

# Learning a Saliency Evaluation Metric Using Crowdsourced Perceptual Judgments

Changqun Xia, Jia Li, Jinming Su and Ali Borji

arXiv:1806.10257v1 [cs.CV] 27 Jun 2018

**Abstract**—In the area of human fixation prediction, dozens of computational saliency models are proposed to reveal certain saliency characteristics under different assumptions and definitions. As a result, saliency model benchmarking often requires several evaluation metrics to simultaneously assess saliency models from multiple perspectives. However, most computational metrics are not designed to directly measure the perceptual similarity of saliency maps so that the evaluation results may be sometimes inconsistent with the subjective impression. To address this problem, this paper first conducts extensive subjective tests to find out how the visual similarities between saliency maps are perceived by humans. Based on the crowdsourced data collected in these tests, we conclude several key factors in assessing saliency maps and quantize the performance of existing metrics. Inspired by these factors, we propose to learn a saliency evaluation metric based on a two-stream convolutional neural network using crowdsourced perceptual judgements. Specifically, the relative saliency score of each pair from the crowdsourced data is utilized to regularize the network during the training process. By capturing the key factors shared by various subjects in comparing saliency maps, the learned metric better aligns with human perception of saliency maps, making it a good complement to the existing metrics. Experimental results validate that the learned metric can be generalized to the comparisons of saliency maps from new images, new datasets, new models and synthetic data. Due to the effectiveness of the learned metric, it also can be used to facilitate the development of new models for fixation prediction.

**Index Terms**—Visual saliency, Evaluation metric, Crowd-sourced perception judgements

## I. INTRODUCTION

IN the past decades, hundreds of visual saliency models have been proposed for fixation prediction. Typically, these models are designed to reveal certain characteristics of visual saliency under different assumptions and definitions, such as Local Irregularity [2]–[4], Global Rarity [5]–[7], Temporal Surprise [8]–[11], Entropy Maximization [12]–[16] and Center-bias [17]–[19]. Consequently, to conduct a comprehensive and fair comparison, it is necessary to evaluate saliency models from multiple perspectives.

Toward this end, researchers have proposed various metrics. For example, the Area Under the ROC Curve, referred

C. Xia, J. Li and J. Su are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China.

J. Li is also with the International Research Institute for Multidisciplinary Science, Beihang University, Beijing, China.

A. Borji is with the Center for Research in Computer Vision, Computer Science Department, University of Central Florida, Orlando, Florida.

An earlier version of this work has been published in ICCV [1].

Correspondence author: Jia Li. E-mail: jiali@buaa.edu.cn.

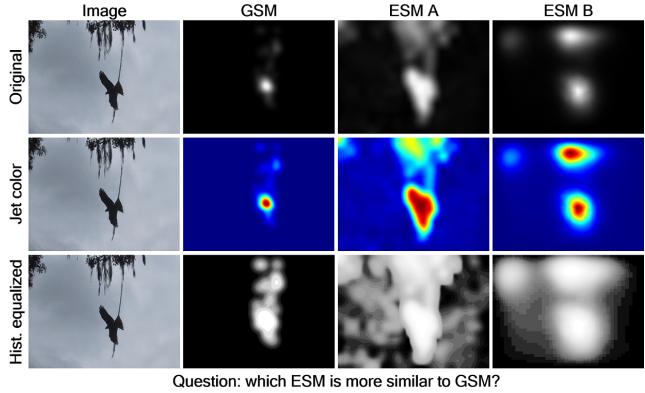


Fig. 1. Three different ways of visual comparison of saliency maps. The rows show the original fixation distribution, the jet color maps, and the histogram equalized maps, respectively. In our subjective tests, we design a set of questions each of which displays the original image, one ground-truth saliency map (GSM) and two estimated saliency maps (ESMs), where both GSM and ESMs are alternatively visualized with the histogram equalized maps as shown in the third row. Subjects are asked to determine which ESM is more similar to the displayed GSM. In this manner, the crowdsourced data collected from subjective tests contain useful cues about the human perception of saliency maps, which can be used to train a evaluation metric for measuring perceptual similarity of saliency maps .

to as **AUC**, is a popular metric to measure the tradeoff between true and false positives at various discrimination thresholds applied to the saliency map (e.g., [11], [20], [21]); the Normalized Scanpath Saliency (**NSS**) aims to measure visual saliency at fixated locations so as to be invariant to linear transformations [22]; the Similarity metric (**SIM**) is introduced to indicate the intersection between pairs of normalized saliency distributions [23]; Kullback-Leibler Divergence (**KLD**) is regarded as a measure of the information loss by evaluating saliency with a probabilistic interpretation [24], [25]; and the Earth Movers’s Distance (**EMD**) measures the cost transforming one distribution to another. All of these metrics evaluate saliency models quantitatively by calculating the performance scores [26].

Beyond these quantitative metrics, the qualitative visual comparison from multiple models is also adopted by almost all previous works. In Fig. 1, three different visualization ways of direct visual comparison of saliency maps are illustrated. Beginning with the influential model of [27], the original fixation distributions are displayed to qualitatively compare the performance of the corresponding saliency models (e.g., [20], [28]–[30]), but small values in the original fixation maps are not perceptible so that it can not always enter into the comparison process, such as the

tree branches shown in the first row of Fig. 1. In view of this point, the jet color maps shown in the second row are utilized to make a better perceptual judgment (*e.g.*, [31]–[33]). Moreover, Bruce *et al.* [34] propose a comparison based on histogram equalization as shown in the third row, wherein the spread of saliency values produced by each model is mapped as closely as possible into the same space. Actually, the human perception of visual similarity underlying these direct visual comparisons is very helpful to design new metrics that qualitatively assess saliency maps as humans do.

To investigate the key factors that influence the human perception of visual similarity in comparing the saliency maps, we conduct subjective tests. As shown in Fig. 1, we ask multiple subjects to determine which of the two estimated saliency maps (ESMs) is more similar to the ground-truth saliency map (GSM). How to visualize GSM and ESMs plays an important role in perceptual judgment of maps. It is difficult to make direct comparisons using the original fixation maps as shown in the first row since the maps vary greatly in their amount of salient pixels. Different from our previous work [1] which displays only ESMs and GSMs with jet colors as shown in the second row in Fig. 1, here we display histogram equalized maps in the third row so that small values also become perceptible and enter into the visual comparison process. In addition, the original image is displayed as well to facilitate the visual comparison process. We collect 134,400 binary annotations from 16 subjects. Through the analyses of the comparison process, we find four key factors that may affect the evaluation of saliency maps. Most importantly, there indeed exists consistency among the human subjective judgments.

Based on these findings and the crowdsourced data, we propose to learn a CNN-based saliency evaluation metric using crowdsourced perceptual judgements, and such a metric is abbreviated as **CPJ**. To optimize the parameters of **CPJ**, we design a two-stream CNN architecture, within which each stream corresponds to the same **CPJ** metric (*i.e.*, two CNNs with cloned parameters). The whole architecture takes two ESMs (denoted as  $A$  and  $B$ ) and one GSM (denoted as  $G$ ) as the input, while its two streams focus on predicting the relative saliency scores of  $(A, G)$  and  $(B, G)$ , respectively. After that, the difference between the scores given by the two streams is expected to approximate the crowdsourced perceptual judgement, which is represented by a numerical score to depict the relative performance of  $A$  and  $B$  in approximating  $G$ . Finally, each stream can be viewed as a evaluation metric that assigns a numerical performance score to reveal the perceptual similarity between an ESM and the corresponding GSM, which is the same as existing classic metrics. Compared with the ten representative metrics, experimental results show that the CNN-based metric has the highest consistency with the human perception in comparing the visual similarity between ESMs and GSM, making it a good complement to existing metrics. Besides, due to the effectiveness and characteristic of the learned metric, it can be used to develop new saliency models.

