)	0.58 (0.003)	0.42 (0.000)	5.57 (0.001)
DeepGaze II [37]	0.86	0.85	**0.83**	1.61	1.36	0.47	0.34	5.89
GazeGAN [39]	0.86	0.84	0.81	2.17	1.21	0.57	0.48	5.97
SAM-ResNet [40]	**0.88**	0.78	0.76	2.37	1.61	0.65	0.54	5.68
SalGAN [116]	**0.88**	0.84	0.82	2.06	1.02	0.63	0.50	**5.08**
ML-Net [67]	0.85	0.77	0.75	2.06	0.88	0.59	0.50	5.31
BMS [117]	0.78	0.77	0.74	1.21	0.34	0.36	0.35	6.01
SROD [118]	0.76	0.75	0.72	1.06	0.17	0.32	0.32	6.12
GBVS [119]	0.82	0.81	0.76	1.34	0.54	0.42	0.38	5.86
IttiKoch [58]	0.75	0.73	0.70	0.96	0.18	0.29	0.33	6.12

TABLE 7
Results on the OSIE Dataset

	AUC-Judd↑	AUC-Borji↑	sAUC↑	NSS↑	IG↑	CC↑	SIM↑	KL↓
CASNet II (ours)	0.89 (0.001)	0.86 (0.001)	0.85 (0.002)	2.49 (0.009)	**2.41** (0.017)	0.78 (0.001)	0.59 (0.006)	**4.91** (0.016)
CASNet I [31]	0.89 (0.002)	0.86 (0.003)	0.84 (0.003)	2.33 (0.029)	2.20 (0.030)	0.74 (0.011)	0.60 (0.007)	5.03 (0.020)
N-CASNet	0.88 (0.002)	0.86 (0.002)	0.85 (0.002)	2.33 (0.010)	2.12 (0.012)	0.73 (0.001)	0.50 (0.003)	5.12 (0.008)
EML-NET [36]	**0.90** (0.001)	0.84 (0.003)	0.83 (0.003)	2.71 (0.007)	1.59 (0.120)	**0.80** (0.002)	**0.67** (0.001)	5.38 (0.075)
MSI-Net [38]	**0.90** (0.002)	0.86 (0.002)	0.85 (0.002)	2.45 (0.012)	2.10 (0.050)	0.78 (0.003)	0.64 (0.004)	5.08 (0.033)
SALICON [115]	0.89 (0.001)	0.87 (0.001)	0.85 (0.001)	2.23 (0.015)	2.20 (0.005)	0.72 (0.003)	0.53 (0.001)	5.06 (0.004)
DeepGaze II [37]	**0.90**	**0.89**	**0.88**	1.87	1.81	0.60	0.44	5.35
GazeGAN [39]	0.88	0.86	0.84	2.17	1.06	0.70	0.59	5.77
SAM-ResNet [40]	0.89	0.77	0.77	**2.68**	1.45	0.77	0.65	5.46
SalGAN [116]	0.89	0.85	0.84	2.29	2.17	0.74	0.62	5.05
ML-Net [67]	0.89	0.78	0.77	2.53	2.10	0.75	0.62	5.06
BMS [117]	0.83	0.81	0.79	1.41	1.68	0.46	0.43	5.46
SROD [118]	0.81	0.80	0.78	1.33	1.50	0.44	0.40	5.58
GBVS [119]	0.81	0.80	0.76	1.30	1.61	0.44	0.42	5.49
IttiKoch [58]	0.76	0.75	0.72	1.02	1.39	0.34	0.39	5.65

ASPP structure) significantly outperforms CASNet I (without ASPP structure) on all five benchmark datasets. CASNet II is consistently better on all three AUC metrics, NSS, and CC. These results demonstrate the efficacy of ASPP structure in learning contextual saliency. Meanwhile, using a single stream framework with ASPP structure, CASNet II also has higher computing efficiency compared to the dual-stream based CASNet I—it only

