A Data-driven Metric for Comprehensive Evaluation of Saliency Models

Jia Li[1,2*], Changqun Xia[1], Yafei Song[1], Shu Fang[3], Xiaowu Chen[1*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University

[2]International Research Institute for Multidisciplinary Science, Beihang University

[3]Cooperative Medianet Innovation Center, School of EE & CS, Peking University

Abstract

In the past decades, hundreds of saliency models have been proposed for fixation prediction, along with dozens of evaluation metrics. However, existing metrics, which are often ***heuristically designed****, may draw conflict conclusions in comparing saliency models. As a consequence, it becomes somehow confusing on the selection of metrics in comparing new models with state-of-the-arts. To address this problem, we propose a data-driven metric for comprehensive evaluation of saliency models. Instead of heuristically designing such a metric, we first conduct extensive subjective tests to find how saliency maps are assessed by the human-being. Based on the user data collected in the tests, nine representative evaluation metrics are directly compared by quantizing their performances in assessing saliency maps. Moreover, we propose to* ***learn a data-driven metric*** *by using Convolutional Neural Network. Compared with existing metrics, experimental results show that the data-driven metric performs the most consistently with the human-being in evaluating saliency maps as well as saliency models.*

1. Introduction

Due to the booming of visual saliency models in the past decades, model benchmarking has become a popular topic in the field of computer vision (*e.g.*, [3, 4, 7]). Usually, such large-scale benchmarking efforts require several evaluation metrics so as to simultaneously assess a saliency model, especially a fixation prediction model, from multiple perspectives. However, the performance of a saliency model may change remarkably when different *heuristically designed* metrics are used. As a consequence, it becomes somehow confusing on which metrics should be used and which models should be compared with in designing new saliency models.

Actually, this phenomenon has already been noticed by many researchers, and a lot of efforts have been spent on refining existing metrics. For each metric, many variants have been proposed to refine its performance, which, unfortunately, raise new problems. For example, the metric Kullback-Leibler Divergence (**KLD**) can be computed as the relative entropy between: 1) saliency histograms at recorded fixations and random points [3, 17, 18, 36]; 2) saliency histograms at recorded fixations and shuffled fixations randomly gathered from different images [29, 34, 38]; 3) saliency distributions over the estimated saliency map (ESM) and the ground-truth saliency map (GSM) [32]. Note that the computation of **KLD** can also take either the symmetric form [3] or the asymmetric form [32]. Without knowing the implementation details of these variants, it becomes difficult to directly compare saliency models, even though their performances on the same dataset have been reported by using the same metric (*e.g.*, **KLD**).

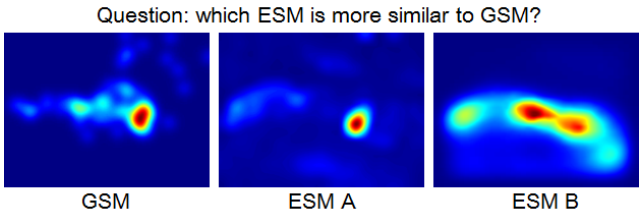


Figure 1. A representative question in the subjective tests. Multiple subjects are asked to determine which estimated saliency map (ESM) is more similar to the ground-truth saliency map (GSM).

To address this problem, this paper proposes a data-driven metric for comprehensive evaluation on saliency models. Instead of heuristically designing such a metric, we first conduct extensive subjective tests to find how saliency maps are assessed by the human-being. As shown in Fig. 1, subjects are asked to determine which of the two ESMs performs better in approximating the GSM (*i.e.*, the fixation density map). By collecting 50,400 binary annotations from 22 subjects, the performances of nine representative metrics are now quantized, which enables the direct comparisons between metrics. Based on the user data, we also propose to learn a comprehensive evaluation metric by us-

* Correspondence should be addressed to Jia Li and Xiaowu Chen. Email: {jiali, chen}@buaa.edu.cn

ing the Convolutional Neural Network (CNN). Compared with the heuristically designed metrics, experimental results show that the data-driven metric performs the most consistently with the human-being in evaluating saliency maps as well as saliency models. Moreover, we also provide the ranking lists of state-of-the-art saliency models by using the learned CNN-based metric.

The main contributions of this paper include: 1) We collect massive user data through subjective tests, which we promise to release so as to facilitate the design of robust and effective metrics for saliency model evaluation; 2) The performances of nine representative metrics are quantized for direct comparisons; 3) We propose a data-driven metric for saliency model evaluation, which performs the most consistently with the human-being.

2. A Brief Review of Evaluation Metrics

In the literature, there already exist many surveys of visual saliency models and evaluation metrics (*e.g.*, [5, 8, 32]). Thus we only briefly introduce nine representative metrics that are widely used in existing studies without elaborating their implementation details. Let **S** be an ESM and **G** be the corresponding GSM, some metrics select a set of positives and/or negatives from **G** so as to validate the "predictions" in **S**. Representative metrics that adopt such an evaluation methodology include $\phi_1 - \phi_5$.

Area Under the ROC Curve (**AUC**, ϕ_1). **AUC** is a classic metric that is widely used in many works (*e.g.*, [13, 20, 25]). It first selects all the fixated locations as positives and takes all the other locations as negatives. Multiple thresholds are then applied to **S**, and the numbers of true positives, true negatives, false positives and false negatives are computed at each threshold. Finally, the ROC curve can be plotted according to the true positive rate and false positive rate at each threshold. Perfect **S** leads to an **AUC** of 1.0, while random prediction has an **AUC** of 0.5.