The main contributions of this paper include:

1) We collect massive crowdsourced data through subjective tests, based on which the performance of evaluation metrics in approximating human perception of saliency maps are directly quantized and compared.

2) The performance of ten representative metrics are quantized for direct comparison at image and model levels. Both perspectives prove that there still exists a large gap between the inherent characteristics of existing metrics and the human perception of visual similarity.

3) We propose a CNN-based metric that agrees better with perceptual similarity judgments of saliency maps. Experimental results show that the learned metric can be utilized for the assessment of saliency maps from new images, new datasets, new models and synthetic data.

The rest of this paper is organized as follows: Section II discusses related works. Section III presents details in subjective tests and analyzes ten representative metrics. Section IV proposes how to learn a comprehensive evaluation metric with CNNs. Extensive experiments are conducted in Section V to validate the learned metric. Finally, the paper is concluded in Section VI.

## II. RELATED WORK

In the literature, there already exist several surveys of saliency models and evaluation metrics (*e.g.*, [35]–[37]). Here, we first briefly introduce ten representative metrics that are widely used in existing studies. Besides, metric analyses are discussed in some references.

### A. Classic Evaluation Metrics

Let  $S$  be an ESM and  $G$  be the corresponding GSM, some metrics select a set of positives and/or negatives from  $G$ , which are then used to validate the predictions in  $S$ . Representative metrics that adopt such an evaluation methodology include  $\phi_1$  to  $\phi_5$ , explained below.

**Area Under the ROC Curve (AUC,  $\phi_1$ ).** AUC is a classic metric widely used in many works (*e.g.*, [11], [20], [21]). It 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, while random prediction has an AUC of 0.5. In this study, we adopt the implementation from [21] to compute AUC.

**Shuffled AUC (sAUC,  $\phi_2$ ).** Since fixated locations often distribute around image centers (*i.e.*, the center-bias effect), AUC favors saliency models that emphasize central regions or suppress peripheral regions. As a result, some models gain a remarkable improvement in AUC by simply using center-biased re-weighting or border-cut. To address this problem, sAUC selects negatives as the fixated locations shuffled from other images in the same benchmark (*e.g.*,

[30], [38], [39]). Since both fixated and non-fixated locations are both center-biased, simple center-biased re-weighting or border-cut will not dramatically change the performance in **sAUC**. In this study, we adopt the implementation from [30] to compute **sAUC**.

**Resampled AUC (rAUC,  $\phi_3$ ).** 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, and such “upper-bound” may change on different images. To address this problem, Li *et al.* [18] 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**. In this study, we adopt the implementation of **rAUC** from [2], [40], which selects the same amount of positives and negatives from each image.

**Precision (PRE,  $\phi_4$ ).** Metrics such as **AUC**, **sAUC** and **rAUC** mainly focus on the ordering of saliency and ignore the magnitude [41], [42]. To measure the saliency magnitudes at positives, **PRE** was proposed in [2], [43] to measure the ratio of energy assigned only to positives (*i.e.*, fixated locations, and **PRE** is denoted as Energy-on-Fixations in [40]). A perfect  $S$  leads to a **PRE** score of 1, and a “clean”  $S$  usually has a higher **PRE** score. In this study, we select positives and negatives as those used in computing **rAUC**, which is also the solution of [40].

**Normalized Scan-path Saliency (NSS,  $\phi_5$ ).** To avoid selecting negatives, **NSS** only selects positives (*i.e.*, fixated locations [44], [45]). 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 representative metrics propose to directly compare  $S$  and  $G$  by taking them as two probability distributions. Representative metrics that adopt such an evaluation methodology include  $\phi_6$  to  $\phi_{10}$ , explained below.

**Similarity (SIM,  $\phi_6$ ).** As stated in [14], **SIM** can be computed by summing up the minimum value at every location of the saliency maps  $S$  and  $G$ , while  $S$  and  $G$  are both normalized to sum up to one. From this definition, **SIM** can be viewed as the intersection of two probability distributions, which falls in the dynamic range of  $[0, 1]$ . Larger **SIM** scores indicate better ESMs.

**Correlation Coefficients (CC,  $\phi_7$ ).** **CC** describes the linear relationship between two variables [46], [47]. It has a dynamic range of  $[-1, 1]$ . Larger **CC** indicates a higher similarity between  $S$  and  $G$ .

**Information Gain (IG,  $\phi_8$ ).** **IG**, as an information theoretic metric, is defined as the entropic difference between the prior and the posterior distribution [48], [49]. **IG** is like

**KLD** but baseline-adjusted. In [50], **IG** over a center prior baseline provides a more intuitive way to interpret model performance relative to center bias.

**Kullback-Leibler Divergence (KLD,  $\phi_9$ ).** **KLD** is an entropy-based metric that directly compares two probability distributions. In this study, we combine the **KLD** metrics in [31] and [37] to compute a symmetric **KLD** according to the saliency distributions over  $S$  and  $G$ . In this case, smaller **KLD** indicates a better performance.

**Earth Mover’s Distance (EMD,  $\phi_{10}$ ).** The **EMD** metric measures the minimal cost to transform one distribution to the another [42], [44]. Compared with  $\phi_1 - \phi_9$ , the computation of **EMD** is often very slow since it requires complex optimization processes. Smaller **EMD** indicates a better performance.

Most existing works on fixation prediction adopted one or several metrics among  $\phi_1 - \phi_{10}$  for performance evaluation. In these works, we notice that the resolutions of  $S$  and  $G$ , as well as the interpolation methods to down-sample or up-sample  $S$  and  $G$  to the same resolutions, may change the scores of some metrics. Therefore, when using  $\phi_1 - \phi_{10}$ , we up-sample or down-sample  $S$  to the same size of  $G$  by bilinear interpolation. After that, we normalize  $S$  and  $G$  to have the minimum value 0 and the maximum value 1.

## B. Metric Analysis

Given these representative metrics, there also exist some works in metric analysis. Wilming *et al.* [51] explored how models of fixation selection can be evaluated. Through deriving a set of high-level desirable properties for metrics through theoretical considerations, they analyzed eight common measures with these requirements and concluded that no single measure can capture them all. Then both **AUC** and **KLD** were recommended to facilitate comparison of different models and to provide the most complete picture of model capabilities. Regardless of the measure, they argued that model evaluation was also influenced by inherent properties of eye-tracking data.

Judd *et al.* [23] provided an extensive review of the important computational models of saliency and quantitatively compared several saliency models against each other. They proposed a benchmark data set, containing 300 natural images with eye tracking data from 39 observers to compare the performance of available models. For measuring how well a model predicted where people look in images, three different metrics including **AUC**, **SIM** and **EMD** were utilized to conduct a more comprehensive evaluation. Besides, they showed that optimizing the model blurriness and bias towards the center to models improves performance of many models.

Emami *et al.* [36] proposed to identify the best metric in terms of human consistency. By introducing a set of experiments to judge the biological plausibility of visual saliency models, a procedure was proposed to evaluate nine metrics for comparing saliency maps using a database of human fixations on approximately 1000 images. This procedure was then employed to identify the best saliency

comparison metric as the one which best discriminates between a human saliency map and a random saliency map, as compared to the ground truth map.

Riche *et al.* [37] investigated the characteristics of existing metrics. To show which metrics are closest to each other and see which metric should be used to do an efficient benchmark, Kendall concordance coefficient is used to compare the relative rank of the saliency models according to the different metrics. Based on the nonlinear correlation coefficient, it is shown that some of the metrics are strongly correlated leading to a redundancy in the performance metrics reported in the available benchmarks. As a recommendation, **KLD** and **sAUC** are most different from the other metrics, including **AUC**, **CC**, **NSS**, and **SIM**, which formed a new measure.