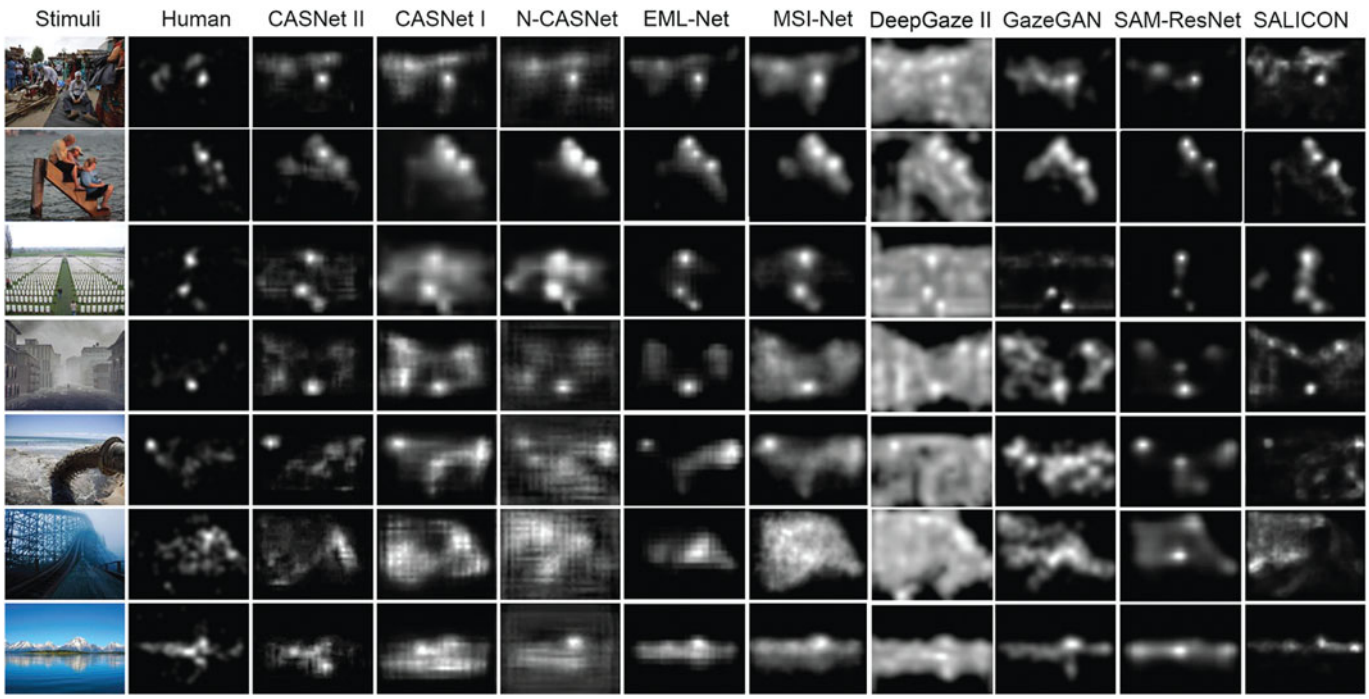


Fig. 8. Qualitative results generated by our saliency model in comparison with state-of-the-art methods. Our model (CASNet II) outperforms other models in both location and order, by taking into consideration contextual information (e.g., encoding relative importance of occurring faces in the first two rows, objects in the third and fourth row, and highlighting areas of interest in scene images in the last three rows). Due to space limit, we only show examples from nine DNN-based models, which are top performers on EMOd dataset.

takes 0.09 s for CASNet II to process one image whereas CASNet I needs 0.25 s on the same NVIDIA 1080Ti GPU. We further tested the model with only the ASPP structure (i.e., without the channel-weighting subnetwork) on five benchmark datasets. Result show that the ASPP structure alone is able to raise the saliency prediction performance, but it achieves its best performance when accompanied with the channel weighting subnetwork (see Tables S6 - S10 in the supplementary material), available online.

To better demonstrate the contribution of the ASPP structure and channel weighting subnetwork, we compute an average score over all metrics for different models in the ablation study using the same approach as described in Section 5.6. Table 9 reports a summary of the means of different models on five datasets. As seen from the table, both ASPP structure and channel weighting subnetwork boost performance (with a 100% or above increment on averaged metrics) in saliency prediction on all five benchmarks except the OSIE dataset. As most OSIE images have clear focal objects with a clean background [35], the ASPP structure and subnetwork do not contribute much. Readers can refer to the supplementary material for details, available online.

5.8 Visualizations and Discussion

In this subsection, we explore how the models encode emotion prioritization through quantitative analyses and visualizations.

Emotion Prioritization. Do the models exhibit emotion prioritization like humans do? To see, we perform the same analyses as in Section 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 II. The result (Fig. 10) is similar to Fig. 3. 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) show that emotion type significantly influences the predicted saliency score $F(2, 2534) = 81.22$, $p < .001$, supporting the emotion prioritization effect of CASNet II.

We repeat the computation process of Fig. 10 for CASNet I and N-CASNet, as well as on the three best performing comparison methods (DeepGaze II, EML-NET

TABLE 8
Normalized Means of all Z-Scored Metrics (AUC-Judd, AUC-Borji, sAUC, NSS, IG, CC, SIM, KL)

Model	EMOd	NUSEF	CAT2000	MIT1003	OSIE
CASNet II	**0.97**	**0.96**	**0.87**	**0.90**	**0.84**
CASNet I	0.73	0.58	0.50	0.71	0.64
N-CASNet	0.45	0.48	-0.43	0.18	0.74
EML-NET	0.08	-0.25	-0.20	0.66	0.47
MSI-Net	0.71	0.59	0.58	0.89	0.74
SALICON	0.66	0.68	0.44	0.46	0.56
DeepGazeII	0.16	-0.06	0.86	-0.05	0.20
GazeGAN	0.01	0.30	0.21	0.26	-0.15
SAM-ResNet	-0.32	-0.38	0.38	0.41	-0.07
SalGAN	0.53	0.46	0.32	0.72	0.60
ML-Net	0.15	-0.15	-0.46	0.07	0.23
BMS	-0.83	-0.65	-0.53	-1.20	-0.82
SROD	-1.34	-1.00	-0.67	-1.58	-1.10
GBVS	-0.42	-0.02	0.03	-0.68	-1.08
IttiKoch	-1.53	-1.54	-1.90	-1.74	-1.80

Our model (CASNet II) achieves the top performance on all five benchmark datasets.

