Supplementary Material for Paper "Emotion-Aware Human Attention Prediction"

Macario O. Cordel II[1]*, Shaojing Fan[2], Zhiqi Shen[2] Mohan S. Kankanhalli[2]
[1]De La Salle University – Manila, [2]National University of Singapore

1. Object Sentiment Classification

Sentiment classification is an important part of the proposed Emotion-Aware Saliency model (EASal). In the design of the object sentiment classifier, we collected the positive and negative emotion-evoking, and emotionally neutral objects from EMOtional attention Dataset (EMOd) [4] and COCO attributes [11] dataset based on their object-level attributes. EMOd is a human attention dataset focusing on emotional images. The objects in EMOd have sentiment labels (either positive, negative, or neutral). The COCO attribute dataset (COCO attributes) provides 169 object-level visual attributes.

1.1. Data preparation

We selected a set of COCO attributes that are strongly related to positive and negative sentiments and transformed them to positive and negative sentiment labels (see Table 1, second column). The rest of the COCO attributes are transformed to neutral sentiment labels. Since each object in COCO attributes has more than one labels, the dominant sentiment label was used as the final sentiment label of this object. In rare cases where the number of positive sentiment label equals the number of negative sentiment label, a neutral label was assigned as the final sentiment label. We classified attributes showing mild expressions into neutral sentiment labels, as initial tests show that using the annotations which strongly express emotion only, rather than including attributes showing mild expression, provides more desirable results in sentiment classification.

Fig. 1 shows some examples on how object labels in the COCO attributes dataset are transformed into object sentiment label. As shown in Fig. 1, attributes such as unhappy in the top-left right image is transformed to negative label and is used as the sentiment label for the object. For the top right image, the joyful and smiling are transformed to positive label. In total, 3828 objects were collected, 3009 of which were used in the training set and the rest were used for the validation set (see Table 2 for the summary). We performed oversampling on the training set using flipping, shifting and cropping in four different directions, resulting in 15045 training samples.

Table 1: COCO attributes dataset provides a set of attributes that relate to emotions. We classify those strongly expressing emotions (second column) into positive and negative object sentiment labels, respectively.

Emotion	strong expression	mild expression
Related to positive emotion	happy, smiling, enjoying, laughing, celebrating, joyful, excited, cute/adorable	clean, peaceful, young, calm, warm, fresh, strong, elegant
Related to negative emotion	unhappy, angry, scared, sad, annoyed	bored, anxious, dull, confused, dirty, dangerous, lazy, old

Figure 1: Example images of how we classified attributes in COCO attributes dataset into either positive, negative, or neutral sentiment label.

1.2. DNN Architecture and Training

We considered six deep learning architectures to perform the emotion classification which include the CaffeNet [8],

*This work was done when the first author was an intern at National University of Singapore. Email: macario.cordel@dlsu.edu.ph

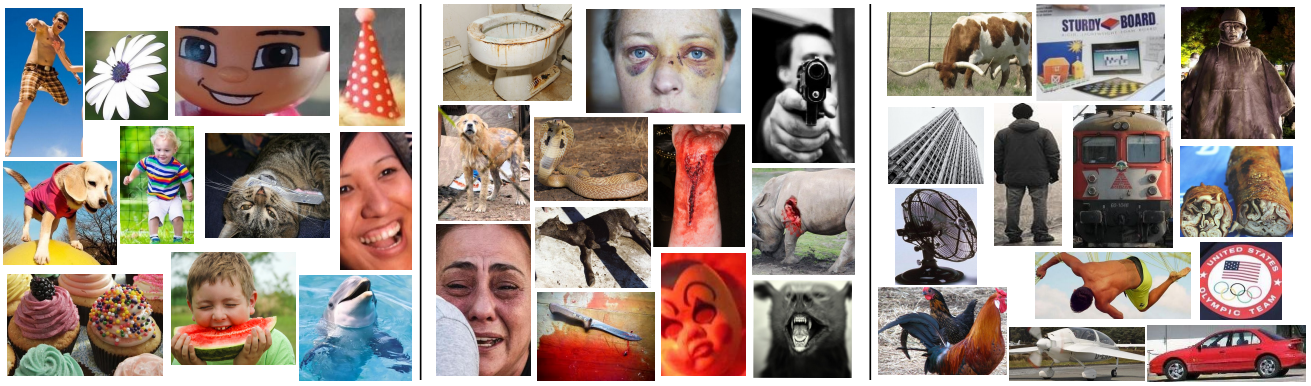


Figure 2: Sample objects labeled by the sentiment classifier as positive-evoking objects (left), negative-evoking objects (middle) and neutral objects (right).

