Structure-measure: A New Way to Evaluate Foreground Maps

Deng-Ping Fan[1] Ming-Ming Cheng[1] * Yun Liu[1] Tao Li[1] Ali Borji[2]
[1] CCCE, Nankai University [2] CRCV, UCF
http://dpfan.net/smeasure/

Abstract

*Foreground map evaluation is crucial for gauging the progress of object segmentation algorithms, in particular in the field of salient object detection where the purpose is to accurately detect and segment the most salient object in a scene. Several widely-used measures such as Area Under the Curve (AUC), Average Precision (AP) and the recently proposed F_β^ω(Fbw) have been used to evaluate the similarity between a non-binary saliency map (SM) and a ground-truth (GT) map. These measures are based on pixel-wise errors and often ignore the structural similarities. Behavioral vision studies, however, have shown that the human visual system is highly sensitive to structures in scenes. Here, we propose a novel, efficient, and easy to calculate measure known as structural similarity measure (**Structure-measure**) to evaluate non-binary foreground maps. Our new measure simultaneously evaluates region-aware and object-aware structural similarity between a SM and a GT map. We demonstrate superiority of our measure over existing ones using 5 meta-measures on 5 benchmark datasets.*

1. Introduction

The evaluation of a predicted foreground map against a ground-truth (GT) annotation map is crucial in evaluating and comparing various computer vision algorithm for applications such as object detection [6, 8, 24, 40], saliency prediction [5, 20, 42], image segmentation [41], content-based image retrieval [12, 15, 22], semantic segmentation [21,45,46] and image collection browsing [10,14,29]. As a specific example, here we focus on salient object detection models [4, 6, 7, 16], although the proposed measure is general and can be used for other purposes. It is necessary to point out that the salient object is not necessary to be foreground object [18].

The GT map is often binary (our assumption here). The foreground maps are either non-binary or binary. As a result, evaluation measures can be classified into two types. The first type is the binary map evaluation with the common measures being F_β-measure [2,13,33] and PASCAL's VOC segmentation measure [17]. The second type is the non-binary map evaluation. Two traditional measures here include AUC and AP [17]. A newly released measure known as Fbw [36] has been proposed to remedy flaws of AP and AUC measures (see Sec. 2). Almost all salient objection detection methods output non-binary maps. Therefore, in this work we focus on non-binary map evaluation.

It is often desired that the foreground map should contain the entire structure of the object. Thus, evaluation measures are expected to tell which model generates a more complete object. For example, in Fig. 1 (first row) the blue-border

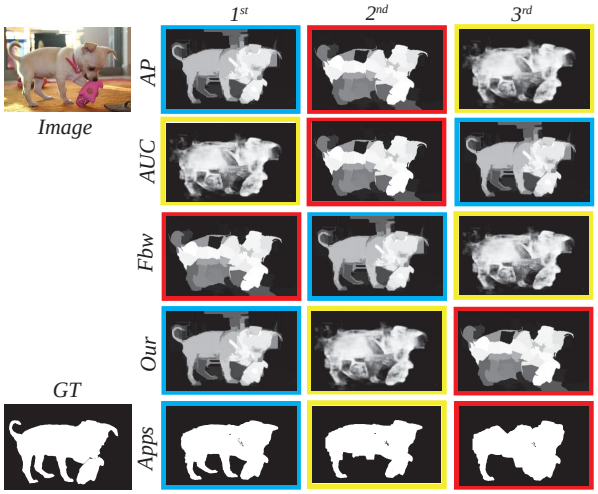


Figure 1. **Inaccuracy of existing evaluation measures.** We compare the ranking of saliency maps generated by 3 state-of-the-art salient object detection algorithms: DISC [11], MDF [27], and MC [48]. According to the application's ranking (last row; Sec. 5.1), the blue-border map ranks first, followed by the yellow- and red-border maps. The blue-border map captures the dog's structure most accurately, with respect to the GT. The yellow-border map looks fuzzy although the overall outline of the dog is still present. The red-border map almost completely destroyed the structure of the dog. Surprisingly, all of the measures based on pixel-wise errors (first 3 rows) fail to rank the maps correctly. Our new measure (4th row) ranks the three maps in the right order.

*M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

GT	0.9984	0.9915	0.9867	0.9504	0.9438	0.8951	0.8136	0.6605	0.6358	0.5127
	(a)rfcn [43]	(b)DISC [11]	(c)DHS [32]	(d)ELD [26]	(e)MC [48]	(f)DCL [28]	(g)DRFI [42]	(h)DSR [30]	(i)MDF [27]	(j)ST [34]

Figure 2. Structure-measure ($\lambda = 0.25, K = 4$) for the outputs of SalCut [13] algorithm (2nd row) when fed with inputs of 10 saliency detection algorithms (1st row).

map better captures the dog than the red-border map. In the latter case, shape of the dog is drastically degraded to a degree that it is difficult to guess the object category from its segmentation map. Surprisingly, all of the current evaluation measures fail to correctly rank these maps (in terms of preserving the structure).