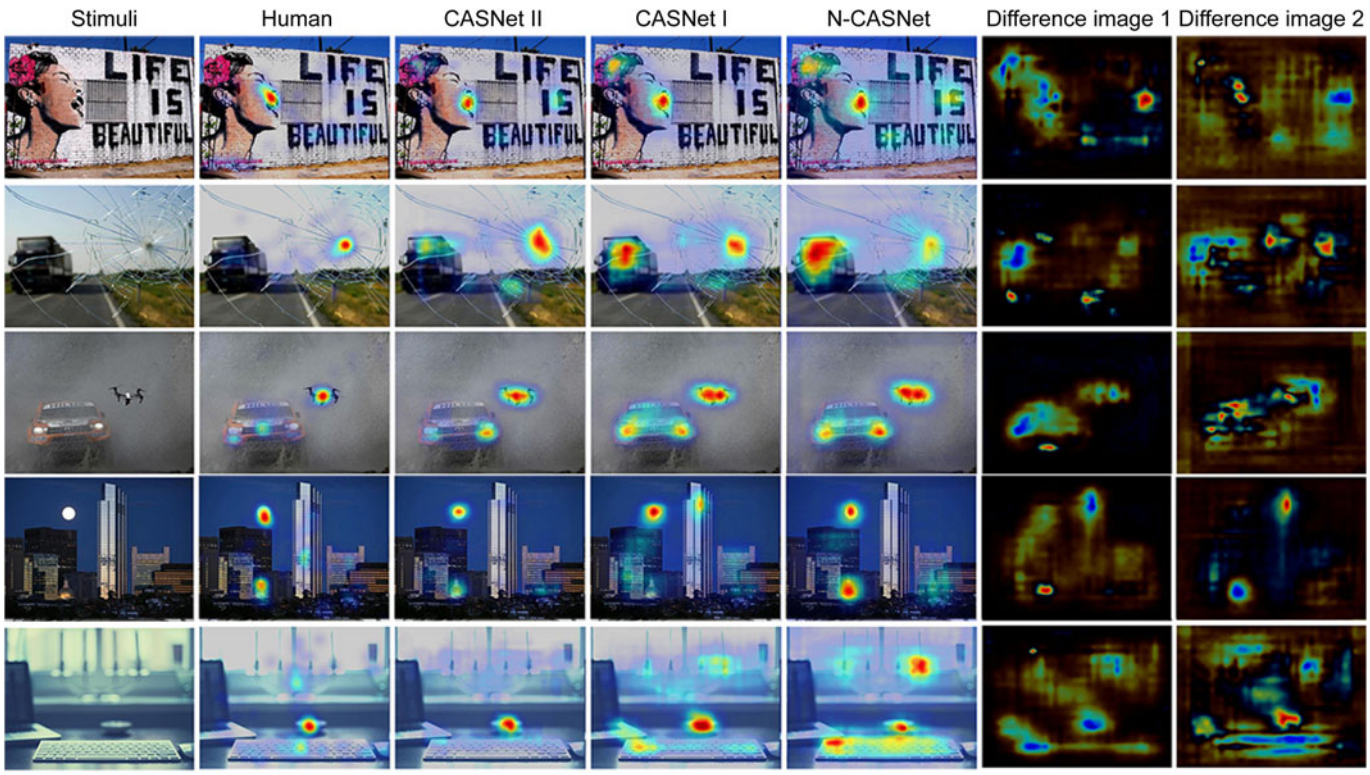


Fig. 9. Examples of how our models gradually improve the relative saliency among different objects in a scene, to closely match human emotion prioritization. The last two columns visualize the difference between predictions from CASNet II and CASNet I (difference image 1), and CASNet I and N-CASNet (difference image 2): colors close to orange/red indicate increased saliency after applying the subnetwork for contextual saliency, whereas colors close to blue/green indicate decreased saliency.

and MSI-Net). An ANOVA (object saliency scores as the dependent variable, object emotion types as the independent variable) for each model indicates that the comparison models have a similar behavior to prioritize emotional objects ($Fs \geq 80.21$), but such effect is not as strong as CASNet II (indicated by a larger ANOVA F-value of CASNet II over other methods, CASNet II: 129.26, CASNet I: 92.17, N-CASNet: 87.95, DeepGaze II: 80.21, EML-Net: 90.28, MSI-Net: 84.87).

Finally, we perform similar analyses as Fig. 3 for a) images with both emotion-eliciting and neutral objects distributed over large image regions, and b) images with gazing cues, for CASNet II, CASNet I and N-CASNet. Results indicate that CASNet II has the strongest emotion prioritization effect in the above images, suggesting the advantage of the ASPP structure in capturing the emotion-eliciting characteristics in scenes with interacting objects or spread-out objects. The detailed statistics and visualizations are reported in the supplementary material, available online.

TABLE 9
Normalized Means of all Z-Scored Metrics (AUC-Judd, AUC-Borji, sAUC, NSS, IG, CC, SIM, KL) of Four Model Versions on Five Benchmark Datasets

Model	EMOd	NUSEF	CAT2000	MIT1003	OSIE
CASNet II	**0.72**	**1.04**	**0.80**	**0.51**	0.35
CASNet I	0.36	-0.08	0.30	0.10	0.25
CASNet	0.07	0.13	0.28	0.15	**0.36**
N-CASNet	-1.15	-1.09	-1.38	-0.77	-0.95

CASNet II contains both channel-weighting subnetwork and ASPP structure, CASNet I has channel-weighting subnetwork only, CASNet only has ASPP, and N-CASNet contains neither channel-weighting subnetwork nor ASPP structure.

