Predicting When Saliency Maps are Accurate and Eye Fixations Consistent

Anna Volokitin[1]   Michael Gygli[1]   Xavier Boix[2,3]

[1] Computer Vision Laboratory, ETH Zurich, Switzerland

[2] Department of Electrical and Computer Engineering, National University of Singapore

[3] CBMM, Massachusetts Institute of Technology, Cambridge, MA

Abstract

Many computational models of visual attention use image features and machine learning techniques to predict eye fixation locations as saliency maps. Recently, the success of Deep Convolutional Neural Networks (DCNNs) for object recognition has opened a new avenue for computational models of visual attention due to the tight link between visual attention and object recognition. In this paper, we show that using features from DCNNs for object recognition we can make predictions that enrich the information provided by saliency models. Namely, we can estimate the reliability of a saliency model from the raw image, which serves as a meta-saliency measure that may be used to select the best saliency algorithm for an image. Analogously, the consistency of the eye fixations among subjects, i.e. *the agreement between the eye fixation locations of different subjects, can also be predicted and used by a designer to assess whether subjects reach a consensus about salient image locations.*

1. Introduction

Gaze shifting allocates computational resources by selecting a subset of the visual input to be processed, *c.f.* [33]. Computational models of visual attention provide a reductionist view on the principles guiding attention. These models are used both to articulate new hypotheses and to challenge the existing ones. Machine learning techniques that can make predictions directly from the image have facilitated the study of visual attention in natural images. Also, these models have found numerous applications in visual design, image compression, and some computer vision tasks such as object tracking.

Many computational models of attention predict the image location of eye fixations, which is represented with the so called saliency map. The seminal paper by Koch and Ullman introduced the first computational model for saliency prediction [21]. This model is rooted in the feature integration theory, that pioneered the characterisation

of many of the behavioural and physiological observed phenomena of visual attention [32]. Since then, a rich variety of models have been introduced to extract the saliency map, *e.g.* [11, 14, 17, 18, 35].

Some authors stressed the need to predict properties of the eye fixations beyond the saliency map to study different phenomena of visual attention and to allow for new applications, *e.g.* [15, 24, 27]. Since visual attention is strongly linked to object recognition, the advent of near-human performing object recognition techniques based on DCNNs opens a new set of possibilities for models of visual attention. In this paper, we analyze two ways to augment the eye fixation location information delivered by saliency models by using features extracted from DCNNs trained for object recognition.

Firstly, inspired by machine learning techniques that provide an estimate of their own accuracy, we show that the accuracy of a saliency model for a given image can be predicted directly from image features. Our results show that whether predicting the location of the human eye fixations is possible depends on the object categories present in the image.

Secondly, we show that the consistency of eye fixation locations among subjects can also be predicted from features based on object recognition. In Fig. 1 we show images with different degrees of consistency among subjects, that illustrate that eye fixation consistency varies depending on the image. There is a plethora of results in the literature showing that consistency varies depending on the group the subjects belong to. There are marked differences between subjects with autism spectrum disorders and those without [6, 20], between subjects from different cultures [5], and between fast and slow readers [19]. Yet, the causes of eye fixation inconsistencies among individual subjects rather than for groups may be difficult to explain in natural images, especially because natural images are not designed to isolate a specific effect.

The model we introduce to predict the eye fixation consistency substantially improves the performance of a previous attempt [24], and it shows that the eye fixation consis-
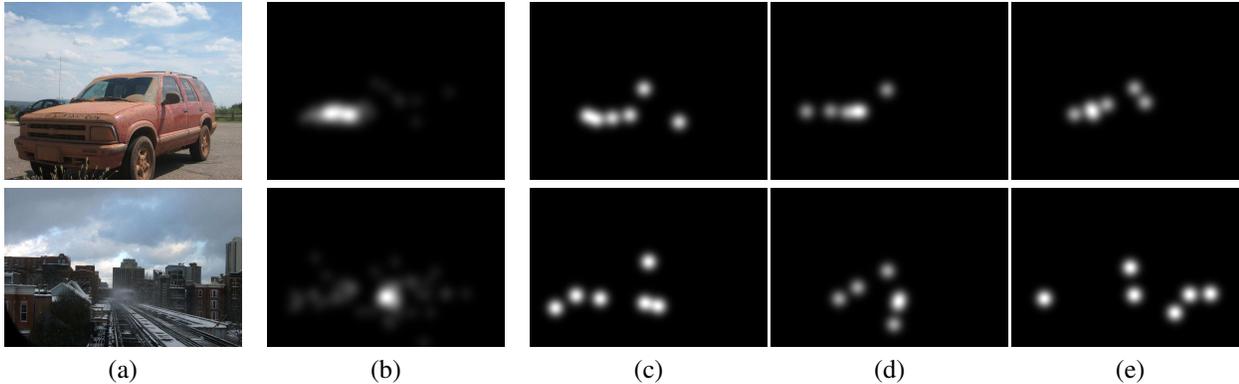
Figure 1: *Fixations from individual subjects.* (a) the raw image, (b) averaged fixation map, (c) - (e) individual fixations from subjects. The top row shows an image where fixations are highly consistent, and the bottom shows one where the fixations are inconsistent.

tency depends on the object categories present in the image. Also, we show that current saliency models and our eye fixation consistency model describe complementary aspects of viewing behaviour, and should be used in conjunction for a more complete characterisation of viewing patterns.