We employed 10 state-of-the-art saliency detection models to obtain 10 saliency maps (Fig. 2; first row) and then fed these maps to the SalCut [13] algorithm to generate corresponding binary maps (2th row). Finally, we used our **Structure-measure** to rank these maps. A lower value for our measure corresponds to more destruction in the global structure of the man (columns e to j). This experiment clearly shows that our new measure emphasizes the global structure of the object. In these 10 binary maps (2rd row), there are 6 maps with Structure-measure below 0.95, *i.e.*, with percentage 60%. Using the same threshold (0.95), we found that the proportions of destroyed images in four popular saliency datasets (*i.e.*, ECSSD [47], HKU-IS [27], PASCAL-S [31], and SOD [37]) are 66.80%, 67.30%, 81.82% and 83.03%, respectively. Using the F_β measure to evaluate the binary maps, these proportions are 63.76%, 65.43%, 78.32% and 82.67%, respectively. This means that our measure is more restrictive than the F_β-measure on object structure.

To remedy the problem of existing measures (*i.e.*, low sensitivity to global object structure), we present a structural similarity measure (**Structure-measure**) [1] based on two observations:

- **Region** perspectives: Although it is difficult to describe the object structure of a foreground map, we notice that the entire structure of an object can be well illustrated by combining structures of constituent object-parts (regions).
- **Object** perspectives: In the high-quality foreground maps, the foreground region of the maps contrast sharply with the background regions and these regions usually have approximately uniform distributions.

Our proposed similarity measure can be divided into two parts, including a region-aware structural similarity measure and an object-aware structural similarity measure. The region-aware measure tries to capture the global structure information by combining the structural information of all the object-parts. The structural similarity of regions has been well explored in the image quality assessment (IQA) community. The object-aware similarity measure tries to compare global distributions of foreground and background regions in SM and GT maps.

We experimentally show that our new measure is more effective than other measures using 5 meta-measures (a new one introduced by us) on 5 publicly available benchmark datasets. In the next section, we review some of the popular evaluation measures.

2. Current Evaluation Measures

Saliency detection models often generate non-binary maps. Traditional evaluation measures usually convert these non-binary maps into multiple binary maps.

Evaluation of binary maps: To evaluate a binary map, four values are computed from the prediction confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These values are then used to compute three ratios: True Positive Rate (TPR) or Recall, False Positive Rate (FPR), and Precision. The Precision and Recall are combined to compute the traditional F_β-measure:

$$F_\beta = \frac{(1+\beta^2) Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (1)$$

Evaluation of non-binary maps: AUC and AP are two universally-agreed evaluation measures. Algorithms that produce non-binary maps apply three steps to evaluate the

[1] Source code and results for this measure on the entire datasets are available at the project page: http://dpfan.net/smeasure/.

agreement between model predictions (non-binary maps) and human annotations (GT). First, multiple thresholds are applied to the non-binary map to get multiple binary maps. Second, these binary maps are compared to the binary mask of the GT to get a set of TPR & FPR values. These values are plotted in a 2D plot, which then the AUC distils the area under the curve.

The AP measure is computed in a similar way. One can get a Precision & Recall curve by plotting Precision $p(r)$ as a function of Recall r. AP measure [17] is the average value of $p(r)$ over the evenly spaced x axis points from $r = 0$ to $r = 1$.

Recently, a measure called Fbw [36] has offered an intuitive generalization of the F_β-measure. It is defined as:

$$F_\beta^\omega = \frac{(1+\beta^2) Precision^\omega \cdot Recall^\omega}{\beta^2 \cdot Precision^\omega + Recall^\omega} \quad (2)$$

The authors of Fbw identified three causes of inaccurate evaluation of AP and AUC measures. To alleviate these flaws, they 1) extended the four basic quantities TP, TN, FP, and FN to non-binary values and, 2) assigned different weights (w) to different errors according to different location and neighborhood information. While this measure improves upon other measures, sometimes it fails to correctly rank the foreground maps (see the 3rd row of the Fig. 1). In the next section, we will analyze why the current measures fail to rank these maps correctly.

3. Limitations of Current Measures

Traditional measures (AP, AUC and Fbw) use four types of basic measures (FN, TN, FP and TP) to compute Precision, Recall and FPR. Since all of these measures are calculated in a pixel-wise manner, the resulting measures (FN, TN, FP and TP) cannot fully capture the structural information of predicted maps. Predicted maps with fine structural details are often desired in several applications. Therefore, evaluation measures sensitive to foreground structures are favored. Unfortunately, the aforementioned measures (AP, AUC and Fbw) fail to meet this expectation.

A typical example is illustrated in Fig. 3 (a) which contains two different types of foreground maps. In one, a black square falls inside the digit while in the other it touches the boundary. In our opinion, SM2 is favored over SM1 since the latter destroys the foreground maps more seriously. However, the current evaluation measures result in the same order. This is contradictory to our common sense.

