

# A Locally Weighted Fixation Density-Based Metric for Assessing the Quality of Visual Saliency Predictions

Milind S. Gide and Lina J. Karam, *Fellow, IEEE*

## Abstract

With the increased focus on visual attention (VA) in the last decade, a large number of computational visual saliency methods have been developed over the past few years. These models are traditionally evaluated by using performance evaluation metrics that quantify the match between predicted saliency and fixation data obtained from eye-tracking experiments on human observers. Though a considerable number of such metrics have been proposed in the literature, there are notable problems in them. In this work, we discuss shortcomings in existing metrics through illustrative examples and propose a new metric that uses local weights based on fixation density which overcomes these flaws. To compare the performance of our proposed metric at assessing the quality of saliency prediction with other existing metrics, we construct a ground-truth subjective database in which saliency maps obtained from 17 different VA models are evaluated by 16 human observers on a 5-point categorical scale in terms of their visual resemblance with corresponding ground-truth fixation density maps obtained from eye-tracking data. The metrics are evaluated by correlating metric scores with the human subjective ratings. The correlation results show that the proposed evaluation metric outperforms all other popular existing metrics. Additionally, the constructed database and corresponding subjective ratings provide an insight into which of the existing metrics and future metrics are better at estimating the quality of saliency prediction and can be used as a benchmark.

## Index Terms

Visual Attention, Saliency, Quality Assessment, Visual Attention Models.

M. S. Gide and L. J. Karam are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, 85287-5706 USA (e-mail: mgide, karam@asu.edu)

## I. INTRODUCTION

Visual attention (VA) is the broad area of research that aims to explain the mechanisms by which the human visual system (HVS) filters the vast amount of visual information captured by the retina. VA has applications in a large number of diverse areas like object recognition, image segmentation, compression, selective reduced-power visual processing, to name a few. As a result, there has been a lot of focus recently on developing computational VA models. The VA mechanism is considered to be a combination of instantaneous pre-attentive, bottom-up processes that depend on low-level cues, and much slower, top-down, cognitive processes that depend on high-level precepts. Most of the existing models are based on bottom-up concepts and output what is known as a saliency map that gives the visual importance of each pixel location. Given the large number of VA models to choose from, it is necessary to evaluate these models. VA models are traditionally evaluated by comparing the saliency maps with eye-tracking data that is obtained from human observers. Several performance metrics that objectively quantify the match between the predicted saliency map and eye-tracking data have been introduced over the past decade for evaluating VA models (see [1] for a detailed review). A number of studies like those by Toet [2], Borji et al. [3], and Judd et al. [4] have evaluated state-of-the-art VA models using a subset of these metrics. However, none of these studies have evaluated the effectiveness of existing metrics in assessing the quality of VA models and ignore certain notable flaws in them. The motivation for the proposed metric is to provide a more accurate assessment of the quality of visual saliency prediction than existing metrics, which can aid in a better comparative evaluation of VA models. The proposed metric can also act as an improved measure of cost for training VA models that use machine learning. Yet another application for the proposed metric is faithful objective comparison of eye-tracking equipments. Given the importance of accurate evaluation of VA models, there have been a few papers in recent years that discuss metrics. LeMeur and Baccino [5] gave an overview of existing performance metrics in literature and discussed their strengths and weaknesses. Riche et al. [6] provided a taxonomy for existing metrics and also studied the correlation between the metrics. They showed that each metric alone is not sufficient to evaluate a VA model and suggested the use of a combination of metrics to get a better estimate of performance. Recently, Gide et al. [7] discussed known flaws in existing metrics through examples and proposed a metric,  $sNSS$ , that resolves the center-bias problem in the Normalized Scanpath Saliency ( $NSS$ ) metric [8] through shuffling. However, none of

TABLE I: Evaluated Metrics.

Metric Name	Category	Ground-truth
$AUC_{Borji}$ [3]	Location-based	Fixation Points
$AUC_{Judd}$ [4]	Location-based	Fixation Points
$sAUC$ [3]	Location-based	Fixation Points
$WF_{\beta}$ [10]	Location-based	Fixation Density Map
$NSS$ [3]	Value-based	Fixation Points
$sNSS$ [7]	Value-based	Fixation Points
$CC$ [3]	Distribution-based	Fixation Density Map
$SIM$ [4]	Distribution-based	Fixation Density Map
$EMD$ [4]	Distribution-based	Fixation Density Map
$MAE$ [11]	Distribution-based	Fixation Density Map

these works provide a common benchmark to compare the performance metrics.

The key contribution of this paper is to propose a novel metric that assigns locally adaptive weights to fixation points based on local fixation density and thus gives more importance to the visually relevant fixations in the ground-truth eye-tracking data. We also address the problem of a lack of a benchmark for evaluating existing and future performance metrics by constructing a subjective database in which ratings on a 5-point categorical scale by human observers are used to rate saliency maps of several VA models based on their visual resemblance to ground-truth saliency maps. The average ratings or mean opinion scores (MOS) are then correlated with the performance metric scores to evaluate the metrics.

This paper is organized as follows. In Section II we highlight the known problems in existing popular metrics [9] through illustrative examples. We then propose a new metric that uses locally adaptive weights for fixation points in Section III. The details of the subjective study are presented in Section IV, and the correlation results for the existing and proposed metrics are presented in Section V. Finally, we conclude the paper in Section VI and also provide directions for future research.

## II. EXISTING METRICS AND THEIR SHORTCOMINGS

Existing metrics can be classified into the following three major categories: value-based, location-based, and distribution-based, depending on the type of similarity measure used to compare the predicted saliency map to the eye-tracking data [6]. The value-based metrics focus on the predicted saliency map values at fixation points, the location-based metrics focus on how

well the salient regions in the predicted saliency maps match with the locations of the fixation points, and the distribution-based metrics focus on the differences in the statistical distributions of the predicted saliency maps and fixation points. In addition to these categories, metrics can also be classified based on the type of ground-truth used. Some metrics use only the fixation locations from the eye-tracking data whereas others use a ground-truth saliency map (GSM) which is obtained by convolving a 2D Gaussian with the fixations and normalizing the resulting map. Several recent studies have used different types metrics to benchmark VA models. Toet [2] evaluated several VA models by using the Spearman’s correlation coefficient. More recently, Borji et al. [3] used the  $AUC_{Borji}$ ,  $CC$  and  $NSS$  measures and Judd et al. [4] used the  $AUC_{Judd}$ ,  $Similarity$  and  $EMD$  metrics to evaluate several VA models. The MIT saliency benchmark project [9] is an up-to-date online benchmarking resource that lists the performance of all the recent state-of-the-art VA models using seven popular evaluation metrics that are a combination of those used in [3] and [4]. The metrics used by the MIT Saliency Benchmark [9] along with recently proposed metrics  $WF_{\beta}$  [10] and  $sNSS$  [7] in addition to a baseline metric  $MAE$  [11] are listed in Table I along with the categories they belong to, as well as the type of ground-truth used.