Shuffled AUC (**sAUC**, ϕ_2). Since fixated locations often distribute around image centers (*i.e.*, the center-bias effect), the classic **AUC** favors saliency models that emphasize center regions or suppress peripheral regions. To address this problem, **sAUC** selects negatives as the fixated locations shuffled from other images in the same benchmark (*e.g.*, [15, 27, 37]). In this study, we adopt the implementation from [37] to compute **sAUC**.

Resampled AUC (**rAUC**, ϕ_3). One drawback of **sAUC** is that label ambiguity may arise when adjacent locations in images are simultaneously selected as positives and negatives (*e.g.*, locations from the same object). Due to the existence of such ambiguity, even the GSM **G** cannot reach a **sAUC** of 1.0, and such "upper-bound" may change on different images. To address this problem, Li *et al.* [24] proposed to re-sample negatives from non-fixated locations (*i.e.*, regions in **G** with low responses) according to the fixation distribution over the whole dataset. In this manner, the selected positives and negatives have similar distributions, and the ambiguity can be greatly alleviated in computing **rAUC**. Note that we re-implement this metric to enforce that the same number of positives and negatives are selected from each image.

Precision (**PRE**, ϕ_4). Metrics such as **AUC**, **sAUC** and **rAUC** only focus on the ordering of saliency [28, 39], while the saliency magnitude is ignored. To measure the saliency magnitudes at positives, **PRE** was proposed in [26] to measure the ratio of energy assigned only to positives (*i.e.*, fixated locations). In our implementation, we select positives and negatives as those used in computing **rAUC**.

Normalized Scan-path Saliency (**NSS**, ϕ_5). To avoid the selection of negatives, **NSS** only selects positives (*i.e.*, fixated locations [9, 31]). By normalizing **S** to zero mean and unit standard deviation, **NSS** computes the average saliency value at selected positives. Note that **NSS** is a kind of Z-score without explicit upper and lower bounds. The larger **NSS**, the better **S**.

Instead of explicitly selecting positives and/or negatives, some metrics propose to directly compare **S** and **G** as two probability distributions. Representative metrics that adopt such an evaluation methodology include $\phi_6 - \phi_9$.

Similarity (**SIM**, ϕ_6). As stated in [14], **SIM** can be computed by summing up the minimum saliency value at every location of **S** and **G**, while **S** and **G** are both normalized to sum up to one. **SIM** can be viewed as the intersection of two probability distribution, which falls in the dynamic range of $[0, 1]$. Larger **SIM** indicates a better **S**.

Correlation Coefficients (**CC**, ϕ_7). **CC** describes the linear relationship between two variables [1, 21]. It has a dynamic range of $[-1, 1]$. Larger **CC** indicates a higher similarity between **S** and **G**.

Kullback-Leibler Divergence (**KLD**, ϕ_8). **KLD** is an entropy-based metric that directly compares two probability distributions. In this study, we combine the **KLD** metrics in [3] and [32] to compute a symmetric **KLD** according to the saliency distributions over **S** and **G**. In this case, smaller **KLD** implies a better performance.

Earth Mover's Distance (**EMD**, ϕ_9). **EMD** measures the minimal cost to transform one distribution to the other one [9, 39]. Compared with $\phi_1 - \phi_8$, the computation of **EMD** is often very slow since it requires complex optimization processes. Smaller **EMD** indicates a better performance.

Most existing saliency models adopted $\phi_1 - \phi_9$ for performance evaluation. There also exist some works for metric analysis. For example, Riche *et al.* [32] investigated the correlation between metrics and provided several ranking

lists of saliency models. Emami *et al.* [8] adopted human fixation to identify the best metric. These works usually built upon a latent assumption that existing metrics are consistent with human perception, which, however, may not always hold (*e.g.*, in the extensive subjective tests we conducted). Therefore, it is still difficult to find the best saliency models to date unless the performances, or reliabilities, of various metrics can be quantized and directly compared.

3. Subjective Tests for Metric Analysis

In this section, we conduct extensive subjective tests to find how saliency maps are assessed by the human-being. Based on the user data collected in these tests, we carry out image-level and model-level analysis to quantize and compare the performance of nine representative metrics $\phi_1 - \phi_9$.

3.1. Subjective Tests

To conduct the subjective tests, we select 300 images from two image fixation datasets, including 120 images from **Toronto** [6] and 180 images from **MIT** [20]. For each image, we generate 7 ESMs with 7 saliency models, including $\mathcal{M}_0$ (AVG, which simply outputs the average fixation density map from **Toronto** or **MIT**, see Fig. 2), $\mathcal{M}_1$ (IT [19]), $\mathcal{M}_2$ (GB [13]), $\mathcal{M}_3$ (CA [11]), $\mathcal{M}_4$ (BMS [37]), $\mathcal{M}_5$ (HFT [22]) and $\mathcal{M}_6$ (SP [24]). For each of the 300 images, these 7 ESMs form $C_7^2 = 21$ ESM pairs.

