Effects of Object Sentiment on Human Attention

Anonymous CVPR submission

Paper ID 1867

Abstract

Despite the recent success in face recognition and object classification, in the field of human gaze prediction, computer models are still struggling to accurately mimic human attention. One main reason is that visual attention is a complex human behavior influenced by multiple factors, ranging from low-level features (e.g., color, contrast) to high-level human perception (e.g., objects interactions, object sentiment), making it difficult to model computationally. In this work, we investigate the relation between object sentiment and human attention. We first introduce a new evaluation metric (AttI) for measuring human attention that focuses on human fixation consensus. A series of empirical data analyses with AttI indicate that emotion-evoking objects receive attention favor, especially when they co-occur with emotionally-neutral objects, and this favor varies with different image complexity. Based on the empirical analyses, we design a deep neural network for human attention prediction which allows the attention bias on emotion-evoking objects to be encoded in its feature space. Experiments on two benchmark datasets demonstrate its superior performance, especially on metrics that evaluate relative importance of salient regions. This research provides the clearest picture to date on how object sentiments influence human attention, and it makes one of the first attempts to model this phenomenon computationally.

1. Introduction

Predicting where humans will look in a scene (*i.e.*, saliency prediction) has attracted a significant amount of research because of its potential applications such as in social advertising and robot vision. Classic bottom-up methods like the Itti-Koch model [18] and Graph-Based Visual Saliency (GBVS) [15] use low-level features such as color, intensity, and orientations. Recently, numerous saliency models based on deep neural network (DNN) have been proposed with largely improved performance [17, 6, 20, 7, 22]. Although these DNN-based models are trained with massive amounts of labeled data [8, 19] and

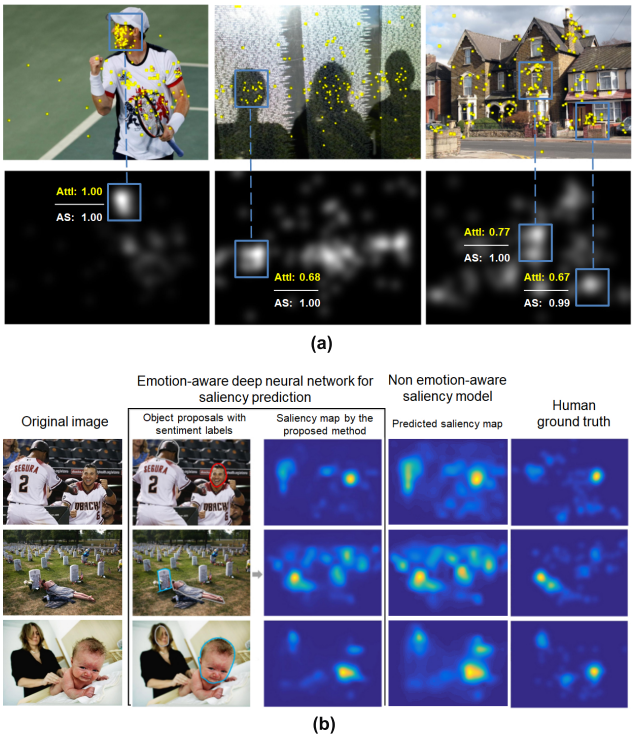

Figure 1. (a) We first introduce a new metric (AttI) for evaluating human attention. Analyses show that AttI can better reflect human attention consensus on images of various complexity compared with previously used Attention Score (AS). (b) Motivated by our empirical data analyses, we propose the Emotion-Aware saliency model (EASal) that incorporates information of object proposals with sentiment labels (red, gray, blue indicate positive, neutral, and negative sentiments, respectively).

equipped with superior object recognition ability, there is still a large gap between their predictions and ground truth human fixations. Aiming to address this limitation, Bylinskii and colleagues [4] re-examined current saliency models, and argued that to approach human-level performance, saliency models need to discover high-level image concepts, such as text or motion, and reason about the relative importance of image regions.

One possible high-level image concept is emotion, a fac-

tor known to influence human attention [12, 35]. Behavioral observations show that people pay attention to affective rather than neutral stimuli, and this commonly happens spontaneously [23]. In a visual search task, the object can be easily found if it contains affective value [35], *e.g.*, a snake among flowers. Initial work has been introduced to explore saliency prediction with emotion information [24, 11]. However, the work in [24] did not analyze how exactly emotion influences human attention, and [11] did not identify emotion-evoking objects or regions. Indeed, more efforts are needed to study how emotion can be used in predicting human attention.

Our work is motivated by the above research but we go further. Aiming to have a deeper look at emotional attention, we first evaluate quantitatively how different emotion-eliciting objects impact human attention. We propose a new evaluation metric (AttI) which focuses on human fixation consensus and study it under various image complexity. With AttI, we discover that emotion-evoking objects are prioritized in human attention, and such prioritization effect is modulated by image complexity. Moreover, it is most significant when they co-occur with emotionally neutral objects. Based on the human findings, we propose a DNN-based model that identifies emotion-evoking objects in an image and incorporates such information in saliency prediction. Results on two benchmark datasets show that the proposed emotion-aware saliency model outperforms other state-of-the-art methods, especially in terms of predicting the relative importance of salient regions within an image (see Fig. 1). Our main contributions are as follows:

1. *We introduce a new metric to evaluate human attention*. The new metric Attention Index (AttI), which focuses on human consensus of fixation, allows us to better examine the relation between attention and object sentiments.