The first notable and well-analyzed problem with existing metrics is the problem of center-bias [1], [7], [5]. This problem arises due to an inherent tendency of images and photographs to contain objects of interest in central regions as compared to peripheral regions. Most metrics that do not factor the center-bias in their formulation tend to incorrectly reward models that independent of content assign higher importance to central regions and lower importance to peripheral regions. One way of tackling this issue is through “shuffling” in which ground-truth fixations for all other images in the dataset are randomly sampled and high saliency predictions at such locations are penalized. Consequently, models that blindly reward central regions are penalized to a greater extent by the shuffling process, and receive a much lower score than more discriminative models [1]. An illustration of the effect of center-bias on shuffled and non-shuffled metrics is shown in Figure 1. As shown in Figure 1, the non-shuffled metrics like  $AUC_{Borji}$ ,  $AUC_{Judd}$ ,  $CC$ ,  $EMD$ ,  $SIM$ , and  $NSS$  tend to give higher scores to models that assign higher saliency to central regions as compared to the boundaries. As a result, these incorrectly result in higher performance scores for a centered Gaussian blob (Figure 1(c)) as compared to a saliency map from a VA model (Figure 1(d)). On the other hand, the shuffled metrics assign a better score to the AIM [12] saliency map over the centered Gaussian map.

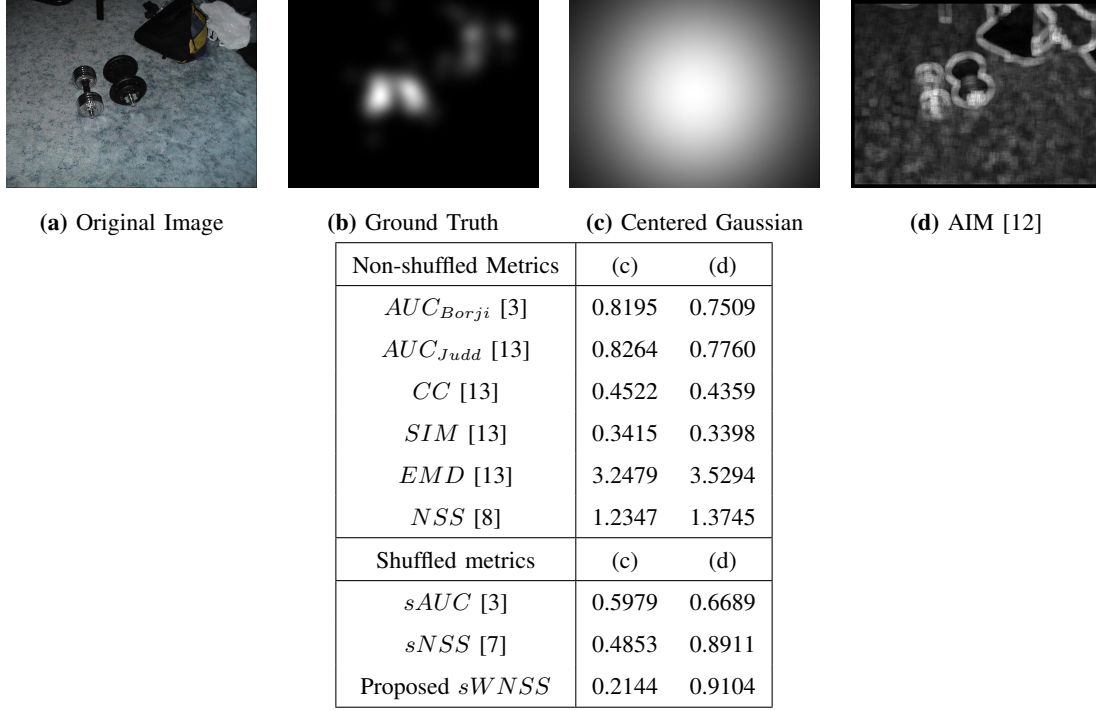


Fig. 1: Center bias problem in existing metrics that is rectified by the shuffled metrics. For  $EMD$ , a lower score indicates better performance; for the other metrics, a higher score indicates better performance.

However, the AUC metrics including the  $sAUC$  suffer from another notable flaw known as the interpolation flaw (described in detail in [10]). As seen in Figure 2,  $AUC_{Borji}$ ,  $AUC_{Judd}$  and  $sAUC$  are less sensitive to false-positives. As a result, a “fuzzy” ground-truth saliency map created by increasing the background activity in the neighborhood of a true-positive peak incorrectly gets higher or almost similar scores than the actual ground-truth saliency when using the AUC-based metrics. The other metrics  $NSS$  [8],  $CC$  [3],  $EMD$  [9] and  $SIM$  [9] do not exhibit the interpolation flaw but do suffer from the center-bias problem as seen in Figure 1. Of these metrics, only  $NSS$  is a viable candidate to be shuffled to tackle the center-bias issue as suggested in [7]. This metric termed Shuffled NSS or  $sNSS$  for short is given by

$$sNSS = NSS(p) - NSS(r) \quad (1)$$

where  $p$  and  $r$  denote, respectively, the ground-truth fixation points for the image and the randomly sampled non-fixation points from the set of fixation points for other images in the

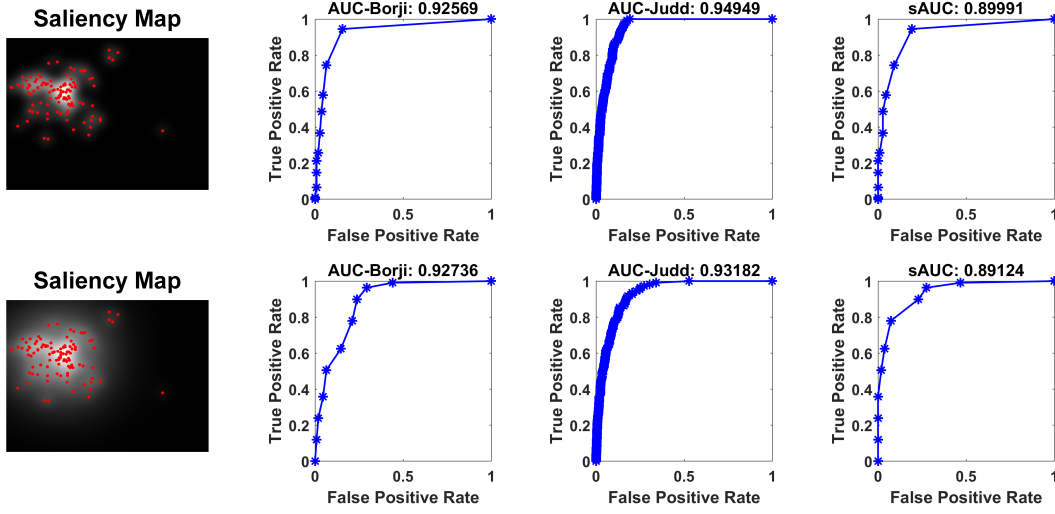


Fig. 2: AUC Interpolation flaw. Top Row: (Left to Right) Ground-truth saliency map with ground-truth fixations and corresponding ROC curves for  $AUC_{Borji}$ ,  $AUC_{Judd}$  and  $sAUC$ . Bottom Row: (Left to Right) Fuzzier version of the ground-truth saliency map with ground-truth fixations and corresponding ROC curves for  $AUC_{Borji}$ ,  $AUC_{Judd}$  and  $sAUC$ .