DNN Visualization. We first visualize our network to analyze the efficacy of the ASPP structure. For CASNet II (with channel-weighting subnetwork and ASPP structure), CASNet I (with the channel-weighting subnetwork only), and N-CASNet (without the channel-weighting subnetwork or ASPP structure), we identify the five feature maps before the last fully convolutional layer with the highest weights for each image. Results show that the feature maps from CASNet II are more refined than those of CASNet I and are closer to human groundtruth. This advantage is more obvious for images with multiple focal points, and images with relatively small focal areas (see Fig. 11). For a better visualization, we combine each feature map with the original stimuli to form an interim heatmap. Due to space limit, only 3 interim heatmaps with from each model are shown. These observations suggest that the ASPP structure in CASNet II allows for larger receptive fields and more resolution scales, thus enabling the model to learn the contextual saliency within a larger area in the image and model human attention more precisely for the whole scene. Meanwhile, the channel-weighting subnetwork help re-direct the attention to the emotional areas.

We perform additional visualizations to examine the models' ability in emotion prioritization. For each image, we extract the top 5 patches with highest and lowest responses, respectively, after the fully connected convolutional layer for

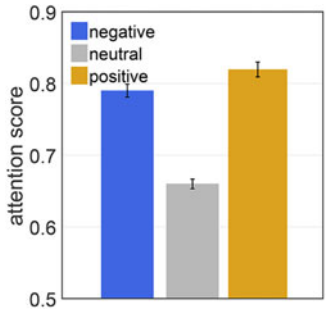


Fig. 10. Emotional objects are predicted as being more salient than neutral objects by CASNet II, which is consistent with the emotion prioritization effect of human observers.

CASNet II, CASNet I and N-CASNet. As illustrated in Fig. 12, the highest response patches of CASNet II and CASNet I show stronger emotions (e.g., emotional faces, scenes of sunrise) than those of N-CASNet, and there is a larger difference in emotion-eliciting content between the highest and lowest patches in CASNet II than in CASNet I and N-CASNet. These observations suggest that the ASPP structure and channel-weighting subnetwork in CASNet II more effectively directs the model's attention to emotion-eliciting content.

We corroborate Fig. 12 with quantitative analyses. For the three model versions (CASNet II, CASNet I and N-CASNet), we calculate the means of saliency scores of positive, negative, and neutral objects that fall within the selected patches. ANOVAs indicate a significant effect of model type on the saliency scores on all three types of objects, $F(2, 3054)$s $\geq 46.85, p$s $< .001$. Post hoc Tukey tests suggest that the highest patches of CASNet II have higher saliency scores than those of CASNet I and N-CASNet (ps $< .001$) for both emotion-eliciting and emotionally-neutral objects. Separate paired samples t-tests within each model show a significantly higher average saliency score for the patches with the highest response compared to those with the lowest response (ps $< .001$) for CASNet II and CASNet I, but not for N-CASNet. The above analyses suggest that CASNet II has higher emotion prioritization ability than CASNet I and N-CASNet. They also show that the advantages of the ASPP structure and channel weighting subnetwork can be generalized to all objects (i.e., not limited to emotion-eliciting content.)

Cross Datasets Performance. We perform additional experiments to test the performance across emotional and non-emotional datasets. We explore if existing approaches trained on EMOd dataset will improve their emotion prioritization performance. More specifically, we first train our three models and three comparison methods (EML-NET, MSI-Net, and SALICON, whose codes are publicly available) on SALICON dataset and fine-tune them on EMOd. We then test them on the NUSEF dataset which focuses on affective content. We further test the above models on non-emotional datasets MIT1003 and OSIE. The results are reported in Tables S11-S13 in the supplementary material, available online. There is no evident performance boost in general ($|t|$s ≤ 1.56, ps $\geq .16$). However, paired samples t-tests indicate a significant increase on NSS for CASNet II, CASNet I and SALICON on MIT1003 ($t(2)$s ≥ 4, ps $< .05$) and OSIE ($t(2)$s ≥ 17, ps $\leq .003$) datasets, suggesting that fine-tuning on EMOd helps these models do better on attention prioritization [71]. This is especially useful for certain applications like content-aware image re-targeting and image rendering, where a high NSS is preferred [71]. Readers can refer to the supplementary material for more discussions, available online.

6 Conclusion

Selective attention is intrinsic to human vision. In this paper we propose EMOd—a new emotional-attention dataset for research on selective attention due to emotion-eliciting content. Analyses on EMOd show that eye fixations correlate with human affective responses to the visual content of the