Bruce *et al.* [34] argued that there existed room for further efforts addressing some very specific challenges for assessing models of visual saliency. Rather than considering fixation data, annotated image regions and stimulus patterns inspired by psychophysics to assess the performance benchmark of saliency models under varying conditions, they aimed to present a high level perspective in computational modeling of visual saliency with an emphasis on human visual behavior and neural computations. They analyzed the shortcomings and challenges in fixation-based benchmarking motivated by the spatial bias, scale and border effect. Besides, they further discussed the biological plausibility of models with respect to behavioral findings.

Kümmerer *et al.* [52] argued that a probabilistic definition is most intuitive for saliency models and explored the underlying reason why existing model comparison metrics give inconsistent results. They offered an information theoretic analysis of saliency evaluation by framing fixation predication models probabilistically and introduced the notion of information gain. Toward this end, they proposed to compare models by jointly optimizing factors such as scale, center bias and spatial blurring so as to obtain consistent model ranking across metrics. Besides, they provided a method to show where and how saliency models fail.

Bylinskii *et al.* [50] aimed to understand the essential reason why saliency models receive different ranks according to different evaluation metrics. Rather than providing tables of performance values and literature reviews of metrics, they offered a more comprehensive explanation of 8 common evaluation metrics and present visualization of the metric computations. By experimenting on synthetic and natural data, they revealed the particular strengths and weaknesses of each metric. Besides, they analyzed these metrics under various conditions to study their behavior, including the treatment of false positives and false negatives, systematic viewing biases, and relationship between metrics by measuring how related the rankings of the saliency models are across metrics. Building on the results of their analyses, they offered guidelines for designing saliency benchmarks and choosing appropriate metrics.

Although these works provide some insights into the advantages and disadvantages of existing representative quantitative metrics, it is still important to obtain a quantita-

tive metric that performs more consistently with humans in assessing visual saliency models. In our previous work [1], we conducted extensive subjective tests and proposed a data-driven metric using convolutional neural networks. Compared with existing metrics, the data-driven metric performs most consistently with the humans in evaluating saliency maps as well as saliency models. However, the jet colormap chosen for subjective tests may fall very far from having distances in colorspace that are a good match for perceptual distances. Besides, the evaluation depends on performing pairwise comparisons, which adds additional complications to assess the saliency model benchmarking.

Toward this end, we consider more reasonable perceptual factors and propose to directly quantify the performance of metrics by using the crowdsourced data collected from extensive subjective tests. Based on the crowdsourced perceptual judgements, a saliency evaluation metric is then learned by using the CNNs, which can qualitatively measure human perceptual similarity of saliency maps.

### III. SUBJECTIVE TESTS FOR METRIC ANALYSIS

In this section, we conduct subjective tests to study how saliency map are compared by the humans. Based on the crowdsourced data collected in these tests, we carry out extensive image-level and model-level analyses to quantify and compare the performance of existing metrics in terms of measuring the human perception of visual similarity.

#### A. Subjective Tests

In subjective tests, we select 400 images from three datasets, including 120 images from **Toronto** [16], 180 images from **MIT** [21] and 100 images from **ImgSal** [53]. Human fixations on these images are collected by different eye-tracking configurations, leading to lower dataset bias. For each image, we generate seven ESMs with seven saliency models, including  $\mathcal{M}_0$  (AVG),  $\mathcal{M}_1$  (IT [27]),  $\mathcal{M}_2$  (GB [20]),  $\mathcal{M}_3$  (CA [54]),  $\mathcal{M}_4$  (BMS [30]),  $\mathcal{M}_5$  (HFT [55]) and  $\mathcal{M}_6$  (SP [18]). Note that AVG simply outputs the average fixation density map from **Toronto**, **MIT** and **ImgSal** (see Fig. 2). For each image, the 7 ESMs form  $C_7^2 = 21$  ESM pairs. Based on the ESM pairs, we carry out subjective tests with  $400 \times 21 = 8,400$  questions.

Typically, the choice of colormap can impact significantly the human perception of visual similarity and can also play a role in shaping how saliency maps are judged, and what information is discernible. In our previous work [1], each question only consisted of two ESMs and one GSM displayed in jet color. Subject needs to determine which ESM is more similar to GSM, without knowing the models that generate the ESMs. In total, 22 subjects (17 males and 5 females, aged from 22 to 29) participated in the tests. Note that each question is presented to exactly 8 subjects, and all subjects know the meaning of colors in ESMs and GSMS (i.e., which colors correspond to the most salient locations and which colors indicate background regions). In the subjective tests, there is no time limitation

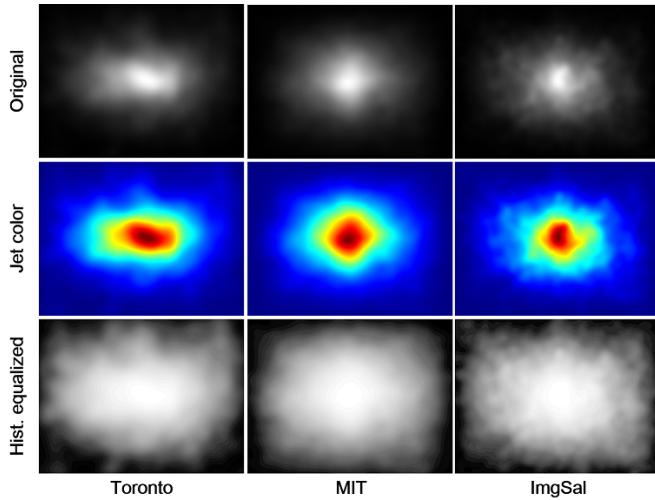


Fig. 2. Average fixation density maps of three datasets. The first row shows the original fixation distribution and the second row shows the jet color map, while the third row displays the histogram equalized maps to get a better perceptual comparisons [23].

for a subject in answering a question. Finally, we obtain  $8400 \times 8 = 67,200$  answers (i.e., binary annotations).

Based on the annotations, the learned metric obtained an impressive performance. However, such an experimental setting has three deficiencies. First, the distance in the jet colorspace may fall a little far from perceptual distances. Second, the original image is not displayed to facilitate the visual comparison, which is actually a necessary supplement when both ESMs are not very visually similar to the GSM. Third, it is hard to judge whether the number of subjects presented to each question is sufficient to make proper annotations.

To address these, we adopt a new setting in the subjective tests by displaying the original image in each question and visualizing histogram-equalized gray saliency maps to make better perceptual comparisons, as suggested in [34]. Finally, each of the 8,400 questions conducted in the current subjective tests consists of one color image as well as one GSM and two ESMs visualized as histogram-equalized gray saliency maps (see the third row of Fig. 1).

In total, 16 subjects (15 males and 1 female, aged from 21 to 24) participated in the tests. All subjects had normal or corrected to normal vision. In the tests, each subject was requested to answer all questions with at least 5 seconds per question. Finally, we obtained  $8,400 \times 16 = 134,400$  answers (binary annotations) under this setting. For the sake of simplification, we represent the crowdsourced data as

$$\{(A_k, B_k, G_k), l_k, c_k, r_k | k \in \mathbb{I}\}, \quad (1)$$

where  $\mathbb{I} = \{1, \dots, 8400\}$  is the set of question indices.  $A_k$  and  $B_k$  are two ESMs being compared with the GSM  $G_k$  in the  $k$ th question.  $l_k \in [0, 1]$  is computed as the percentage of subjects that choose  $A_k$  in answering the  $k$ th question. We can see that one ESM clearly outperforms the other one when  $l_k$  equals to 0 or 1, while  $l_k = 0.5$  indicates two ESMs with similar performance. For the sake of simplification, we present a variable  $c_k \in [0, 1]$  which

is computed as  $2 \times |l_k - 0.5|$  to quantize the confidence that one ESM clearly outperforms the other one. Besides, relative saliency score  $r_k \in [-1, 1]$  is the likelihood that  $A_k$  outperforms  $B_k$ , which is computed as the percentage of difference between subjects that choose  $A_k$  and  $B_k$  (i.e.,  $r_k = 2 \cdot l_k - 1$ ).