Based on the ESM pairs, we carry out subjective tests with $300 \times 21 = 6,300$ questions in total. As shown in Fig. 1, each question consists of a pair of ESMs and the corresponding GSM (*i.e.*, the fixation density map). A subject needs to determine which ESM is more similar to GSM, without knowing which two models are actually being compared. In total, 22 subjects (17 males and 5 females, aged from 22 to 29) participate in the tests. 4 subjects (3 males and 1 female) will answer all the questions, while the rest 18 subjects (14 males and 4 females) will answer a random number of questions. Note that each question will be presented to exactly 8 subjects, and all subjects know the meaning of colors in ESMs and GSMs. Finally, we obtain $6,300 \times 8 = 50,400$ answers (*i.e.*, binary annotations). For the sake of simplification, we represent the user data as

$$\{(\mathbf{S}_k^g, \mathbf{S}_k^p, \mathbf{G}_k), n_k | k \in \mathbb{I}\}, \quad (1)$$

where $\mathbb{I} = \{1, \ldots, 6300\}$ is the set of indices of all questions. $\mathbf{S}_k^g$ and $\mathbf{S}_k^p$ are the two ESMs being compared in the kth question, and the "good" ESM $\mathbf{S}_k^g$ performs better than or comparable to the "poor" ESM $\mathbf{S}_k^p$ in approximating $\mathbf{G}_k$. The integer label $n_k \in \{4, 5, 6, 7, 8\}$ records how many subjects (among the eight subjects) select the "good" ESM $\mathbf{S}_k^g$ in the kth question. Larger n_k implies higher confidence that $\mathbf{S}_k^g$ outperforms $\mathbf{S}_k^p$.

In the tests, subjects also report the reasons why they think certain saliency maps are "good" or "poor." By investigating these explanations, we find the following key factors that may affect the evaluation of saliency maps:

1) The most salient locations. In most cases, both ESMs unveil visual saliency to some extent, and the most salient regions play a critical role in determining which ESM performs better. In particular, the overlapping ratio of the most salient regions in ESM and GSM is the most important factor in assessing saliency maps.

2) Energy distribution. The compactness of salient locations is an important factor for assessing saliency maps as well. ESMs that only pop-out object borders are often considered to be unsatisfactory. Moreover, the cleanness of background is also taken into account in the evaluation.

3) Number and shapes of salient regions. A perfect ESM should contain exactly the same number of salient regions as in the corresponding GSM. Moreover, salient regions with simple and regular shapes are preferred.

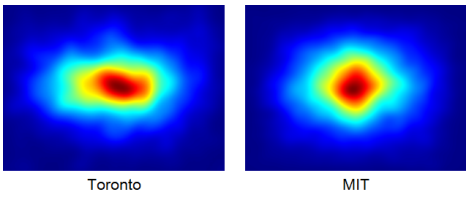


Figure 2. The average fixation density maps of **Toronto** and **MIT**.

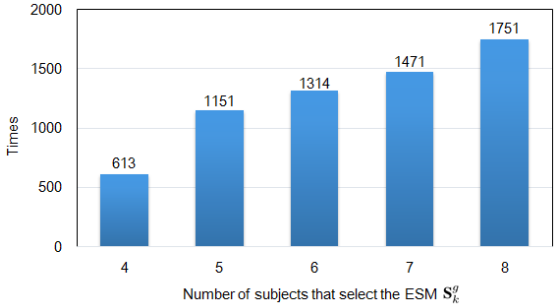
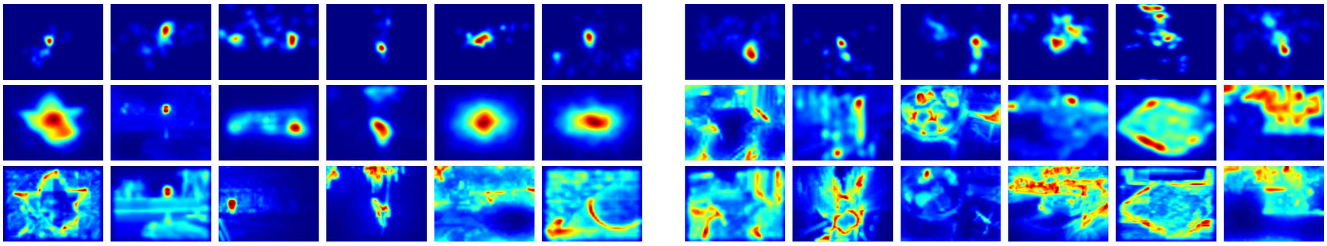


Figure 3. The histogram of user annotations over all the 6,300 questions. In most cases, the majority of subjects perform consistent in determining which ESM is better (*i.e.*, $n_k = 6, 7, 8$).

3.2. Statistics of User Data

Given the user data obtained from tests, we first address a concern that may arise: whether the annotations from various subjects stay consistent? Therefore, we show the distribution of $\{n_k | k \in \mathbb{I}\}$ in Fig. 3, from which we find that the majority of subjects act consistent in most cases. To further clarify that, we define two types of annotations, including:

1) Consistent annotations. In $4,536$ questions (72.0%), at least 6 subjects select the same ESMs (*i.e.*, $n_k = 6, 7, 8$). Such annotations often occur when one ESM performs significantly better than the other one (see Fig. 4 (a)).



(a) Consistent annotations (b) Ambiguous annotations

Figure 4. Representative examples of consistent and ambiguous annotations (1st row: GSM; 2nd and 3rd rows: ESMs). In (a), ESMs at the 2nd row outperform those at the 3rd row. Note that the ESMs from AVG, in certain cases, outperform the ESMs generated by $\mathcal{M}_1 - \mathcal{M}_6$ (the last two columns of (a)).