dataset and

$$NSS(x) = \frac{1}{N} \sum_{x \in X} \frac{S(x) - \mu_s}{\sigma_s}. \quad (2)$$

In (2),  $\mu_s$  and  $\sigma_s$  are, respectively, the mean and standard deviation of the predicted saliency map  $S$  and  $N$  is the number of points in the set  $X$ . The random sampling for the non-fixation points  $r$  is repeated a number of times, typically 100, and the final result is the average of scores obtained for each of these trials. The  $sNSS$  metric improves on the  $sAUC$  scores by correctly assigning a better score to the saliency map in Figure 3(d) as compared to the one in Figure 3(c). It also improves upon  $NSS$  by giving the centered Gaussian map in Figure 3(e) a low score.

For the  $sAUC$ , the locations used for determining false-positives are sampled from the distribution of fixations for all other images. Because of the center-bias inherent in most eye-tracking datasets, these locations tend to be in the central portion of the image. As a result, if false-positives crop up in regions away from the center,  $sAUC$  is not able to penalize them. In contrast, because of the zero-mean unit-standard deviation normalization in  $sNSS$ , blurrier maps are penalized as a result of which  $sNSS$  is able to correctly assign a lower score to fuzzy maps such as map (c) in Figure 3. However, a drawback of the  $NSS$  and  $sNSS$  metrics is that in

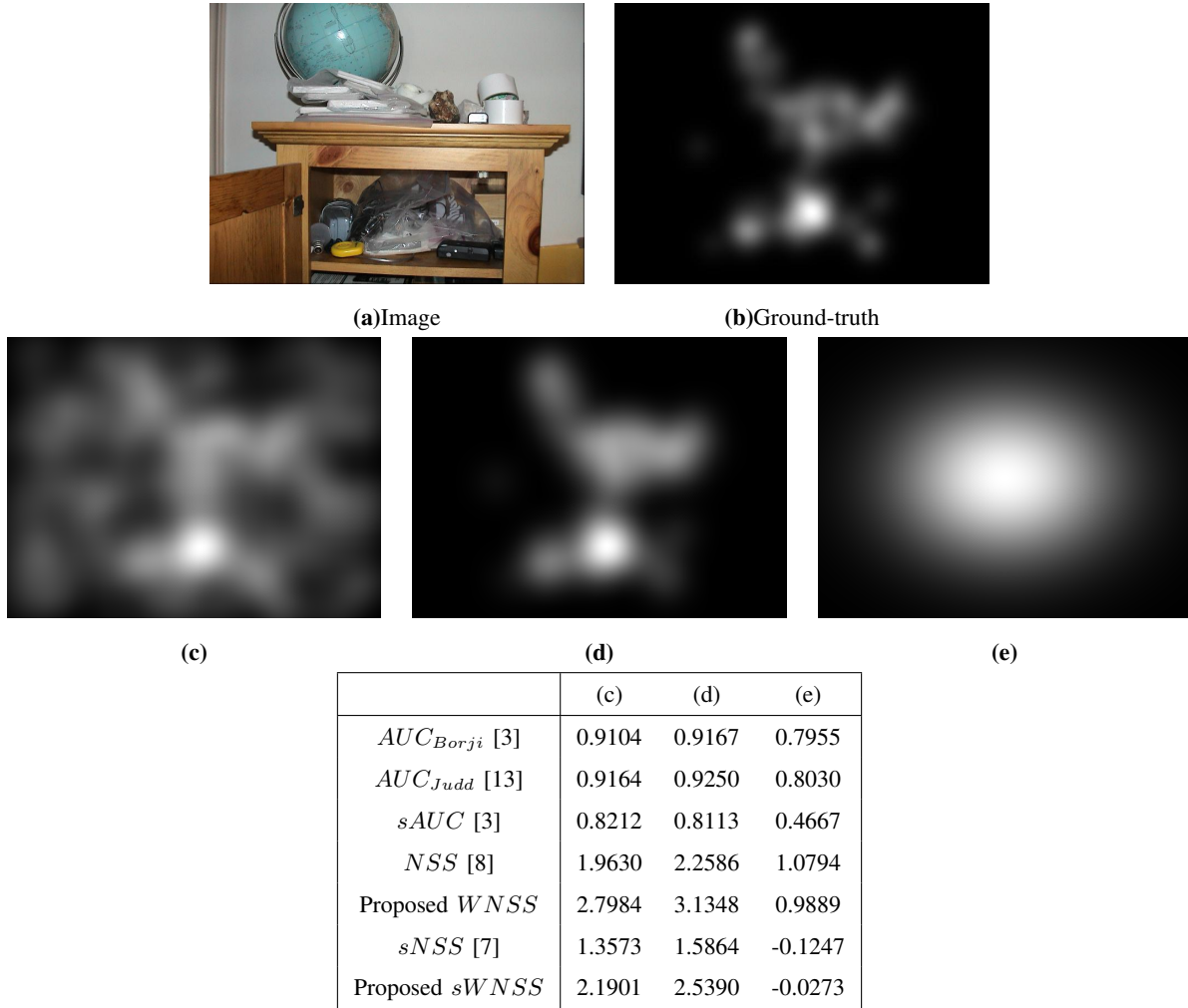


Fig. 3: The two saliency maps (c) and (d) have nearly identical  $sAUC$ ,  $AUC_{Borji}$  and  $AUC_{Judd}$  scores, however it is clear that (d) is a much “better” saliency map.  $NSS$ ,  $sNSS$  and the proposed  $WNSS$  and  $sWNSS$  metrics do not have this problem as they assign a significantly higher score to (d) than (c). A centered Gaussian blob (e) will perform well using  $NSS$  and the proposed  $WNSS$  metrics, however using the  $sNSS$  and the proposed shuffled  $sWNSS$  the same Gaussian blob receives a low score as expected.

their computation all fixations are given equal weights and fixation density is ignored. Figure 4 illustrates this drawback through two created saliency maps. Though map (d) is much better than map (e) in Figure 4, it gets lower  $NSS$  and  $sNSS$  scores than map (e). This happens because in the  $NSS$  formulation, when the normalized saliency values are averaged, each fixation location contributes equally to the average.