Finally, our results reveal that, like memorability [13] and interestingness [10], eye fixation consistency is an attribute of natural images that can be predicted.

2. Eye Fixation and Saliency Maps

In this section, we introduce the datasets we use and review how to build the eye fixation and saliency maps. This will serve as the basis for the rest of the paper.

Datasets We use the MIT [17] dataset, which includes 1003 images with everyday indoor and outdoor scenes. All images are presented to 15 subjects for 3 seconds. This dataset is one of the standard benchmarks to evaluate the prediction of eye fixation locations in natural images. To show the generality of our conclusions, we also report results on the PASCAL saliency dataset [26]. This dataset uses the 850 natural images of the validation set of the PASCAL VOC 2010 segmentation challenge [8], with the eye-fixations during 2 seconds of 8 different subjects.

Eye Fixation Maps An eye fixation map is constructed for each subject by taking the set of locations where the eyes are fixated for a certain period of time (conventionally taken to be 50ms). The fixation map is a probability distribution over salient locations in an image, and ideally would be computed by taking an average over infinite subjects. In practice, the eye fixation map is computed by summing eye fixation maps of the individual subjects (which are binary images, with ones at fixation locations and zeroes elsewhere). The result is smoothed with a Gaussian of width dependent on the eye tracking set up (1 degree of visual angle in the MIT dataset, and $\sigma = 0.03 \times$ image width in the PASCAL dataset). Finally, the map is normalised to sum to one.

Saliency Maps A saliency map is the prediction of the eye fixation map by an algorithm. We use seven state-of-the-art models to predict the saliency maps: Boolean Map based Saliency (BMS) [37], Adaptive Whitening Saliency Model (AWS) [9], Graph-based Visual Saliency (GBVS) [11], Attention based on Information Maximization (AIM) [3], Saliency using Natural Statistics (SUN) [38], the original saliency model developed by Itti *et al*. (IttiKoch) [14], and a new DCNN-based model called SALICON [12]. We use the standard procedure (with code from [16]) to optimise these saliency maps for the MIT dataset. For a complete review of these algorithms, we refer the reader to the thorough analysis of Borji *et al*. [2] and Judd *et al*. [16].

3. Enriching Saliency Maps

In this section we introduce the two estimates we use to add to a saliency map: the estimate of the saliency map accuracy, and the consistency of the eye fixations among subjects. We introduce the computational model to predict them in section 4.

3.1. Predicting the Saliency Map Accuracy

We explore whether the saliency model accuracy can be predicted by features extracted directly from the image, *i.e.* before computing the saliency map. This prediction depends on the algorithm for saliency prediction, and also, it depends on the metric used to evaluate the accuracy of the saliency map.

Metric of the Saliency Map Accuracy Since there is no consensus among researchers about which metric best captures the accuracy of the saliency map (*c.f.* [30]), we follow the lead of [16] and report 3 metrics. Now we briefly define the metrics used in this paper, and refer the reader to [30] for a more complete treatment. Under all of these metrics a higher score indicates better performance. Below, M_F is the map of eye fixation map (ground truth) and M_S is the (predicted) saliency map:

- *Similarity (Sim).* The similarity metric is also known as the histogram intersection metric, and it is defined as $S = \sum_x \min(M_F(x), M_S(x))$.

- *Cross Correlation (CC).* This metric quantifies to what extent there is a linear relationship between the two maps. It is defined: $CC = cov(M_F, M_S)/(\sigma_{M_F}\sigma_{M_S})$, where σ_M is the standard deviation of the map M.

- *Shuffled Area under the Curve (sAUC).* The saliency map is treated as a binary classifier to separate positive from negative samples at various intensity thresholds. It is called shuffled because the points of the saliency map are sampled from fixations on other images to discount the effect of center bias. This metric can take values between 0.5 and 1. Although the previous two metrics are symmetric, meaning the two maps are interchangeable, this one is not.

Applications Our goal is to predict the evaluation metrics of a saliency model we just introduced, for a given image. Providing such estimate of the saliency model accuracy may be used to select the best saliency algorithm for a specific image. Additionally, the accuracy estimate could be used as a meta-saliency measure that indicates the quality of the saliency map to the user.

3.2. Predicting the Eye Fixations Consistency

The second estimate we provide to enrich the saliency map is the eye fixation consistency among subjects, *i.e.* the amount of inter-subject variability in viewing the image. To do this, we first measure the true eye fixation consistency given the eye fixations of individual subjects, adapting a procedure used in [31], which we introduce next.