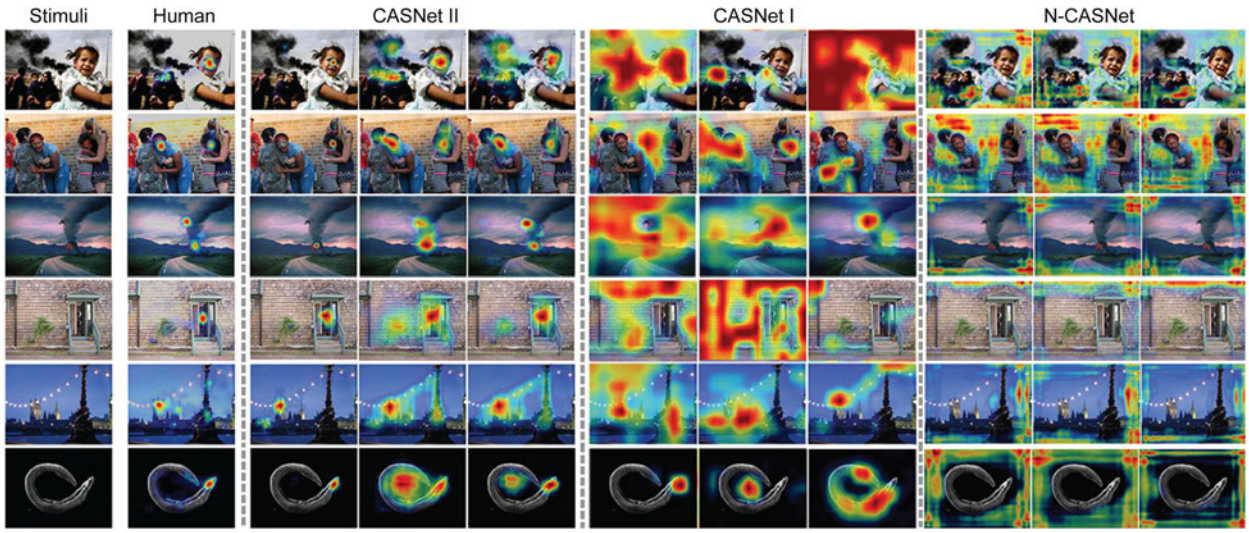


Fig. 11. This figure shows the interim heatmaps with the highest weights from CASNet II, CASNet I and N-CASNet before the last fully convolutional layer. The heatmaps from CASNet II are closer to human groundtruth than CASNet I on images with multiple human focuses (first three rows), and images with relatively small focused areas (last three rows). This suggests that the ASPP structure in CASNet II allows for larger receptive fields with more resolution scales, thus enabling the model to learn the contextual saliency within a larger area in the image and capture human attention more precisely for the whole scene. The results of CASNet I are closer to human groundtruth than N-CASNet, suggesting the channel-weighting subnetwork help re-direct the attention to the emotional areas.

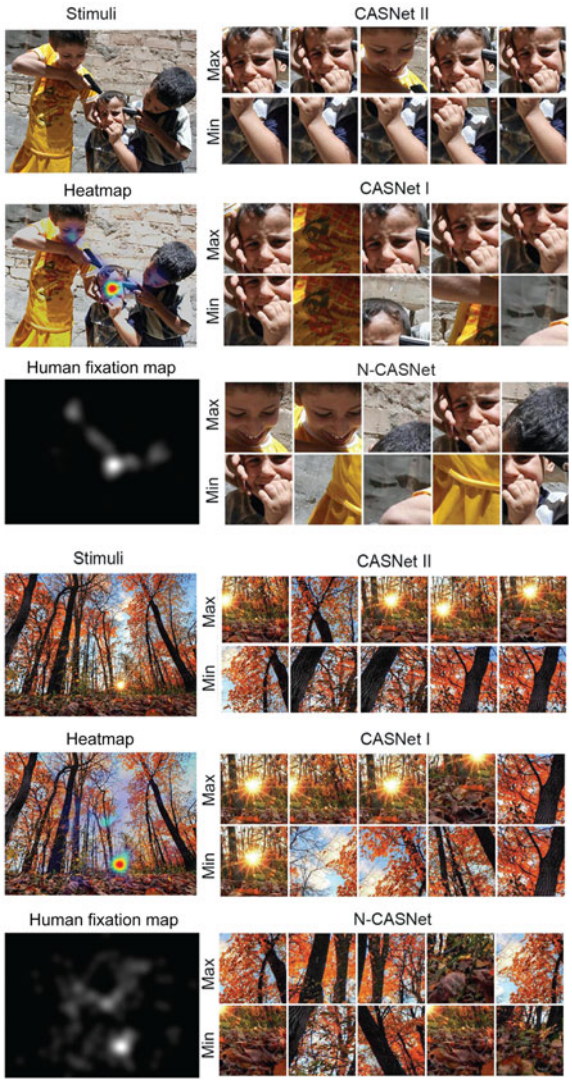


Fig. 12. Examples of neuron responses before the last fully connected convolutional layer. For each image, the 5 patches on the first row are the high activations of the channels with the largest weights, and those on the second row are the high activations of the channels with minimum weights. The highest response patches of CASNet II show stronger emotions (e.g., emotional faces, scenes of sunrise) than those of CASNet I and N-CASNet. The differences in emotion-eliciting content between the highest and lowest patches increases from N-CASNet to CASNet II. These observations suggest that the ASPP structure and channel-weighting subnetwork effectively direct CASNet II's attention to emotion-eliciting content.

images at both object- and scene-levels. We design a deep learning model (CASNet II) to computationally model the human attention behavior. The model, with a much simpler structure but carefully designed to encode emotion prioritization, achieves the top performance on five benchmark datasets when evaluated by the normalized mean of all metrics. This suggests that understanding human behavior helps create simple yet effective computational models.

Our research distinguishes itself from other investigations into human attention by its comprehensive analyses on the relationships among human affective responses and visual attention on complex scenes, with a DNN model that effectively mimics human attention in this context. The analysis framework and the resulting findings not only provide unique contributions toward understanding human visual attention, but also have a variety of related applications, such as improving computer vision deep learning models, emotion-aware robots, and online advertising.

