Revisiting Saliency Metrics: Farthest-Neighbor Area Under Curve

Sen Jia
Ryerson University
sen.jia@ryerson.ca

Neil D. B. Bruce
Ryerson University, Vector Institute
bruce@ryerson.ca

Abstract

In this paper, we propose a new metric to address the long-standing problem of center bias in saliency evaluation. We first show that distribution-based metrics cannot measure saliency performance across datasets due to ambiguity in the choice of standard deviation, especially for Convolutional Neural Networks. Therefore, our proposed metric is AUC-based because ROC curves are relatively robust to the standard deviation problem. However, this requires sufficient unique values in the saliency prediction to compute AUC scores. Secondly, we propose a global smoothing function for the problem of few value degrees in predicted saliency output. Compared with random noise, our smoothing function can create unique values without losing the existing relative saliency relationship. Finally, we show our proposed AUC-based metric can generate a more directional negative set for evaluation, denoted as Farthest-Neighbor AUC (FN-AUC). Our experiments show FN-AUC can measure spatial biases, central and peripheral, more effectively than S-AUC without penalizing the fixation locations. The generated negative samples are available at:https://github.com/SenJia/Farthest-Neighbor-AUC.

1. Introduction

Extensive studies have been proposed to predict the most salient region within an image. Saliency methods can be roughly grouped into two categories, bottom-up and top-down. The former considers the visual stimuli of an image to determine the Regions Of Interest(ROIs); while the latter one assumes the ROI is task-dependent, prior knowledge plays a significant part in saliency prediction. With the recent development of Convolutional Neural Networks (CNNs), saliency prediction heavily relies on model-based algorithms which can be trained in an end-to-end fashion. A new question has emerged regarding what type of saliency features are the best design for applications, top-down or bottom-up? Hand-crafted or CNN-learned?

The measurement of saliency is still challenging because the definition of "saliency" varies depending on the vision task [2], so saliency algorithms can also be grouped by different taxonomies for different purposes [10, 16, 6]. In this paper, we follow the common problem setting in [29, 45], computational models are trained to predict the most salient region to the Human Visual System(HVS). Not only is this a common assumption to all the CNN-based methods [27, 9, 28, 34, 13, 12, 31, 24, 25], but it also can be used in an extremely wide range of applications [48, 1, 47, 44, 26, 22, 35]. However, it is still difficult to comprehensively evaluate saliency models due to the bias of each metric. For instance, the challenge of Large-Scale Scene Understanding [27] uses seven saliency metrics, Shuffled Area Under Curve (S-AUC), Information Gain (IG) [32], Normalized Scanpath Saliency (NSS) [38], Pearson's Correlation Coefficient (CC), AUC-Judd, SIMilarity (SIM) and Kullback–Leibler Divergence (KLD) [30]. Another public saliency benchmark, MIT300 [28], applies eight saliency metrics, AUC-Judd, AUC-Borji, S-AUC, NSS, CC, KLD, SIM and Earth Mover's Distance (EMD) [40]. The use of multiple measures indicates that it is difficult to evaluate a model from only one angle. Previous studies intended to categorize and compare those metrics, e.g., [39, 11] grouped the metrics into location-based and distribution-based. As shown in Figure 1, the location-based measure consists of a set of fixation locations captured by an eye tracker or mouse click. While the distribution of saliency is normally considered as a post-process on the raw data by applying a Gaussian filter.

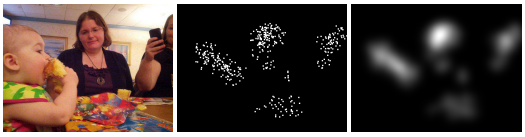


Image Fixation Distribution

Figure 1. A saliency sample from the SALICON dataset.

To evaluate a saliency model, one solution is to overcome the "disagreement" among the metrics. Kummerer *et al.* [32] proposed to optimize the saliency scale, center bias and spatial blurring jointly. But their post-process can hardly satisfy all the metrics simultaneously, the process requires all the compared models and optimization only uses the loss of IG. Later this idea was extended to a metric-

tailored design [33], saliency models and saliency maps are decoupled so that one model can output different metric specific maps. However, their solutions are based on the assumption that all the metrics are able to evaluate saliency reasonably, the output should be optimized separately and specifically for each metric. In this work, we are more interested in investigating the differences among those metrics in theory. We believe some of the metrics may contain inherent drawbacks so that not necessarily all the metrics should be considered. The first contribution of this work is that we revisit the widely used saliency metrics based on [11], showing that the "balanced" metrics, NSS and CC, still have limitations in evaluating modern CNN-based systems. Saliency datasets have been created using their own choices of Gaussian standard deviation and a CNN model learns to fit this biased distribution, see Section 2.

Center bias is a long-standing problem in saliency evaluation, simply placing a Gaussian distribution at the center could outperform a well-designed system on most of the metrics [46]. S-AUC is a common solution used by the SALICON and the MIT benchmarks, a negative set is sampled based on the positive from other images within the same dataset. The study [11] shows a centered Gaussian distribution can only achieve an S-AUC score of 0.5. However, S-AUC has a strong bias that it only considers negatives near the center, peripherally-favored systems can achieve higher scores [8]. In this paper, we propose a new saliency metric which introduces one more constraint on the spatial relationship between the positive and negative. We show that the distribution of all fixations can be interpreted as a 2D probability density distribution. Our method builds the negative set for each image by searching its farthest neighbors according to the distribution similarity, denoted as FN-AUC. We also propose a fast version of our FN-AUC in case the size of the dataset is too large. To compare with S-AUC, we propose a strategy to measure the quality of the sampled negative set, which takes the spatial relationship of both the center bias and the positive into account, see Section 3.2. Our experiment shows FN-AUC can draw a more reasonable negative set in order to penalize the center bias only without undermining the true positives, Section 4.3.