A more realistic example is shown in Fig. 3 (b). The blue-border map here better captures the pyramid than the red-border map, because the latter offers a fuzzy detection map that mostly highlights the top part of the pyramid while ignoring the rest. From an application standpoint (3th row; the output of the SalCut algorithm fed with saliency maps

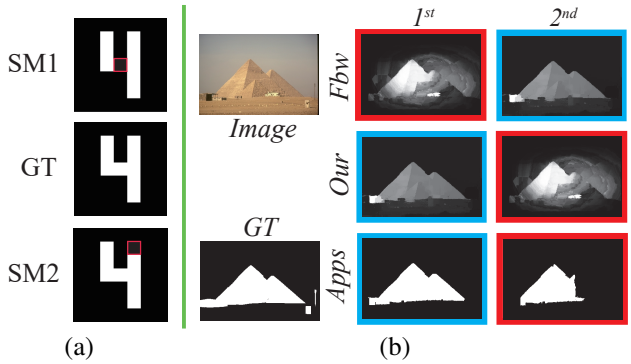


Figure 3. **Structural similarity evaluation**. In subfigure (a), two different foreground maps result in the same FN, TN, FP, and TP scores. In subfigure (b), two maps are produced by two saliency models DSR [30], and ST [34]. According to the application's ranking and our user-study (last row; Sec. 5.1), the blue-border map is the best, followed by the red-border map. Since Fbw measure does not account for the structural similarity, it results in a different ranking. Our measure (2th row) correctly ranks the blue-border map as higher.

and ranked by our measure, *i.e.*, the 2nd row), the blue-border map offers a complete shape of the pyramid. Thus, if the evaluation measure cannot capture the object structural information, it cannot provide reliable information for model selection in applications.

4. Our measure

In this section, we introduce our new measure to evaluate foreground maps. In image quality assessment (IQA) field, a measure known as structural similarity measure (SSIM) [44] has been widely used to capture the structural similarity of the original image and a test image.

Let $x = \{x_i | i = 1, 2, \cdots, N\}$ and $y = \{y_i | i = 1, 2, \cdots, N\}$ be the SM and GT pixel values, respectively. The $\bar{x}$, $\bar{y}$, σ_x, σ_x are the mean and standard deviations of x and y. σ_{xy} is the covariance between the two. Then, SSIM can be formulated as a product of three components: luminance comparison, contrast comparison and structure comparison.

$$ssim = \frac{2\bar{x}\bar{y}}{(\bar{x})^2 + (\bar{y})^2} \cdot \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \cdot \frac{\sigma_{xy}}{\sigma_x\sigma_y} \quad (3)$$

In Equ. (3), the first two terms denote the luminance comparison and contrast comparison, respectively. The closer the two (*i.e.*, $\bar{x}$ and $\bar{y}$, or σ_x and σ_y), the closer the comparison (*i.e.*, luminance or contrast) to 1. The structures of the objects in an image are independent of the luminance that is affected by illumination and the reflectance. So the design of a structure comparison formula should be independent of luminance and contrast. SSIM [44] associate two unit vectors $(x - \bar{x})/\sigma_x$ and $(y - \bar{y})/\sigma_y$ with the structure of the two images. Since the correlation between these two

vectors is equivalent to the correlation coefficient between x and y, the formula of structure comparison is denoted by the third term in Equ. (3).

In the field of salient object detection, researchers are concerned more about the foreground object structures. Thus, our proposed structure measure combines both region-aware and object-aware structural similarities. The region-aware structural similarity performs similar to [44], which aims to capture object-part structure information without any special concern about complete foreground. The object-aware structural similarity is designed to mainly capture the structure information of the complete foreground objects.

4.1. Region-aware structural similarity measure

In this section, we investigate how to measure region-aware similarity. The region-aware similarity is designed to assess the object-part structure similarity against the GT maps. We first divide each of the SM and GT maps into four blocks using a horizontal and a vertical cut-off lines that intersect at the centroid of the GT foreground. Then, the subimages are divided recursively like the paper [25]. The total number of blocks is denoted as K. A simple example is shown in Fig. 4. The region similarity $ssim(k)$ of each block is computed independently using Equ. (3). We assign a different weight (w_k) to each block proportional to the GT foreground region this block covers. Thus, the region-aware structural similarity measure can be formulated as

$$S_r = \sum_{k=1}^{K} w_k * ssim(k) \tag{4}$$

According to our investigation, our proposed S_r can well describe the object-part similarity between a SM and a GT map. We also tried to directly use SSIM to assess the similarity between SM and GT at the image level or in the sliding window fashion as mentioned in [44]. These approaches fail to capture region-aware structure similarities.

4.2. Object-aware structural similarity measure

Dividing the saliency map into blocks helps evaluate the object-part structural similarity. However, the region-aware measure (S_r) cannot well account for the global similarity. For high-level vision tasks such as salient object detection, the evaluation of the object-level similarity is crucial. To achieve this goal, we propose a novel method to assess the foreground and background separately. Since, the GT maps usually have important characteristics, including sharp foreground-background contrast and uniform distribution, the predicted SM is expected to possess these properties. This helps easily distinguish foreground from the background. We design our object-aware structural similarity measure with respect to these two characteristics.