After the tests, subjects are also requested to explain the key factors they adopted in making decisions. By investigating their explanations, we find the following key factors that may affect the evaluation of saliency maps.

**1) The most salient and non-salient locations.** In most cases, both ESMs can unveil visual saliency to some extent, and the most salient and non-salient regions play a critical role in determining which ESM performs better. In particular, the overlapping ratio of the most salient and non-salient regions between 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. ESMs that only pop-out object borders are often considered to be poor. Moreover, the background cleanliness also influences the evaluation since fuzzy ESMs usually rank lower in the pair-wise subjective comparisons.

**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 by most subjects.

**4) Salient targets in the original image.** When it is difficult to judge which ESM performs better in approximating the GSM, subjects may refer to the original image and check what targets actually pop-out in both ESMs.

### B. Statistics of User Data

Given the crowdsourced data collected from subjective tests, an important concern is whether the annotations from various subjects are consistent. To answer this question, we defined two types of annotations in our previous work [1] and found that there indeed existed consistency in most cases. In this work, we show the distribution of confidence scores  $\{c_k | k \in \mathbb{I}\}$  in Fig. 3 (a). From Fig. 3 (a), we find that the majority of subjects act highly consistently in answering about 79.5% of questions with confidence scores no less than 0.25 (i.e., at least 10 out of the 16 subjects select the same ESM in answering a question), indicating that there do exist some common clues among different subjects in assessing the quality of saliency maps. We define this type of annotations as consistent annotations. As shown in Fig. 4 (a), such annotations often occur when one ESM performs significantly better than the other one. Meanwhile, we also notice that in 20.5% of questions the annotations are quite ambiguous with confidence scores below 0.25, which are defined as ambiguous annotations, making it difficult to distinguish which ESM performs better. As shown in Fig. 4 (b), both ESMs in most of these questions perform unsatisfactory and it is difficult to determine which ESM is better.

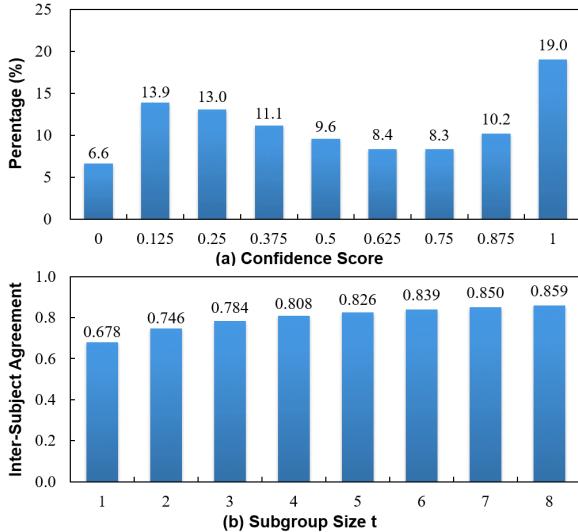


Fig. 3. Statistics of user data. (a) Distribution of confidence scores over 8,400 questions. (b) Inter-subject agreement between different subgroups.

Beyond the consistency, another concern is that whether 16 subjects presented to each question are enough or not. To answer this, we compute the inter-subject agreement between different groups of subjects. We divide the 16 subjects into a set of subgroups  $\mathbb{U}_t$  with group size  $t \in \{1, \dots, 8\}$ , and the inter-subject agreement at group size  $t$ , denoted as  $\alpha_t$ , can be computed as

$$\alpha_t = 1 - \frac{\sum_{u_1, u_2 \in \mathbb{U}_t} \xi(u_1 \cap u_2 = \emptyset) \cdot \sum_{k \in \mathbb{I}} |l_k^{u_1} - l_k^{u_2}|}{\sum_{u_1, u_2 \in \mathbb{U}_t} \xi(u_1 \cap u_2 = \emptyset) \cdot |\mathbb{I}|} \quad (2)$$

where  $u_1$  and  $u_2$  are two subgroups formed by different subjects.  $l_k^{u_1}$  and  $l_k^{u_2}$  are the likelihood that the ESM  $A_k$  outperforms  $B_k$  in the subgroups  $u_1$  and  $u_2$ , respectively. By enumerating the pair-wise combination all subgroups, we show the inter-subject agreement in Fig. 3 (b). We can see that when the size of subgroup grows, the decision made by different subgroups gradually becomes more consistent. In particular, the growth in the inter-subject agreement stays almost stable (*i.e.*, from 0.850 to 0.859) even when the group size varies from  $t = 7$  to  $t = 8$ , implying that 16 subjects are sufficient to provide stable visual comparison results in the tests we conducted.

### C. Analysis of Ten Representative Metrics

Given the crowdsourced data, we can quantize the performance of  $\phi_1 - \phi_{10}$  so as to directly compare them in terms of measuring the human perception of visual similarity. The comparisons are conducted from two perspectives, including image-level and model-level. In image-level comparison, we aim to see if existing metrics can correctly predict which ESM acts better similar to humans. Given a metric  $\phi_i$ , its accuracy in predicting the ordering of ESMs can be computed as:

$$P_i = \frac{1}{\sum_{k \in \mathbb{I}} c_k} \cdot \begin{cases} \sum_{k \in \mathbb{I}} [cs_k > 0]_I \cdot c_k, & i = 1, \dots, 8 \\ \sum_{k \in \mathbb{I}} [cs_k < 0]_I \cdot c_k, & i = 9, 10 \end{cases} \quad (3)$$

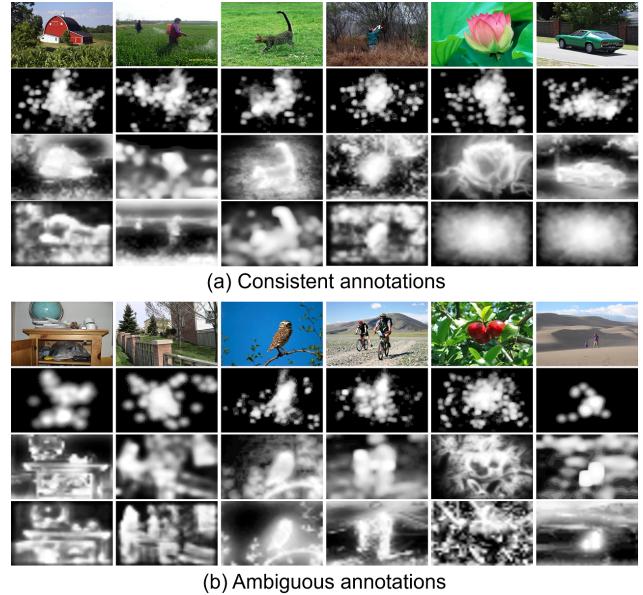


Fig. 4. Representative examples of consistent and ambiguous annotations (1st row: Original image; 2nd row: GSM; 3rd and 4th rows: ESMs). In (a), ESMs at the 3rd row outperform those at the 4th row. In most of the cases, ESMs from  $\mathcal{M}_1 - \mathcal{M}_6$  outperform ESMs from AVG (the last two columns of (a)).

where  $cs_k = (\phi_i(A_k) > \phi_i(B_k)) \cdot (l_k - 0.5)$ , and  $[\mathbf{e}]_I = 1$  if the event  $\mathbf{e}$  holds, otherwise  $[\mathbf{e}]_I = 0$ . Note that in (3) we omit  $G$  in  $\phi_i(S, G)$ .

The accuracies of ten heuristic metrics are shown in Fig. 5. We find that the top two metrics that perform most consistently with human perception are  $\phi_2$  (**sAUC**) and  $\phi_3$  (**rAUC**) and the lowest prediction accuracy is only 42.9% from  $\phi_9$  (**KLD**). However, the best metric, **sAUC**, only reaches an accuracy of 72.8% in comparing all the ESM pairs, while random prediction achieves an accuracy of 50% in addressing such binary classification problems. Actually, there still exists a huge gap between these existing metrics and the human perception of visual similarity. Note that *the low accuracy does not mean that the metric is not suitable for assessing saliency models*. Instead, such a low-accuracy metric works complementary to the direct visual comparisons of histogram equalized saliency maps.