Another contribution of this work is that we propose a global smoothing strategy in computing AUC metrics. Based on the MIT benchmark, it is problematic to compute the receiver operating characteristic (ROC) property when many locations share the same value magnitude, which could result in a lower performance as well. This is a scenario that is common for CNN based models which can produce near binary outputs. One solution could be jittering a map by adding small random noise, but this may break the relative saliency rank. We propose to apply a Gaussian filter using a relatively large standard deviation, the output is expected to cover all regions within a map. In this way, our method can generate a map with unique values for an AUC metric, meanwhile retaining the relative saliency relationship, see Section 2.3.

Dataset	Size (Width by Height)	σ	CC
Toronto [9]	681×511	20	.998
MIT1003 [29]	(Min - Max)$405 - 1024$	24	.998
CAT2000 [3]	1920×1080	41	.998
SALICON [27]	640×480	19	.999

Table 1. The attributes of the four saliency datasets. Gaussian processes with different standard deviations are used to generate the distribution ground-truth.

2. Revisiting Saliency Metrics

In this section, we first revisit the widely used saliency metrics based on previous studies [39, 8, 11]. We further investigate the impact of applying the balanced saliency metrics across datasets, NSS and CC.

2.1. Distribution-based Metrics

The distribution-based saliency metrics, SIM, CC, EMD and KLD, consider each saliency map as a distribution then the similarity between two maps can be measured based on a probabilistic view. The drawbacks of IG, SIM and KLD have already been studied in that they focus more on FNs than FPs[1] which leads to a biased evaluation. While the EMD metric is sensitive to the sparsity of the map, a lower score can be obtained due to fewer bins requiring moving. The CC measure is recommended for penalizing FPs and FNs equally [11]. However, when comparing two distributions, the "shape" of each distribution also plays an important role (the choice of the Gaussian sigma σ). Especially for CNN-based saliency systems, a CNN model is designed to learn the distribution information from the training set, a lower performance may be achieved only because the test set is drawn from a different distribution. This is a common problem in practice because there is no standard on how to build the ground-truth, each dataset was built using a different σ value. The Gaussian filter can be written as:

$$g(m,n) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{m^2+n^2}{2\sigma^2}} \quad (1)$$

where m and n represent the distance from the current location and σ denotes the standard deviation. We search the σ value used by the four saliency datasets, Toronto[9], MIT1003[29], CAT2000[3] and SALICON [27], based on the highest CC score achieved. As we can see in Table 1, the σ value used varies across those datasets. A high performance can be achieved simply because the distributions of the training and test sets are similar and vice versa. We believe the location of ROI should be considered more for saliency instead of the distribution, a good model should

[1]False Postive(FP), False Negative(FN), True Positive(TP), True Negative(TN).

capture the correct region but the shape (or contrast) is less important. One can imagine that all the distribution-based metrics suffer from the inherent drawback of shape sensitivity and it is difficult to avoid the center bias problem, see Section 4.

2.2. Location-based Metrics

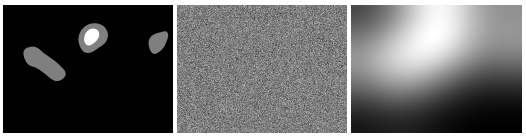
Location-based metrics do not rely on the distribution built by the Gaussian process(Equation 1). Similar to CC, NSS is the recommended metric due to its equal penalty on FPs and FNs [11]. However, NSS essentially considers all the fixation locations as positive and the others as negative, more FPs will be introduced when a larger σ is applied on the training set. Our experiment validates this hypothesis by showing the highest NSS score is achieved when training on the setting of the smallest σ, see Section 4.1. This bias of NSS also makes it challenging to evaluate models across datasets.

The family of AUC metrics was criticized for ignoring FPs with small values [11]. But the FPs will be ignored only when its value is smaller than the smallest threshold (value at fixations). That is, in practice, the ignored FPs would be *relatively* small and this relative saliency is considered to be more important than absolute magnitudes [8]. One study [11] has shown that AUC metrics are robust to σ, but this happens only when the highest value of prediction is a TP. Our experiment shows the AUC metric is also slightly affected by the choice of σ, Section 4.1, but they are relatively more robust than CC and NSS. Nevertheless, the AUC metric is still the most promising way to overcome the center bias issue because we can directly sample negatives rather than considering distribution properties. Before discussing the center bias and demonstrating our method, we first show a potential problem in computing AUC metrics and our proposed global smoothing function in the next section.

2.3. Global Smoothing Strategy

As shown in Figure 1, there are different ways to represent the fixation ground-truth. Both of the maps can be considered as matrices, the distribution map can also be interpreted as a 2D probability function and the fixation map can be converted into a set of positive locations. In this work, we demonstrate our method using all interpretations of saliency. For clarity, we denote a matrix as $\mathbf{X}$ or $\mathbf{Y}$ (an image or the fixation map, the first and second graphs in Figure 1), a set of coordinates as $\mathcal{P}$ or $\mathcal{N}$ and the probability function as $f_{\mathbf{X}}$ (density or distribution maps, third graph in Figure 1) and they can be converted to each other, see Section 3.2.

All the AUC-based metrics are computed by applying various thresholds on the map to draw an ROC curve. This can be problematic when different positive locations share



(a) Quantized (b) Random Noise (c) Global Gaussian

Figure 2. (a) A quantized example map to show the problem of limited value degrees. (b) A random noise map, $\mathbf{O}$, to jitter the output. (c) Our proposed global Gaussian map, $\mathbf{G}$.