Metric of the Eye Fixation Consistency The eye fixation consistency metric tests whether the fixation map computed from a subset of subjects can predict the fixation map computed from the rest of the subjects. Let $\mathcal{O}$ be the set of all subjects (*e.g.* 15 in MIT dataset), and $\mathcal{H}$ be the subset of K subjects held out for testing. We compute two eye fixation maps: $M_{\mathcal{H}}$ from $\mathcal{H}$, and $M_{\mathcal{O}\setminus\mathcal{H}}$ from $\mathcal{O}\setminus\mathcal{H}$ (the remaining $15 - K$ subjects). We define the consistency score to be the score of $M_{\mathcal{H}}$ in predicting $M_{\mathcal{O}\setminus\mathcal{H}}$ using any of the standard metrics for evaluating saliency prediction algorithms (introduced previously in section 3.1). To be consistent in our evaluation of consistency, $M_{\mathcal{H}}$ is treated as the saliency map, and $M_{\mathcal{O}\setminus\mathcal{H}}$ as the eye fixation map, as it is computed from more subjects than $M_{\mathcal{H}}$.

In the experiments we analyse several properties of this metric of the eye fixation consistency, such as the dependency on the number of subjects, K, and that the metric is independent on the subjects chosen to build the eye fixation map $M_{\mathcal{O}\setminus\mathcal{H}}$ and $M_{\mathcal{H}}$. The results show that our metric can generalise to different subjects, but that we need to evaluate different values of K and ways to compare the saliency maps (*e.g.* Sim, sAUC and CC) because the results highly depend on these parameters.

A possible alternative to the metric we use is the Shannon entropy of the eye fixation map. In fact, the Shannon entropy has been used as an alternative consistency measure in [17]. This measure makes the assumptions that inconsistent viewing patterns will yield a flat fixation map, while consistent ones will yield a map with sharp peaks. The entropy is high in the first case, and low in the second. There are a few cases where these assumptions do not hold. The eye fixation map might have several sharp peaks, and thus low entropy, but the subjects can be inconsistent by each only looking at a subset of the peaks. This is the kind of viewing behaviour sometimes exhibited on natural images with several salient regions (*e.g.* in Fig. 6 in the image in the bottom right, there is a person standing near the edge of the image that not all subjects notice). Thus, the entropy is not equivalent to consistency. In the experiments, we corroborate this point by showing that the entropy of the eye fixation maps and the consistency measure we use are correlated but up to a certain extent.

Applications We aim at estimating the eye fixation consistency among subjects from the raw image, by predicting the value of the aforementioned eye fixation consistency metric. The prediction of the eye fixation consistency can be used to enrich the information provided by the saliency map, because current saliency models have no measure of the consistency of the eye fixations of the subjects, and in this sense are incomplete. The reader may object that since the entropy of the eye fixation map is related to the consistency, it could be that the entropy of the saliency map is also related to the eye fixation consistency. Then saliency models would already provide an estimate of the consistency through the entropy of the saliency map. Our results discard this hypothesis by showing that the entropy of the saliency maps are uncorrelated with the eye fixation consistency.

Applications that make use of saliency maps, such as visual design, could incorporate eye fixation consistency information to create designs with a greater consensus of fixations in the location of the designer's choice. Since advertisement needs to have maximal effect in minimal time, it is desirable to have viewers consistently attending to specific locations.

Also, as mentioned in the introduction, some groups in the population have distinct viewing patterns. To better study these phenomena in the laboratory, it might be advantageous to use images in which the individual viewing pattern variability is controlled for. Thus, our computational model could be used to determine *a priori* which images naturally produce very variable viewing patterns in all subjects, to deliberately include or exclude such images in the dataset.

4. Computational Model

To predict both accuracy and consistency we train a regressor between the features extracted from the image and the response variable. The features and learner are the same same for both applications. We use a Support Vector Regressor [34] with the χ^2 kernel. We introduce several image features to test the hypothesis that the saliency accuracy and consistency can be predicted from the spatial distribution and the categories of the objects in the image. The splits are done taking randomly 60% of images for training and the rest for testing. The learning parameters are set with a 10 fold cross-validation using LIBSVM to determine the cost C (range 2^{-4} to 2^6) and ϵ (2^{-8} to 2^{-1}) of the ϵ-SVR.

Deep Convolutional Neural Networks To capture the spatial distribution and category of the objects in the image, we use features taken from the layers of a DCNN. A DCNN is a feedforward neural network with constrained connections between layers, that take the form of convolutions or spatial pooling, besides other possible non-linearities, *e.g.* [25, 22, 29]. We use the DCNN called AlexNet [22] with trained parameters in ImageNet, which achieved striking results in the task of object recognition. It consists of eight layers, the last three of which are fully connected.

Let $\mathbf{y}_l$ be a two-dimensional matrix that contains the responses of the neurons of the DCNN at the layer l. $\mathbf{y}_l$ has size $s_l \times d_l$, that varies depending on the layer. The first dimension of the table indexes the spatial location of the center of the neuron's receptive field, and the second dimension indexes the patterns to which the neuron is tuned. The response of a neuron, $y_l[i][j]$, has a high response when pattern j is present at location i. Neural responses at higher layers in the network encode more meaningful semantic representations than at lower layers [36], but the spatial resolution at the last layers is lower than at the first layers.

The neural responses from the top of each layer $\mathbf{y}_l$ are used as features.