In addition to the image-level comparison, we also compare these ten metrics at the model-level. That is, we generate a ranking list of the seven models with each metric. The model rankings generated by various metrics, as well as the numbers of inconsistently predicted model pairs, can be found in Fig. 6. We find that  $\phi_2$  (**sAUC**) and  $\phi_3$  (**rAUC**) still perform the best. These results are almost consistent with those in the image-level comparison. In particular, these representative evaluation metrics have their respective different ranking lists. This implies that different metrics capture different characteristics of visual saliency in assessing saliency models.

From above, we find that most of existing metrics demonstrate a large inconsistency with humans. Therefore, it is necessary to develop a new metric that better captures human perception of saliency.

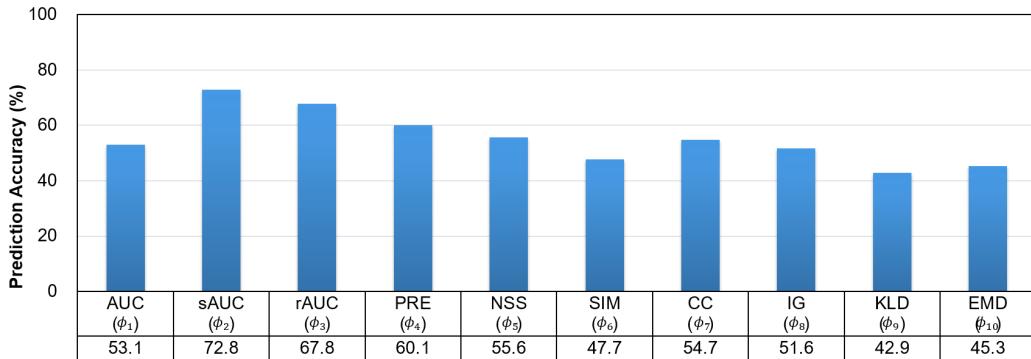


Fig. 5. Prediction accuracies of ten representative evaluation metrics, which indicate the consistency degree in predicting the ordering of all ESMs in terms of measuring the human perception of visual similarity.

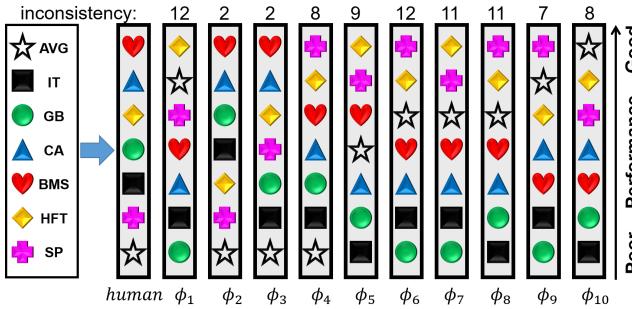


Fig. 6. Ranking lists of seven models generated by 10 representative evaluation metrics. The number above each bar indicates how many pairs of models are inconsistently ranked.

#### IV. LEARNING A SALIENCY EVALUATION METRIC USING CROWDSOURCED PERCEPTUAL JUDGEMENTS

According to the quantized performance, we need a good complement to existing metrics from the perspective of the human perception of visual similarity. Toward this end, we propose to learn a saliency evaluation metric  $\phi_{CPJ}(S, G)$  from the ESM pairs judged by 16 subjects under a new setting. Different from our previous work [1] that focuses on the ordering of the ESM pair, we treat the learned metric as a regressor rather than a binary classifier. That is, the learned metric assigns each saliency map a numerical performance score similar to existing classic metrics.

The architecture used in optimizing the parameters of **CPJ** is shown in Fig. 7. It starts with two streams with cloned parameters and ends with the relative saliency performance score. Each stream is initialized from the VGG16 network [56], which contains five blocks and three full connected layers. In particular, blocks  $B_1$  and  $B_2$  contain two convolutional layers with  $3 \times 3$  kernels. Subsequently, each of the blocks  $B_3$ ,  $B_4$  and  $B_5$  consists of three convolutional layers with  $3 \times 3$  kernels. Note that we use rectified linear unit (ReLU) activation functions [57] in the final convolutional layers of each block. Then, three fully connected layers  $F_6$ ,  $F_7$  and  $F_8$  are followed. It is worth noting that we replace the last softmax layer with the sigmoid activation functions. Besides, different from the original VGG16 architecture, an ESM and the

corresponding GSM are combined as the input for each stream. Note that both ESMs and GSM are resized to the same resolution of  $128 \times 128$  through bilinear interpolation and are then normalized to the dynamic range of  $[0, 1]$ . In this manner, the revised VGG16 architecture in each stream takes a  $128 \times 128$  two-channel image as the input and outputs a performance score of perceptual similarity. Then, two separate perceptual similarity scores from the two streams are computed (*i.e.*,  $\phi_{CPJ}(A, G)$  and  $\phi_{CPJ}(B, G)$ ), whose difference is then computed to approximate the relative saliency score (*i.e.*,  $r_k \in [-1, 1], \forall k$ ) that indicates the likelihood that  $A$  outperforms  $B$  in approximating  $G$ . Here, we regard either of the two streams as our final metric  $\phi_{CPJ}(S, G)$  since they are the same.

To optimize the parameters in **CPJ**, we adopt the crowdsourced data collected from all questions (*i.e.*,  $\{(A_k, B_k, G_k) | k \in \mathbb{I}\}$ , with relative saliency score  $r_k$ ). Unlike our previous work [1] that adopts the crowdsourced data only with consistent annotations, we utilize all crowdsourcing data that takes into account all possibilities of the performance scores so as to make the learned metric more precise. In our experiments, the crowdsourced data  $(A_k, B_k, G_k)$  are split into  $(A_k, G_k)$  and  $(B_k, G_k)$  and are fed into the two streams, respectively. Moreover, we expand the training data by swapping  $A_k$  and  $B_k$  (*i.e.*,  $\{(B_k, A_k, G_k) | k \in \mathbb{I}\}$ , with relative saliency score  $-r_k$ ). The mean square loss is utilized to optimize the final relative saliency scores.

Further, to set the upper and lower bound of the learned metric, we add an extra pair for each image, that is,  $(G_k, R_k, G_k)$ , with the setting of relative saliency score to 1. Specifically,  $R_k$  is a synthetic saliency map with random pixel values uniformly sampled between  $[0, 255]$ , as shown in the training data of Fig. 7. Here, during the training phase we set additional loss functions for pairs  $(G_k, G_k)$  and  $(R_k, G_k)$ , that is,  $\frac{(1 - \phi_{CPJ}(G_k, G_k))^2}{2}$  and  $\frac{(\phi_{CPJ}(R_k, G_k))^2}{2}$ , respectively. Then, for these pairs three loss values are summed directly. In this way, the performance score of pairs  $(G_k, G_k)$  can reach to 1 and  $(R_k, G_k)$  to 0.