Table 2: Number of extracted objects from the training set used for fine-tuning the sentiment classification module.

Dataset source	Positive objects	Negative objects	Neutral objects	Total no. of objects
EMOd [4]	705	1002	661	2368
Coco attributes [11]	630	100	730	1460
Total	1335	1102	1391	3828

AlexNet [10], GoogleNet [15], VGG-16 [13], ResNet-50 and ResNet-100 [6]. To determine which architecture best captures the feature space for predicting the object sentiment, we fine-tune these networks using update momentum equal to 0.9, base learning rate equal to 0.001 and weight decay equal to 0.005. The learning rate is then dropped by a factor of 0.96 every after 10^4 iterations for CaffeNet, AlexNet, GoogleNet and VGG and after 5×10^4 for ResNet-50 and ResNet-100. Due to different number of layers in each architecture, CaffeNet, AlexNet and GoogleNet are trained for 80 epochs, the VGG for 100 epochs and the ResNet-50 and ResNet-100 for 300 and 350 epochs, respectively.

1.3. Results

Using the dataset described in Table 2, we found out that the best architecture for the sentiment classifier is GoogleNet achieving 71.91% classification accuracy. Table 3 summarizes the average performance of each network on the five-fold cross validation. Sample images classified as positive-evoking, negative-evoking and neutral objects are shown in Fig. 2.

Table 3: Average five-fold cross validation accuracy of emotion classifier after fine-tuning the different networks. The main difference of GoogleNet from other deep networks is in its inception architecture that essentially represents the multi-resolution information.

Networks	Accuracy
CaffeNet	60.01%
GoogleNet	**71.91%**
AlexNet	58.68%
ResNet-50	60.31%
ResNet-101	62.26%
VGG-16	63.12%

2. Architectures for integrating emotion to saliency model

The integration of emotion to the feature space of saliency model is composed of two modules, the sentiment mask generation module and the semantic feature extraction module. We considered three architectures, as shown in Fig. 4, for the final design of the EASal.

2.1. Sentiment Mask Generation Module

The sentiment mask generation module has object proposal submodule (we used Mask R-CNN [5]) and object sentiment classifier submodule, discussed in the previous section. It interfaces the discrete outputs of the object sentiment classifier submodule to the semantic feature generation module by converting the discrete sentiment labels into a three-channel mask. Each channel contains the corresponding object sentiment mask. That is, the first channel is for the positive sentiment, the second channel is for the negative sentiment, and the third channel is for the neutral sentiment. The generated three-channel mask is then used as additional feature in saliency prediction. The propagated

Figure 3: Sample generated masks from the object proposals. For illustration, red pixels corresponds to the positive emotion-evoking object mask, blue pixels corresponds to the negative emotion-evoking object mask and gray pixels corresponds to the neutral objects.

masks shown in Fig. 3 are some of the outputs of the sentiment masks generation module and is used accordingly in the semantic feature extraction module.

2.2. Semantic Feature Extraction Module

The incorporation of emotion to saliency model is based on our empirical study that emotion attracts attention. We fine-tune all architectures considered using SALICON dataset. The semantic feature extraction branch of the saliency model has two VGG-16 branches, corresponding to the fine and coarse branches to account for the selective human attention at different viewing resolution. Increasing the number of branches further to consider more viewing resolutions, according to [7], does not improve the performance of the system. In combining the sentiment mask with the semantic feature extraction branch, we considered three fusing architectures shown in Fig. 4.