2) Ambiguous annotations. In $1,764$ questions (28.0%), the answers of eight subjects becomes ambiguous or even conflict (*i.e.*, $n_k = 4, 5$). Usually, both ESMs in most of these questions perform unsatisfactory and it is difficult to determine which ESM is better (see Fig. 4 (b)).

In the following studies, we will mainly rely on the user data obtained from the $4,536$ questions with consistent annotations. The indices of these questions are denoted as $\mathbb{C}$. Based on the ordered pairs in $\{(\mathbf{S}_k^q, \mathbf{S}_k^p, \mathbf{G}_k) \,|k \in \mathbb{C}\}$, we count the times that one model outperforms the other six models. The results of such one-vs-all comparisons are shown in Tab. 1, from which we obtain a subjective ranking list of the seven models:

$$\mathcal{M}_5 > \mathcal{M}_6 > \mathcal{M}_4 > \mathcal{M}_0 > \mathcal{M}_3 > \mathcal{M}_2 > \mathcal{M}_1. \quad (2)$$

We can see that state-of-the-arts (HFT, SP and BMS) outperform AVG and classic models (IT, GB and CA). Surprisingly, the average fixation density maps outperform ESMs from saliency models in 35.1% subjective comparisons. This indicates that it is unreasonable to heuristically design a metric that assigns the lowest score to AVG, since AVG outperforms many "poor" ESMs given by existing saliency models (see the last two columns of Fig. 4 (a)).

3.3. Analysis of Nine Representative Metrics

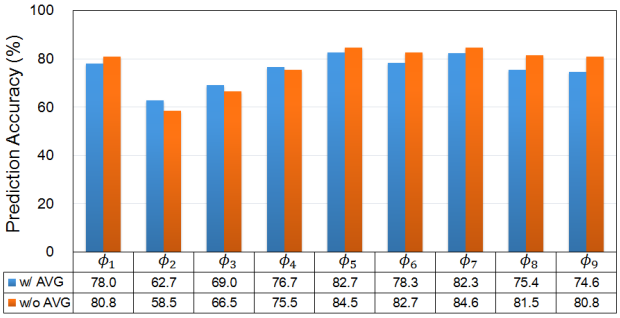
Based on the user data, we quantize the performance of $\phi_1 - \phi_9$ so as to directly compare them. The comparisons are conducted from two perspectives, including image-level and model-level comparisons. In image-level comparison, we aim to see if existing metrics can correctly predict which ESM acts better. Given a metric ϕ_i, its accuracy in predicting the ordering of ESMs can be computed as:

$$P_i = \frac{1}{|\mathbb{C}|} \cdot \begin{cases} \sum_{k\in\mathbb{C}} [\phi_i(\mathbf{S}_k^q) > \phi_i(\mathbf{S}_k^p)]_{\mathrm{I}}\,, & i = 1, ..., 7 \\ \sum_{k\in\mathbb{C}} [\phi_i(\mathbf{S}_k^q) < \phi_i(\mathbf{S}_k^p)]_{\mathrm{I}}\,, & i = 8, 9 \end{cases} \quad (3)$$

where $[\mathbf{e}]_{\mathrm{I}} = 1$ if the event $\mathbf{e}$ holds, otherwise $[\mathbf{e}]_{\mathrm{I}} = 0$. $|\mathbb{C}| = 4,536$ is the number of such ESM pairs with consistent annotations. Note that in (3) we omit the GSM $\mathbf{G}$ in $\phi_i(\mathbf{S}, \mathbf{G})$. The accuracies of nine metrics are shown in Fig. 5, from which we can draw several conclusions:

Table 1. The times and probability that a model outperforms all the other six models in $4,536$ questions

	# comparison	# winner	win rate (%)
$\mathcal{M}_0$	1,211	425	35.1
$\mathcal{M}_1$	1,320	256	19.4
$\mathcal{M}_2$	1,301	315	24.2
$\mathcal{M}_3$	1,257	370	29.4
$\mathcal{M}_4$	1,285	933	72.6
$\mathcal{M}_5$	1,339	1,124	83.9
$\mathcal{M}_6$	1,359	1,113	81.9



	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9
w/ AVG	78.0	62.7	69.0	76.7	82.7	78.3	82.3	75.4	74.6
w/o AVG	80.8	58.5	66.5	75.5	84.5	82.7	84.6	81.5	80.8

Figure 5. Prediction accuracy of nine representative evaluation metrics. Note that two accuracies are reported, which is computed with or without considering the ESMs generated by AVG.

1) The best metrics. The top three metrics that perform the most consistently with the human-being are ϕ_5 (**NSS**), ϕ_7 (**CC**) and ϕ_6 (**SIM**). Note that we also test the performances of these metrics when the ESMs generated by AVG are ignored in the evaluation. In this case, ϕ_7 (**CC**), ϕ_5 (**NSS**) and ϕ_6 (**SIM**) are still the best three metrics, which further validates their effectiveness and robustness. However, the best metric, **NSS**, only reaches an accuracy of 82.7% in comparing all the ESM pairs, while random prediction achieves an accuracy of 50% in such binary classification problems.

2) The worst metrics to handle AVG. When the ESMs from AVG are excluded in computing the prediction accuracy, the prediction accuracy of ϕ_8 (**KLD**) and ϕ_9 (**EMD**) increase by 6.1% and 6.2%, respectively. This may imply

that these two metrics are sensitive to post-processing operations such as center-bias Gaussian re-weighting.