All models are trained and tested with Caffe [58] on a single NVIDIA TITAN XP. In training **CPJ**, we optimize the parameters for up to 20,000 iterations. On average,

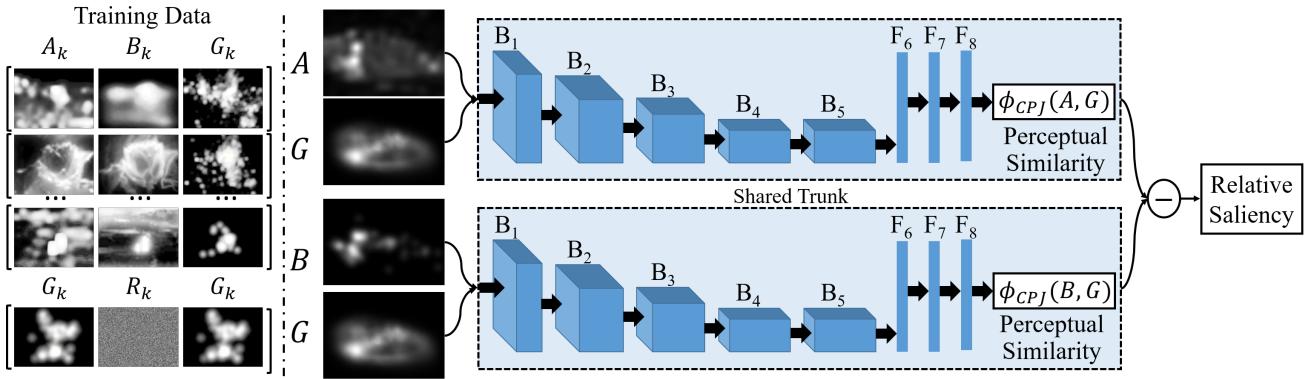


Fig. 7. Architecture of the Saliency Evaluation Metric based on Convolutional Neural Network for Measuring Perceptual Similarity (CPJ). Here, The network consists of two-stream CNNs that share the same parameters and takes one ESM ( $A$  or  $B$ ) and one GSM ( $G$ ) as the input. Each of the trunks assigns each saliency map a numerical performance score of perceptual similarity. Two separate scores after the sigmoid layers are computed by the eltwise layer with the **SUM** operation, where the coeffs are set to be 1 and -1, respectively.

TABLE I  
ACCURACIES (%) OF TEN REPRESENTATIVE METRICS AND **CPJ** ON 50 IMAGES FROM **TORONTO** AND **MIT**.

| Metrics  | <b>AUC</b> ( $\phi_1$ ) | <b>sAUC</b> ( $\phi_2$ ) | <b>rAUC</b> ( $\phi_3$ ) | <b>PRE</b> ( $\phi_4$ ) | <b>NSS</b> ( $\phi_5$ ) | <b>SIM</b> ( $\phi_6$ ) | <b>CC</b> ( $\phi_7$ ) | <b>IG</b> ( $\phi_8$ ) | <b>KLD</b> ( $\phi_9$ ) | <b>EMD</b> ( $\phi_{10}$ ) | <b>CPJ</b>  |
|----------|-------------------------|--------------------------|--------------------------|-------------------------|-------------------------|-------------------------|------------------------|------------------------|-------------------------|----------------------------|-------------|
| Accuracy | 56.5                    | 71.6                     | 67.3                     | 63.9                    | 57.9                    | 50.9                    | 55.9                   | 52.8                   | 48.1                    | 47.9                       | <b>88.8</b> |

\* **CPJ** is trained on the other 250 images from **Toronto** and **MIT** with crowdsourced perceptual judgements of 5,500 questions.

it takes about 83.2s per 100 iterations. We use stochastic gradient descent with a mini-batch size of 32 images and initial learning rate of 0.001, multiplying the learning rate by 0.1 when the error plateaus. Moreover, a momentum of 0.9 and a weight decay of 0.0005 are used. The testing speed of **CPJ** is much faster since it only involves convolution, pooling and connection operations. On the same GPU platform, **CPJ** takes about  $8.5 \times 10^{-3}$  s to compare an ESM with the corresponding GSM (preloaded into memory).

## V. EXPERIMENTS

In this section, we conduct several experiments to validate the effectiveness of the learned metric on new images, new datasets, new models and synthetic data.

### A. Validation of the Learned Metric

To validate the effectiveness of **CPJ**, we need to test whether it can be generalized for the comparison of ESMs from new images, new datasets or new models. In addition, the rationality on synthetic data should be tested as well. Notably, in the testing phase there is an additional  $(G_k, R_k, G_k)$  pair except for 21 pairs of 7 methods for each image. That is, there are 22 testing pairs per image. Toward this end, we design the following experiments:

**Performance on new images.** In the first experiment, we train **CPJ** with the user data obtained on 250 images randomly selected from **Toronto** and **MIT** (*i.e.*,  $250 \times 22 = 5500$  training instances). The learned metric is then tested on all user data obtained on the remaining 50 images from **Toronto** and **MIT** (*i.e.*,  $50 \times 22 = 1100$  testing instances). Note that, the sequence of images selected for training and testing is the same as the setting in our previous work [1].

The main objective is to validate the effectiveness of the learned metric on new images whose GSMSs are obtained under the same configurations in eye-tracking experiments. The performance of **CPJ** and the representative metrics  $\phi_1 - \phi_{10}$  are shown in Table I.

From Table I, we can see that **CPJ** reaches an accuracy of 88.8%, while **sAUC** and **rAUC** perform the best among the metrics  $\phi_1 - \phi_{10}$  with the accuracy scores of 71.6% and 67.3%, respectively. This result proves that **CPJ** can be generalized to the model comparisons on new images when human fixations are recorded using the same eye-tracking configurations (*e.g.*, the rest 823 images from **MIT** that are not used in the subjective tests).

Beyond the overall performance, Table II shows the accuracies of **CPJ** by using different resolutions of ESMs and GSMSs. We can see that the best performance is reached at the resolution  $128 \times 128$ . This can be explained by the fact that when the resolutions reduce to  $64 \times 64$  and  $32 \times 32$ , many important details in ESMs and GSMSs are missing, while such details may facilitate the measurement of similarity between ESMs and GSMSs. On the contrary, when the resolution increases to  $256 \times 256$ , severe over-fitting risk may arise as the training data obtained from subjective tests may be somehow insufficient to train the rapidly growing parameters. Therefore, we use the resolution of  $128 \times 128$  in training **CPJ**.

Moreover, Figure 8 shows the accuracies of **CPJ** when different numbers of feed-forward and back-propagation iterations are reached in the training stage. From Fig. 8, we can see that the prediction accuracy of **CPJ** reaches up to 88.8% when 10,000 iterations are reached. After that, it stays almost stable even with more iterations. Therefore, we use 10,000 iterations in training **CPJ**.

TABLE II  
ACCURACIES (%) OF **CPJ** AT DIFFERENT RESOLUTIONS.

| Resolution | $256 \times 256$ | $128 \times 128$ | $64 \times 64$ | $32 \times 32$ |
|------------|------------------|------------------|----------------|----------------|
| Accuracy   | 87.8             | <b>88.8</b>      | 87.7           | 84.6           |

\* **CPJ** is trained on 250 images from **Toronto** and **MIT** and tested on the remaining 50 images from **Toronto** and **MIT**. The structure of **CPJ** is automatically fine-tuned for each resolution.

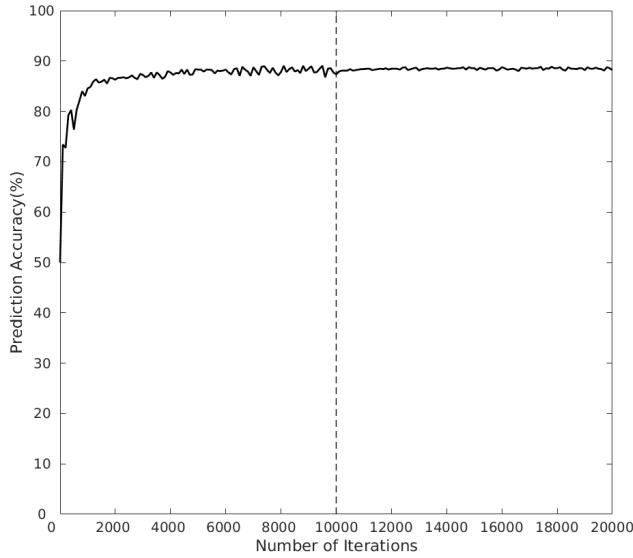


Fig. 8. The accuracies of **CPJ** trained with different numbers of feed-forward and back-propagation iterations.