2. *We provide a comprehensive picture on how objects with different sentiments compete for human attention under different scene complexity*.

3. *We introduce an emotion-aware DNN model for predicting human attention that utilizes object sentiment information*. With a control signal based on detected image complexity and context, the new model conditionally integrates the predicted emotion information in the final saliency map.

2. Related Work

Emotional attention: In psychology, human attention is considered as a state of arousal, during which human brains selectively concentrate on a discrete aspect of information, while ignoring others [1]. Due to their evolutionary salience, threat- and reward-related stimuli, such as snakes, angry faces, and delicious food constitute a special class of stimuli believed to capture human attention in a rapid, or even involuntary manner [25]. This "automatic" capture of attention is supported by research in neuroscience, which has unraveled neural pathways for emotional stimuli processing [36]. The above works lead us to look at emotion in saliency prediction.

Predicting human attention: The legacy approaches of saliency prediction [18, 15] are based on Feature Integration Theory of attention which suggests that features are registered automatically and in parallel across the visual field. In recent DNN-based saliency models, human attention at different resolutions is assembled in SALICON [17] and DeepFix [20]. Some models incorporate the human central fixation bias in their system either by superimposing the center priors [21, 6] or by learning [7]. These models took advantage of the representational power of the semantic-rich DNN feature detectors trained on ImageNet [8]. Although these models have largely boosted the performance for saliency prediction, they are mainly trained on the datasets and learn weights as a whole, enabling few insights on how different objects in an image compete for human attention. Different from existing networks, our study focuses on the relative importance of salient regions.

Predicting emotional regions in images: Peng et al. [28] and Sun et al. [33] both introduced systems which predict the affective regions in an image. The authors in [28] proposed the prediction of Emotion Stimuli Map which estimates the pixel-wise contribution to evoked emotion of an image. The work in [33] used object proposals and emotion score to determine an affective region. More recently, Yang et al. [38] presents a system for automatic identification of the affective region using image-level label. These studies demonstrate that emotion-evoking region are predictable. However, how the resulting emotion-evoking maps/regions reflect visual saliency remains unclear. Our work aims to bridge the two types of knowledge through empirical data analyses and computational modeling.

Attention prediction with emotion Saliency researchers have made initial attempts to incorporate emotion in attention prediction, such as the human fixation datasets featuring emotional contents [29, 11], saliency models that identifies emotional objects, such as faces [31, 24], injury, worm and snake [24]. The work closest to ours is [11], in which the authors made a preliminary study on the relation between image sentiment and visual saliency, and reported an emotion prioritization effect for emotion-eliciting content. Their work is insightful but it also leads to the following unresolved questions: (1) the authors used "object attention score" (the maximum fixation-map value inside the objects contour) to measure an object's attention level. This is inadequate as the fixation map was normalized, leading to a situation that each object receives an attention score close to 1 in a very scattered fixation map (see Fig. 1 (a)); (2)

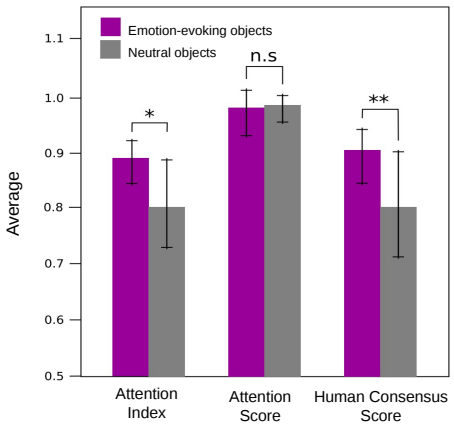

Figure 2. The proposed evaluation metric Attention Index (AttI) which focuses on human consensus is able to identify more intricate differences than attention score (AS). For images containing fewer than 4 labeled objects in EMOd dataset, AttI indicates a significant difference between the emotion-evoking objects and neutral objects whereas AS shows none. The difference on AttI is mainly due to the difference on Human Consensus Score. In all figures in this paper, error bars represent the standard error of means. The asterisks are denoted as following: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.0001$.

the proposed saliency model in [11] did not identify specific object sentiments, thus it is hard to say if the performance improvement is due to emotion or due to other factors (e.g. more information learned about semantics, spatial location, and so on). In our work, we focus on human consensus in measuring each object's attention level. We show that factoring human consensus allows us to better describe the relative attention level of emotion-evoking objects and emotionally-neutral objects under various image complexities. We use our human findings to guide the design of an emotion-aware, DNN-based saliency model with object sentiment masks.

3. Empirical data analyses on emotion and attention

In this section, we first introduce a new evaluation metric (AttI) for measuring human attention. We then conduct a series of statistical analyses using this metric. The analyses were performed on the EMOtional attention dataset (EMOd) proposed in [11]. EMOd contains 1019 emotional images each with detailed object sentiment label (positive, negative and neutral) and has eye fixation data collected from 16 human subjects.