We first investigate the early fusion architecture which adds sentiment masks as new channels at the input of the saliency model as illustrated in Fig. 4-(a). The positive mask, negative mask and neutral mask are concatenated with the RGB input, generating an input with six channels. The first convolution layer for both the fine and coarse branches are modified for the six-channel input. The resulting size (number of outputs $\times$ number of input channels $\times$ 2D filter size) of the first convolution layer, for both the fine and coarse branches, changes $64\times3\times3\times3$ to $64\times6\times3\times3$. The advantage of the early fusion scheme is the minimal increase in the complexity of the saliency model.

We next evaluate late fusion architecture which incorporates the sentiment mask at the concatenation layer. However, as the concatenation layer has 1024 feature maps while the sentiment mask has only three layers (negative, positive and neutral), the concatenation layer overwhelms the three-channel sentiment mask during linear combination. To address this, the 1024 feature maps are transformed to three feature maps by modifying the last layer from size $1\times1024\times1\times1$ convolution to size $3\times1024\times1\times1$ convolution (see Fig. 4-(b)). A second concatenation layer is then introduced, combining the three-channel feature map output of the first concatenation layer to the three-channel sentiment mask. Other than complexity, the advantage of the late fusion scheme is that it can directly connect saliency prediction and object sentiment using the filter weights for the emotion layers.

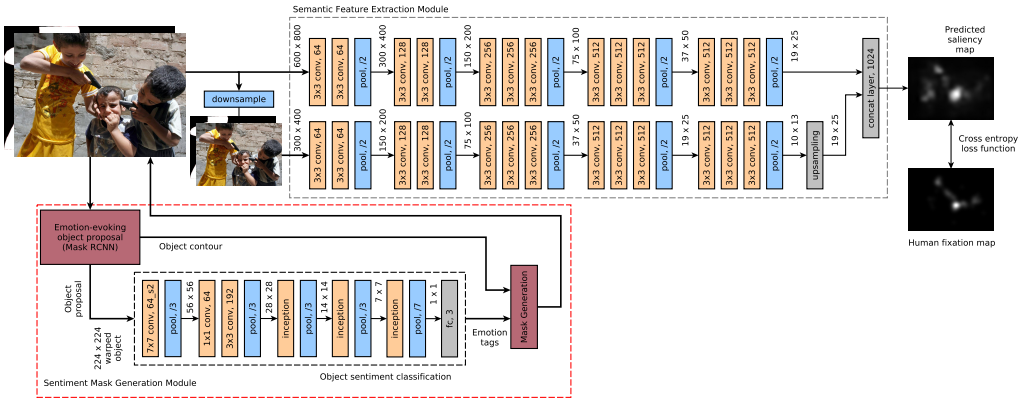
Finally, we test the intermediate fusion architecture. The feature maps from the semantic feature extraction branch in the last layer are copied to the sentiment mask generation branch. The copied feature maps are then multiplied element-wise to the sentiment masks via the 1×1 convolution filter, as illustrated in Fig. 4-(c). As the semantic feature maps are copied to the sentiment mask generation module to combine with the sentiment mask, the sentiment masks are down-sampled to 19×25 so that the mask size becomes equal to the size of the semantic feature maps. At the output is the concatenation layer, whose length was changed to 4096 feature maps from the original size of 1024. To reduce the dimension of the output to a single 19×25, 1×1 convolution is added at the last layer.