**Performance on new datasets.** In the second experiment, we train **CPJ** with all crowdsourced data obtained on the images from any two of the three datasets (*i.e.*, **Toronto**, **MIT** and **ImgSal**) and test the learned metric on the images of the remaining dataset. The main objective is to validate the effectiveness of **CPJ** on new datasets constructed using different configurations in eye-tracking experiments. The performance of **CPJ** and the ten representative metrics  $\phi_1 - \phi_{10}$  are shown in Table III.

From Table III, we find that **CPJ** still performs the best accuracy in all three cases and outperforms the metrics  $\phi_1 - \phi_{10}$  by at least 8.1% in IM\_T, even up to 14.1% in IT\_M. Interestingly, under these three different training and testing settings, the performance ranks of existing metrics is basically the same as that of Table I. Such consistency in performance ranks and low accuracy scores of representative metrics on different datasets further proves that the existing metrics still lack certain human visual cognition consistency. On the contrary, **CPJ** outperforms the best representative metric on datasets with different eye-tracking configurations. As the performance scores and ranks of **CPJ** stay much more stable across different datasets, it has better capability to assess visual saliency models than the ten representative metrics.

**Performance on new models.** The third experiment validates whether **CPJ** trained on the crowdsourced data obtained in comparing some specific models can be applied

to the assessment of new models. In this experiment, we adopt a leave-one-out strategy in training and testing **CPJ**. That is, we remove one of the three representative models (*i.e.*, BMS, HFT and SP) from the training stage and use only the instances obtained from the subjective comparisons between the remaining six models for training **CPJ** (about 71.4% instances). The trained metric is then tested on the instances obtained from subjective comparisons between the removed model and the six models (about 28.6% instances). The experiment is repeated three times and the results are shown in Tab. IV.

From Tab. IV, we find that **CPJ** still outperforms the ten representative metrics in comparing a new model with existing ones. Given the new model BMS, HFT or SP, **CPJ** outperforms the best representative metric (*i.e.*, **SAUC** by 7.5%, 8.5% and 16.2%, respectively. As a result, we can safely assume that **CPJ** can be applied to the assessment of newly developed bottom-up (*e.g.*, BMS), spectral (*e.g.*, HFT) and learning-based (*e.g.*, SP) saliency models.

**Performance on synthetic data.** In the last experiment, we validate the rationality of **CPJ** in assessing saliency maps. We re-train **CPJ** on all the crowdsourced perceptual judgements obtained on the 400 images from all the three datasets, and test the metric on the same number of synthesized data (*i.e.*,  $\{(G_k, A_k, G_k) | k \in \mathbb{I}\}$  and  $\{(G_k, B_k, G_k) | k \in \mathbb{I}\}$ ). Intuitively, the GSM  $G_k$  should always outperform better than either of ESM  $A_k$  and  $B_k$  in subjective tests. The objective of this experiment is to see whether **CPJ** can perfectly capture this attribute, and we find that accuracy of **CPJ** reaches 99.6%.

Beyond comparing with the GSM that can be viewed as an “upper bound”, we also test **CPJ** on synthesized data  $\{(A_k, R_k, G_k) | k \in \mathbb{I}\}$  and  $\{(B_k, R_k, G_k) | k \in \mathbb{I}\}$ . Similarly, comparing with the “lower-bound”  $R_k$ , either of ESM  $A_k$  and  $B_k$  in subjective tests should always be “good”. In this case, **CPJ** also achieves an accuracy of 94.3%. These results ensure that the fixation density maps always achieve the best performance and the random predictions always perform the worst, even though such synthesized data are not used during training **CPJ**.

**Discussion.** From these four experiments, we find that **CPJ** is effective in assessing saliency maps on new images, new datasets, new models and synthetic data. Over different experimental settings, **CPJ** outperforms the ten representative metrics by learning from a variety of subjects assessments. Actually, although different subjects may have different biases in visually comparing the saliency distributions of ESMs and GSMS, the CNN-based framework can automatically infer the most commonly adopted factors shared by various subjects in assessing saliency maps. Therefore, **CPJ** can be viewed as a crowdsourced metric that performs consistently with most of the 16 subjects. Due to this characteristic of the human perception of visual similarity, **CPJ** can evaluate the saliency models from another perspective, making the learned metric a good complement to the existing metrics.

TABLE III  
ACCURACIES (%) OF TEN REPRESENTATIVE METRICS AND CPJ.

| Metrics | AUC ( $\phi_1$ ) | sAUC ( $\phi_2$ ) | rAUC ( $\phi_3$ ) | PRE ( $\phi_4$ ) | NSS ( $\phi_5$ ) | SIM ( $\phi_6$ ) | CC ( $\phi_7$ ) | IG ( $\phi_8$ ) | KLD ( $\phi_9$ ) | EMD ( $\phi_{10}$ ) | CPJ         |
|---------|------------------|-------------------|-------------------|------------------|------------------|------------------|-----------------|-----------------|------------------|---------------------|-------------|
| MT_I    | 45.8             | 75.3              | 68.6              | 54.1             | 50.7             | 41.1             | 51.1            | 46.8            | 34.0             | 38.6                | <b>86.0</b> |
| IM_T    | 60.5             | 77.4              | 71.1              | 65.8             | 62.9             | 56.8             | 60.8            | 56.9            | 51.2             | 52.8                | <b>85.5</b> |
| IT_M    | 57.0             | 71.2              | 68.4              | 63.6             | 58.0             | 50.7             | 57.3            | 55.6            | 48.0             | 48.4                | <b>85.3</b> |

\* MT\_I means CPJ is trained on the 300 images from MIT and Toronto and tested on the 100 images from ImgSal. IM\_T means CPJ is trained on the 280 images from ImgSal and MIT and tested on the 120 images from Toronto. IT\_M means CPJ is trained on the 220 images from ImgSal and Toronto and tested on the 180 images from MIT.

TABLE IV

ACCURACIES (%) OF TEN REPRESENTATIVE METRICS AND CPJ WHEN DIFFERENT MODELS ARE EXCLUDED FROM THE TRAINING STAGE.

| Metrics | AUC ( $\phi_1$ ) | sAUC ( $\phi_2$ ) | rAUC ( $\phi_3$ ) | PRE ( $\phi_4$ ) | NSS ( $\phi_5$ ) | SIM ( $\phi_6$ ) | CC ( $\phi_7$ ) | IG ( $\phi_8$ ) | KLD ( $\phi_9$ ) | EMD ( $\phi_{10}$ ) | CPJ         |
|---------|------------------|-------------------|-------------------|------------------|------------------|------------------|-----------------|-----------------|------------------|---------------------|-------------|
| BMS     | 59.7             | 76.9              | 72.5              | 64.5             | 62.3             | 49.8             | 60.5            | 55.2            | 44.9             | 45.9                | <b>84.4</b> |
| HFT     | 55.1             | 69.5              | 66.4              | 56.1             | 56.6             | 45.2             | 55.5            | 50.2            | 43.3             | 46.7                | <b>78.4</b> |
| SP      | 45.9             | 70.8              | 63.6              | 46.0             | 46.0             | 38.2             | 45.8            | 42.2            | 32.5             | 38.6                | <b>87.0</b> |

\* In each row, instances involving the model in the first column are used for testing, while the remaining ones are used to train CPJ.

## VI. DISCUSSION AND CONCLUSION

In visual saliency estimation, hundreds of models have been proposed to reveal certain characteristics of visual saliency under different assumptions and definitions. Therefore various evaluation metrics are utilized to simultaneously assess the performance of saliency models from multiple perspectives. Usually, the existing metrics are designed to quantitatively compute the performance score of each model without considering perceptual factors. Inspired by the fact that most saliency papers provide representative ESMs and GSMS for qualitative comparisons, the human perception of visual similarity underlying direct visual comparison is very helpful to design new metrics that assess saliency maps as the humans do.