3.1. Attention Index: a new metric for evaluation human attention

In the field of saliency research, most of the previous studies [10, 37, 11] used the attention score (hereafter *AS*) of an object defined as the maximum value of the normalized fixation map inside the object's contour. However, when comparing the attention levels of objects among different images each with various complexities[1], the relative importance of different objects are concealed due to the normalization procedure during the fixation map generation. For example, let's consider two images. The first image with several objects has rather scattered human fixations, and each object receives an AS score close to 1 after the normalization of the fixation map. In contrast, the second image has a single object standing out among others, catching most human attention, thus it will also have an AS score close to 1. The objects in the two images receive different level of human attention but have similar high AS score due to fixation map normalization. In summary, AS alone is inadequate to reflect the human attention level for objects among various images.

To address the above limitation, we propose a new metric to measure human attention, we name it Attention Index (AttI). To do this, we first define *human consensus of fixation score (HCS)*, which measures the consensus of observers' fixation on an object. We adopt the agreement ratio of the Fleiss' kappa [13] for the ith subject as the HCS of n observers for the ith object provided in Eqs. 1 and 2.

$$HCS = \frac{P_i - P_{\text{min}}}{1 - P_{\text{min}}} \tag{1}$$

$$P_i = \frac{1}{n(n-1)}[(\sum_{j=1}^{k} n_{ij}^2) - n], \text{for } n > 1 \tag{2}$$

where P_{min} represents the P_i value for complete disagreement (50% fixated and 50% not fixated), *i.e.*, $\frac{\frac{n^2}{2} - n}{n(n-1)}$. $k = 2$ represents the non-fixated and fixated categories. That is, n_{ij} for $j = 1$ is the number of observers who do not fixate on the ith object and n_{ij} for $j = 2$ is the number of observers who fixate on the ith object. If all observers agree to fixate or do not fixate on an object, HCS is equal to 1.0 and a complete disagreement is equal to 0.0. Using HCS alone is inadequate to indicate an object's attention level, as it generates high scores for objects in both cases when they are fixated or not fixated by most of the observers. To address this issue, we multiply HCS with AS scores for the final Attention Index (AttI), as shown in Eq. 3. The range of AttI values is from 0.0 to 1.0.

$$AttI = HCS \times AS \tag{3}$$

Fig. 2 illustrates the advantage of AttI over AS in practice. We computed the average AttI, AS and HCS of

[1] In our work, image complexity is determined by the number of objects in the image.

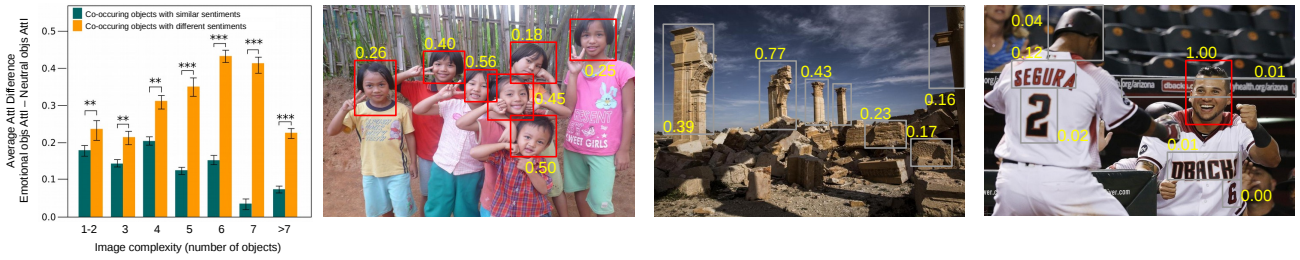

Figure 3. (a) The prioritization effect of emotion-evoking objects is modulated by image complexity, and is most significant when they co-occur with emotionally neutral objects. (b-c) Sample images with co-occurring objects of positive and neutral sentiments, respectively. (d) Sample image with co-occurring objects with different sentiments. Red and gray indicate positive and neutral sentiments, respectively.

emotion-evoking and neutral objects in all EMOd images with low complexity (*i.e.* containing fewer than four labeled objects). Independent-samples t-test indicate no significant difference on the AS of emotional objects and neutral objects ($p = 0.162$). However, independent-samples t-test on AttI reveals that emotion-evoking objects indeed have a higher AttI than neutral objects ($p = 0.010$). As shown in Fig. 2, the significant difference on AttI is mainly carried by the difference on HCS ($p = 0.008$), suggesting the efficacy of considering human consensus.

3.2. Effect of emotion under different image complexity

With the new evaluation metric AttI, we are able to explore in detail how emotion-evoking and neutral objects compete for attention under varying image complexity. We performed a series of analyses using inferential statistics. More specifically, we considered two groups of images: (1) images with co-occurring objects with similar sentiments (*e.g.*, all negative or all positive); (2) images with co-occurring objects with different sentiments (*i.e.*, positive/negative objects co-occur with neutral objects[2]). For the first group, there are 50, 94 and 387 images which contain only positive, negative, neutral objects, respectively. For the second group, there are 137 images which contain both positive and neutral objects, and 342 images which contain both negative and neutral objects. In each group, we classified images into seven subgroups based on the number of labeled objects contained in the image (see Fig. 3). We then computed the average AttI of objects with positive, neutral, and negative sentiment labels, respectively.