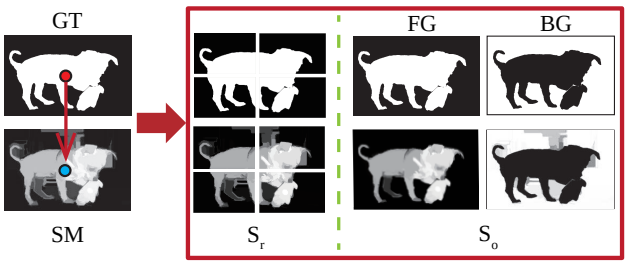


Figure 4. **Framework of our Structure-measure.**

Sharp foreground-background contrast. The foreground region of the GT map contrasts sharply with the background region. We employ a formulation that is similar with the luminance component of SSIM, to measure how close the mean probability is between the foreground region of SM and the foreground region of GT. Let x_{FG} and y_{FG} represent the probability values of foreground region of SM and GT, respectively. $\bar{x}_{FG}$ and $\bar{y}_{FG}$ denote the means of x_{FG} and y_{FG}, respectively. The foreground comparison can be represented as,

$$O_{FG} = \frac{2\bar{x}_{FG}\bar{y}_{FG}}{(\bar{x}_{FG})^2 + (\bar{y}_{FG})^2}. \tag{5}$$

Equ. (5) has several satisfactory properties:

- Swapping the value of $\bar{x}_{FG}$ and $\bar{y}_{FG}$, O_{FG} will not change the result.
- The range of O_{FG} is [0,1].
- If and only if $\bar{x}_{FG} = \bar{y}_{FG}$, we will get $O_{FG} = 1$.
- The most important property, however, is that the closer the two maps, the closer the O_{FG} to 1.

These properties make Equ. (5) suitable for our purpose.

Uniform saliency distribution. The foreground and background regions of the GT maps usually have uniform distributions. So, it is important to assign a higher value to a SM with salient object being uniformly detected (*i.e.*, similar saliency values across the entire object). If the variability of the foreground values in the SM is high, then the distribution will not be even.

In probability theory and statistics, the coefficient of variation which is defined as the ratio of the standard deviation to the mean ($\sigma_x/\bar{x}$) is a standardized measure of dispersion of a probability distribution. Here, we use it to represent the dispersion of the SM. In other words, we can use the coefficient of variation to compute the distribution of dissimilarity between SM and GT. According to Equ. (5), the total dissimilarity between SM and GT in object level can be written as,

$$D_{FG} = \frac{(\bar{x}_{FG})^2 + (\bar{y}_{FG})^2}{2\bar{x}_{FG}\bar{y}_{FG}} + \lambda * \frac{\sigma_{x_{FG}}}{\bar{x}_{FG}} \tag{6}$$

where λ is a constant to balance the two terms. Since the mean probability of the GT foreground is exactly 1 in practice, the similarity between SM and GT in object level can be formulated as,

$$O_{FG} = \frac{1}{D_{FG}} = \frac{2\bar{x}_{FG}}{(\bar{x}_{FG})^2 + 1 + 2\lambda * \sigma_{x_{FG}}} \quad (7)$$

To compute background comparison O_{BG}, we regard the background as the complementary component of foreground by subtracting the SM and GT maps from 1 as shown in Fig. 4. Then, O_{BG} can be similarly defined as,

$$O_{BG} = \frac{2\bar{x}_{BG}}{(\bar{x}_{BG})^2 + 1 + 2\lambda * \sigma_{x_{BG}}} \quad (8)$$

Let μ be the ratio of foreground area in GT to image area ($width * height$). The final object-aware structural similarity measure is defined as,

$$S_o = \mu * O_{FG} + (1 - \mu) * O_{BG} \quad (9)$$

4.3. Our new structure-measure

Having region-aware and object-aware structural similarity evaluation definitions, we can formulate the final measure as,

$$S = \alpha * S_o + (1 - \alpha) * S_r, \quad (10)$$

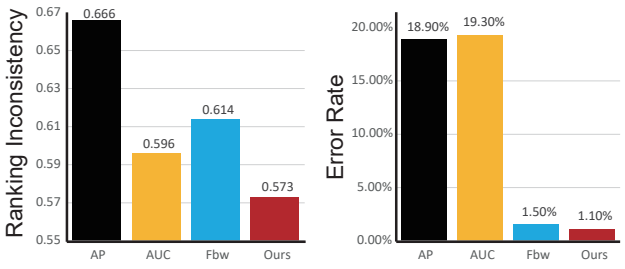
where $\alpha \in [0, 1]$. We set $\alpha = 0.5$ in our implementation. Using this measure to evaluate the three SM maps in Fig. 1, we can correctly rank the maps consistent with the application rank.

5. Experiments

In order to test the quality of our measure, we utilized 4 meta-measures proposed by Margolin *et al*. [36] and 1 meta-measures proposed by us. These meta-measures are used to evaluate the quality of evaluation measures [39]. To conduct fair comparisons, all meta-measures are computed on the ASD (a.k.a ASD1000) dataset [1]. The non-binary foreground maps (5000 maps in total) were generated using five saliency detection models including CA [19], CB [23], RC [13], PCA [35], and SVO [9]. We assign $\lambda = 0.5$ and $K = 4$ in all experiments. When using a single CPU thread (4 GHz), our Matlab implementation averagely takes 5.3 ms to calculate the structure measure of an image.

5.1. Meta-Measure 1: Application Ranking

An evaluation measure should be consistent with the preferences of an application that uses the SM as input. We assume that the GT map is the best for the applications. Given a SM, we compare the application's output to that of the GT output. The more similar a SM is to the GT map, the closer its application's output should be to the GT output.