3) Classic AUC is flawed. The classic **AUC** (ϕ_1) only ranks the 4th place, which may be caused by the fact that it relies solely on the interpolated ROC curve without considering the distribution of thresholding points [28]. In particular, if the ESMs from AVG are ignored in (3), **AUC** will even decrease to the 5th rank.

4) Shuffled metrics perform unsatisfactory. Shuffled metrics such as ϕ_2 (**sAUC**), ϕ_3 (**rAUC**) and ϕ_4 (**PRE**) perform inconsistent with subjects. Actually, it is insufficient to provide comprehensive evaluation if both positives and negatives are from center regions (imagine an ESM with wrongly popped-out regions at image corners). Moreover, ESMs generated by AVG often obtain extremely low scores when the shuffled metrics are used, which, however, outperform 35.1% ESMs generated by other models in subjective tests (see Tab. 1 and the last two columns of Fig. 4 (a)).

Beyond the image-level comparison, we also compare these nine representative metrics at model-level. That is, we generate a ranking list of the seven saliency models with each metric, and compare them with the ranks reported in (2). The ranking lists of models generated by various metrics, as well as the numbers of erroneously predicted model pairs, can be found in Fig. 6. We find that ϕ_5 (**NSS**), ϕ_7 (**CC**) and ϕ_1 (**AUC**) perform the best, while ϕ_2 (**sAUC**) and ϕ_3 (**rAUC**) still perform the worst. These results are almost consistent with those in the image-level comparison. In particular, we find recent models, such as HFT, SP and BMS, outperform AVG and classic models such as IT, GB and CA by using certain metrics. This implies the performances of saliency models keep on improving in the past decades, even though the evaluation metrics are imperfect.

4. Learning a Comprehensive Evaluation Metric with Convolutional Neural Network

From the quantized performances, we find that the nine representative metrics perform somehow inconsistent with subjects. The best metric **NSS**, which successfully ranks all the seven models, achieves only 82.7% agreement with subjects in assessing saliency maps. Therefore, it is necessary to develop a metric which can assess saliency maps as the human-being does. Toward this end, we propose to *learn* a comprehensive metric $\phi_L(\mathbf{S}^1, \mathbf{S}^2, \mathbf{G})$ from the user data. Different from existing metrics, the learned metric focuses on the *ranking* of $\mathbf{S}^1$ and $\mathbf{S}^2$. In other words, we treat the CNN-based metric as a binary classifier and optimize its parameters so as to maximize its accuracy on classifying the correlation of two ESMs. By using this metric, the comparison between two saliency models can be conducted by measuring the times (and probability) that ESMs from one model outperform those from the other model.

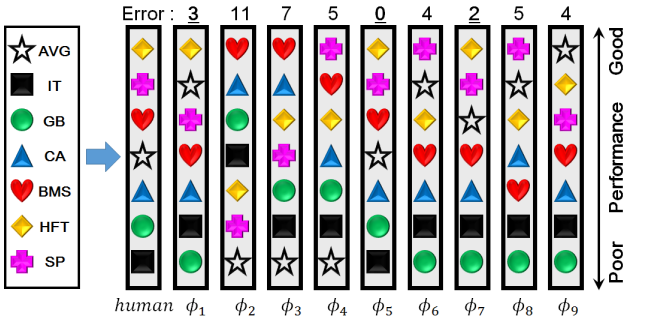


Figure 6. The ranking lists of seven models generated by nine representative evaluation metrics. The number above each bar indicates how many pairs of models are wrongly ranked.

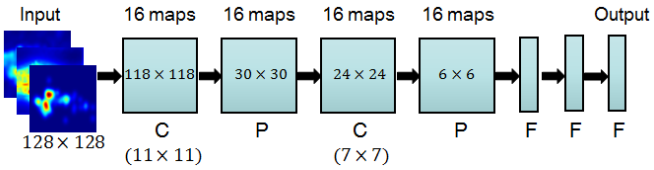


Figure 7. The structure of the Convolutional Neural Network (C: Convolutional Layer; P: Max pooling layer; F: Full connection layer). The CNN takes two ESMs ($\mathbf{S}^1$ and $\mathbf{S}^2$) and one GSM ($\mathbf{G}$) as input, which are resized to the resolution of 128×128. A soft-max function is applied to the output of CNN so as to infer a binary label, which equals to 1 if $\mathbf{S}^1$ outperforms $\mathbf{S}^2$, and 0 otherwise.

The structure of CNN is shown in Fig. 7, which consists of 8 layers in total. The input layer takes two ESMs $\mathbf{S}^1$ and $\mathbf{S}^2$ and one GSM $\mathbf{G}$ as the input. Note that both ESMs and GSM are resized to the same resolution of 128×128 through bilinear interpolation. The 2nd and 4th layers are convolutional layers, and the sizes of convolutional kernels are 11×11 and 7×7, respectively. Note that we use rectified linear unit (ReLU) activation function [30] in all convolutional layers. The 3rd and 5th layers are max pooling layers that sub-sample the input over each 4×4 non-overlapping window. The last three layers are full connection layers, and the CNN will output a 2D feature vector. Finally, a soft-max function is adopted to generate a binary label, which equals to 1 if $\mathbf{S}^1$ outperforms $\mathbf{S}^2$, and 0 otherwise.