As shown in Fig. 3 (a), for images with co-occurring objects with similar sentiments, when the image complexity increases, the difference of AttI between the emotion-evoking objects and the neutral objects is more significant in less complex images. For more complex images (*e.g.*, images with more than 6 objects), the advantage of the emotion-evoking objects for human attention is reduced. This is understandable as when there are too many stimuli that catch human eyes, the effect from an individual stimuli will be weakened. For images with co-occurring emotion-evoking objects and neutral objects, the AttI difference between emotion-evoking objects and neutral objects remains large even when the image complexity increases (see Fig. 3 (d)). This suggests that emotion-evoking objects are most advantageous when they co-occur with neutral objects, and such priority is manifested on human attention consistency regardless of image complexity.

To summarize, in this section we propose a new evaluation metric AttI for measuring human attention, which takes into account human consensus under different image complexity. Our empirical data analyses with AttI indicate that emotion-evoking objects are prioritized in human attention. Such prioritization effect is modulated by image complexity, and is most significant when they co-occur with emotionally neutral objects. Our findings are consistent with previous studies on emotional attention [25, 11], but provide a more nuanced evidence on how the emotion prioritization effect is influenced by image context and complexity. Our findings also guide us in the design of an emotion-aware DNN saliency model, as described in the next section.

4. Emotion-aware saliency prediction

Based on our human findings, we design an emotion-aware DNN model that integrates object sentiment information in saliency prediction. Experiments on two benchmark datasets demonstrate the efficacy of the emotion-aware mechanisms, especially on metrics that measure the relative importance of salient regions.

4.1. DNN architecture

We propose an Emotion-Aware Saliency model (hereafter EASal), as illustrated in Fig. 4. EASal is composed of two branches: (1) semantic feature extraction branch and

[2]There are only 9 images in EMOd that contain both positive and negative objects, which is a very small sample size for us to reasonably detect an effect. Thus, we exclude this case in our statistical analyses.

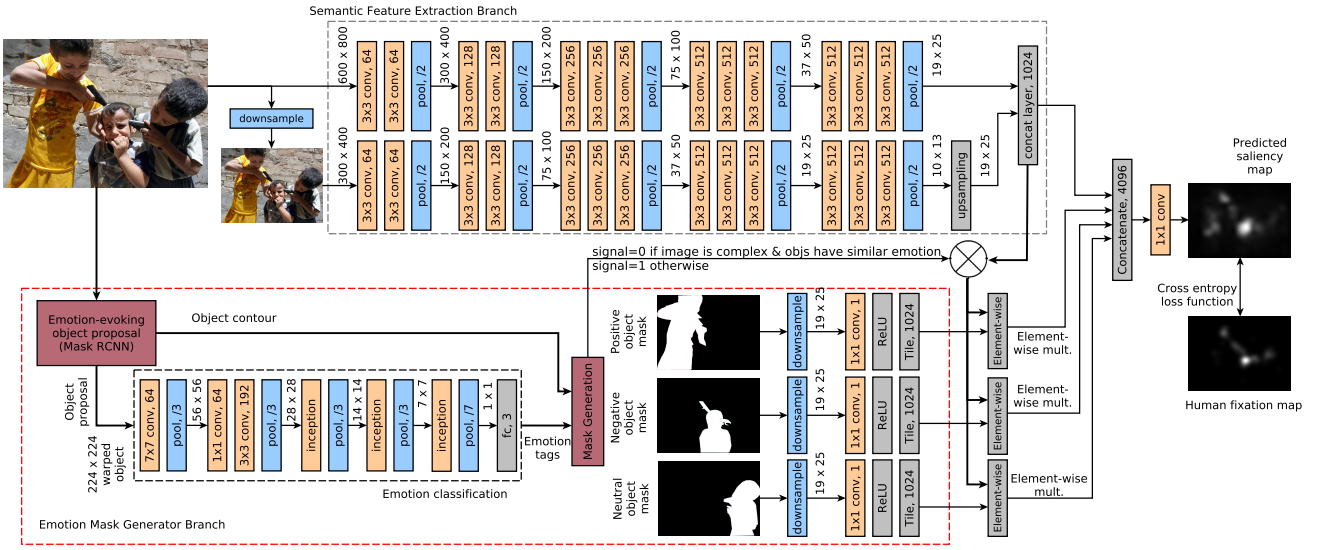

Figure 4. Our proposed Emotion-Aware Saliency model (EASal) is composed of two branches: (1) the semantic feature extraction branch which learns semantic information from the input image, and (2) the object-level emotion mask generation branch which generates and incorporates the objects' emotion masks to the feature maps from the semantic branch. We use a control signal to determine if the two branches should be combined based on detected image complexity and object sentiments. If the signal is on, all feature maps will be combined via the last convolution filter block.

(2) emotion mask generation branch. The *feature extraction branch* learns image semantics and multi-scale information to form semantic feature maps. The *emotion mask generation branch* detects and localizes possible emotion-evoking objects, predicts the emotions evoked by these detected objects, and adjusts the predicted saliency of the corresponding affective regions in the feature maps when a certain condition is met. The combination of the two branches is controlled by a signal which takes into consideration the image complexity and image context (*i.e.*, number of sentiment types co-occurring in the image). Below we describe the details of the two branch and the combination mechanism, which is informed by the empirical data analyses.