(a) Meta-Measure 1 (b) Meta-Measure 2

Figure 5. **Meta-measure 1&2-results.**

To quantify the accuracy in ranking, we use the SalCut [13] as the application to perform this meta-measure.

Here, we utilize 1-Spearman's ρ measure [3] to evaluate the ranking accuracy of the measures, where lower values indicates better ranking consistency. Comparison between different measurements are shown in Fig. 5 (a), which indicates that our structure measure produces best ranking consistency among other alternative methods.

5.2. Meta-Measure 2: State-of-the-art vs. Generic

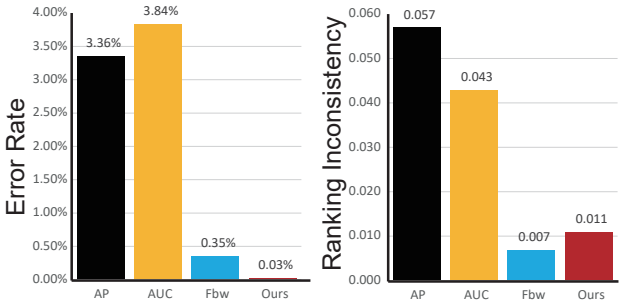
The second meta-measure is that a measure should prefer the output achieved by a state-of-the-art method over generic baseline maps (*e.g*., centered Gaussian map) that discard the image content. A good evaluation measure should rank the SM generated by a state-of-the-art model higher than a generic map.

We count the number of times a generic map scored higher than the mean score generated by the five state-of-the-art models (CA [19], CB [23], RC [13], PCA [35], SVO [9]). The mean score provides an indication of model robustness. The results are shown in Fig. 5 (b). The lower the value here, the better. Over 1000 images, our measure has only 11 errors (*i.e*., generic winning over the s.t.a). Meanwhile, the AP and AUC measures are very poor and make significantly more mistakes.

5.3. Meta-Measure 3: Ground-truth Switch

The third meta-measure specifies that a good SM should not obtain a higher score when switching to a wrong GT map. In Margolin *et al*. [36], a SM is considered as "good" when it scores at least 0.5 out of 1 (when compared to the original GT map). Using this threshold (0.5), top 41.8% of the total 5000 maps were selected as "good" ones. For a fair comparison, we follow Margolin *et al*. to select the same percentage of "good" maps. For each of the 1000 images, 100 random GT switches were tested. We then counted the percentage of times that a measure increases a saliency map's score when an incorrect GT map was used.

The Fig. 6 (a) shows the results. The lower the score, the higher capability to match to the correct GT. Our measure performs the best about 10 times better compared to the second best measure. This is due to the fact that our



(a) Meta-Measure 3 (b) Meta-Measure 4

Figure 6. **Meta-measure 3&4-results.**

(a) GT (b) GT (c) Dif (d) SM1 (e) SM2

Figure 7. **Meta-measure 4: Annotation errors.** (a) ground-truth map, (b) morphologically changed version of a, (c) difference map between a and b, (d) saliency map1, (e) saliency map2.

measure captures the object structural similarity between a SM and a GT map. Our measure will assign a lower value to the "good" SM when using a random selected GT since the object structure has changed in the random GT.

5.4. Meta-Measure 4: Annotation errors

The fourth meta-measure specifies that an evaluation measure should not be sensitive to slight errors/inaccuracies in the manual annotation of the GT boundaries. To perform this meta-measure, we make a slightly modified GT map by using morphological operations. An example is shown in Fig. 7. While the two GT maps in (a) & (b) are almost identical, measures should not switch the ranking between the two saliency maps when using (a) or (b).

We use 1-Spearman's Rho measure to examine the ranking correlation before and after the annotation errors were introduced. The lower the score, the more robust an evaluation measure is to annotation errors [36]. The results are shown in Fig. 6 (b). Our measure outperforms both the AP and the AUC but not the best. Inspecting this finding, we realized that it is not always the case that the lower the score, the better an evaluation measure. The reason is that sometimes "slight" inaccurate manual annotations can change the structure of the GT map, which in turn can change the rank. We examined the effect of structure change carefully. Major structure change often corresponds to continuous large regions in the difference map between GT and its morphologically changed version. We try to use the sum of corroded version of the difference map as measure of major structure change and sort all GT maps.

Among top 10% least change GT maps, our measure and Fbw have the same MM4 scores (same rank). When the topology of GT map does not change, our measure and Fbw measure keep the original ranking. We can see from the example Fig. 9 (a). While ground truth maps (GT and Morphologic GT) differ slightly, both Fbw and our measure keep the ranking order of the two saliency maps, depending on the GT used.