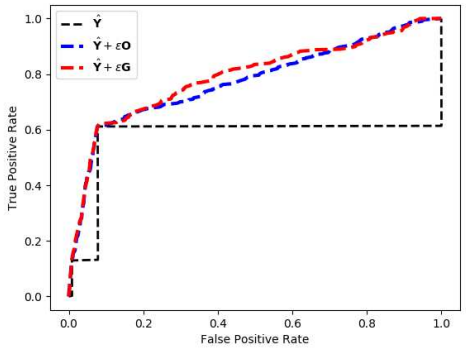


Figure 3. ROC curves of the three maps, the AUC scores are: Quantized(black): 0.573, Jittered with Random Noise(blue): 0.774, Jittered with Global Gaussian(red): 0.790

the same magnitude value. Let's denote an output map as $\hat{\mathbf{Y}}$ and we quantize the map into three value degrees to demonstrate this problem, $\{0, 0.5, 1\}$. It is worth noting that this problem is not artificial since the output image format only contains 255 value degrees in most cases. One naive solution[2] is to add a noise matrix $\mathbf{O}$ to the prediction map, $\hat{\mathbf{Y}} + \varepsilon\mathbf{O}$, $O_i \sim \mathcal{U}(0,1), \mathbf{O} \in R^{h \times w}$, where O_i is ith element of the noise matrix and ε is a small number. But this operation may break the relative saliency relationship due to the randomness.

In this work, we propose a global smoothing process to solve this problem. Instead of introducing random noise, we build a global Gaussian map, $\mathbf{G}$, by applying Equation 1 using a relatively large standard deviation to *diversify* the value range, e.g., $\sigma = min(h/4, w/4)$. As shown in Figure 2, the noise matrix $\mathbf{O}$ used by [28] is similar to white Gaussian noise. Our global Gaussian map, $\mathbf{G}$, covers most areas of the image for tie-breaking. From Figure 3, we can see that our proposed map $\mathbf{G}$ achieves a larger AUC compared with the random noise $\mathbf{O}$. Note that this improvement is model-agnostic in evaluation, but we are hoping it results in a more meaningful prediction for a fair model comparison. This operation can be also combined with the widely used Gaussian post-process, which utilizes a small value of σ to achieve "local" smoothness. Given the small performance differences between models, this operation may arguable help to disambiguate which are in fact the best performing.

[2]According to the MIT benchmark https://github.com/cvzoya/saliency/tree/master/code_forMetrics

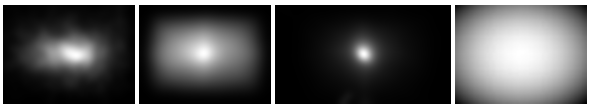
3. Center Bias and Spatial Metrics

3.1. Center Bias

The measurement of saliency has suffered from the center bias problem for a long time. There exist various causes for center bias, e.g., viewing strategy, initial orbital positions or motor bias. The main reason behind this may also be the photographer bias, humans tend to place the most interesting object or region near the center of an image [43, 36, 37, 42]. Therefore this tendency makes it difficult to show how good a saliency model is, a "faked" high performance may result from centrally biased methods. An early saliency study [36] has shown the stimuli in an image, e.g., color, intensity and orientation, are important in guiding attention. They also showed a discrepancy that the predictions are uniformly distributed within an image while the fixations are more likely to be near the center. This leads to a hypothesis that the low-level visual features have an indirect effect on attention, while the resulting 'objectness' is more significant [15]. Later an alternative explanation to their work was proposed by [4], the good performance achieved is because the objectness corresponds more with the center bias.

We show the distribution maps by applying Equation 1 on all the fixation locations within each dataset in Figure 4. The center bias is intrinsic to HVS across the datasets such that a synthetic center bias map, fourth map in Figure 4 denoted as $\mathbf{CB}$, can achieve a decent performance by covering most of the fixation locations. We believe the objective of saliency prediction is to model the mechanism of HVS regardless of what types of features should be used or what the bias could be. The metric applied is expected to differentiate a good system from a synthetic map. Most of the metrics suffer from this problem because they are not designed for spatial biases, especially for the distribution-based case. In contrast, the location-based metric seems more promising on this issue.

The standard AUC metric was originally used for statistical analysis, and later was introduced to measure saliency performance by [7], also known as AUC-Judd [29], which considers all the non-fixated locations as negative. AUC-Borji [5] proposed to randomly sample negatives from all the non-fixated locations, which can be considered as a subset of AUC-Judd. But these two variants of AUC are not designed for the problem of center bias. S-AUC is a widely used metric specifically for the center bias, which samples negatives based on positive locations from other images within the same dataset. The assumption behind this is that the positives are also subject to a central Gaussian distribution so that they can be used to penalize the synthetic center map $\mathbf{CB}$, see Figure 4. But this sampling strategy ignores the spatial relationship between the positive and negative, which may result in an "over-penalty" to TPs.



Toronto SALICON CAT2000 Center Bias

Figure 4. Distribution maps generated using all the fixations within each dataset and the center bias map from the MIT benchmark.

Furthermore, S-AUC may favor an "anti-center" prediction, saliency methods [21, 41] that are biased to peripheral regions and can achieve a higher S-AUC score [8].

3.2. Farthest-Neighbor AUC

Before we show our proposed FN-AUC, we first formalize AUC metrics in different representations and discuss their own focuses in evaluation. Let $\mathbf{X} \in R^{h\times w\times 3}$ and $\mathbf{Y} \in R^{h\times w}$ denote the input image and its fixation ground-truth (second image in Figure 1) respectively. The fixation map $\mathbf{Y}$ can be converted into a set of fixation locations, denoted as $\mathcal{F} = \{(m,n) : \mathbf{Y}(m,n) = 1, m = 1\dots w, n = 1\dots h\}$. The set of all the possible locations $\mathcal{S}$ can be formulated as $\mathcal{S} = \{(m,n) : m = 1\dots w, n = 1\dots h\}$, $\mathcal{F} \subseteq \mathcal{S}$. When computing an AUC score, the positive set is $\mathcal{P} = \mathcal{F}$, but the generation of the negative set varies depending on the AUC metric applied. For AUC-Judd, it considers all the non-fixated locations as negative, $\mathcal{N}^J = \{\forall l \in \mathcal{S} : l \notin \mathcal{P}\} \iff \mathcal{S} \setminus \mathcal{P}$. For AUC-Borji, the negative set can be considered as applying $Bernoulli$ sampling on $\mathcal{N}^J$ with a cardinality constraint, $\mathcal{N}^B = \{s \in \mathcal{N}^J : |\mathcal{N}^B| = |\mathcal{P}|\}, \mathcal{N}^B \subseteq \mathcal{N}^J$ (a bijective function could be applied). It is obvious that both the AUC metrics actually focus on the same statistical property, but they can not overcome the center bias problem because no spatial information is utilized.