Acknowledgments

We thank Dr. Tian-Tsong Ng, Dr. Chenyao Shen, and Ms. Juan Xu for their contribution to this work, and all the reviewers for our papers on this topic. We are grateful to Departmental Ethics Review Committee of NUS for their aid on the IRB application. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore. Shaojing Fan and Zhiqi Shen are equal contribution to this work.

References

[1] J. Duncan, "Selective attention and the organization of visual information," *J. Exp. Psychol.: Gen.*, vol. 113, no. 4, 1984, Art. no. 501.
[2] R. Marois and J. Ivanoff, "Capacity limits of information processing in the brain," *Trends Cogn. Sci.*, vol. 9, no. 6, pp. 296–305, 2005.
[3] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Front. Cogn. Neurosci.*, vol. 229, pp. 342–345, 1985.
[4] G. R. Mangun, "Neural mechanisms of visual selective attention," *Psychophysiology*, vol. 32, no. 1, pp. 4–18, 1995.
[5] M. Poletti, M. Rucci, and M. Carrasco, "Selective attention within the foveola," *Nature Neurosci.*, vol. 20, no. 10, 2017, Art. no. 1413.
[6] J. Birulés, L. Bosch, R. Brieke, F. Pons, and D. J. Lewkowicz, "Inside bilingualism: Language background modulates selective attention to a talker's mouth," *Devlop. Sci.*, vol. 22, no. 3, 2019, Art. no. e12755.
[7] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.
[8] M. Spain and P. Perona, "Some objects are more equal than others: Measuring and predicting importance," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 523–536.
[9] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vis. Res.*, vol. 50, no. 22, pp. 2233–2247, 2010.
[10] T. Brosch, G. Pourtois, D. Sander, and P. Vuilleumier, "Additive effects of emotional, endogenous, and exogenous attention: Behavioral and electrophysiological evidence," *Neuropsychologia*, vol. 49, no. 7, pp. 1779–1787, 2011.
[11] A. Borji, D. N. Sihite, and L. Itti, "Objects do not predict fixations better than early saliency: A re-analysis of einhäuser *et al.*'s data," *J. Vis.*, vol. 13, no. 10, pp. 18–18, 2013.
[12] C. Bundesen, S. Vangkilde, and A. Petersen, "Recent developments in a computational theory of visual attention (TVA)," *Vis. Res.*, vol. 116, pp. 210–218, 2015.
[13] A. M. Treisman, "Strategies and models of selective attention," *Psychol. Rev.*, vol. 76, no. 3, 1969, Art. no. 282.
[14] R. J. Compton, "The interface between emotion and attention: A review of evidence from psychology and neuroscience," *Behav. Cogn. Neurosci. Rev.*, vol. 2, no. 2, pp. 115–129, 2003.
[15] M. G. Calvo and P. J. Lang, "Gaze patterns when looking at emotional pictures: Motivationally biased attention," *Motivation Emotion*, vol. 28, no. 3, pp. 221–243, 2004.
[16] P. Vuilleumier, "How brains beware: Neural mechanisms of emotional attention," *Trends Cogn. Sci.*, vol. 9, no. 12, pp. 585–594, 2005.
[17] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1960–1968.
[18] P. Barros, G. I. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: A deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, 2017.
[19] 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," *Psychol., Crime Law*, vol. 19, no. 1, pp. 35–66, 2013.
[20] H. Pashler, *Attention*. East Sussex, U.K.: Psychology Press, 2016.

[21] A. Öhman, A. Flykt, and F. Esteves, "Emotion drives attention: Detecting the snake in the grass," *J. Exp. Psychol.: Gen.*, vol. 130, no. 3, pp. 466–474, 2001.

[22] R. Gupta, "Commentary: Neural control of vascular reactions: Impact of emotion and attention," *Front. Psychol.*, vol. 7, 2016, Art. no. 1613.

[23] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 30–43.

[24] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

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

[26] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9223.

[27] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–14.

[28] S. Shomstein, G. L. Malcolm, and J. C. Nah, "Intrusive effects of task-irrelevant information on visual selective attention: Semantics and size," *Curr. Opin. Psychol.*, vol. 29, pp. 153–159, 2019.

[29] R. T. Wilson, D. W. Baack, and B. D. Till, "Creativity, attention and the memory for brands: An outdoor advertising field study," *Int. J. Advertising*, vol. 34, no. 2, pp. 232–261, 2015.

[30] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2942–2950.

[31] S. Fan *et al.*, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7521–7531.

[32] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Inferring attention shift ranks of objects for image saliency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12133–12143.

[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[34] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[35] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 28–28, 2014.

[36] S. Jia and N. D. Bruce, "EML-NET: An expandable multi-layer network for saliency prediction," *Image Vis. Comput.*, vol. 95, 2020, Art. no. 103887.

[37] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4789–4798.

[38] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder-decoder network for visual saliency prediction," *Neural Netw.*, vol. 129, pp. 261–270, 2020.

[39] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? Dataset and model," *IEEE Trans. Image Process.*, vol. 29, pp. 2287–2300, 2019.

[40] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.

[41] A. S. Cauquil, G. E. Edmonds, and M. J. Taylor, "Is the face-sensitive N170 the only ERP not affected by selective attention?," *Neuroreport*, vol. 11, no. 10, pp. 2167–2171, 2000.

[42] M. L. Furey *et al.*, "Dissociation of face-selective cortical responses by attention," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 4, pp. 1065–1070, 2006.