To train the CNN-based metric, we adopt the user data obtained in the 4,536 questions with consistent annotations (*i.e.*, $\{(\mathbf{S}_k^g, \mathbf{S}_k^p, \mathbf{G}_k) \,|k \in \mathbb{C}\}$, with binary label 1). Moreover, we expand the training data by swapping $\mathbf{S}_k^g$ and $\mathbf{S}_k^p$ (*i.e.*, $\{(\mathbf{S}_k^p, \mathbf{S}_k^g, \mathbf{G}_k) \,|k \in \mathbb{C}\}$, with binary label 0). To alleviate the over-fitting risk, we adopt the dropout technique by setting the output of each hidden neuron in the full connection layers to zero with probability 0.5. In this manner, we enforce the learning of more robust features that best fit for the comparison of ESMs. In the experiments, we optimize the CNN parameters with 80 feed-forward and back-propagation iterations, and it takes about 21.6s per iteration on a GPU platform (NVIDIA GTX 980). The testing speed

of the learned metric is much faster (*i.e.*, 0.085s to compare 100 pairs of ESMs preloaded into memory), since it only involves convolution, pooling and connection operations.

5. Experiments

In this section, we first conduct several experiments to validate the effectiveness of the learned metric. After that, we benchmark 23 saliency models with the data-driven metric to find the best saliency models.

5.1. Validation of the Learned Metric

We conduct three experiments to validate the effectiveness of the learned metric. In the first experiment, we train the CNN-based metric with the user data obtained on 250 randomly selected images (*i.e.*, $3783 \times 2 = 7566$ training instances with consistent annotations), and test the metric with the user data obtained on the rest 50 images (*i.e.*, $753 \times 2 = 1506$ testing instances with consistent annotations). The main objective is to validate the effectiveness, especially the generalization ability, of the learned metric. The prediction accuracies of the learned metric, when different numbers of iterations are reached in training CNN, are shown in Fig. 8. From Fig. 8, we can see that the prediction accuracy reaches up to 90.2% when 80 iterations are reached in training CNN. Note that on these testing images, the best heuristically designed metric, ϕ_5 (**NSS**), reach only an accuracy of 84.9% on these testing instances. This indicates that the data-driven metric performs the most consistently with human perception and can be generalized to the comparison of ESMs from new images.

In the second experiment, we re-train the CNN-based metric on all the user data with 80 iterations (*i.e.*, $4536 \times 2 = 9072$ training instances with consistent annotations), and test the metric on 9,072 synthesized data, including $\{(\mathbf{G}_k, \mathbf{S}_k^p, \mathbf{G}_k) \,|k \in \mathbb{C}\}$ (with binary label 1) and $\{(\mathbf{S}_k^p, \mathbf{G}_k, \mathbf{G}_k) \,|k \in \mathbb{C}\}$ (with binary label 0). Intuitively, the GSM $\mathbf{G}_k$ should always outperform the ESM $\mathbf{S}_k^p$ which performs worse than $\mathbf{S}_k^g$ in subjective tests. The objective of this experiment is to see whether the learned metric well captures this attribute. On the synthesized data, we find that prediction accuracy reaches 99.6%. This ensures that the fixation density maps always achieve the best performance, even though such synthesized data are not involved in training CNN.

In the third experiment, we test whether the CNN-based metric trained in the second experiment can be generalized to a completely new dataset. Toward this end, we select 100 images from the **ImgSal** dataset [22] and have the 21 ESM pairs on each image assessed by 8 subjects (6 males, 2 females, only two of them participated in the tests conducted on **MIT** and **Toronto**). Finally, we obtain $1,342$ consistent annotations and generate $1,342 \times 2 = 2,684$ testing instances. On these testing instances, the learned metric has an accuracy of 87.6%, while **NSS** reaches only 79.5%. These results further validates the generalization ability of the CNN-based metric, implying that it can be re-used on new datasets without being re-trained.

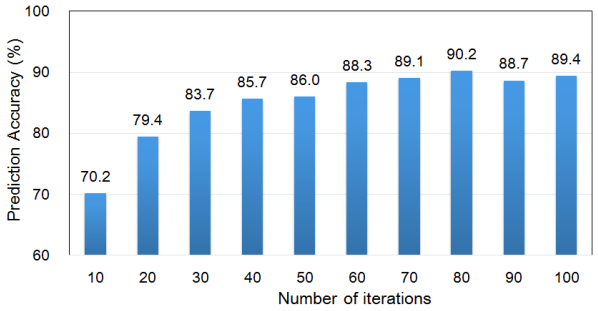


Figure 8. The performance of the data-driven metric when different numbers of iterations are reached in training CNN.

5.2. Benchmarking State-of-the-arts

Given the metric learned from all user data, we use it to benchmark state-of-the-art saliency models. In total, 23 models are involved, including three baselines: RND, AVG and GND. RND outputs a random ESM, while GND outputs the GSM of each image. The other 20 saliency models can be categorized into three groups, including:
1) The first group contains 8 bottom-up saliency models, including IT [19], GB [13], CA [11], RARE [33], AWS [10], LG [2], COV [9] and BMS [37].
2) The second group contains 7 models that utilize the prior knowledge obtained through unsupervised or supervised learning, including: AIM [6], SUN [38], ICL [17], JUD [20], SER [36], BST [1] and SP [24].
3) The third group contains 5 models that estimate visual saliency in the frequency domain, including: SR [16], PFT and PQFT [12], QDCT [35] and HFT [23].