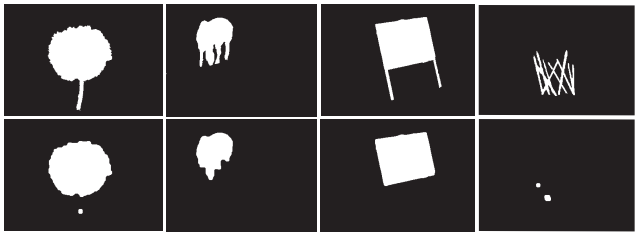
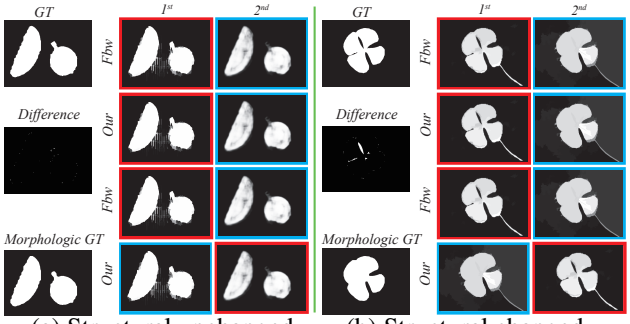


Figure 8. **Structural changes examples.** The first row are GT maps. The second row are its morphologically changed version.



(a) Structural unchanged (b) Structural changed

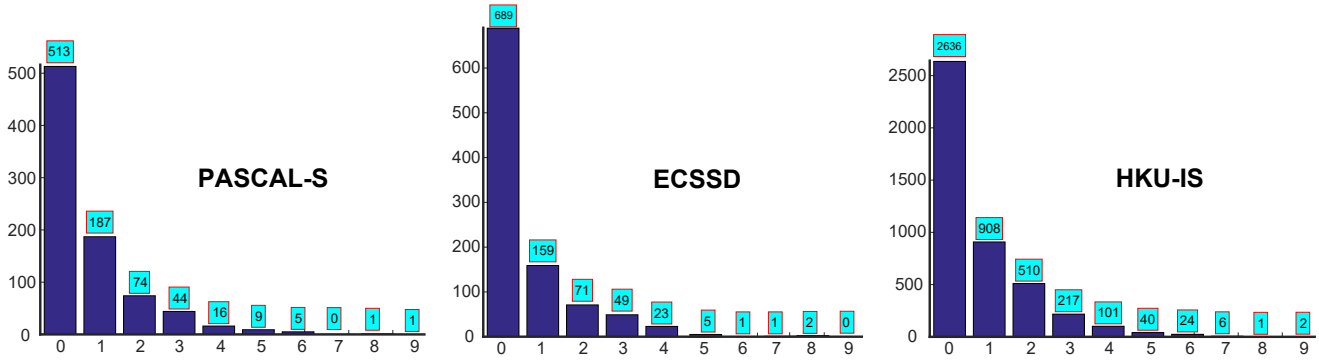
Figure 9. **Structural unchanged/changed.** (a) Both of our and Fbw measures are not sensitive to a inaccuracies (structural unchanged) in the manual annotation of the GT boundaries. (b) The ranking of an evaluation measure should be sensitive to the structural changes. Surprisingly, the current best measure (Fbw) cannot adaptive to the structural changes. Using our measure, we can change the rank correctly. Best viewed on screen.

Among top 10% most changes GT maps, we asked 3 users to judge whether the GT maps have major structure change. 95 out of 100 GT maps were considered to have major structure change, (similar to Fig. 8, such as small bar, thins legs, slender foot and minute lines in each group), for which we believe that keeping rank stability is not good. Fig. 9 (b) demonstrates this argument. When we use the GT map as the reference, Fbw and our measure rank the two maps properly. However, when using Morphologic GT as the reference, ranking results are different. Clearly, the blue-border SM is visually and structurally more similar to the Morphologic GT map than the red-border SM. The measure should rank the blue-border SM higher than red-border SM. So the ranking of these two maps should be changed. While the Fbw measure fails to meet this end, our measure gives the correct order.

Above-mentioned analysis suggests that this meta-measure is not very reliable. Therefore, we do not include it in our further comparison on other datasets.

Table 1. Quantitative comparison with current measures on 3 meta-Measures. The best result is highlighted in **blue.** MM:meta-Measure.

	PASCAL-S [31]			ECSSD [47]			SOD [37]			HKU-IS [27]		
	MM1	MM2(%)	MM3(%)	MM1	MM2(%)	MM3(%)	MM1	MM2(%)	MM3(%)	MM1	MM2(%)	MM3(%)
AP	0.452	12.1	5.50	0.449	9.70	3.32	0.504	9.67	7.69	0.518	3.76	1.25
AUC	0.449	15.8	8.21	0.436	12.1	4.18	0.547	14.0	8.27	0.519	7.02	2.12
Fbw	0.365	7.06	1.05	0.401	**3.00**	0.84	0.384	16.3	0.73	0.498	0.36	0.26
Ours	**0.320**	**4.59**	**0.34**	**0.312**	3.30	**0.47**	**0.349**	**9.67**	**0.60**	**0.424**	**0.34**	**0.08**



(a) (b) (c)

Figure 10. **The rank distance between Fbw and our measure.** The (a)-(c) is the three datasets that present the rank distance between Fbw and our **Structure-measure**. The y axis of the plot is the number of the images. The x axis is the rank distance.

5.5. Further comparison

The results in Fig. 5 & Fig. 6 (a) show that our measure achieves the best performance using 3 meta-measures over the ASD1000 dataset. However, a good evaluation measure should perform well over almost all datasets. To demonstrate the robustness of our measure, we further performed experiments on four widely-used benchmark datasets.