[43] M. G. Calvo, A. Gutiérrez-García, and A. Fernández-Martín, "Time course of selective attention to face regions in social anxiety: Eye-tracking and computational modelling," *Cogn. Emotion*, vol. 33, no. 7, pp. 1481–1488, 2019.

[44] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annu. Rev. Neurosci.*, vol. 23, no. 1, pp. 315–341, 2000.

[45] T. Brosch, G. Pourtois, and D. Sander, "The perception and categorisation of emotional stimuli: A review," *Cogn. Emotion*, vol. 24, no. 3, pp. 377–400, 2010.

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

[47] V. Yanulevskaya *et al.*, "In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 349–358.

[48] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher, "Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes," *J. Vis.*, vol. 14, no. 3, pp. 31–31, 2014.

[49] H. Zheng, T. Chen, Q. You, and J. Luo, "When saliency meets sentiment: Understanding how image content invokes emotion and sentiment," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 630–634.

[50] L. McCay-Peet, M. Lalmas, and V. Navalpakkam, "On saliency, affect and focused attention," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 541–550.

[51] S. D. McCrackin, S. K. Soomal, P. Patel, and R. J. Itier, "Spontaneous eye-movements in neutral and emotional gaze-cuing: An eye-tracking investigation," *Heliyon*, vol. 5, no. 4, 2019, Art. no. e01583.

[52] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[53] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[54] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Process.*, vol. 93, no. 6, pp. 1401–1407, 2013.

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

[56] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, 2018, Art. no. 38.

[57] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

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

[59] 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 Vis.*, vol. 13, no. 2, pp. 201–214, 2000.

[60] 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, Psychophys.*, vol. 71, no. 6, pp. 1337–1352, 2009.

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

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

[63] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3183–3192.

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

[65] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," 2015, *arXiv:1510.02927*.

[66] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

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

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

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

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

[71] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 809–824.

[72] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1147–1154.

[73] E. B. Roesch, D. Sander, C. Mumenthaler, D. Kerzel, and K. R. Scherer, "Psychophysics of emotion: The quest for emotional attention," *J. Vis.*, vol. 10, no. 3, pp. 4–4, 2010.

[74] 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 *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 729–732.

[75] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 512–528.

[76] X. Ding, L. Huang, B. Li, C. Lang, Z. Hua, and Y. Wang, "A novel emotional saliency map to model emotional attention mechanism," in *Proc. Int. Conf. Multimedia Model.*, 2016, pp. 197–206.

[77] M. O. Cordel, S. Fan, Z. Shen, and M. S. Kankanhalli, "Emotion-aware human attention prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4026–4035.

[78] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.

[79] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2755–2766, Nov. 2020.

[80] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232.

[81] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 1–4.

[82] Z. Bylinskii *et al.*, "Mit saliency benchmark," 2019. [Online]. Available: http://saliency.mit.edu/

[83] P. Lebreton, T. Mäki, E. Skodras, I. Hupont, and M. Hirth, "Bridging the gap between eye tracking and crowdsourcing," in *Proc. Hum. Vis. Electron. Imag.*, 2015, Art. no. 93940W.

[84] S. Eraslan, Y. Yesilada, and S. Harper, "Crowdsourcing a corpus of eye tracking data on web pages: A methodology," in *Proc. Measuring Behav.*, 2018, pp. 267–273.

[85] B. Ni *et al.*, "Touch saliency: Characteristics and prediction," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1779–1791, Oct. 2014.

[86] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

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

[88] K. Krafka *et al.*, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2176–2184.

[89] N. W. Kim *et al.*, "BubbleView: An interface for crowdsourcing image importance maps and tracking visual attention," *ACM Trans. Comput.-Hum. Interaction*, vol. 24, 2017, Art. no. 36.

[90] M. Othman, T. Amaral, R. McNaney, J. D. Smeddinck, J. Vines, and P. Olivier, "CrowdEyes: Crowdsourcing for robust real-world mobile eye tracking," in *Proc. 19th Int. Conf. Hum.- Comput. Interaction Mobile Devices Serv.*, 2017, Art. no. 18.

[91] A. Duchowski, *Eye Tracking Methodology: Theory and Practice*, vol. 373. Berlin, Germany: Springer Science & Business Media, 2007.

[92] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, no. 3/4, pp. 169–200, 1992.

[93] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[94] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[95] S. Fan, T.-T. Ng, B. L. Koenig, M. Jiang, and Q. Zhao, "A paradigm for building generalized models of human image perception through data fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5762–5771.

[96] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 419–426.

[97] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 105–108.

[98] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 229–238.

[99] S. Fan, T.-T. Ng, J. S. Herberg, B. L. Koenig, C. Y.-C. Tan, and R. Wang, "An automated estimator of image visual realism based on human cognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4201–4208.

[100] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Univ. Florida, Gainesville. Tech. Rep. A-8, 2008.

[101] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 101–104.

[102] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 20th ACM Int. Conf. Multimedia*, 2010, pp. 83–92.

[103] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.

[104] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[105] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, 2012, pp. 40–46.

[106] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[107] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2429–2437.

[108] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *Int. J. Med. Educ.*, vol. 2, 2011, Art. no. 53.

[109] O. L. Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behav. Res. Methods*, vol. 45, no. 1, pp. 251–266, 2013.