Spatial Distribution of Objects We introduce two different features to capture the spatial distribution of the objects without describing their object categories. The first feature

is based on the DCNNs previously introduced. We take the neural responses in a layer, $\mathbf{y}_l$, and convert them into a feature that has one response for each location that corresponds to the presence of a pattern or object detected by the CNN (it has dimensions $s_l \times 1$). To do so, we discard information about which pattern is present at a certain location and simply take the highest response among the patterns. Thus, the image feature is $f_l[i] = \max_j y_l[i][j]$. This corresponds to max pooling over the pattern responses.

A second feature we introduce is based on the objectness, or the likelihood that a region of an image contains an object of any class [1]. Objectness is based on detecting properties that are general for any object, such as the closedness of boundaries. We use the code provided by [4] to generate bounding boxes ranked by the probability that they contain an object. We take the top 500 boxes to create a heatmap. The intensity of each pixel in this heatmap is proportional to the number of times it has been included in an objectness proposal[1]. We divide the heatmap into sub-regions at four different levels of resolution and evaluate the L_2 energy in each sub-region, creating a spatial pyramid [23]. This feature gives an indication of how objects are located in the image. We call this feature PyrObj.

Object Categories For each not fully connected layer of the DCNN, we construct a feature with only semantic information analogously to the feature with only spatial information. This image feature is $f_l[j] = \max_i y_l[i][j]$, and is of dimension $1 \times d_l$. This corresponds to max pooling over space. The last layers of the DCNN already capture object categories, as they transform the neural responses to object classification scores that contain little to no information about the location of the objects in the image.

Gist of the scene This descriptor of length 512, introduced by [28], gives a representation of the structure of real world scenes where local object information is discarded. Scenes belonging to the same semantic categories (such as streets, highways and coasts) have similar GIST descriptors.

5. Experiments

We now report results on the MIT benchmark [17] and PASCAL saliency dataset [26] (introduced in section 2).

5.1. Predicting the Saliency Map Accuracy

Performance of the Predictor of the Saliency Map Accuracy Fig. 2 shows the results for predicting saliency

---

[1]This heatmap, when normalised, was also evaluated as a saliency map. Interestingly, it achieved results close to the Judd *et al.* model [16] on AUC and NSS metrics (objectness heatmap achieves AUC = 0.83, NSS = 1.23; Judd model achieves AUC = 0.81, NSS = 1.18). This could be explained by the fact that objects predict fixations better than low level features [7].
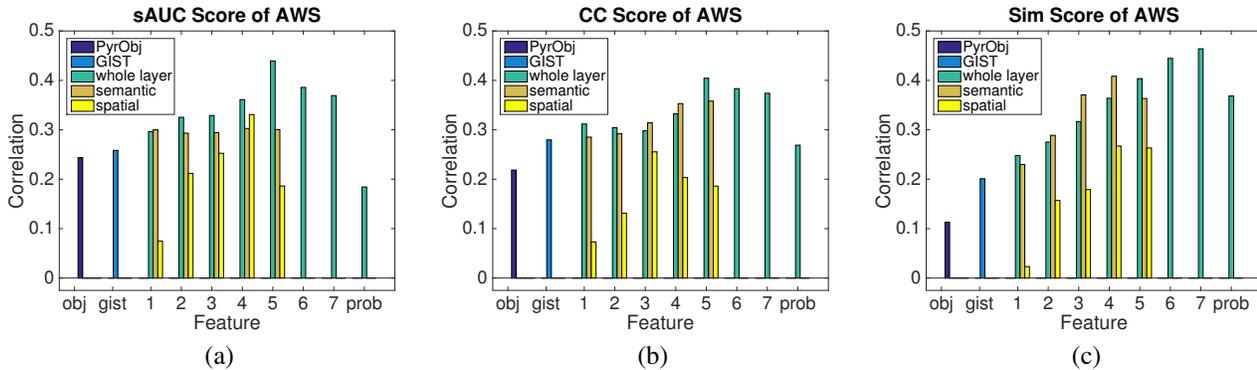
Figure 2: *Evaluation of the Prediction of the Saliency Accuracy.* The correlation between the predicted accuracy and the true accuracy of the saliency map is evaluated using different input features (including each of the 7 layers of the DCNN). The metric used to evaluate the accuracy of the saliency map is (a) sAUC, (b) CC, and (c) Sim. The trend has the same shape for all methods.

model accuracy for the different features we have introduced. We report the Spearman correlation between the true and predicted values. The results show that the PyrObj objectness feature can partially describe the object distribution and performs similarly to the spatial features of the DCNN. In general, Gist performs better than PyrObj, on par with the best spatial feature. Interestingly, we see that the semantic feature is much more informative for predicting consistency than the spatial feature, which suggests that semantic information has a greater contribution to predicting saliency map accuracy than information about the distribution of the objects.

Also, we can observe that some of the differences between the performance of the features are due to overfitting. Features with fewer dimensions (highest layers, or features with only semantic or spatial information) achieve better generalisation than features with more dimensions, until saturation (and under-fitting), and this point depends on the metric (predicting CC suffers less over-fitting than sAUC). In fact, the performance decreases significantly (< 0.10) when training using a concatenation of neural activations from all layers.