Two steps were implemented in the emotion mask generation branch. First, we use Mask-RCNN [16] to generate the contours of the object proposals. The object proposals are then passed to the *emotion classification* to infer the object sentiment (Fig. 4, dashed block in the lower left). The emotion classification network is based on GoogleNet [34] architecture. The emotion mask generation branch outputs three types of important information for human attention: object contour, object position, and object sentiment.

Our empirical data analyses show that the emotion prioritization effect depends on image complexity and image context. Based on this finding, we design a control signal to determine whether the information from emotion mask generation branch should be incorporated in the final saliency map. The signal is set as on by default, but it is automatically turned off when both of the two following situations are met: (1) the image is complex (*i.e.*, more than 6 object proposals are detected within the same image); and (2) the image contains only one type of object sentiment (*i.e.* all detected objects sentiment labels were the same). We determined the threshold for the number of objects in (1) using our empirical data analyses. Experiments on EMOd show that for images with similar object sentiments and contain greater than 6 objects, integrating emotion information will not improve their saliency prediction (see the supplementary material for details).

Parallel to the emotion mask generation branch, the feature extraction branch consists of two VGG-16 [32] modules that capture the object semantics and multi-resolution information. The input image for each module is of size 600×800 and 300×400, respectively. We tested three integration architectures, namely early fusion, intermediate fusion, and late fusion to determine the best way to combine the two branches. We finally selected intermediate fusion as experiments show that it has best performance (refer to the supplementary material for details). During the intermediate fusion, we removed the classifier of the VGG networks such that each module has 512 feature maps as output. The 1024 feature maps are then copied to each element-wise block of the positive, negative and neutral emotion, and multiplied by a scaled version of the emotion masks (1×1 convolution filters) to provide saliency level correction to the corresponding emotion-evoking regions (see the element-wise block in the lower right of Fig. 4). A concatenation layer and a 1×1 convolution filter are used to

Table 1. Quantitative comparison of EASal and other saliency models on EMOd and the affective category of CAT2000. The best score in each metric are highlighted in bold. (↑) indicates higher values are better. (↓) indicates lower values are better.

	Metrics	NSS↑	KL↓	IG↑	EMD↓	AUC-Judd↑	sAUC↑	CC ↑	SIM↑
EMOd	EASal (Proposed)	**1.85**	**5.50**	**1.65**	**2.55**	**0.83**	**0.78**	**0.66**	0.57
	N-EASal	1.78	5.54	1.59	2.69	0.82	0.77	0.63	0.56
	CASNet[11]	1.75	5.54	1.58	2.66	**0.83**	**0.78**	**0.66**	**0.58**
	SALICON[17]	1.69	5.60	1.52	2.75	0.82	0.76	0.62	0.56
	SalGAN[26]	1.74	5.82	1.15	2.63	0.82	0.76	0.64	**0.58**
	SROD[5]	0.98	6.04	0.92	4.31	0.72	0.69	0.33	0.43
	BMS[39]	0.81	6.97	0.64	3.95	0.70	0.65	0.29	0.41
	GBVS[15]	1.18	5.86	1.17	3.27	0.77	0.73	0.45	0.48
	IttiKoch2[18]	0.99	5.98	0.98	3.96	0.73	0.69	0.35	0.44
Affective CAT2000	EASal (Proposed)	**2.27**	**0.65**	**29.36**	4.94	**0.86**	0.67	**0.72**	**0.59**
	N-EASal	2.09	0.70	29.27	5.10	**0.86**	0.67	0.66	0.57
	CASNet[11]	2.02	0.73	29.29	**4.10**	0.85	0.67	0.68	**0.59**
	SALICON[17]	2.08	0.71	29.20	4.50	**0.86**	0.67	0.69	**0.59**
	SalGAN[26]	2.05	0.94	28.83	5.27	**0.86**	**0.68**	0.69	0.58
	SROD[5]	1.32	1.04	28.69	6.87	0.81	0.64	0.46	0.45
	BMS[39]	1.16	1.86	28.56	5.97	0.78	0.59	0.39	0.44
	GBVS[15]	1.49	0.90	28.89	6.08	0.83	0.60	0.52	0.48
	IttiKoch2[18]	1.26	1.02	28.72	7.37	0.80	0.61	0.44	0.44

combine the 4096 feature maps.

4.2. Training and testing

The training and testing are implemented using Caffe framework. The feature extraction branch was first fine-tuned using SALICON dataset [17] with momentum of 0.9 and initial learning rate of 10^{-5}. The learning rate decreases by a factor of 0.1 every 8000 iterations. As the SALICON training dataset has no ground truth emotion mask, the semantic feature extraction branch is separately fine-tuned.

The trained saliency prediction branch is then combined with the emotion mask generation branch for fine-tuning. Except for the first two layers whose filter weights were fixed, all filter weights were fine-tuned with momentum of 0.9 and initial learning rate of 10^{-5}. For the emotion mask generation branch, the three 1×1 convolution filters' initial kernel weights are set to 1 and all the kernel biases is fixed to 0, so as to force the system to use the emotion information. The learning rate multiplier is set to 10^{-3} and the bias multiplier set to 0. The emotion classification module in the emotion mask generation branch is trained separately. The continuous fixation maps were used as the ground truth. We trained EASal using GeForce GTX TITAN X.