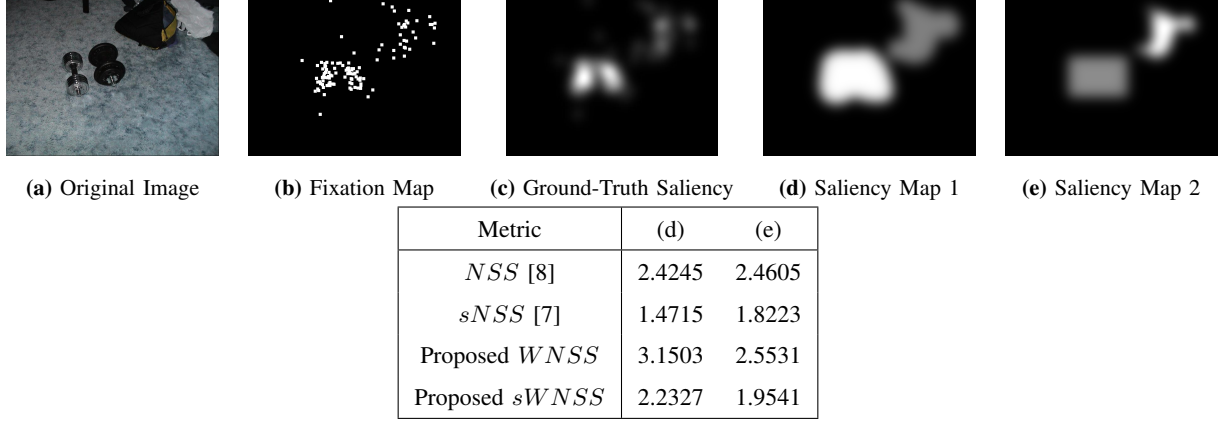


Fig. 4: Problem with  $NSS$  and  $sNSS$ : The  $NSS$  and  $sNSS$  metrics give equal weight to every fixation point and ignore density. As a result, they incorrectly give higher scores to Saliency Map 2 (e) as compared to Saliency Map 1 (d). The proposed  $WNSS$  and its shuffled variant  $sWNSS$  weight the fixations based on their local density and assign a higher score to map (d) as expected.

### III. PROPOSED METRIC

Figure 5 illustrates that fixations that are closely clustered together lie on actual objects in the scene and are most important for identifying salient regions as compared to others that are scattered around and lie on background areas. However,  $sNSS$  and  $NSS$  both do not discriminate between relevant fixations that belong to a dense cluster and represent objects, from fixations that are sparse and usually fall on background regions and could be considered as outliers. One way to remedy this is to assign weights to each fixation point based on its importance. If  $W(p)$  is the weight assigned to the fixation point  $p \in P$ , where  $P$  is the set of all fixation points, the proposed metric termed as weighted  $NSS$  or  $WNSS$  for short is defined as

$$WNSS = \frac{1}{\sum_{p \in P} W(p)} \sum_{p \in P} \frac{W(p)(S(p) - \mu_s)}{\sigma_s} \quad (3)$$

where  $\mu_s$  and  $\sigma_s$  are the mean and standard deviation, respectively, of the predicted saliency map  $S$ , and  $P$  denotes the set of ground-truth fixation points for the image. To obtain appropriate weights for each fixation, we use the fact that fixations that are in higher density clusters are more important and should be weighted higher than those in low density clusters. For this purpose, we use the density-based spatial clustering of applications with noise (DBSCAN) algorithm [14] for clustering the fixations based on their density to obtain fixation clusters. We then assign



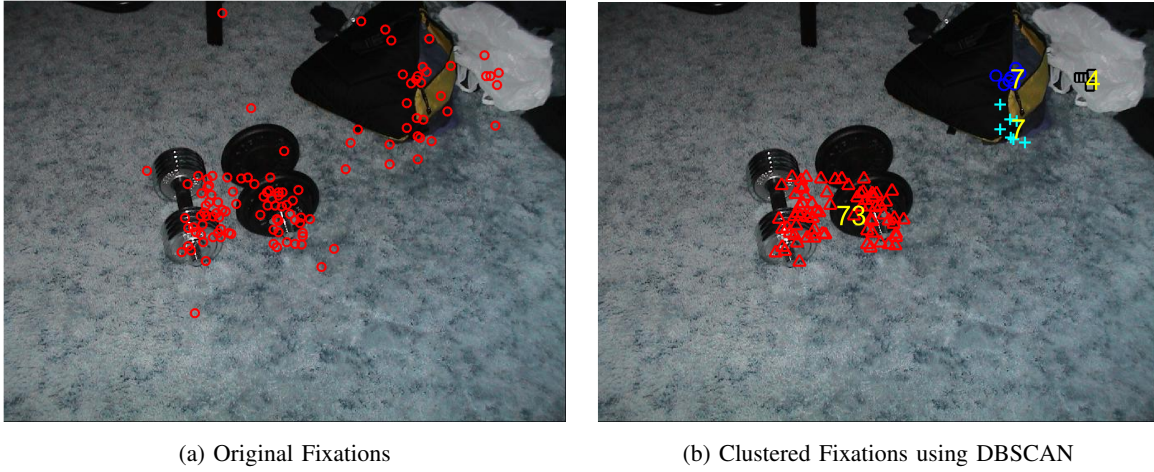


Fig. 5: (a) Importance of fixation density for identifying relevant fixations. The more important fixations are those that are clustered tightly together as they lie on the actual salient regions. The sparsely clustered fixations tend to lie on less salient background regions. (b) Fixation weights assigned based on the number of fixations in each cluster. Different symbols and colors represent different clusters with weights for each cluster shown.

every fixation in a particular cluster a weight equal to the number of fixations in that cluster. Mathematically, if  $P$  is the set of all fixation points, and if  $C = \{c_1, c_2, \dots, c_N\}$  is the set of clusters obtained after applying the DBSCAN algorithm, such that  $C$  is a partition of set  $P$ , then the weights are given by

$$W(p) = |c_i|, \forall p \in c_i \quad (4)$$

where  $|\cdot|$  represents the  $l_0$  norm corresponding to set cardinality (or the number of elements in the set). The DBSCAN algorithm has two parameters, the minimum distance  $\epsilon$  within which points are considered as belonging to the same cluster and the minimum number of points required to form a dense cluster  $minPts$ . We choose the  $\epsilon$  parameter such that it is equal to the diameter of a circle that is subtended by one degree of visual angle for the eye-tracking setup for the Toronto dataset [12], [15]. The  $minPts$  parameter is chosen to be 3 so that isolated clusters with 2 or less points are rejected. For rejected clusters, the weights are considered to be zero (as they represent clusters with zero points) and saliency values at such outlier fixations are ignored during the score computation. An illustration of the weighting scheme is shown in Figure 5(a) where the different clusters of fixation points are shown using different colors, and the weight

for each cluster is shown. On comparison with the original fixations seen in Figure 5(b) one can see that most of the outlier fixations on the carpet that are distant from the objects of interest are rejected and hence do not influence the score.