For S-AUC, we build the positive set for the entire dataset by $\mathcal{P}_{all} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_N\}$, assuming there are N samples in the dataset. The positive set of S-AUC is the same as AUC-Judd and AUC-Borji, $\mathcal{P} = \mathcal{F}, \mathcal{P} \subseteq \mathcal{P}_{all}$. For the negative set, S-AUC draws samples according to $\mathcal{N}^S = \{s \in \mathcal{P}_{all} : s \notin \mathcal{P}, |\mathcal{N}^S| = |\mathcal{P}|\}$. It is interesting to see that both S-AUC and AUC-Borji sample negatives from other sets, $\mathcal{P}_{all}$ and $\mathcal{N}^J$ respectively using the $Bernoulli$ sampling process. But S-AUC implicitly assumes that the set $\mathcal{P}_{all}$ is spatially subject to a centered Gaussian distribution (empirically validated by Figure 4) so that the sampled negatives can be used to penalize the center bias. But AUC-Judd and AUC-Borji do not make use of this spatial information.

Given the size of the image (h, w), we can easily "vectorize" the total positive set $\mathcal{P}_{all}$ into a fixation map by first initializing a zero matrix, $\mathbf{Y_{all}} = \mathbf{0} \in R^{h\times w}$, then $\mathbf{Y_{all}}(m,n) = 1 : \forall(m,n) \in \mathcal{P}_{all}$, let $v(\cdot)$ denote this "vectorization" conversion. A Gaussian filter (Equation 1) is applied on the map $\mathbf{Y_{all}}$ to build the distribution map $f_{\mathbf{Y_{all}}}$ for each dataset as shown in Figure 4. Al-

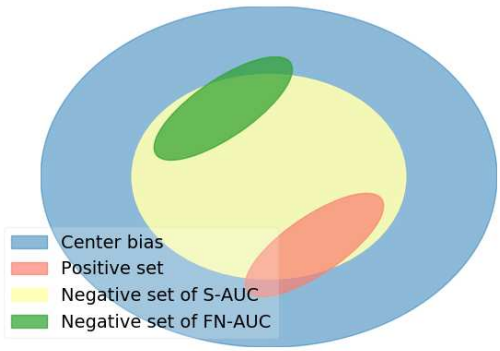


Figure 5. Diagram of our proposed FN-AUC vs S-AUC, our method aims at sampling a more directional negative set.

Figure 6. Distribution maps of the negative set sampled by S-AUC and FN-AUC, with different numbers of neighbors on the Toronto dataset.

though the distribution map is a matrix, it can be interpreted as a 2D probability density function given the constraint $\sum_{m=1}^{h}\sum_{n=1}^{w} f_{\mathbf{Y_{all}}}(m,n) = 1$. In this way, the density maps in Figure 4 can be viewed as the probability function of the S-AUC sampling process. Although each element in the negative set $\mathcal{N}^S \subseteq \mathcal{P}_{all}$ is drawn by $Bernoulli$ sampling, spatially it can be interpreted as a $Poisson$ sampling process such that the sampled elements have a higher probability to be located near the center (equal probability for sampling).

Given the synthetic center bias map $\mathbf{CB}$ (as shown in Figure 4) and its probability function $f_{\mathbf{CB}}$, we can reformulate the center bias problem in terms of the distribution similarity between $f_{\mathbf{CB}}$ and $f_{\mathbf{Y_{all}}}$. The synthetic map is designed to "mimic" the distribution of the total fixations as a baseline such that the distance between the distributions should be low or minimized, $arg\,min(d(f_{\mathbf{CB}}, f_{\mathbf{Y_{all}}}))$. The S-AUC metric can penalize the center bias because the probability distribution of $f_{v(\mathcal{N}^S)}$ is similar to $f_{\mathbf{CB}}$, given that $\mathcal{N}^S$ is sampled from $\mathcal{P}_{all}$.

The S-AUC metric only considers global position information that the negative sample should be near the center, but it ignores the relative spatial relationship between the positive and negative. The positive set is also a subset of the total $\mathcal{P} \subseteq \mathcal{P}_{all}$, which means spatially the probability function of $f_{v(\mathcal{N}^S)}$ also overlaps with $f_{v(\mathcal{P})}$. This may lead to an over-penalty on the TP rate and also explains why S-AUC blindly favors peripherally-focused methods [8]. To solve this problem, we propose to not only make use of the global information, but also take the relative spatial relationship into account. The sampled negative set should be able to penalize the center bias map $\mathbf{CB}$ meanwhile without affecting the positive set. It is easy to formulate this constraint in the representation of a probability function, $arg\,max(d(f_{v(\mathcal{P})}, f_{v(\mathcal{N}^{FN})}))$, the sampled negative set by FN-AUC should be far apart from the positive locations. We visually show the relationship between the negative sets drawn by S-AUC and our method in Figure 5. The negative set of S-AUC is near the center, which overlaps with the positive. While our method intends to avoid the positive locations but still sample within the area of the synthetic map $\mathbf{CB}$.

Algorithm 1 Farthest-Neighbor AUC

Input: $(\mathbf{X_i}, \mathbf{Y_i})$, ith Data Sample in the dataset. $\mathcal{P}_{all} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$, a set contains all the fixation locations within the dataset.

Output: $\mathcal{N}_i^{FN}$ the negative set for ith sample.