2.3. Training

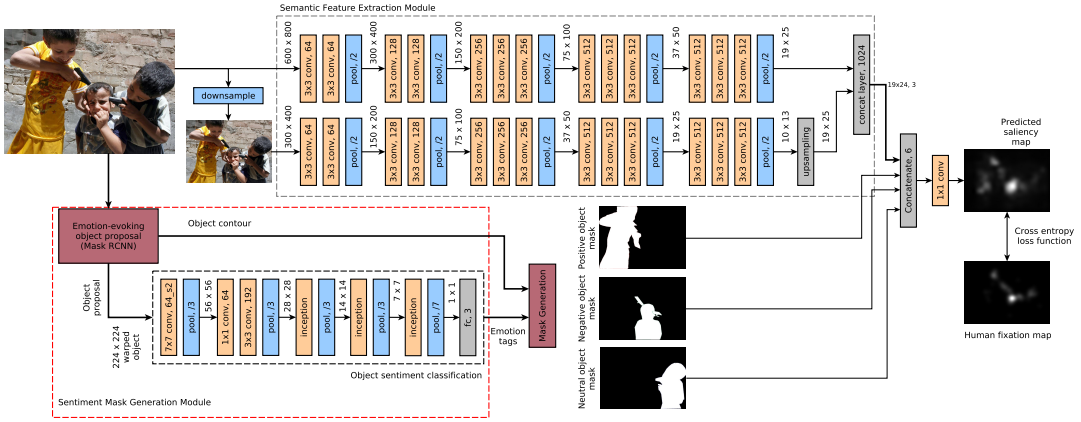
In the initial evaluation to determine which of the three architectures is the best, we used 776 images for fine-tuning the proposed architectures described in Fig. 4. The input image was transformed into 300×400 image and 600×800 image as input to the coarse and fine channel, respectively. The corresponding sentiment mask of early fusion was also converted to the appropriate input size. For the late and intermediate fusion architectures, the corresponding ground truth mask of the input image is converted to 19×25 which is the size of the feature maps in the concatenation layer.

The feature extraction branch is first fine-tuned using SALICON dataset [9] with momentum of 0.9 and initial learning rate of 1e-5. The learning rate decreases by a factor of 0.1 every 8000 iterations. As the SALICON training dataset has no ground truth sentiment mask, the semantic feature extraction branch is separately fine-tuned. The trained saliency prediction branch is then combined with the sentiment information, correspondingly using the three architectures, for fine-tuning. A subset of EMOD (776 images) is used as the training set. 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 1e-5.

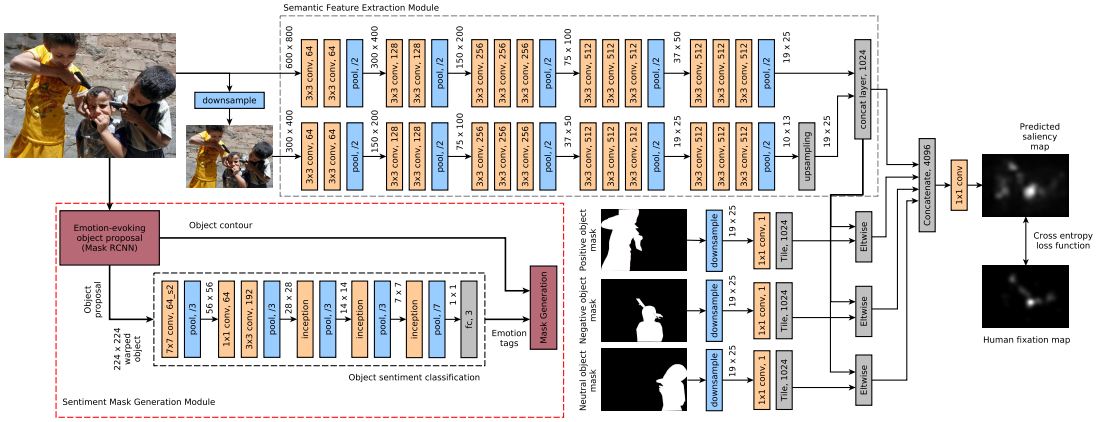
For the sentiment mask generation branch of the intermediate fusion architecture, the three 1×1 convolution filters are initialized to 1.0 and the biases are fixed to 0.0, to force the whole network to use the sentiment information in saliency prediction. The learning rate multiplier is set to 10^{-3} and the bias multiplier set to 0.0. The object sentiment classification module in the sentiment mask genera-



(a) Early fusion



(b) Late fusion



(c) Intermediate fusion

Figure 4: We considered three architectures for incorporating the object sentiment information in saliency prediction. The first architecture in (a) is called early fusion. It adds a fourth layer which corresponds to the sentiment layer, to the RGB input data. The second architecture in (b) is called the late fusion. It incorporates the sentiment information at the output of the concatenation layer, thus adding second concatenation layer. Finally in (c) is called the intermediate fusion. It adds a new channel for the masks (parallel of three convolution filters followed by a regularization function ReLU). The output at the last layer is then multiplied with the semantic feature maps from the semantic feature extraction module. For the three architectures, other parameters, such as the number of layers and filters in the semantic feature maps, are retained.