[110] R. A. Bailey, *Design of Comparative Experiments*, vol. 25. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[111] W. Becker and A. F. Fuchs, "Further properties of the human saccadic system: Eye movements and correction saccades with and without visual fixation points," *Vis. Res.*, vol. 9, no. 10, pp. 1247–1258, 1969.

[112] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool, "The interestingness of images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1633–1640.

[113] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.

[114] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.

[115] C. Thomas, "OpenSALICON: An open source implementation of the salicon saliency model," 2016, *arXiv:1606.00110*.

[116] J. Pan *et al.*, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*.

[117] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 153–160.

[118] H. Tang, C. Chen, and X. Pei, "Visual saliency detection via sparse residual and outlier detection," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1736–1740, Dec. 2016.

[119] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

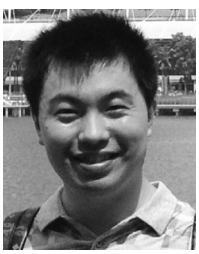
[120] M. Kümmerer *et al.*, "Mit/tübingen saliency benchmark," 2022. [Online]. Available: https://saliency.tuebingen.ai/

[121] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. Baileys Harbor, WI, USA: Peninsula Pub., 1988.

[122] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" 2016, *arXiv:1604.03605*.
[123] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, 2005.
[124] A. Das, R. K. Kumar, D. R. Kisku, and G. Sanyal, "Attention identification via relative saliency of localized crowd faces," in *Proc. 10th Int. Conf. Informat. Syst.*, 2016, pp. 101–106.
[125] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.
[126] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.
[127] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
[128] J. M. Joyce, "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science*, Berlin, Germany: Springer, 2011, pp. 720–722.
[129] M. Kümmerer, T. S. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 52, pp. 16054–16059, 2015.

Shaojing Fan received the BE and ME degrees in communication and information system from the South China University of Technology, China, and the DPhil degree from Institute for Infocomm Research, Singapore, and Ningbo University, China. She is a postdoctoral research fellow with the School of Computing, National University of Singapore (NUS), Singapore. Prior to joining NUS, she was a senior research engineer with Institute for Infocomm Research, part of Singapore's Agency for Science, Technology, and Research. Her main research interests includes cognitive vision, computer vision, and experimental psychology.

Zhiqi Shen received the BE degree in network engineering from the Ningbo University of Technology, China, in 2015. He is currently working toward the fourth-year PhD degree with the School of Computing, National University of Singapore, Singapore. He was an intern student with Institute for Infocomm Research, part of Singapore's Agency for Science, Technology, and Research from 2014 to 2015. His research interest lies in deep learning for computer vision and pattern recognition.



Ming Jiang (Member, IEEE) received the BE and ME degrees in computer science from Zhejiang University, China, and the PhD degree in electrical and computer engineering from the National University of Singapore, Singapore. He is a postdoctoral associate with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota. He is interested in computer vision, cognitive vision, psychophysics and computational neuroscience. His studies focus on computational models of visual attention.

Bryan L. Koenig received the BA degree with a major in psychology and a minor in Latin, from St. John's University, New York, in 1998, the MA degree in general/experimental psychology from the College of William and Mary, Williamsburg, Virginia, in 2005, and the PhD degree in social psychology with a minor in experimental statistics from New Mexico State University, Las Cruces, New Mexico, in 2009. For three years he then worked as a research scientist with the Institute of High Performance Computing, part of Singapore's Agency for Science, Technology, and Research. Until recently, he was an adjunct instructor with Washington University in St. Louis, St. Louis, Missouri. He is currently an assistant professor with the Department of Psychology, Southern Utah University, Cedar City, Utah. He does research on social perception, emotions, morality, and evolutionary psychology.

Mohan S. Kankanhalli (Fellow, IEEE) received the BTech degree from IIT Kharagpur, India, and MS and PhD degrees from the Rensselaer Polytechnic Institute, Troy, New York. He is provost's chair professor of computer science with the National University of Singapore (NUS), Singapore. He is the dean of NUS School of Computing and he also directs N-CRiPT (NUS Centre for Research in Privacy Technologies) which conducts research on privacy on structured as well as unstructured (multimedia, sensors, IoT) data. His research interests are in multimedia computing, computer vision, information security & privacy and image/video processing. He has made many contributions in the area of multimedia & vision – image and video understanding, data fusion, visual saliency as well as in multimedia security – content authentication and privacy, multi-camera surveillance.

Qi Zhao (Member, IEEE) received the PhD degree in computer engineering from the University of California, Santa Cruz, in 2009. She is an associate professor with the Department of Computer Science and Engineering, University of Minnesota, Twin Cities. Her main research interests include computer vision, machine learning, cognitive neuroscience, and healthcare. She was a postdoctoral researcher with the Computation & Neural Systems, California Institute of Technology from 2009 to 2011. Before joining the University of Minnesota, she was an assistant professor with the Department of Electrical and Computer Engineering and the Department of Ophthalmology, National University of Singapore, Singapore. She has published more than 100 journal and conference papers, and edited a book with Springer, titled Computational and Cognitive Neuroscience of Vision, that provides a systematic and comprehensive overview of vision from various perspectives. She serves as an associate editor for *IEEE Transactions on Neural Networks and Learning Systems* and *IEEE Transactions on Multimedia*, as a program chair for WACV'22, and as an organizer and/or area chair for CVPR and other major venues in computer vision and AI regularly.