Initialize an empty list, denoted as L.
for $j = 1$ to N **do**
  if $i \neq j$ **then**
    $d_j = d(f_{v(\mathcal{F}_i)}, f_{v(\mathcal{F}_j)})$
    add $(d_j, \mathcal{F}_j)$ to L.
Sort L in descending order based on d_j. ▷ suppose $d(\cdot,\cdot)$ is a similarity measure.
Add the associated fixation set to the $\mathcal{N}_i^{FN}$ based on the top K elements in L, $\mathcal{N}_i^{FN} = \{\mathcal{F}_k : (d_k, \mathcal{F}_k) \in L, k = 1 \ldots K\}$.
return $\mathcal{N}^{FN}$

We show how our FN-AUC samples the negative set $\mathcal{N}^{FN}$ in Algorithm 1. The negative set consists of fixations from the neighbors that are least similar to the positive set, i.e. the farthest neighbors. Then we can sample from this negative set to have the same cardinality as the positive, $|\mathcal{N}^{FN}| = |\mathcal{P}|$. It is obvious that the FN-AUC sampling process has a complexity of $O(n)$. It is feasible to apply FN-AUC on a small dataset, e.g., Toronto, but it becomes problematic for large-scale datasets, e.g., SALICON. We also propose a fast version of FN-AUC for better scalability. Normally the number of fixations of each image is similar within one dataset. We can select only one farthest neighbor, $K = 1$ in Algorithm 1 and omitting the cardinality constraint. More importantly, we can set an empirical threshold to select the first matched element without iterating over the entire dataset, e.g., a CC score below zero (inversely related). One extreme case could be that there exists one sample whose positive set is near the corner, every other sample may select it as the farthest neighbor such that FP rate is always zero. In this case, increasing the number of neighbors K can deliver a more robust sampling process. However, we did not experience this problem on the datasets even applying $K = 5$. When K equals to the total number of the images within a dataset($K = N - 1$), FN-

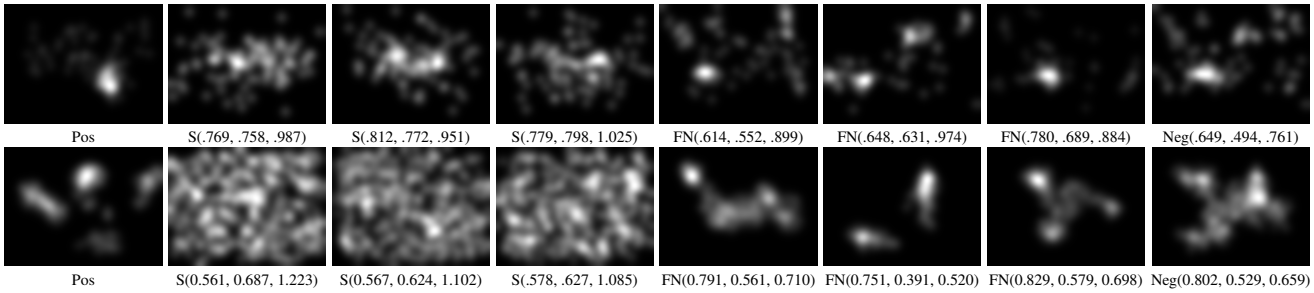


Figure 7. The negative set sampled by S-AUC and FN-AUC. The first column is the distribution map of positive locations. Columns 2-4 are negative maps by S-AUC, column 5-7 are negative maps by FN-AUC. The last column is the final output map of FN-AUC, sampled from columns 5-7. The top row is a sample from the Toronto dataset drawn using Algorithm 1, while the bottom is from the SALICON dataset drawn using the fast version of FN-AUC($K = 3$, $CC < 0$). The annotation represents (β, γ, γ/β).

AUC will reduce to S-AUC by sampling from $\mathcal{P}_{all}$. That is the larger value of K, the more we penalize the center bias map, we can see the effect of this trade-off of K in Figure 6.

To compare with S-AUC, we propose to measure the quality of the negative set in two terms: 1) considering the negative set as positive locations and the center bias map $\mathbf{CB}$ as a prediction to measure the performance, hoping that a high score (e.g., CC or AUC) can be achieved so that the negative set can penalize $\mathbf{CB}$. 2) considering the negative set as a prediction map ($f_{v(\mathcal{N})}$) to measure its performance on the ground-truth $\mathcal{P}$, a low score is expected so that the negative set has little impact on the positive. Let's denote these two measures as $\uparrow \beta$ (the higher the better) and $\downarrow \gamma$ (the lower the better) respectively, we also show the ratio $\downarrow \gamma/\beta$ as an indicator of the quality.

We can visually check the negative sets of $\mathcal{N}^S$ and $\mathcal{N}^{FN}$ in Figure 7 (presented in distribution for visualization). Unsurprisingly, the random samples from S-AUC, columns 2-4, tend to locate near the center, which leads to a higher β value. But S-AUC dose not take the relative spatial relationship into account, the sampled negative map also largely overlaps with the positive, which results in a higher γ value as well. While the three negative candidates drawn by FN-AUC(columns 5-7) try to avoid penalizing the positive, therefore a lower γ can be achieved. Moreover, the negative set still intersects with the center bias map, because the samples are drawn from $\mathcal{P}_{all}$ whose 2D distribution is similar to $f_{\mathbf{CB}}$. The final output negative set (the last column) sampled from the neighbors also achieves low scores of γ and γ/β. It is interesting to see that our proposed FN-AUC has a more significant effect when evaluating on the SALICON dataset. The maps of S-AUC(bottom row) have ratio values of $\gamma/\beta > 1$, which indicates the negative set is penalized more on the positive over the center bias. The reason behind this is that the fixations of this dataset are more spread-out covering almost the whole image, see Figure 4. While our method can still generate more directional negative sets achieving low γ/β values. (More examples are shown in Figure 1 in the supplementary material.)