We evaluated EASal on two publicly available datasets rich in emotion-evoking objects. The first dataset is EMOd [11] which consists of 1019 emotional images. We divided EMOd into training set containing 776 images and test set containing 243 images. The second dataset is the CAT2000 [2] which is composed of 2000 images from different categories. We used the affective category of CAT2000.

4.3. Evaluation methods and results

We compared EASal with respect to other saliency models on the two aforementioned benchmark datasets. We chose three other DNN-based models with available implementation/code and no center-bias, namely SALICON [17], SalGAN [26], and CASNet [11]. SALICON and SalGAN are commonly used benchmarking models for several saliency models e.g. [7, 20], while CASNet [11] focuses on the relative saliency within an image. We also used two recent non DNN-based models SROD [5] and BMS [39] and two classical algorithms GBVS [15] and Itti-Koch [18]. We further compared EAsal with a similar model but without the emotion mask generation branch (hereafter N-EASal).

The performance is measured using standard saliency metrics (see [3] for details). Area Under the ROC curve (AUC) is the most commonly used metric for saliency evaluation. AUC measures the salient object detection capability of the model. The shuffled AUC (sAUC) is similar to AUC but it penalizes models which incorporate center bias in their design. CC and SIM treat saliency predictions as valid distribution. CC equally penalizes both false positive and false negatives while SIM is sensitive to false positive. EMD and KL are distance-based metrics thus, lower values reflect better performance. IG is based on information theory which measures the ability of the model to make predictions above baseline mode of center bias. It sums the information gain for each saliency map pixel such that IG for images with different size are incomparable. NSS is a discrete approximation of CC. NSS, KL and IG take into consideration the range of saliency map during evaluation

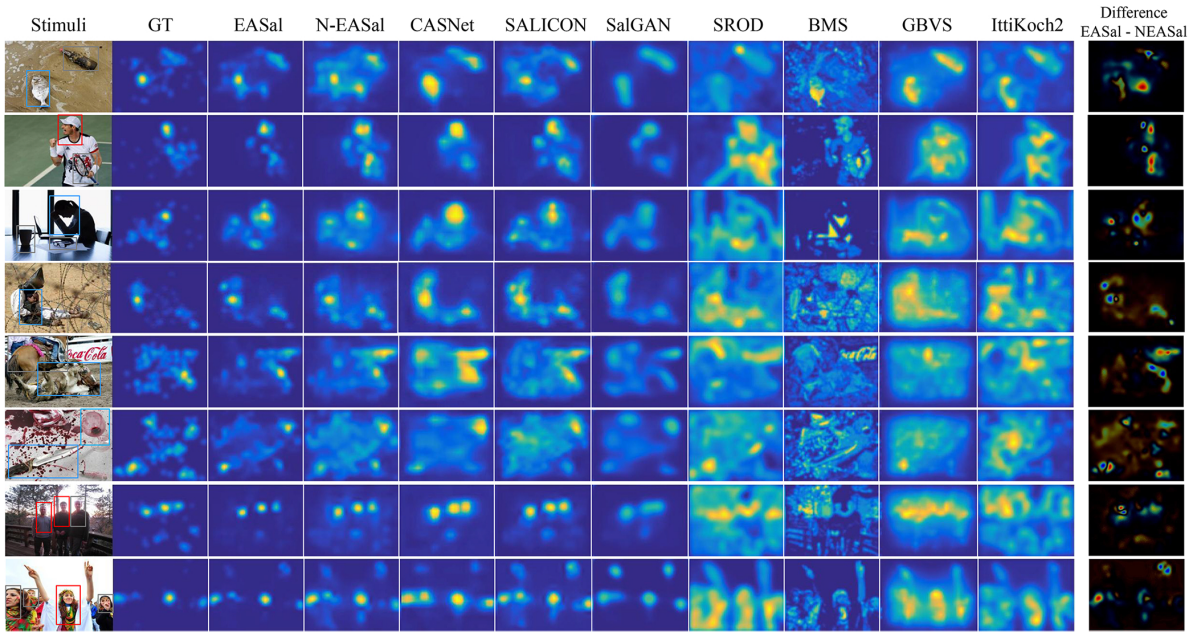

Figure 5. Qualitative comparison of predicted saliency maps. The detected object proposals by EASal are outlined in the first column. Red, gray, and blue color indicates positive, neutral, and negative sentiment, respectively. The last column shows the difference of the saliency maps between EASal and N-EASal. EASal better emphasizes the detected emotional objects when compared with other models.

thus capturing the relative importance of image regions.

The quantitative results are reported in Table 1. EASal demonstrates state-of-the-art performance on metrics reflecting relative saliency [4], *i.e.*, NSS, KL, and IG. EASal shows marginal improvement on AUC, sAUC, CC and SIM. Notably, EASal outperforms N-EASal on almost all metrics, suggesting the efficacy of the emotion mask generation branch and the combination mechanism which is motivated by our empirical data analyses.

The qualitative results are shown in Fig. 5. The emotion labels of the object proposals are indicated using red, blue and gray marks to signify positive, negative and neutral objects. When compared with other saliency models, EASal is more effective in assigning relative importance for the labeled objects, either in less complex images (*e.g.*, first four images in Fig. 5) or more complex images (*e.g.*, last four images in Fig. 5). When compared with N-EASal, EASal yields better relative saliency prediction. A visualization of the corrected image locations in N-EASal is shown in the last column of Fig. 5. Higher saliency values are assigned to emotion-evoking objects and lower saliency values are assigned to emotionally-neutral objects. The results demonstrate that EASal better embodies the phenomenon that emotion-evoking objects attract human attention.