Table 4: Comparison of saliency performance for late, early and intermediate fusion architectures using the sentiment mask. The intermediate fusion architecture shows the best performance among the three fusion types, especially in terms of metrics that measure relative saliency of image regions, i.e. NSS and IG. The dataset used is EMOd.

Fusion types	NSS↑	KL↓	IG↑	EMD↓	AUC-Judd↑	sAUC↑	CC↑	SIM↑
Early (Fig. 4 a)	1.68	5.64	1.41	2.74	0.81	0.73	0.65	0.58
Late (Fig. 4 b)	0.90	6.83	0.18	4.28	0.73	0.62	0.33	0.45
Intermediate (Fig. 4 c)	1.83	5.54	1.58	2.61	0.82	0.73	**0.66**	**0.59**
Intermediate with control signal (Fig. 5)	**1.85**	**5.50**	**1.65**	**2.55**	**0.83**	**0.78**	**0.66**	0.57

tion branch is trained separately.

For all three candidate architectures, the continuous fixation maps were used as the ground truth. The training and testing are implemented using Caffe framework and GeForce GTX TITAN X.

3. Evaluation

3.1. Metrics

The saliency metric scores reported in this paper are the AUC-Judd, sAUC, NSS, SIM, KL and IG. The area under the ROC curve (AUC) score is the most commonly-used metric for saliency evaluation. It measures the trade-off between the true and false positives at various discrimination thresholds. It is invariant to contrast and monotonic transformation such that it is particularly good in detection applications. The normalized saliency scanpath (NSS) and the correlationa coefficient (CC) are highly related saliency metrics because of their analogous computation. NSS measures the correspondence between the saliency map and the ground truth fixation. It is sensitive to false positives, relative differences in saliency across the image and monotonic transformation. Similarly, CC measures the how correlated or dependent two variables are. As opposed to NSS, CC equally penalizes false positive and false negatives such that the increase in CC alone can not distinguish whether the improvement is due to false positives or false negatives.

The Kullback-Leibler divergence (KL), information gain (IG) and similarity (SIM) metrics rank differently the saliency maps, as opposed to NSS and CC, for the reason that these metrics are extremely sensitive to false positives. KL is a dissimilarity metric which evaluates the loss of information when the saliency map is used to approximate the ground truth fixation map. IG on the other hand, measures the average information gain of the saliency map over the center prior baseline at fixated locations. Lastly, SIM measures the similarity between the saliency map and the ground truth fixation map.

Note that recently, AUC scores, SIM and CC have come to a situation when no significant increase in value is seen regardless of the saliency model [1, 2]. These metrics cannot differentiate between models and cannot measure model performances at a finer-grained level.

3.2. Results and optimization

As shown in Table 4, the most desirable performance among early, late and intermediate fusion types is from the intermediate fusion architecture in terms of NSS and IG scores. Thus, we use the intermediate fusion as the base configuration for the emotion-aware saliency prediction model (EASal). We then variably introduce emotion information to saliency prediction inspired by our empirical data analyses.

Our empirical data analyses show that the emotion prioritization effect depends on image complexity and image context. Based on this finding, we initially design a control signal (as illustrated in Fig. 5) 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. Results show (see Tab. 4 last row) that saliency prediction improves when emotion prioritization effect based on our empirical data analyses is included in the design. We present the automatic learning of this control signal in our final design presented in the main paper.

4. Discussion

As opposed to the emotion-evoking regions (ER) introduced in [12, 14, 17], the emotion-evoking objects sentiment masks have detailed object contours. With the detailed contours and object-based nature, the sentiment masks have the following advantages over ERs: i) They provide more precise information for saliency prediction. ii) They are more explainable and indicative of image complexity, which echoes with our human findings and contributes to the EASal's control signal subnetwork. Additional experiments by replacing the sentiment masks with

Table 5: Performance of EASal with sentiment masks (ours) versus EASal with emotion-evoking regions (ERs) on EMOd.