Different from existing metrics that measure model performance by averaging the scores over all images, we adopt the one-vs-all comparisons to provide a comprehensive evaluation of saliency models (*i.e.*, the way we adopted in generating the subjective ranking list of models). We count the times and the "win rate" that a model outperforms all the other 22 models on all images. The more frequently a model outperforms other models, the better it is. As shown in Fig. 9, we think this one-vs-all ranking methodology can provide a more comprehensive evaluation of saliency models than directly using the average performance scores.

According to the times and win rate that a model outperforms all the other models, we provide three ranking lists of 23 models, as shown in Fig. 10. We can see that on all the testing images, HFT, SP and RARE are the top three models, and these three models also perform the best on both **Toronto** and **MIT**. Surprisingly, IT slightly outperforms GB by 0.7% due to the fact that IT wins in more

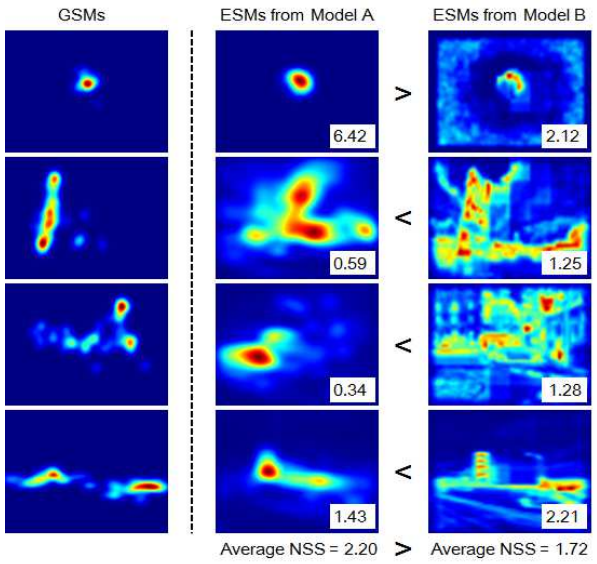


Figure 9. Models are compared according to the times of winning the competition instead of the average performance scores. From the **NSS** score of each ESM, model B outperforms model A three times on the four images. However, its average **NSS** score is still lower than that of model A, indicating a worse performance.

cases than GB in the comparisons with many other models beyond $\mathcal{M}_0 - \mathcal{M}_6$. Moreover, even the best model, HFT, only outperforms the other models in 80.7% comparisons (*i.e.*, 83.7% if RND and GND are excluded). This implies that existing saliency models perform far from satisfactory, and there is still a long way to go in the area of visual saliency estimation.

6. Discussion and Conclusion

In visual saliency estimation, many researchers have noticed that it is often insufficient to use only one metric in model comparison. Usually, only marginal improvements can be achieved with some metrics, while the scores given by certain metrics can be easily improved to a large extent by using simple tricks (*e.g.*, re-parameterizations, Gaussian smoothing, center-biased re-weighting and border cut). This makes the selection of evaluation metric a much confusing step in developing new saliency models. Therefore, it is necessary to address a long-standing concern: how to measure the performance (in other words, the reliability) of a metric in evaluating saliency maps and saliency models?

To compare various metrics, we conduct extensive subjective tests to find how saliency maps are assessed by subjects. By assuming that human performs the best in assessing saliency maps, we can thus provide a quantitative performance score for each metric. Among nine representative metrics, we find that **NSS** performs the most consistently with the human-being, while the classic **AUC** only ranks the $4th$ place with the prediction accuracy of 78.0%. That also explains the reason why human often thinks existing models are far from perfect in real-world applications, even though some models can achieve extremely high **AUC** scores. Moreover, shuffled metrics such as **sAUC** and **rAUC**, which are frequently used in recent studies, perform inconsistent with subjects. This is due to the fact that shuffled metrics are designed to alleviate the center-bias effect. However, it is often insufficient to focus only on the center regions. For instance, a distractor wrongly popped-out at the corner of ESM will be ignored in computing **sAUC**, since shuffled fixations distribute around image center as well. Therefore, the feasibility of using shuffled fixations in the evaluation should be carefully instigated.

In this study, we propose a data-driven metric for comprehensive evaluation of saliency models. This metric differs from existing metrics in two aspects: First, it is learned from user-data other than being heuristically designed. Second, it focuses on predicting the ordering of ESMs other than assigning each ESM a real-valued score. Experimental results show that this CNN-based metric outperforms nine representative metrics in assessing saliency maps. We also provide three ranking lists of 23 models to reveal the best saliency models. In the future work, we will incorporate eye-tracking devices so as to discover the latent mechanisms in assessing saliency maps. We will also explore the feasibility of designing new saliency models under the guidance of such a CNN-based metric.

Acknowledgement. This work was supported by grants from National Natural Science Foundation of China (61370113, 61421003 & 61325011), National Hightech R&D Program of China (2013AA013801), and Fundamental Research Funds for the Central Universities.