4. Experiment

Implementation Details: SALICON is the largest saliency dataset. Its training set SALICON-train contains $10,000$ images with a resolution of 480×640 for training. We train CNN models on SALICON-train and report result on SALICON-val and the other datasets as shown in Table 1. The Toronto dataset has a similar image size to SALICON, so we simply resize the Toronto dataset during evaluation. But for MIT1003 and CAT2000, the ratio of the image size is completely different from SALICON. We apply the padding strategy used in [13, 12], each image has been resized and padded to keep the same ratio ($3/4$) as the input.

The network used is the ResNet-50 model [20], which has been pretrained on the ImageNet dataset. We simply apply the multi-level strategy used in [12, 31, 25] on the model, the side outputs from $\{conv1, conv10, conv22, conv40, conv49\}$ are combined to generate the final prediction. The initial learning rate was set to 0.1 with a weight decay of $1e^{-4}$. The total number of training epochs was 10 and we reduced the learning rate every three epochs by multiplying by a factor of 0.1. The batch size was set to 8 and stochastic gradient descent was used to update the model after computing the mean squared error as the loss.

Negative Set of FN-AUC: To compute FN-AUC scores, we first build the negative set $\mathcal{N}^{FN}$ for each dataset. The metric of CC was used to compare the two distributions however other similarity measures can also be applied. To compare with S-AUC, we build a more directional negative set for FN-AUC in this experiment, $K = 5$, and the final negative set is randomly drawn from the neighbors such that the number of elements is the same as the positive set. For Toronto, MIT1003 and CAT2000, we used the standard procedure as shown in Algorithm 1. For SALICON, we applied the fast version of FN-AUC due to its large size with selected candidates chosen according to the first $K = 5$ neighbors satisfying the requirement of $CC < 0$. An optimal way to choose K could be based on the average ratio of γ/β across the entire dataset.

Dataset	Metric	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	$\sigma = 40$	$\sigma = 50$	Deviation
Toronto $\sigma = 20$	CC	0.684	**0.694**	0.684	0.669	0.647	0.016
	NSS	**2.016**	1.938	1.839	1.754	1.665	0.125
	AUC-J	0.852	**0.856**	0.855	0.853	0.854	0.001
	AUC-B	0.810	0.828	0.837	0.840	**0.842**	0.011
	S-AUC	0.713	**0.717**	0.713	0.705	0.694	0.008
	FN-AUC	0.805	0.817	**0.824**	**0.824**	0.820	0.006
CAT2000 $\sigma = 41$	CC	0.539	0.556	0.556	0.558	**0.559**	0.007
	NSS	1.688	**1.697**	1.668	1.655	1.641	0.020
	AUC-J	0.846	0.853	0.854	0.859	**0.864**	0.005
	AUC-B	0.814	0.837	0.844	0.852	**0.860**	0.015
	S-AUC	0.702	0.716	0.719	0.725	**0.732**	0.010
	FN-AUC	0.704	0.716	0.716	**0.719**	0.717	0.005
MIT1003 $\sigma = 24$	CC	**0.623**	0.610	0.586	0.558	0.532	0.033
	NSS	**2.497**	2.330	2.175	2.024	1.906	0.210
	AUC-J	0.897	**0.899**	0.898	0.897	0.895	0.001
	AUC-B	0.860	0.873	0.879	**0.880**	**0.880**	0.007
	S-AUC	0.807	**0.816**	**0.816**	0.812	0.809	0.003
	FN-AUC	0.758	0.778	**0.788**	**0.788**	0.783	0.011
SALICON $\sigma = 19$	CC	0.843	**0.863**	0.860	0.839	0.812	0.018
	NSS	**1.895**	1.832	1.757	1.668	1.580	0.112
	AUC-J	0.858	**0.859**	0.857	0.853	0.848	0.004
	AUC-B	0.811	0.834	0.843	**0.845**	0.842	0.012
	S-AUC	0.781	0.799	**0.805**	0.804	0.799	0.008
	FN-AUC	0.833	0.857	**0.868**	0.867	0.861	0.012

Table 2. CNN models trained on ground-truth maps built using different σ values. The σ value used by each test set is shown. The highest score of each row is highlighted in bold. The standard deviation shows how robust each metric is to the change of σ. (AUC-J for AUC-Judd and AUC-B for AUC-Borji.)

CC	0.018	NSS	0.116	AUC-J	0.002
AUC-B	0.011	S-AUC	0.007	FN-AUC	0.008

Table 3. The average deviation for each metric across datasets.

4.1. The choice of σ

As discussed in Section 2, the distribution map of ground-truth varies according to the choice of σ. In this experiment, we show how this problem affects CNN-based systems when evaluating across datasets. Five different σ values $\{10, 20, 30, 40, 50\}$ are applied on SALICON-train to build ground-truth for training using Equation 1. Then five CNN models with the same architecture are trained on each of the created ground-truth distributions. We evaluate the five models on different test sets using different metrics to show how sensitive those metrics are to the choice of σ. A metric can be considered *biased* if a score for one model clearly outperforms or underperforms compared to the other models for the same metric since that the model architecture is the same.

We report the results of the five models in Table 2. We can see that a high score of CC tends to be achieved by the matched distribution when similar σ values are applied. The NSS metric favors a small σ value and thus, less FPs are produced (a more sparse prediction). This experiment validates our discussion in Section 2. NSS is sensitive to the σ value applied on the training set while CC relies more on the σ difference between the training and test set. A potential risk that may further limit the practical use of distribution-based metrics is that interpolation operations (e.g., bi-linear) can also affect distribution properties. As shown in Table 2, the AUC metrics are also sensitive to the choice of σ. For the AUC metrics, AUB-Borji, S-AUC and FN-AUC, there exists a sampling process which may deliver slight randomness into evaluation. We report the deviation of each setting in the last column of Table 2 and its average score across datasets in Table 3. We can see that the sensitivity of AUC metric is relatively smaller than CC and NSS. Even though AUC-Judd seems correlated with σ from Table 2, a low deviation score denotes that it does not show which setting has a clear advantage. Larger deviations of CC and NSS may result from their value range or the computation process and due to this, they may deliver false intuitions with respect to the quality of the model.