Method	NSS↑	KL↓	IG↑	EMD↓	AUC-Judd↑	sAUC↑	CC↑	SIM↑
Sentiment masks (ours)	**1.85**	**5.50**	**1.65**	**2.55**	**0.83**	**0.78**	**0.66**	**0.57**
Emotion-evoking regions [17]	1.61	6.41	1.41	2.95	0.83	0.71	0.59	0.54

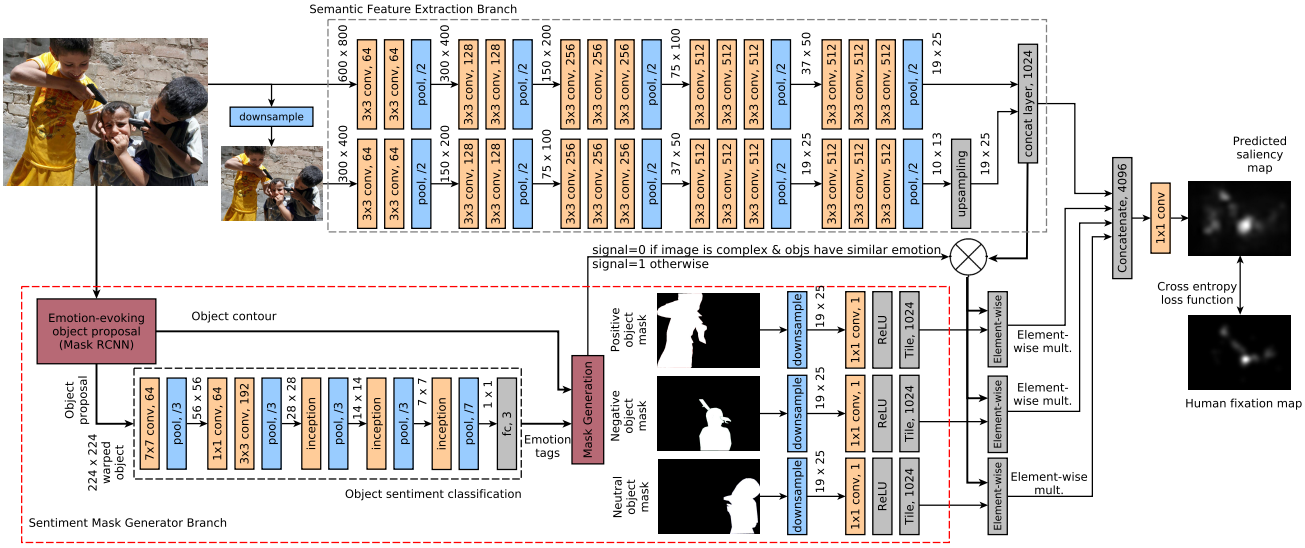


Figure 5: The initial design of the proposed saliency model is composed of two branches: (1) the semantic feature extraction branch which learns semantic information from the input image, and (2) the object-level sentiment mask generation branch which generates and incorporates the objects' sentiment 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.

the ERs from [17], shown in Table 5, support our claim.

Through emotion information, EASal was able to correct the relative saliency prediction of image regions by identifying emotion-evoking objects and providing higher saliency prediction on their locations (through the 1×1 convolution filters at the sentiment mask generation branch). These can be observed when we compare the top 5 activated neurons from the output of EASal and N-EASal as shown in Fig. 6.

The design of EASal is based on our empirical study on the interrelation of emotion and attention, taking into consideration the semantic complexity of an image. The metric used by other studies e.g. in [3, 16, 4], which is the Attention Score (AS) or the maximum value of the fixation map inside the object's contour, may not be enough in studying emotion and attention across different image complexities.