Finally, note that the best performing feature is the whole layer of the DCNN, achieving a ρ of above 0.4 for all metrics. To show that this performance is also obtained in other datasets, we evaluate our model on the PASCAL dataset and summarise our results in Table 1. The accuracy prediction of the saliency models performs similarly or better on PASCAL dataset, with a maximum correlation of 0.80 *vs* 0.52 on MIT. This results show that our method provides a useful prediction to automatically assess the quality of saliency. Also, when this prediction is used to select the best algorithm for saliency prediction per image, we find a (modest) absolute improvement of about 1%.

Fig. 3 shows some examples of images for which

	MIT			PASCAL		
	sAUC	CC	Sim	sAUC	CC	Sim
SALICON	0.39	0.52	0.49	0.33	0.52	0.49
BMS	0.44	0.43	0.42	0.38	0.62	0.72
GBVS	0.46	0.39	0.43	0.53	0.64	0.61
AIM	0.50	0.36	0.43	0.44	0.72	0.62
AWS	0.48	0.41	0.42	0.41	0.72	0.70
SUN	0.51	0.39	0.45	0.48	0.80	0.71
IttiKoch	0.52	0.43	0.42	0.50	0.62	0.57

Table 1: *Evaluation of the Prediction of the Saliency Accuracy.* Spearman correlation between the predicted accuracy of the saliency map using layer 5 of the DCNN and the ground truth accuracy.

	Consistency			Entropy Fixation Map		
	sAUC	CC	Sim	sAUC	CC	Sim
SALICON	0.63	0.44	−0.10	−0.55	−0.52	−0.04
BMS	0.34	−0.19	0.81	−0.50	0.26	0.94
GBVS	0.27	−0.24	0.81	−0.46	0.31	0.94
AIM	0.32	−0.31	0.82	−0.49	0.40	0.95
AWS	0.32	−0.25	0.82	−0.49	0.34	0.95
SUN	0.18	−0.33	0.83	−0.41	0.45	0.95
IttiKoch	0.22	−0.29	0.82	−0.41	0.37	0.94

Table 2: *Does the Accuracy of the Saliency Map Predict the Eye Fixation Consistency?* Spearman correlation between the accuracy of the saliency map and (left) the consistency ($K = 7$ and $S = 15$, with the same metric consistency and accuracy evaluation), and (right) with the entropy of the fixation map.

saliency model accuracy is high and low *vs* predictable and unpredictable.

Does the Accuracy of the Saliency Map Predict the Eye Fixation Consistency? Now that we have shown that the accuracy of the saliency map can be predicted from our model, the reader may ask whether we really need a dif-
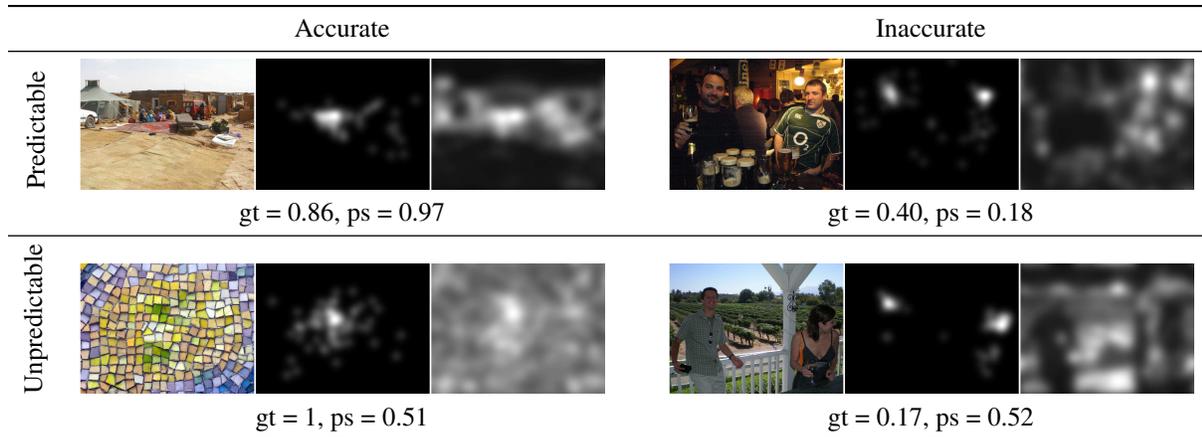
Figure 3: *Qualitative results.* We show images that have accurate and inaccurate AWS saliency maps (under the cross-correlation metric). gt is the ground truth which corresponds to the cross-correlation score of the saliency map, ps is the predicted score. Scores are predicted by the whole fifth layer feature. Both scores and predictions scaled between 0 and 1. The images are place in a row: original, fixation map, saliency map.

ferent model to predict the eye fixation consistency. If the tasks were similar enough there would be no need to introduce two different models. In [16], Judd *et al.* qualitatively analyse the consistency of the eye fixations and the saliency map accuracy, and suggest that there maybe be a relationship, but it depends on the evaluation metric. We extend this result by directly evaluating the Spearman correlation between the accuracy of the saliency map and eye fixation consistency. This is shown in Table 2 for all methods. All the correlations are low for sAUC and CC metrics (around $0.2 - 0.3$), and high for Sim (around 0.82)[2].

Yet, SALICON shows the opposite trend, as the rest of models have added blur for optimal performance, while SALICON remains peaky. If we use the entropy of the fixation map as a consistency metric (Table 2, right) we see the same situation. Thus, in general, a different model is needed to predict the eye fixation consistency independently on the metric used for the accuracy.