Method	CC	NSS	AUC-J	AUC-B	S-AUC	FN-AUC
CB	0.397	0.969	0.802	0.786	0.515	0.607
Itti [23]	0.270	0.820	0.693	0.677	0.638	0.701
AIM [9]	0.312	0.896	0.725	0.720	0.659	0.725
GBVS [19]	0.569	1.519	0.829	0.819	0.636	0.747
SUN [46]	0.215	0.650	0.665	0.652	0.610	0.654
SDSR [41]	0.403	1.096	0.763	0.756	0.697	0.786
CAS [18]	0.449	1.271	0.781	0.768	0.688	0.781
AWS [17]	0.466	1.341	0.787	0.775	0.707	0.789
SWD [14]	0.575	1.523	0.836	**0.828**	0.632	0.741
ImageSig [21]	0.396	1.085	0.762	0.749	0.679	0.753
CNN	**0.694**	**1.938**	**0.856**	**0.828**	**0.717**	**0.817**

Table 4. Comparison of different saliency methods on the Toronto dataset.

4.2. Spatial Biases

The inner workings of NNs are still under exploration but it is established that the features learned by NNs contain high-level objectness, therefore CNN-based methods may share the same spatial bias as discussed in [4]. Our proposed metric aims at solving the spatial bias problem, so we compare different types of "early" visual features(low-level) as well as the center bias map (Figure 4) using our metric. The center bias map, **CB**, is taken from the MIT benchmark. The traditional saliency methods for comparison include: Itti [23], AIM [9], GBVS [19], SUN [46], SDSR [41], CAS [18], AWS [17], SWD [14] and ImageSig(RGB) [21]. It has been shown that those traditional methods utilize various low-level features [8, 4, 46], which leads to relatively different predictions. Some of the hand-crafted methods are considerably time consuming (e.g., CAS needs more than 20 seconds to process each image due to its multi-scale design). Therefore we only focus on the smallest dataset, Toronto, for simplicity. We take the model ($\sigma = 20$) from the last experiment as a CNN baseline because it is the closest to the default settings of Toronto and SALICON. The metrics can be roughly categorized into bias-tolerant (CC, NSS, AUC-J, AUC-B) and bias-sensitive (S-AUC, FN-AUC).

As shown in Table 4 and Figure 8, we can see that the CNN model achieves the best performance on all the metrics, including FN-AUC. While it is trivial to compare high-level vs low-level features in this experiment, we are more interested in how the metrics measure the intrinsic spatial

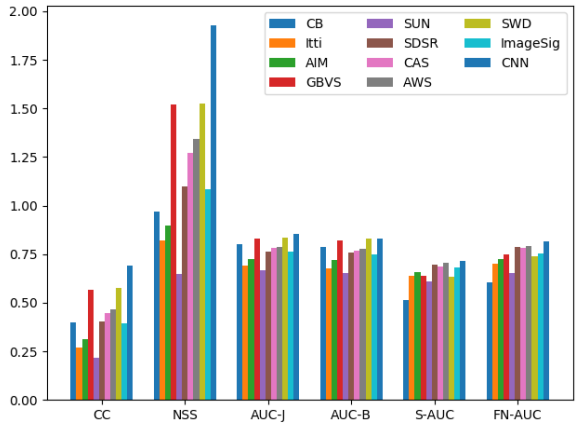


Figure 8. Bar graph of the compared methods using different saliency metrics.

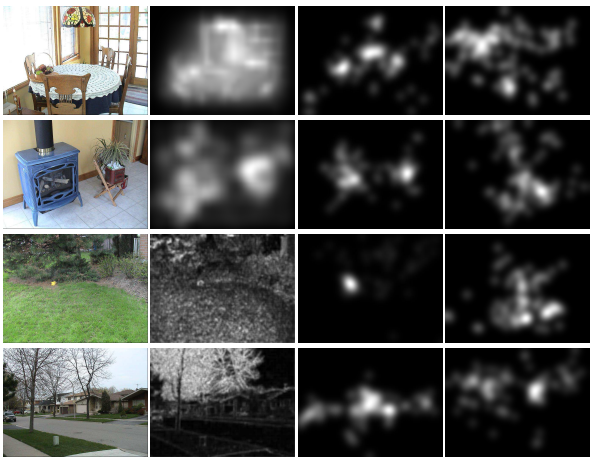


Figure 9. Samples which have a large score difference between S-AUC and FN-AUC. From top to bottom (S-AUC, FN-AUC), 1:(0.573, 0.762), 2:(0.665, 0.867), 3:(0.540, 0.379), 4:(0.445, 0.285). The third column is the ground-truth distribution (positive) and the fourth column is the negative map drawn by FN-AUC.

bias each method has. The first row, **CB**, indicates a Gaussian map can achieve decent scores on most of the metrics. The **CB** map achieves higher FN-AUC (0.607) than S-AUC (0.515) because S-AUC only penalizes center bias as discussed in Section 3.1 (all the methods can outperform random guess 0.5 on S-AUC). Moreover, Itti, AIM and SUN, all achieve lower performances on the bias-tolerant metrics, but our FN-AUC can still distinguish those methods from **CB**. From the study [45], the methods GBVS and AWS are the least and the most spatially consistent algorithms respectively. Therefore GBVS outperforms AWS on the bias-tolerant metrics, but AWS obtains higher scores on both of the bias-sensitive metrics, S-AUC and FN-AUC. We can also see this contrast between GBVS and other relatively consistent methods, SDSR, CAS and ImageSig. Moreover, our experiment shows the SWD method is even more spatially inconsistent than GBVS. When comparing with "early" vision, we are not surprised by the high performance achieved by the CNN due to its ability to learn high-level features. One thing that should be noted is although CNNs may tend to output center-favored maps due to its "objectness" knowledge, the CNN model still outperforms GBVS and SWD by a large margin. Both S-AUC and FN-AUC can penalize the spatial bias and we further investigate the difference between the two metrics in the next section.