References

[1] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *CVPR*, pages 438–445, 2012. 2, 6

[2] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012. 6

[3] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2013. 1, 2

[4] A. Borji, D. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE TIP*, 22(1):55–69, 2013. 1

[5] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, pages 921–928, 2013. 2

[6] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, Vancouver, BC, Canada, 2005. 3, 6

[7] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/. 1

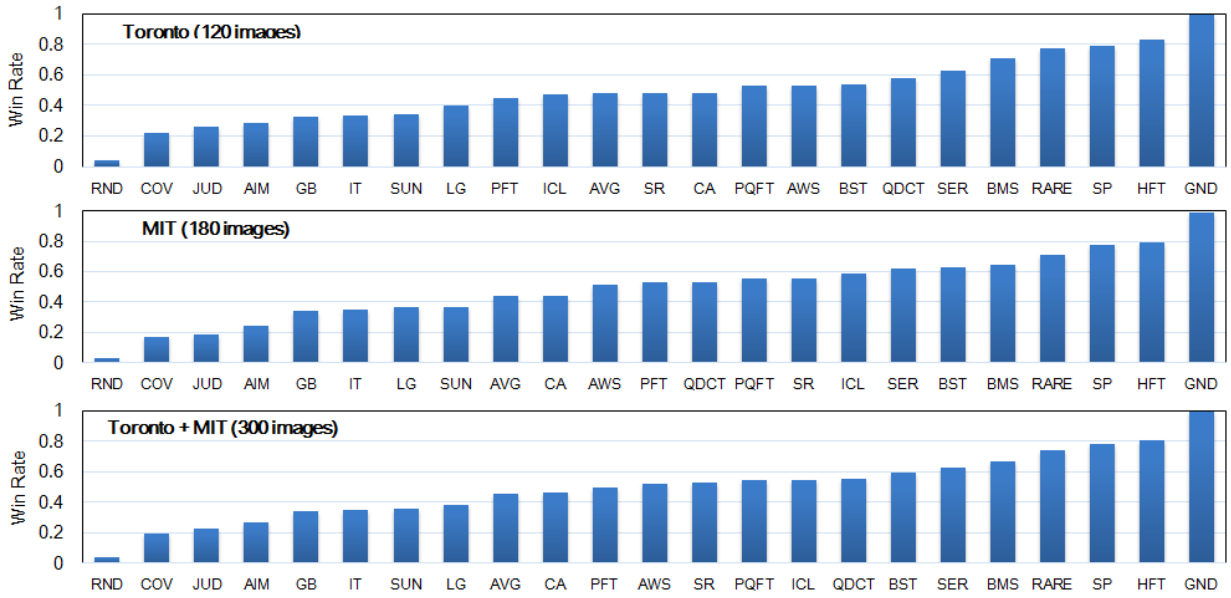


Figure 10. The ranking lists of 23 models obtained by using the data-driven metric learned from user data.

[8] M. Emami and L. L. Hoberock. Selection of a best metric and evaluation of bottom-up visual saliency models. *Image and Vision Computing*, 31(10):796–808, 2013. 2, 3

[9] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *JOV*, 13(4):11, 1–20, 2013. 2, 6

[10] A. Garcia-Diaz, V. Leborn, X. R. Fdez-Vidal, and X. M. Pardo. On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *JOV*, 12(6), 2012. 6

[11] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10):1915–1926, 2012. 3, 6

[12] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *CVPR*, pages 1–8, 2008. 6

[13] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2007. 2, 3, 6

[14] W. Hou, X. Gao, D. Tao, and X. Li. Visual saliency detection using information divergence. *PR*, 46(10):2658 – 2669, 2013. 2

[15] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE TPAMI*, 34(1):194–201, 2012. 2

[16] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8, 2007. 6

[17] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, pages 681–688, 2009. 1, 6

[18] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *NIPS*, pages 547–554, 2005. 1

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 3, 6

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 2, 3, 6

[21] C. Lang, G. Liu, J. Yu, and S. Yan. Saliency detection by multitask sparsity pursuit. *IEEE TIP*, 21(3):1327–1338, 2012. 2

[22] C. Li, J. Xue, N. Zheng, X. Lan, and Z. Tian. Spatio-temporal saliency perception via hypercomplex frequency spectral contrast. *Sensors*, 13(3):3409–3431, 2013. 3, 6

[23] J. Li, M. Levine, X. An, X. Xu, and H. He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE TPAMI*, 35(4):996–1010, 2013. 6

[24] J. Li, Y. Tian, and T. Huang. Visual saliency with statistical priors. *IJCV*, 107(3):239–253, 2014. 2, 3, 6

[25] J. Li, Y. Tian, T. Huang, and W. Gao. Probabilistic multi-task learning for visual saliency estimation in video. *IJCV*, 90(2):150–165, 2010. 2

[26] J. Li, D. Xu, and W. Gao. Removing label ambiguity in learning-based visual saliency estimation. *IEEE TIP*, 21(4):1513–1525, 2012. 2

[27] S. Lu, C. Tan, and J. Lim. Robust and efficient saliency modeling from image co-occurrence histograms. *IEEE TPAMI*, 36(1):195–201, 2014. 2

[28] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *CVPR*, 2014. 2, 5

[29] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR*, pages 433–440, 2011. 1

[30] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010. 5

[31] R. J. Peters and L. Itti. Congruence between model and human attention reveals unique signatures of critical visual events. In *NIPS*, Vancouver, BC, Canada, 2009. 2

[32] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *ICCV*, pages 1153–1160, 2013. 1, 2

[33] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *SPIC*, 28(6):642 – 658, 2013. 6

[34] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94:1–15, 2014. 1

[35] B. Schauerte and R. Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *ECCV*, pages 116–129, 2012. 6

[36] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In *CVPR*, pages 2368–2375, 2010. 1, 6

[37] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, pages 153–160, 2013. 2, 3, 6

[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A Bayesian framework for saliency using natural statistics. *JOV*, 8(7):32, 1–20, 2008. 1, 6

[39] Q. Zhao and C. Koch. Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *JOV*, 12(6):22, 1–15, 2012. 2