5.2. Predicting the Eye Fixations Consistency

Analysis of the Metric of the Eye Fixation Consistency
Recall that the metric we use to evaluate the eye fixation consistency, tests whether the fixation map computed from a subset of subjects ($M_{\mathcal{H}}$) is similar the fixation map computed from the rest of the subjects ($M_{\mathcal{O} \setminus \mathcal{H}}$). Eye fixation consistency may vary depending on the number of subjects

in $\mathcal{H}$, *i.e.* K. For low values of K, consistency is lower than for high values of K because the individual characteristics of each subject have not been averaged with other subjects. In the limiting case of having infinite subjects, increasing K will eventually lead to the consistency score saturating. In fact, in many works on saliency prediction, the value of the evaluation metric at $K \rightarrow \infty$ is used as an upper bound of the achievable prediction score. Thus, to characterise consistency, we need to report results using different numbers of subjects. We test K up to a value equal to half of the number of subjects in the dataset, as when $|\mathcal{O} \setminus \mathcal{H}|$ becomes small, $M_{\mathcal{O} \setminus \mathcal{H}}$ does not represent the totality of the users well anymore. Besides K, consistency also depends on the subjects used to compute $M_{\mathcal{H}}$, which may introduce some bias specific to the group of subjects $\mathcal{H}$. To remove this variability, we evaluate the consistency multiple times with different $\mathcal{H}$, and average the consistency scores. Let S be the number of different $\mathcal{H}$ sets used to compute the average. In Fig. 4a, we show the mean consistency score as a function of K, and we see that it increases because $\mathcal{H}$ becomes more representative as more subjects are added to it. Then, Fig. 4b shows that our metric is not dependent on the particular subject in each group. To show this, we check that the consistency scores do not vary when we choose different subjects in $\mathcal{H}$. We compute the average consistency two times for different groups $\mathcal{H}$, and then, we compute the Spearman correlation between these two consistency scores. This procedure is repeated ten times and the results averaged. Thus, Fig. 4b shows the average correlation between two measures of the eye fixation consistency for different subjects in $\mathcal{H}$ (for different K, and different number of groups averaged to obtain the consistency score, S). We can see that for any K, when S is sufficiently

---

[2]The Sim metric tends to assign higher scores when the eye fixation maps are relatively flat, independently of the saliency map. Recall that the Sim metric calculates intersection distance, *i.e.* the sum for all pixels of the minimum value between the saliency and eye fixation maps. If the eye fixation map is flat (inconsistent eye fixations), most pixel values $\approx \frac{1}{\text{num pixels}}$, while if the map has peaks (consistent eye fixations), the most pixel values (and the minimum) are 0. As a result, a flat (inconsistent) eye fixation map is likelier to have higher Sim than other eye fixation maps.
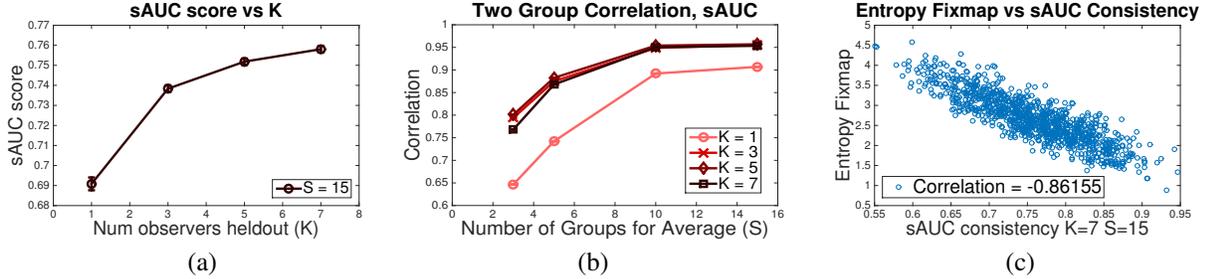
Figure 4: *Analysis of the Metric of the Eye Fixation Consistency*. (a) Consistency metric as the average of the sAUC score between the eye fixation maps of two groups of subjects of K and $15-K$ subjects, respectively. (b) Correlation between two consistency metrics evaluated with different subjects. (c) Correlation between entropy of the eye fixation map and consistency of the eye fixations. The trends are similar for Sim and CC consistency.

large to average the possible individual characteristics of the groups of subjects ($S > 10$), the correlation becomes about 0.95, *i.e.* almost the maximum possible correlation. This shows that the consistency score does not depend on the subjects in $\mathcal{H}$ when S is sufficiently large.

Finally, we check the agreement between the consistency metric we use, and the metric based on the entropy of the eye fixation map used in previous works [17]. Recall from section 3.2, that the entropy may not capture some cases of inconsistency of eye fixations, because the assumptions that inconsistent viewing patterns will yield a flat fixation map, while consistent ones will yield a map with sharp peaks may not always hold (*e.g.* an eye fixation map with few peaks but each subject only looks at a subset of these peaks). To show this point, we plot our consistency metric against the entropy of the eye fixation map in Fig. 4c (using $K = 7$, $S = 15$, and the Sim metric for consistency). We can see that the correlation is negative because the entropy measures inconsistency rather than consistency. Although the correlation is quite high, about 0.85, we see that the entropy does not fully capture the consistency of eye fixations.