4.4. DNN visualization and discussion

An interesting observation in EASal training is the convergence of the three 1×1 convolution filters to the following values: 1.53, 1.32 and -0.90, which serve as multiplier to the positive negative and neutral emotion mask, respectively. Note that there is a regularization function ReLU at the output of these filters such that only the emotion-evoking objects get amplified (*i.e.*, receive higher saliency prediction). In connection to this, we observe that the advantage of EASal over N-EASal is due to the increase in saliency value for emotion-evoking objects, which naturally leads to a decrease in the saliency values of neutral objects when the predicted saliency map is normalized.

In producing the saliency maps, the output feature map is directly mapped to the input image via image resizing. Therefore, each output feature map node encapsulates the predicted saliency of image regions. To visualize how emotion information changes the predicted relative importance of image regions, we show in Fig. 6 the top five nodes from the output feature map of EASal and N-EASal. This is done by selecting the five maximum node values from the 19×25 output feature map and locating their respective sections in the 600×800 input image. These nodes correspond to the predicted top five most important 32×32 regions in the input image. For example, the node located at the first row and first column of the output feature map corresponds to the upper left 32×32 region of the input image. Also, the node located at the last row and last column of the output feature map corresponds to the lower right 32×32 region of the input image.

We further quantify the changes in the saliency values for emotion-evoking objects under various image complexity. As shown in Fig. 7, images with objects of diverse

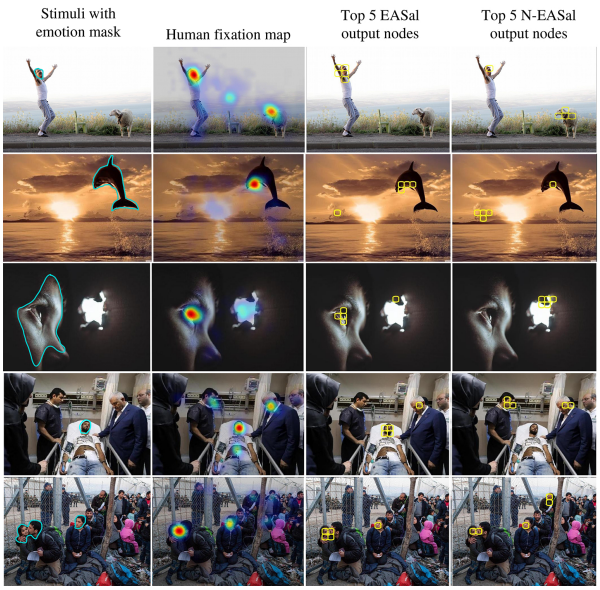

Figure 6. EASal emphasizes emotion-evoking objects. Here we visualize the top five nodes from the output feature map of EASal and N-EASal on five emotional images. The yellow squares in the last two columns indicate the predicted top five most important regions. The top five regions in EASal (3^{rd} column) show stronger emotions than those in N-EASal (4^{th} column), suggesting the efficacy of the proposed emotion integration mechanism.

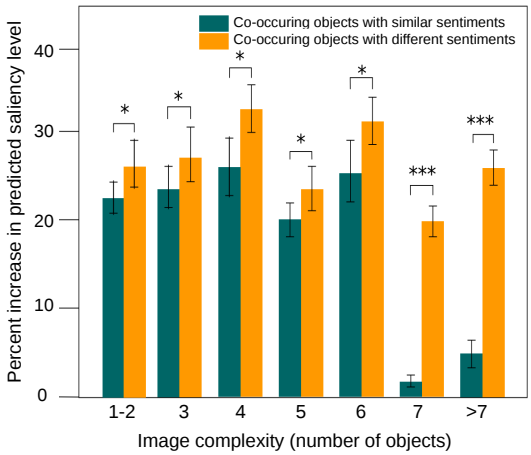

Figure 7. Percent increase in EASal with respect to N-EASal in the predicted saliency values of emotion-evoking objects. Emotion-evoking objects co-occurring with other different sentiments receive higher increase in saliency levels as compared when the objects within an image are of the same sentiment. Such advantage diminishes when image complexity increases, mainly due to the control signal mechanism in EASal, which disables the emotion mask generation branch under high image complexity.

sentiments have higher percent increase in saliency value than images with similar object sentiments. For more complex images (> 6 objects), the increase in saliency level of co-occurring objects with similar sentiments is minimal as compared to co-occurring objects with diverse sentiments. These observations show that EASal successfully encodes our empirical finding that attention favors emotion-evoking objects, especially when they co-occur with objects with different sentiments, regardless of image complexity.

An important part of EASal is the object emotion classifier. In the design of the emotion classifier, besides the labeled emotional and neutral objects from EMOd [11], we further generated objects with sentiment labels from COCO attributes [27] datasets based on their object-level attributes. For example, COCO attributes "happy" and "joyful" are converted to positive sentiments, whereas "unhappy" and "sad" are linked to negative sentiments. We used GoogleNet architecture for the emotion classifier, achieving 71.91% classification performance on EMOd. This suggests future space for improvement for EASal—its performance will be even better if given a perfect emotion classifier. Due to space limit, we describe the details of the emotion classifier in the supplementary material.