A shuffled version of the proposed metric that does not exhibit center-bias can be obtained in a manner similar to the  $sNSS$  metric (1) as follows:

$$sWNSS = WNSS(p) - NSS(r) \quad (5)$$

where equal weights are assumed for the random set of fixation points  $r$ . Equal weights are chosen for the random set of fixation points because, unlike the “good” fixation points, random fixations that are chosen from the set of fixations for other images cannot be weighted by a density based criteria and are treated equally in order to capture the centered distribution of fixations for the database to nullify center bias. As shown in Figures 3 and 4, the proposed  $WNSS$  and  $sWNSS$  metrics give a higher score to the better map. The proposed  $sWNSS$  metric is also able to correctly assign the lowest score to the centered Gaussian blob (e) in Figure 3.

#### IV. SUBJECTIVE EVALUATION OF VA MODELS

Even though a large number of metrics have been proposed in the literature for evaluating VA models, currently, there is no ground-truth subjective database that validates these metrics. To address this need and evaluate the performance of our proposed metric, a Visual Attention Quality (VAQ) database is constructed as part of this work. The constructed database consists of saliency maps that are obtained from state-of-the-art VA models and their corresponding ground-truth saliency maps. A ground-truth saliency map is obtained by first aggregating the fixation locations obtained by eye-tracking for all subjects to get a fixation map. The obtained fixation map is then convolved with a 2D Gaussian kernel with a standard deviation  $\sigma$  proportional to one degree of visual angle followed by normalization [5]. Thus, a ground-truth saliency map represents the likelihood that a pixel will be attended to by a human observer. As a result, ground-truth saliency maps are more suitable for at-a-glance visual comparisons as opposed to fixation points [12]. Subjective ratings are obtained by asking human subjects to rate the similarity of the predicted saliency map to the corresponding ground-truth saliency map on a 5-point scale (5-Excellent, 4-Good, 3-Fair, 2-Poor, and 1-Bad). The two aspects the subjects are asked to focus on are the how well the locations of the highest intensity values in the ground-truth match those

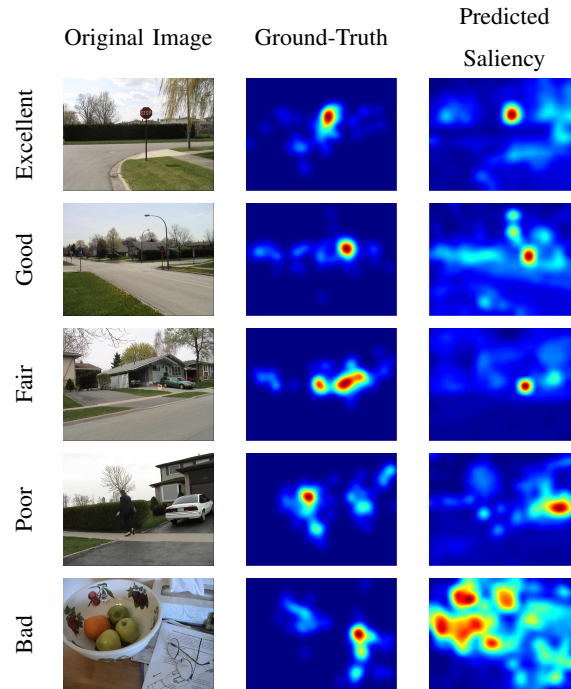


Fig. 6: Training samples from each category shown to subjects before taking the main test.

in the predicted saliency map and the amount of false-positive activity, i.e. high activity in the predicted saliency map that falls on regions of low activity in the ground-truth. The subjects are given a training session and are shown examples of each rating type from excellent to bad. Figure 6 shows samples of the training images (one for each category) shown to the subjects. As seen in Figure 6, in the “Excellent” category the high saliency regions align very well with those in the ground-truth map and have minimal false-positive activity. The misalignment of high saliency regions and amount of false-positive activity increases for the “Good” and “Fair” categories. For the “Poor” to “Bad” categories the highest saliency regions in the ground truth and the predicted saliency maps are totally misaligned and the false positive activity increases from high to very-high, respectively.

The images shown in the training as well as main sessions were taken from the popular Toronto eye-tracking database [12]. The images in that database have all the same size, which makes the computation of the shuffled metrics easier. The images used in the training session were different from those in the main test. To ensure variety in the images shown in the main test, the ground truth maps for all the images were analyzed based on their standard deviation as it is a good measure of the spread of salient regions in an image. Figure 7 shows a histogram

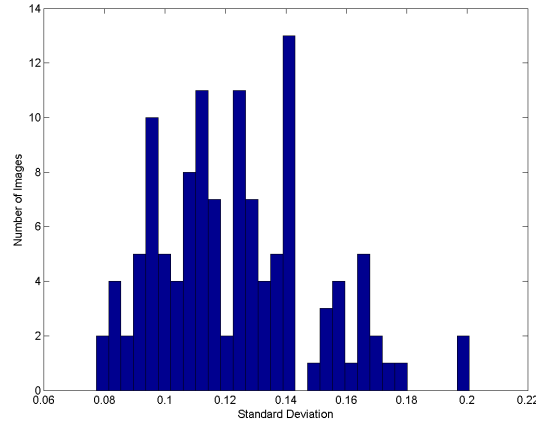


Fig. 7: Histogram of standard deviations for all normalized ground-truth saliency maps in the Toronto dataset [12].