Performance of the Predictor of the Eye Fixation Consistency We now evaluate the performance of the prediction of the eye fixation consistency. We report the Spearman correlation between the true and predicted values in Fig. 5. The same features perform well as in the accuracy prediction task, although the correlation values are higher in this task, achieving a ρ of around 0.5. Subsequent layers outperform the preceding ones, except of the last *prob* layer, which performs slightly worse. This could happen because the *prob* layer has lost all spatial information.

The previous work that also used machine learning to predict the eye fixation consistency [24], reports a Pearson correlation of 0.27 on a set of 27 images they have selected at hand, which shows the challenge of this task. Our results substantially improve over previous work, mainly because we use features based on object recognition. Our results

	Consistency			Entropy
	sAUC	CC	Sim	Fixation Map
SALICON	-0.48	-0.42	-0.44	0.48
BMS	0.09	0.18	0.16	-0.15
GBVS	-0.04	0.06	0.03	-0.01
AIM	0.02	0.09	0.10	-0.10
AWS	0.07	0.15	0.15	-0.14
SUN	-0.05	0.01	-0.01	0.00
IttiKoch	0.15	0.23	0.25	-0.22

Table 3: *Is the Eye Fixation Consistency Predicted by the Entropy of the Saliency Map?* Correlation between the entropy of the saliency map and (left) the eye fixation consistency ($K = 7$ and $S = 15$, with the same metric consistency and accuracy evaluation), and (right) entropy of the fixation map.

reveal that the eye fixation consistency among subjects is an attribute of natural images that can be predicted.

In Fig. 6, we present images that are consistent and inconsistent *vs* predictable and unpredictable. We see several examples of the consistency being at odds with the entropy.

Is the Eye Fixation Consistency Predicted by the Entropy of the Saliency Map? Finally, to make sure that the prediction of the eye fixation consistency enriches the information of the saliency map, we check whether eye fixation consistency information is already encoded in saliency maps. Recall that we showed that the entropy of the eye fixation map is correlated with eye fixation consistency. Thus, if the saliency map predicts eye fixation consistency, this would be encoded in the entropy of the saliency map. In Table 3, we report the correlation between the entropy of the saliency map and the consistency of the fixations based on the three metrics. All methods had a weak correlation (≤ 0.25), except SALICON. We suggest that the leading performance of SALICON on benchmarks is due to it encoding consistency much better than other methods. Note that our computational model can also enrich the saliency map of SALICON, as our computational model predicts the
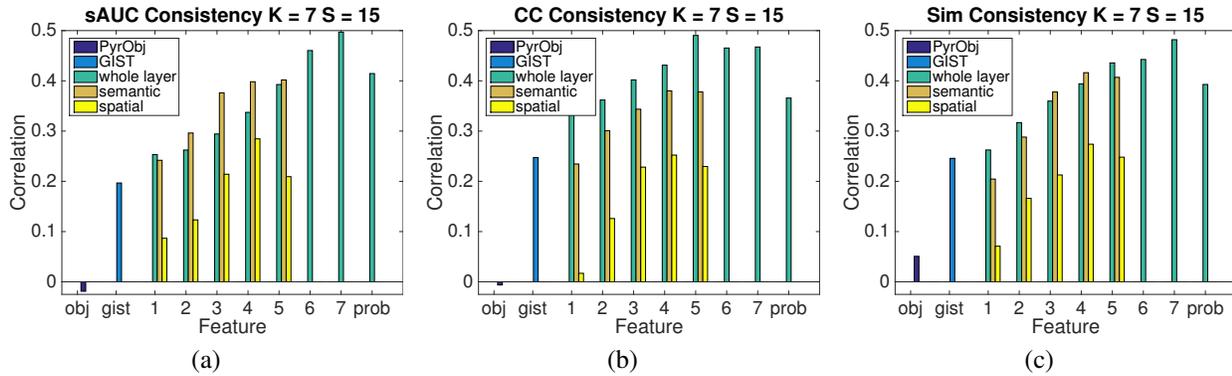
Figure 5: *Evaluation of the Prediction of the Eye Fixations Consistency.* The correlation between the predicted consistency of the eye fixation and the true consistency is evaluated using different input features (including each of the 7 layers of the DCNN). The metric used to evaluate the consistency uses $K = 7$ subjects and $S = 15$ groups and is based on: (a) sAUC, (b) CC, and (c) Sim. The results show a similar trend with different values of K
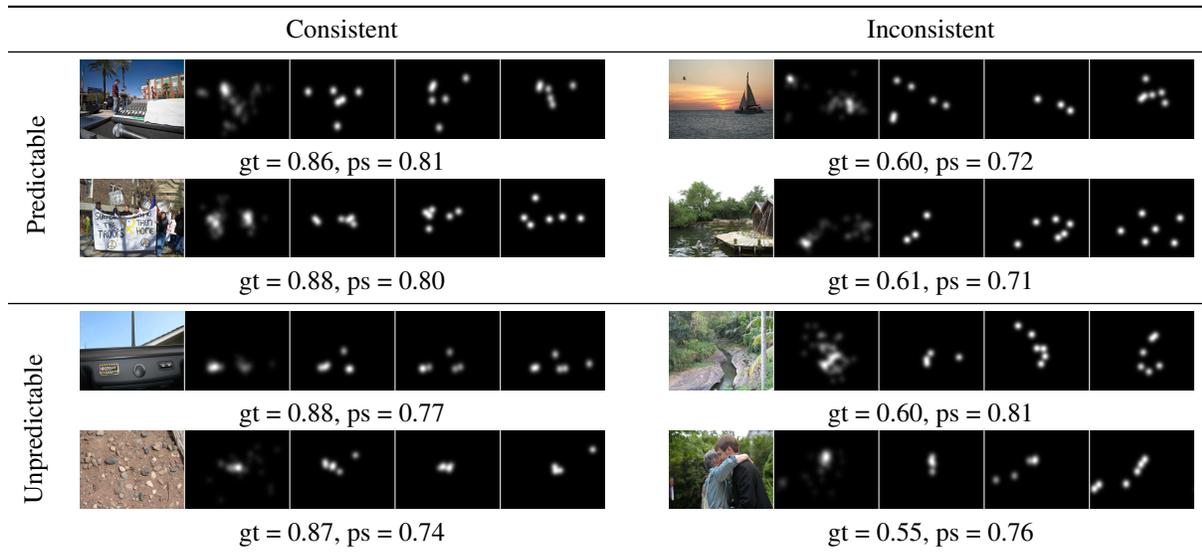