5. Conclusion

In this paper, we propose a new metric (AttI) for evaluating human attention that takes into account human consensus and image context (in terms of object sentiment). AttI enables us to have a comprehensive picture on how emotion-evoking objects complete for human attention under various image context and image complexity. Our statistical analyses show that emotion-evoking objects attract human attention, and such advantage is modulated by image complexity and image context.

Based on the empirical data analyses, we propose EASal—an emotion-aware DNN model for saliency prediction. EASal incorporates object sentiment information, by modulating the final saliency map using the automatically detected object emotion masks. Such modulation mechanism is controlled by a signal that is determined by image complexity and image context. With two benchmark datasets featuring emotional images, EASal demonstrates notable improvement on evaluation metrics that indicate relative importance of salient regions within an image (*i.e.*, NSS, KL, IG), suggesting that incorporating emotion information improves relative saliency prediction.

To the best of our knowledge, this work is a first attempt to quantify an object's attention level while considering human consensus and image complexity. The proposed saliency model is distinctive from existing models in that it conditionally incorporates emotional information, in which the condition is determined by the image context and image complexity. In future work it will be interesting to investigate the relationship between human attention and other measures, *e.g.* object interestingness [14], object memorability [9], and traits of social relation within an image [30].

References

[1] J. R. Anderson. *Cognitive psychology and its implications*. WH Freeman/Times Books/Henry Holt & Co, 1985.

[2] A. Borji and L. Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[3] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. on Pattern Anal. and Mach. Intell*, 2018.

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

[5] C. Chuanbo, T. He, L. Zehua, L. Hu, S. Jun, and S. Mudar. Saliency modeling via outlier detection. *Journal of Electronic Imaging*, 23(5), 2014.

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

[7] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Trans. on Image processing*, 27(10):5124–5154, 2018.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conf on*, 2009.

[9] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem. What makes an object memorable? In *Proceedings of the ieee international conference on computer vision*, pages 1089–1097, 2015.

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

[11] S. Fan, Z. Shen, M. Jiang, B. Koenig, J. Xu, M. Kankanhalli, and Q.Zhao. Emotional attention: A study of image sentiment and visual attention. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conf. on*, 2018.

[12] R. H. Fazio, D. R. Roskos-Ewoldsen, and M. C. Powell. Attitudes, perception and attention. *The Heart's Eye: Emotional Influences in Perception and Attention*, pages 197–216, 2004.

[13] J. Fleiss, B. Levin, and M. Cho Paik. *Statistical Methods for Rates and proportions, Third Edition*, chapter The measurement of interrater agreement. John Wiley and Sons, Inc., 2003.

[14] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1633–1640, 2013.

[15] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007.

[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *Computer Vision (ICCV), IEEE Int. Conf. on*, 2017.

[17] X. Huang, C. Shen, X. Boix, and Q. Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Computer Vision (ICCV), 2015 IEEE Int. Conf. on*, pages 262–270, Santiago, Chile, 2016. IEEE.

[18] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Anal. and Mach. Intell*, 20(11):1254–1259, 1998.

[19] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Computer Vision (ICCV), 2015 IEEE Int. Conf. on*, June 2015.

[20] S. S. Kruthiventi, Srinivas, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixation. *IEEE Trans. on Image Processing*, 26(9):4446–4456, 2017.

[21] M. Kümmerer, L. Theis, and M. Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint 1610.01563v1*, 2016.

[22] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge. Understanding low-and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision*, pages 4799–4808, 2017.

[23] D. Lane, Richard, M.-L. Chua, Phyllis, and J. Dolan, Raymond. Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, 37:989–997, 1999.

[24] H. Liu, M. Xu, J. Wang, T. Rao, and I. Burnett. Improving visual saliency computing with emotion intensity. *IEEE Trans. on Neural Networks and Learning Systems*, 27(6):1201–1213, 2016.

[25] A. Mohanty and T. J. Sussman. Top-down modulation of attention by emotion. *Frontiers in Human Neuroscience*, 7:102, 2013.

[26] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. *arXiv: 1701.01081v2*, Jan. 2017.

[27] G. Patterson and J. Hays. Coco attributes: Attributes for people, animals and objects. In *Computer Vision - ECCV 2016. Lecture Notes in Computer Science*, pages 85–100. Springer, Cham, 2016.

[28] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *Image Processing (ICIP), IEEE Int. Conf. on*. IEEE, 2016.

[29] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision*, pages 30–43. Springer, 2010.

[30] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2475–2482, 2013.

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

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[33] M. Sun, J. Yang, K. Wang, and H. Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. pages 1–6, 07 2016.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conf. on*, pages 1–9, 2015.

[35] P. Vuilleumier. How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12):585–594, 2005.

[36] P. Vuilleumier and J. Driver. Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1481):837–855, 2007.

[37] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 4:1–20, 2014.

[38] J. Yang, D. She, M. Sun, M. Cheng, P. L. Rosin, and L. Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Trans. on Multimedia*, 20(9):2513–2525, Sept 2018.

[39] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *2013 IEEE Int. Conf. on Computer Vision*, pages 153–160, Dec 2013.