As additional illustration, consider Fig. 7. The two sets of images (a) and (b), have objects with AS equal to 1.0 and near 1.0. These objects receive different levels of human attention, in terms of human fixations, but have similar high AS score due to fixation map normalization. The first set of images (a) has few outstanding objects, catching most human attention, thus it will also have an AS score close to 1. The second set of images (b) has several objects with scattered human fixations, but after normalization, these objects will receive AS close to 1. We propose AttI which factors in the consensus of human fixation or HCS to better reflect the human attention level.

References

[1] 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.

[2] 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.

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

[4] 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.

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

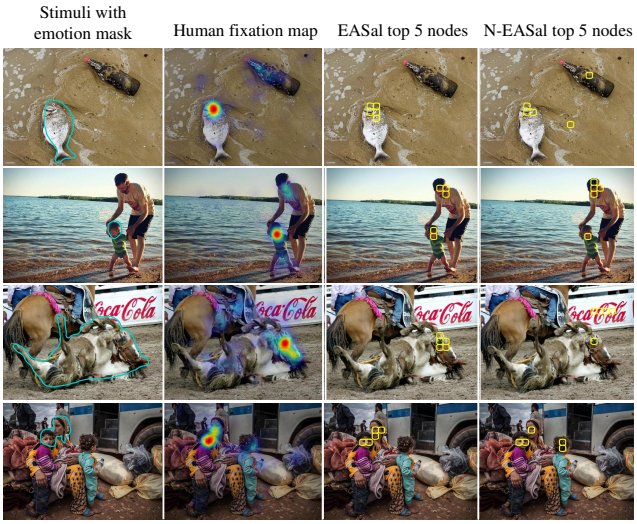


Figure 6: EASal emphasizes emotion-evoking objects. Here we visualize the top five nodes from the output feature map of EASal and N-EASal on four emotional images. The yellow squares in the last two columns indicate the predicted top five most important regions. The top five regions in EASal (3^{rd}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

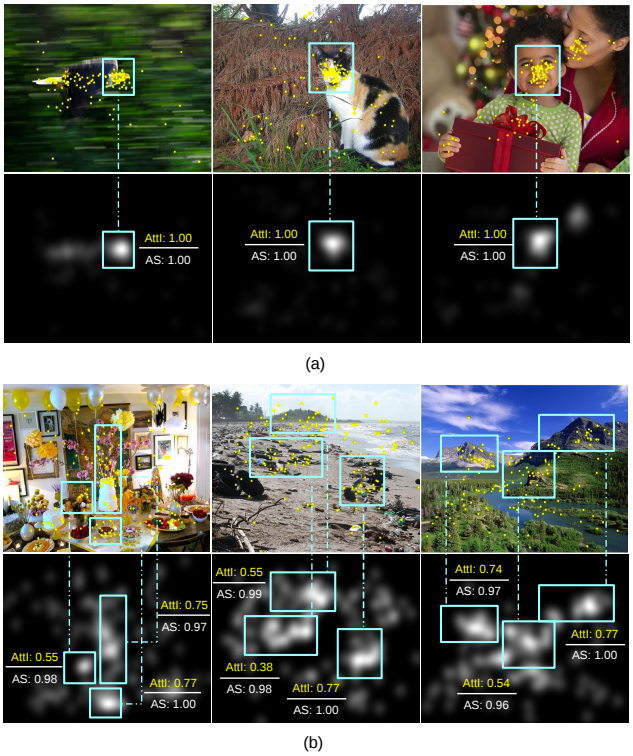


Figure 7: Attention Index (AttI) captures the attention level of objects from different images, as compared with Attention Score (AS). (a) images with one to two outstanding objects which capture human fixations. (b) images with several objects and scattered human fixations. Both sets of images have objects with AS equal to 1.0 despite of the consensus of human fixations. By factoring the consensus of human fixations, AttI better reflects the attention level of these objects.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conf. on*, pages 770–778. IEEE Computer Society, 2016.

[7] 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.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

[10] A. Krizhevsky, I. Sutskeyer, and E. Hinton, Geoffrey. Imagenet classification with deep convolutional neural network. In *NIPS2012 Proceedings of the 25th Int. Conf. on Neural Information Processing Systems*, pages 1097–1105, 2012.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] 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.