Figure 6: *Qualitative results.* We show images that have consistent and inconsistent viewing patterns. gt is ground truth which corresponds to the eye fixation consistency measure, ps is the predicted score.

consistency more accurately than SALICON.

6. Conclusions

We used machine learning techniques and automatic feature extraction to predict the accuracy of saliency maps and the eye fixation consistency among subjects in natural images. This was possible due to the good performance of DCNNs for object recognition, since eye fixations locations are strongly related to the object categories. Our results showed that saliency models can be enriched with the two predictions made from our model, because saliency models themselves do not capture eye fixation consistency among subjects, and their accuracy has not been estimated for a given

image. Also, we observed that the eye fixation consistency among subjects is an attribute of natural images that can be predicted from object categories. We expect that all these results allow for numerous applications in computer vision and visual design.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the object-ness of image windows. *TPAMI*, 2012. 4

[2] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, 2013. 2

[3] N. Bruce and J. Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007. 2

[4] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Bina-rized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014. 4

[5] H. F. Chua, J. E. Boland, and R. E. Nisbett. Cultural variation in eye movements during scene perception. *PNAS*, 2005. 1

[6] K. M. Dalton, B. M. Nacewicz, T. Johnstone, H. S. Schae-fer, M. A. Gernsbacher, H. Goldsmith, A. L. Alexander, and R. J. Davidson. Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 2005. 1

[7] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixa-tions better than early saliency. *Journal of Vision*, 2008. 4

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) chal-lenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2

[9] A. Garcia-Diaz, V. Leboran, X. R. Fdez-Vidal, and X. M. Pardo. On the relationship between optical variability, vi-sual saliency, and eye fixations: A computational approach. *Journal of Vision*, 2012. 2

[10] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. The interestingness of images. In *ICCV*, 2013. 2

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2007. 1, 2

[12] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reduc-ing the semantic gap in saliency prediction by adapting deep neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[13] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *CVPR*, 2011. 2

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

[15] M. Jiang, X. Boix, G. Roig, J. Xu, L. Van Gool, and Q. Zhao. Learning to predict sequences human visual fix-ations. *TNNLS*, 2015. 1

[16] T. Judd, F. Durand, and A. Torralba. A benchmark of com-putational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 2, 3, 4, 6

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009. 1, 2, 3, 4, 7

[18] W. Kienzle, F. Wichmann, B. Schölkopf, and M. Franz. A nonparametric approach to bottom-up visual saliency. In *NIPS*, 2006. 1

[19] R. Kliegl, A. Nuthmann, and R. Engbert. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psy-chology: General*, 2006. 1

[20] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, 2002. 1

[21] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol-ogy*, 1985. 1

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4

[23] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 4

[24] O. Le Meur, T. Baccino, and A. Roumy. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *ACM international conference on Mul-timedia*, 2011. 1, 7

[25] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In D. S. Touretzky, editor, *Advances in Neural Information Process-ing Systems 2*, pages 396–404. Morgan-Kaufmann, 1990. 4

[26] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The se-crets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion*, pages 280–287, 2014. 2, 4

[27] S. Mathe and C. Sminchisescu. Action from Still Images Datasets and Models to Learn Task Specific Human Visual Scanpaths. In *NIPS*, 2013. 1

[28] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 4

[29] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computa-tional biology*, 2009. 4

[30] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Du-toit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *ICCV*, 2013. 3

[31] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Hender-son. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 2006. 3

[32] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 1980. 1

[33] S. Ungerleider. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 2000. 1

[34] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 1995. 4

[35] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 2006. 1

[36] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 4

[37] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, 2013. 2

[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cot-trell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008. 2