Datasets. The used datasets include PASCAL-S [31], ECSSD [47], HKU-IS [27], and SOD [37]. PASCAL-S contains 850 challenging images, which have multiple objects with high background clutter. ECSSD contains 1000 semantically meaningful but structurally complex images. HKU-IS is another large dataset that contains 4445 large-scales images. Most of the images in this dataset contain more than one salient object with low contrast. Finally, we also evaluate our measure over SOD dataset, which is a subset of the BSDS dataset. It contains a relatively small number of images (300), but with multiple complex objects.

Saliency Models. We use 10 state-of-the-art models including 3 traditional models (ST [34], DRFI [42], and DSR [30]) and 7 deep learning based models (DCL [28], rfcn [43], MC [48], MDF [27], DISC [11], DHS [32], and ELD [26]) to test the measures.

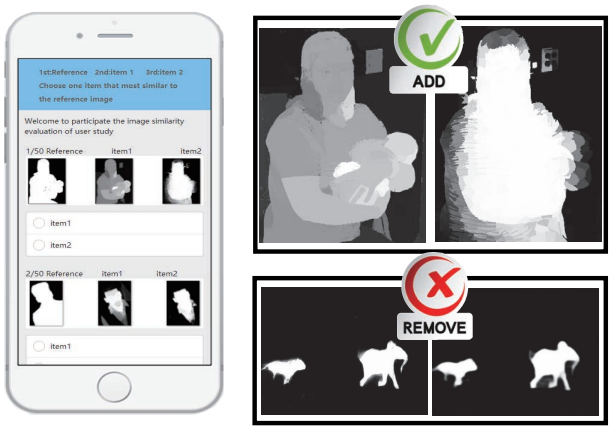
Results. Results are shown in Tab. 1. Our measure performs the best according to the first meta-measure. This indicates that our measure is more useful for applications than others. According to meta-measure 2, our measure performs better than the existing measures, except that ECSSD where it is ranked second. For meta-measure 3, our measure reduces the error rate by 67.62%, 44.05%, 17.81%, 69.23% in PASCAL, ECSSD, SOD and HKU-IS, respectively compared to the second ranked measure. This indicates that our measure has higher capacity to measure the structural similarity between a SM and a GT map. All in all, our measure wins in the majority of cases which clearly demonstrates that our new measure is more robust than other measures.

5.6. Meta-Measure 5: Human judgments

Here, we propose a new meta-measure to evaluate foreground evaluation measures. This meta-measure specifies that the map ranking according to an evaluation measure should agree with the human ranking. It is argued that "a human being is the best judge to evaluate the output of any segmentation algorithm" [38]. However, subjective evaluation over all images of a dataset is impractical due to time and monetary costs. To the best of our knowledge, there is no such visual similarity evaluation database available that meets these requirements.

Source saliency maps collection. The source saliency maps are sampled from the three large scale datasets: PASCAL-S, ECSSD, and HKU-IS. As mentioned above, we use 10 state-of-the-art saliency models to generate the saliency maps in each dataset. Therefore, we have 10 saliency maps for each image. We use Fbw and our measure to evaluate the 10 maps and then pick the first ranked map according to each measure. If the two measures choose the same map, their rank distance is 0. If one measure ranks a map first, but the other ranks the same map in the n-th place, then their rank distance is $|n-1|$. Fig. 10 (a), (b) and (c) show the rank distance between the two measures (*i.e.*, histogram). The blue-box is the number of images for each rank distance. Some maps with rank distance greater than 0 are chosen as candidates for our user study.

User study. We randomly selected 100 pairs of maps



(a) (b)

Figure 11. **Our user study platform**.

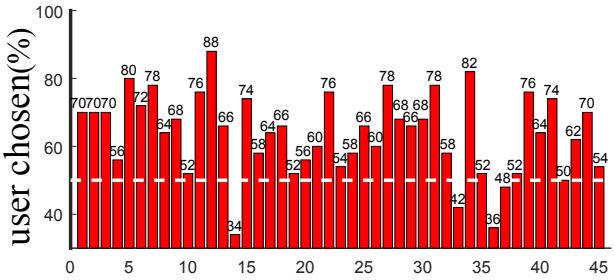


Figure 12. **Results of our user study**. The x axis is the viewer id. The y axis shows the percentage of the trials in which a viewer preferred the map chosen by our measure.

from the three datasets. The top panel in Fig. 11 (b) shows one example trial where the best map according to our measure in the left, and the best map according to the Fbw on the far right. The user is asked to choose the map he/she thinks resembles the most with the GT map. In this example, these two maps are obviously different making the user decide easily. In another example (bottom panel in Fig. 11 (b)), the two maps are too similar making it difficult to choose the one closet to the GT. Therefore, we avoid showing such cases to the subjects. Finally, we are left with a stimulus set of size 50 pairs. We developed a mobile phone app to conduct the user study. We collected data from 45 viewers who were naive to the purpose of the experiment. Viewers had normal or corrected vision. (Age distribution is 19-29 years old; Eduction from undergraduate to Ph.D; 10 different major such as history, medicine and finance; 25 males and 20 females)