To investigate the key factors that influence the human perception of visual similarity in comparing saliency maps, we conduct extensive subjective tests to find out how saliency maps are assessed by subjects. Besides, to analyze the existing metrics, we propose to quantize the performance of evaluation metrics from the crowdsourced data collected from 16 subjects. A latent assumption here is that even though a subject may have certain biases in assessing saliency maps and models, such biases can be greatly alleviated by summing up the annotations from multiple subjects. Given the crowdsourced data, we find that the top metric that performs the most consistently with the humans only reaches an accuracy of 72.8% in comparing all the ESM pairs. This indicates that there still exists a large gap between the existing models and models in terms of saliency prediction.

To obtain a metric that contains the characteristic of the human perception, we propose to learn a saliency evaluation metric using crowdsourced perceptual judgements based on a two-stream convolutional neural network. Similar to existing saliency evaluation metrics, the learned metric assigns the ESM from an ESM pair a numerical performance score and can capture the key factors shared by various subjects in assessing saliency maps and models. Experimental results show that such a CNN-based metric performs more

consistently with human perception of saliency maps and could be a good complement to existing metrics for model comparison and development. In the future work, we will incorporate eye-trackers to study the latent mechanisms by which humans assessing saliency maps. We will also explore the feasibility of building new saliency models under the guidance of our CNN-based metric.

## REFERENCES

- [1] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen, “A data-driven metric for comprehensive evaluation of saliency models,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 190–198.
- [2] J. Li, Y. Tian, X. Chen, and T. Huang, “Measuring visual surprise jointly from intrinsic and extrinsic contexts for image saliency estimation,” *International Journal of Computer Vision*, pp. 1–17, 2016.
- [3] Y. Hu, D. Rajan, and L.-T. Chia, “Adaptive local context suppression of multiple cues for salient visual attention detection,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.
- [4] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit, “Rare: A new bottom-up saliency model,” in *IEEE Conference on Image Processing (ICIP)*, 2012, pp. 641–644.
- [5] S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. Venkatesh Babu, “Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5781–5790.
- [6] A. Borji and L. Itti, “Exploiting local and global patch rarities for saliency detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 478–485.
- [7] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2376–2383.
- [8] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, “Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation,” *IEEE transactions on circuits and systems for video technology*, 2016.
- [9] W. Kim and C. Kim, “Spatiotemporal saliency detection using textural contrast and its applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 646–659, 2014.
- [10] V. Mahadevan and N. Vasconcelos, “Spatiotemporal saliency in dynamic scenes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 1, pp. 171–177, 2010.
- [11] J. Li, Y. Tian, T. Huang, and W. Gao, “Probabilistic multi-task learning for visual saliency estimation in video,” *International Journal of Computer Vision*, vol. 90, no. 2, pp. 150–165, 2010.
- [12] Y. Zhang, F. Zhang, and L. Guo, “Saliency detection by selective color features,” *Neurocomputing*, vol. 203, pp. 34–40, 2016.

- [13] M. Kümmeler, T. Wallis, and M. Bethge, "How close are we to understanding image-based saliency?" *CoRR*, vol. abs/1409.7686, 2014. [Online]. Available: <http://arxiv.org/abs/1409.7686>
- [14] W. Hou, X. Gao, D. Tao, and X. Li, "Visual saliency detection using information divergence," *Pattern Recognition*, vol. 46, no. 10, pp. 2658 – 2669, 2013.
- [15] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2368–2375.
- [16] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2005, pp. 155–162.
- [17] A. Borji and J. Tanner, "Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, pp. 1214–1226, 2016.
- [18] J. Li, Y. Tian, and T. Huang, "Visual saliency with statistical priors," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 239–253, 2014.
- [19] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, pp. 4, 1–16, 2009.
- [20] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 545–552.
- [21] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2106–2113.
- [22] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.
- [23] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.
- [24] U. Rajashekhar, L. K. Cormack, and A. C. Bovik, "Point-of-gaze analysis reveals visual search strategies," in *Human Vision and Electronic Imaging*, vol. 5292, 2004, pp. 296–306.
- [25] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643–659, 2005.
- [26] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.
- [27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [28] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," <http://saliency.mit.edu/>.
- [29] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32, 1–20, 2008.
- [30] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 153–160.
- [31] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.
- [32] A. Borji, D. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.
- [33] A. F. Russell, S. Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision Research*, vol. 94, pp. 1–15, 2014.
- [34] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, "On computational modeling of visual saliency: Examining what's right, and what's left," *Vision research*, vol. 116, pp. 95–112, 2015.
- [35] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 921–928.
- [36] M. Emami and L. L. Hoberock, "Selection of a best metric and evaluation of bottom-up visual saliency models," *Image and Vision Computing*, vol. 31, no. 10, pp. 796–808, 2013.
- [37] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1153–1160.
- [38] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.
- [39] S. Lu, C. Tan, and J. Lim, "Robust and efficient saliency modeling from image co-occurrence histograms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 195–201, 2014.
- [40] J. Li, L.-Y. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the secret of image saliency in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2428–2440, 2015.
- [41] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [42] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *Journal of Vision*, vol. 12, no. 6, pp. 22, 1–15, 2012.
- [43] J. Li, D. Xu, and W. Gao, "Removing label ambiguity in learning-based visual saliency estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1513–1525, 2012.
- [44] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 11, 1–20, 2013.
- [45] R. J. Peters and L. Itti, "Congruence between model and human attention reveals unique signatures of critical visual events," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2009.
- [46] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 438–445.
- [47] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multitask sparsity pursuit," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1327–1338, 2012.
- [48] M. Kümmeler, T. Wallis, and M. Bethge, "How close are we to understanding image-based saliency?" *arXiv*, Sep 2014. [Online]. Available: <http://arxiv.org/abs/1409.7686>
- [49] M. Kümmeler, T. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Science*, vol. 112, no. 52, pp. 16 054–16 059, Oct 2015. [Online]. Available: <http://www.pnas.org/content/112/52/16054.abstract>
- [50] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.
- [51] N. Wilming, T. Betz, T. C. Kietzmann, and P. König, "Measures and limits of models of fixation selection," *PloS one*, vol. 6, no. 9, p. e24038, 2011.
- [52] M. Kümmeler, T. S. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16 054–16 059, 2015.
- [53] J. Li, M. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.
- [54] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
- [55] C. Li, J. Xue, N. Zheng, X. Lan, and Z. Tian, "Spatio-temporal saliency perception via hypercomplex frequency spectral contrast," *Sensors*, vol. 13, no. 3, pp. 3409–3431, 2013.
- [56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.
- [58] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

**Changqun Xia** is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image understanding.



**Jia Li** is currently an associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He received the B.E. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. Before he joined Beihang University in Jun. 2014, he served as a researcher in several multimedia groups of Nanyang Technological University, Peking University and Shanda Innovations. He is the author or coauthor of over 50 technical articles in refereed journals and conferences such as TPAMI, IJCV, TIP, CVPR, ICCV and ACM MM.

His research interests include computer vision and multimedia big data, especially the learning-based visual content understanding. He is a senior member of IEEE and CCF.

**Jinming Su** is currently pursuing the Master degree with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and machine learning.



**Ali Borji** Ali Borji received his BS and MS degrees in computer engineering from Petroleum University of Technology, Tehran, Iran, 2001 and Shiraz University, Shiraz, Iran, 2004, respectively. He did his Ph.D. in cognitive neurosciences at Institute for Studies in Fundamental Sciences (IPM) in Tehran, Iran, 2009 and spent four years as a postdoctoral scholar at iLab, University of Southern California from 2010 to 2014. He is currently an assistant professor at the Center for Research in Computer Vision and

the Department of Computer Science at the University of Central Florida. His research interests include visual attention, active learning, object and scene recognition, and cognitive and computational neurosciences.