4.3. Case Discussion

Our FN-AUC differs from S-AUC in that we also consider the relative relationship with the positive set. It is important to visually check the samples on which the metrics disagree with each other the most. We show samples which achieve a large score difference between S-AUC and FN-AUC in Figure 9. We can see from the scores, the top five rows achieve higher FN-AUC than S-AUC. From the third column, we can see that the ground-truth is near the center, but S-AUC will penalize the prediction regardless of whether it is a reasonable output or not. In contrast, our proposed FN-AUC achieves a higher score because the sampled negative is a more directional, rather than blind penalty. From the bottom four rows of Figure 9, we can see that the S-AUC score is higher than FN-AUC. We can see that the ground-truth is still near the center, but the predicted saliency region is near the periphery. In this case, those predictions should be considered as low-scoring outputs. But the negative set sampled by S-AUC will always be near the center as shown in Figure 4 so that it cannot penalize the FPs. FN-AUC has a higher probability to penalize this type of prediction because it takes the spatial relationship into account. (More examples are shown in Figure 2 in the supplementary material.)

5. Conclusion

In this paper, we have shown that the NSS and CC metrics still suffer from sensitivity to the choice of the σ value. This indicates that they can not fairly evaluate the CNN-based system on commonly used saliency datasets. NSS has been shown to be sensitive to the training set only, while CC is affected by the difference of σ applied on the training and the test sets. The AUC metrics are relatively more robust to the change of σ. We delved into the AUC metrics based on different mathematical representations to show the drawback of S-AUC. Our proposed FN-AUC metric considers the relative position information so that a more directional negative set can be built to penalize the center bias only. Finally, our proposed global smoothing strategy can deliver a more stable AUC computation by retaining the saliency relationship. By no means can our method completely solve the problem of saliency evaluation, but our work sheds new light on the drawbacks of existing metrics as well as introduces a new sampling process.

References

[1] C. O. Ancuti, C. Ancuti, and P. Bekaert. Enhancing by saliency-guided decolorization. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 257–264, Washington, DC, USA, 2011. IEEE Computer Society. 1

[2] Jonathan Boisvert and Neil Bruce. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing*, 207, 05 2016. 1

[3] Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *CoRR*, abs/1505.03581, 2015. 2

[4] Ali Borji, Dicky N. Sihite, and Laurent Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of Vision*, 13(10):18–18, 08 2013. 4, 7

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

[6] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *2013 IEEE International Conference on Computer Vision*, pages 921–928, Dec 2013. 1

[7] Neil D.B. Bruce and John K. Tsotsos. A statistical basis for visual field anisotropies. *Neurocomputing*, 69(10):1301 – 1304, 2006. Computational Neuroscience: Trends in Research 2006. 4

[8] Neil D.B. Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95 – 112, 2015. Computational Models of Visual Attention. 2, 3, 4, 5, 7

[9] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 155–162, Cambridge, MA, USA, 2005. MIT Press. 1, 2, 7

[10] Z. Bylinskii, E.M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J.K. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision Research*, 116:258 – 268, 2015. Computational Models of Visual Attention. 1

[11] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, March 2019. 1, 2, 3

[12] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. *CoRR*, abs/1609.01064, 2016. 1, 6

[13] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, Oct 2018. 1, 6

[14] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. In *CVPR 2011*, pages 473–480, June 2011. 7

[15] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 11 2008. 4

[16] Simone Frintrop, Erich Rome, and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7(1):6:1–6:39, Jan. 2010. 1

[17] A. Garcia-Diaz, X.R. Fdez-Vidal, X.M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012. 7

[18] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, Oct 2012. 7

[19] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 545–552, Cambridge, MA, USA, 2006. MIT Press. 7

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 6

[21] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201, Jan 2012. 4, 7

[22] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. *CoRR*, abs/1809.01125, 2018. 1

[23] 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*, 20(11):1254–1259, Nov 1998. 7

[24] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. *CoRR*, abs/1804.01793, 2018. 1

[25] Sen Jia and Neil D.B. Bruce. Eml-net:an expandable multi-layer network for saliency prediction. *Image and Vision Computing*, page 103887, 2020. 1, 6

[26] S. Jia, Y. Zhang, D. Agrafiotis, and D. Bull. Blind high dynamic range image quality assessment using deep learning. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 765–769, Sept 2017. 1

[27] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[28] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 01 2012. 1, 3

[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, Sep. 2009. 1, 2, 4

[30] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. 1

[31] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *CoRR*, abs/1411.1045, 2014. 1, 6

[32] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015. 1

[33] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 798–814, Cham, 2018. Springer International Publishing. 2

[34] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[35] Shengxi Li, Mai Xu, Yun Ren, and Zulin Wang. Closed-form optimization on saliency-guided image compression for hevc-msp. *Trans. Multi.*, 20(1):155–170, Jan. 2018. 1

[36] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107 – 123, 2002. 4

[37] Derrick Parkhurst and Ernst Niebur. Scene content selected by active vision. *Spatial vision*, 16:125–54, 02 2003. 4

[38] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005. 1

[39] 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 *2013 IEEE International Conference on Computer Vision*, pages 1153–1160, Dec 2013. 1, 2

[40] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, Jan 1998. 1

[41] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, 11 2009. 4, 7

[42] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643 – 659, 2005. 4

[43] Po-He Tseng, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7):4–4, 07 2009. 4

[44] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *CoRR*, abs/1704.06904, 2017. 1

[45] C. Wloka and J. Tstotsos. Spatially binned roc: A comprehensive saliency metric. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 525–534, June 2016. 1, 8

[46] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32, 12 2008. 2, 7

[47] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1266–1278, June 2016. 1

[48] F. Zünd, Y. Pritch, A. Sorkine-Hornung, S. Mangold, and T. Gross. Content-aware compression using saliency-driven image retargeting. In *2013 IEEE International Conference on Image Processing*, pages 1845–1849, Sep. 2013. 1