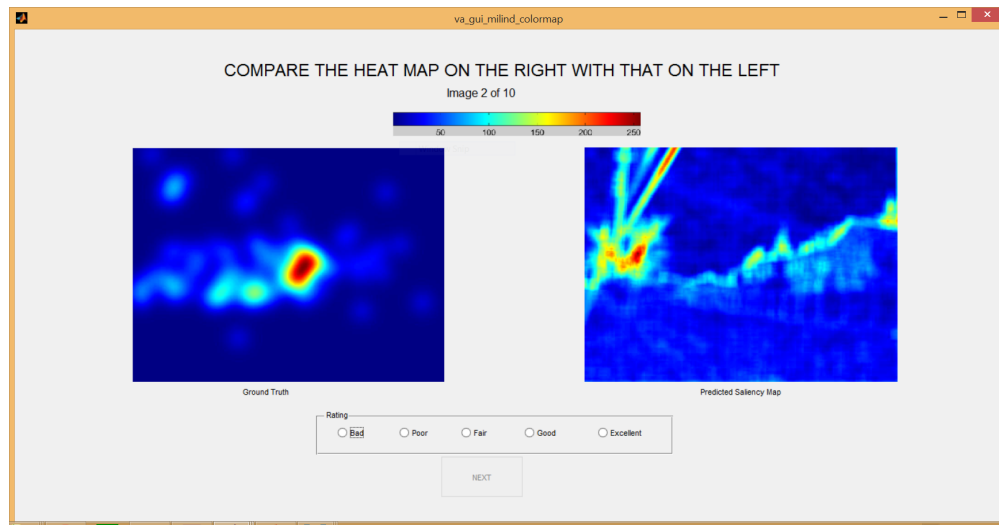


Fig. 8: GUI used for obtaining subjective ratings.

of the standard deviations of the ground-truth saliency maps of the images. Based on the 4 noticeable peaks in the histogram, we cluster the images in the dataset based on their standard deviations by using kmeans. Then, the 3 images with standard deviation nearest to the cluster centroids are chosen for each cluster. This gives us the 12 images that are used in the main test. The GUI used for the subjective testing and the colormap used is shown in Figure 8. Figure 9 shows the images that are chosen and Figure 10 shows their ground-truth saliency maps. We then compute the predicted saliency maps for each of these 12 images using the following 17

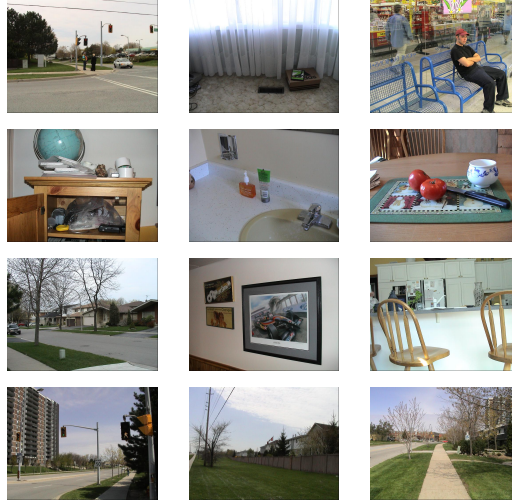


Fig. 9: The 12 chosen images from the Toronto [12] dataset used in the main subjective test.

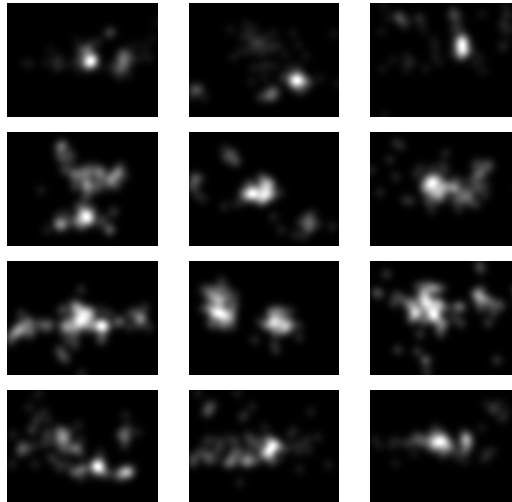


Fig. 10: The ground-truth saliency maps for the 12 chosen images from the Toronto [12] dataset used in the main subjective test.

state-of-the-art VA models: GAFFE [16], ITTI [17], GBVS [18], AIM [12], HouNIPS [19], GR [20], SDSR [21], SUN [22], Torralba [23], FES [24], SigSal [25], SpectRes [26], AWS [27], BMS [28], Context [29], CovSal [30], and RandomCS [31]. We also evaluate the “center” model which is an image independent model that consists of a centered Gaussian blob. In addition, the original ground truth saliency maps are also added to the list of images shown. We expect the “center” model to get lower scores in most cases and the original ground truth saliency maps

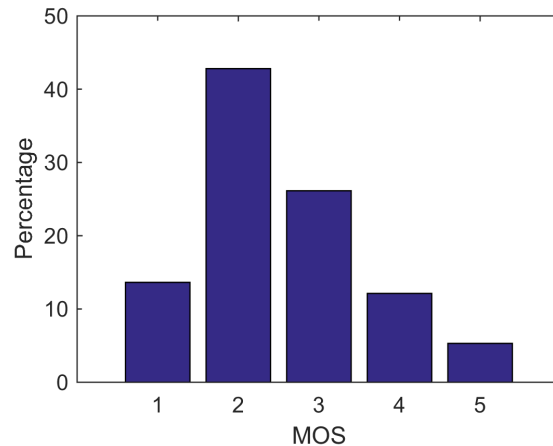


Fig. 11: Distribution of MOS scores.

to get the highest score. The total pairs of ground-truth saliency and test saliency maps shown are 228. These are presented to each subject in a randomized order. Both the ground-truth and predicted saliency maps are shown with the ‘jet’ colormap that indicates high intensity values in red and low intensity values in blue to make it easy for the subjects to assess the maps. The maps were shown to 16 subjects with age ranging between 22 to 33, who were checked for both color blindness as well as visual acuity. Out of the 16 participants, 1 subject was working in the area of visual attention, 7 subjects were working in the area of computer vision but not specifically in the area of visual attention, and 8 subjects were working in areas completely unrelated to computer vision. Out of the 16 subjects, 6 were female and 10 were male. The ratings for each predicted saliency map shown were averaged over 16 subjects to get a mean opinion score. These mean opinion scores (MOS) were then correlated with the scores obtained from popular VA performance metrics in addition to the proposed WNSS and sWNSS metrics. Figure 11 shows the distribution of the MOS scores obtained for the predicted saliency maps given by VA models. It illustrates that in only about 16% of the cases, models received a subjective rating of good or excellent. It also shows that the saliency maps shown cover the entire range of ratings from Excellent to Poor. The VAQ database will be provided online to download for free for the research community to benchmark metrics developed in the future.

Figure 12 shows the ranking of all models in terms of the mean subjective rating obtained for each VA model over all subjects for the VAQ database and shows which VA models are preferred by the human observers for the images in the VAQ database.

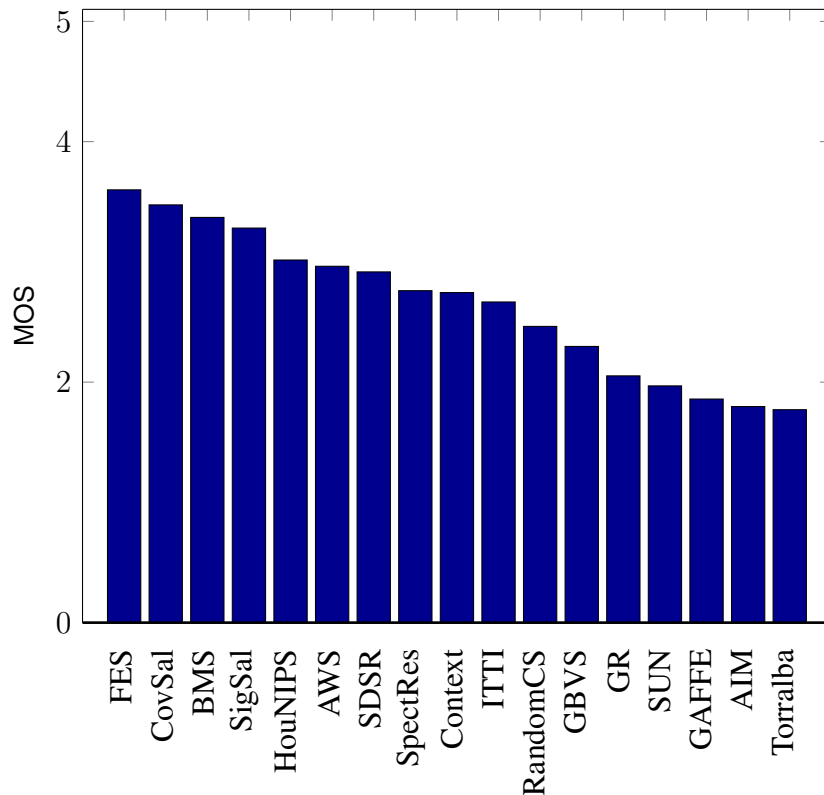


Fig. 12: MOS taken over all predicted saliency maps for each VA model and arranged in descending order.

## V. METRICS PERFORMANCE EVALUATION RESULTS

This section discusses the correlation results between the subjective ratings and the metric scores. To evaluate how good a performance metric is, we compare the scores given by each metric to each of the considered models with the average scores given by the subjects to the same models. To correlate the scores we use the widely used correlation measures of Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KROCC) and Pearson Linear Correlation Coefficient (PLCC). The SROCC and KROCC are rank correlation coefficients, and enable us to compare the ranking given to the VA models by a VA metric with the ranking given by the MOS scores. The PLCC is a linear correlation coefficient that measures how linear the relationship between the metric scores and MOS score is. The metric scores are normalized by the metric score obtained for the ground-truth saliency map that serves as an upper-bound on most metrics before performing the correlation. For the *EMD* metric for which a lower score is better and the best possible score is zero this normalization was not performed,

TABLE II: Correlation results using the VAQ database.

Non-Shuffled Metrics	SROCC	KROCC	PLCC
$AUC_{Borji}$ [3]	0.5617	0.3952	0.5493
$AUC_{Judd}$ [13]	0.5883	0.4222	0.5846
$WF_{\beta}$ [10]	0.3126	0.2113	0.2958
$NSS$ [8]	0.7563	0.5810	0.8297
$EMD$ [13]	0.4470	0.3168	0.5683
$CC$ [13]	0.7461	0.5726	0.8216
$SIM$ [13]	0.5891	0.4246	0.6739
$MAE$ [11]	0.5063	0.3599	0.4803
Proposed $WNSS$	<b>0.7858</b>	<b>0.6178</b>	<b>0.8687</b>
Shuffled Metrics	SROCC	KROCC	PLCC
$sAUC$ [3]	0.5455	0.3871	0.5631
$sNSS$ [7]	0.6526	0.4843	0.7533
Proposed $sWNSS$	<b>0.7624</b>	<b>0.5891</b>	<b>0.8553</b>

and subjective scores were inverted by subtracting the scores from the maximum subject score of 5 to obtain positive correlation scores. For the  $WF_{\beta}$  measure [10], which requires the ground-truth to be a binary mask, we threshold the ground-truth saliency map by its standard deviation as suggested in [32]. For the  $MAE$  metric, instead of using a binary ground-truth map as in [11], we use a real-valued ground-truth saliency map since, by definition, the  $MAE$  can be computed for two real-valued maps. Table II shows the result of the correlations for the VAQ database for all the existing metrics listed in Table I and our proposed metric (the shuffled  $sWNSS$  and non-shuffled  $WNSS$  versions). The results for shuffled and non-shuffled metrics are reported separately because there is no explicit way to remove the center bias effect from the ground-truth saliency maps in the subjective study. As a result, human ratings will tend to be better for the saliency maps that boost central regions over peripheral regions. This leads to non-shuffled metrics like  $NSS$  and  $WNSS$  which tend to reward maps with more central than peripheral activity correlating better with human scores compared to their shuffled versions  $sNSS$  and  $sWNSS$ . Figure 13 shows the scatter plots corresponding to the existing and proposed metrics. From the existing metrics used in the MIT saliency benchmark [9], the  $NSS$  [8] and  $CC$  [3] metrics perform significantly better than the other metrics with the  $NSS$  performing the best among them. The  $AUC_{Borji}$  [3] and its derivative  $sAUC$  [3] metric that suffer from the most number of flaws perform the worst among the MIT benchmark metrics. The  $WF_{\beta}$  [10] metric



also performs poorly. The proposed  $WNSS$  metric gives the best correlation in the non-shuffled metrics and correspondingly the proposed  $sWNSS$  metric gives the best performance among the shuffled metrics.

## VI. CONCLUSION

This paper proposes a locally weighted fixation-density based performance metric for assessing the quality of saliency predictions for VA models. A subjective ground-truth Visual Attention Quality (VAQ) database is created to evaluate the performance of the proposed metric and other existing metrics. Results of the evaluation show that the proposed metrics ( $WNSS$  and its shuffled version  $sWNSS$ ) outperform the widely used  $sAUC$ ,  $AUC_{Borji}$  and  $AUC_{Judd}$  measures as well as other popular metrics used in the MIT Benchmark [9] in terms of their agreement with the subjective ratings. The subjective database is made available online to the research community as a performance metric evaluation benchmark on which future metrics can be tested.

## REFERENCES

- [1] 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, Jan 2013.
- [2] A. Toet, "Computational versus psychophysical Bottom-Up image saliency: A comparative evaluation study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.
- [3] 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, Jan 2013.
- [4] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," *MIT Tech Report*, 2012.
- [5] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, vol. 45, no. 1, pp. 251–66, Mar. 2013. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/22773434>
- [6] 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)*, Dec 2013, pp. 1153–1160.
- [7] M. S. Gide, S. F. Dodge, and L. J. Karam, "The Effect of Distortions on the Prediction of Visual Attention," *ArXiv e-prints*, arXiv:1604.03882, <http://arxiv.org/abs/1604.03882>.
- [8] 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. 2397–416, Aug. 2005.
- [9] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT Saliency Benchmark," <http://saliency.mit.edu/>.
- [10] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 248–255.

- [11] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE CVPR*, 2012, pp. 733–740.
- [12] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *ARVO Journal of Vision*, vol. 9, no. 3, pp. 1–24, Mar. 2009.
- [13] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.
- [14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [15] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2013, pp. 212–217.
- [16] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: a Gaze-Attentive fixation finding engine," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 564–573, Apr. 2008.
- [17] 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, Nov 1998.
- [18] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, vol. 19, pp. 545–552, 2007.
- [19] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2008, pp. 681–688.
- [20] M. Mancas, C. Mancas-thillou, B. Gosselin, and B. Macq, "A rarity-based visual attention map-application to texture description," in *IEEE International Conference on Image Processing*, Oct. 2006, pp. 445–448.
- [21] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, p. 15, Nov. 2009.
- [22] 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. 1–20, Dec. 2008.
- [23] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, May 2001.
- [24] H. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Scandinavian Conference on Image Analysis*. Springer Berlin Heidelberg, 2011, pp. 666–675.
- [25] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 194–201, Jul. 2011.
- [26] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.
- [27] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: a computational approach," *Journal of Vision*, vol. 12, no. 6, pp. 1–22, Jan. 2012.
- [28] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 153–160.
- [29] 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.
- [30] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, p. 11, Mar. 2013, PMID: 23509407.
- [31] T. N. Vikram, M. Tscherepanow, and B. Wrede, "A saliency map based on sampling an image into random rectangular regions of interest," *Pattern Recognition*, vol. 45, no. 9, pp. 3114–3124, 2012.

- [32] A. Mittal, A. Moorthy, and A. Bovik, “Automatic prediction of saliency on JPEG distorted images,” in *2011 Third International Workshop on Quality of Multimedia Experience (QoMEX)*, September 2011, pp. 195 –200.

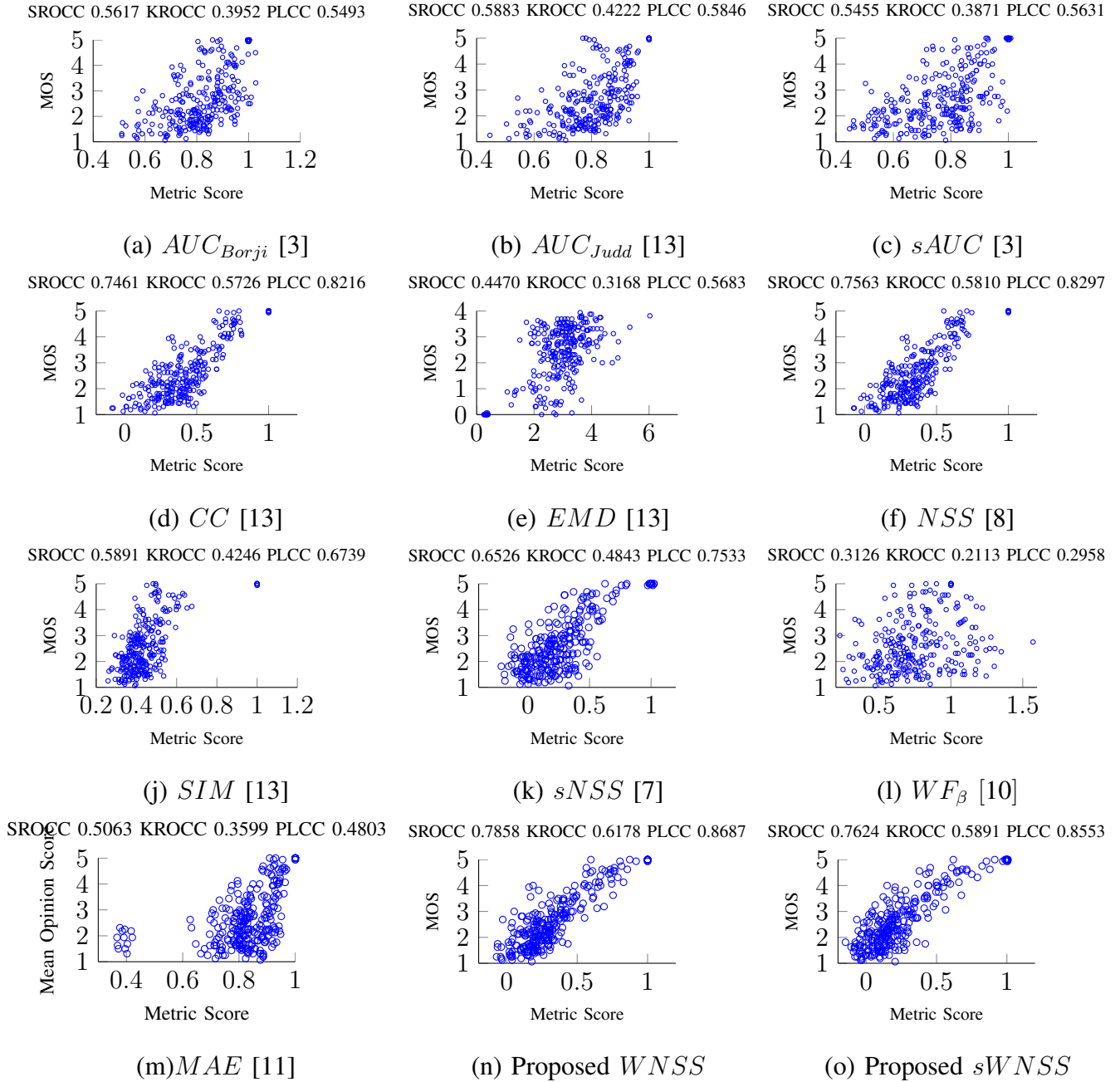


Fig. 13: Scatter plots for all metrics with correlation scores displayed for the VAQ database. Most of the existing metrics along with the widely used AUC based metrics like  $sAUC$  [3],  $AUC_{Borji}$  [3] and  $AUC_{Judd}$  [13] do not show high correlation with subjective scores. Among the non-shuffled metrics, existing metrics  $CC$  [3] and  $NSS$  [3] exhibit better correlation scores with subjective scores and the proposed metric  $WNSS$  performs the best. Amongst the existing shuffled metrics, the recently proposed  $sNSS$  [7] metric performs well with the proposed  $sWNSS$  performing the best.