Results. Results are shown in Fig. 12. The percentage of trials (averaged over subjects) in which a viewer preferred the map chosen by our measure is 63.69%. We used the same way to do another 2 user study experiments (AP compare to our measure, AUC compare to our measure). The results are 72.11% and 73.56% respectively, which means

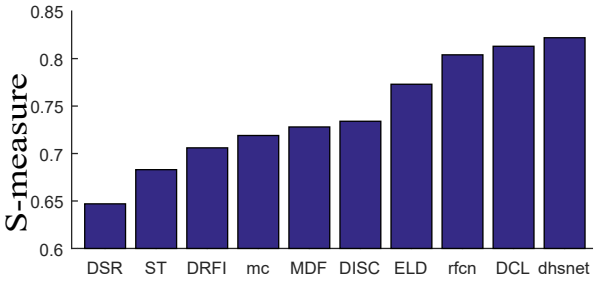


Figure 13. **Ranking of 10 saliency models using our new measure.** The y axis shows the average score on each dataset (PASCAL-S [31], ECSSD [47], HKU-IS [27], SOD [37]).

that our measure correlates better with human judgments.

5.7. Saliency model comparison

Establishing that our Structure-measure offers a better way to evaluate salient object detection models, here we compare 10 state-of-the-art saliency models on 4 datasets (PASCAL-S, ECSSD, HKU-IS, and SOD). Fig. 13 shows the rank of 10 models. According to our measure, the best models in order are dhsnet, DCL and rfcn. Please see the supplementary material for sample maps of these models.

6. Discussion and Conclusion

In this paper, we analyzed the current saliency evaluation measures based on pixel-wise errors and showed that they ignore the structural similarities. We then presented a new structural similarity measure known as **Structure-measure** which simultaneously evaluates region-aware and object-aware structural similarities between a saliency map and a ground-truth map. Our measure is based on two important characteristics: 1) sharp foreground-background contrast, and 2) uniform saliency distribution. Further, the proposed measure is efficient and easy to calculate.

Experimental results on 5 datasets demonstrate that our measure performs better than the current measures including AP, AUC, and Fbw. Finally, we conducted a behavioral judgment study over a database of 100 saliency maps and 50 GT maps. Data from 45 subjects shows that on average they preferred the saliency maps chosen by our measure over the saliency maps chosen by the AP, AUC and Fbw.

In summary, our measure offers new insights into salient object detection evaluation where current measures fail to truly examine the strengths and weaknesses of saliency models. We encourage the saliency community to consider this measure in future model evaluations and comparisons.

Acknowledgement We would like to thank anonymous reviewers for their helpful comments on the paper. This research was supported by NSFC (NO. 61572264, 61620106008), Huawei Innovation Research Program, CAST YESS Program, and IBM Global SUR award.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2011.

[3] D. Best and D. Roberts. Algorithm as 89: the upper tail probabilities of spearman's rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):377–379, 1975.

[4] A. Borji. What is a salient object? a dataset and a baseline model for salient object detection. *IEEE TIP*, 24(2):742–756, 2015.

[5] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*, 2014.

[6] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.

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

[8] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark (2015). 2015.

[9] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *IEEE ICCV*, pages 914–921, 2011.

[10] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: internet image montage. *ACM TOG*, 28(5):124, 2009.

[11] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li. Disc: Deep image saliency computing via progressive representation learning. *IEEE transactions on neural networks and learning systems*, 27(6):1135–1149, 2016.

[12] T. Chen, P. Tan, L.-Q. Ma, M.-M. Cheng, A. Shamir, and S.-M. Hu. Poseshop: Human image database construction and personalized content synthesis. *Visualization and Computer Graphics, IEEE Transactions on*, 19(5):824–837, 2013.

[13] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[14] M.-M. Cheng, Q.-B. Hou, S.-H. Zhang, and P. L. Rosin. Intelligent visual media processing: When graphics meets vision. *Journal of Computer Science and Technology*, 32(1):110–121, 2017.

[15] M.-M. Cheng, N. Mitra, X. Huang, and S.-M. Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014.

[16] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *IEEE ICCV*, pages 1529–1536, 2013.

[17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[18] D. Feng, N. Barnes, S. You, and C. McCarthy. Local background enclosure for rgb-d salient object detection. In *IEEE CVPR*, pages 2343–2350, 2016.

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

[20] Q. Hou, M.-M. Cheng, X. Hu, Z. Tu, and A. Borji. Deeply supervised salient object detection with short connections. In *IEEE CVPR*, 2017.

[21] S.-M. Hu, T. Chen, K. Xu, M.-M. Cheng, and R. Martin. Internet visual media processing: a survey with graphics and vision applications. *The Visual Computer*, 29(5):393–405, 2013.

[22] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang. Joint Salient Object Detection and Existence Prediction. *Front. Comput. Sci.*, 2017.

[23] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, 2011.

[24] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *IEEE CVPR*, pages 2472–2479, 2010.

[25] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE CVPR*, pages 2169–2178, 2006.

[26] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *IEEE CVPR*, pages 660–668, 2016.

[27] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE CVPR*, pages 5455–5463, 2015.

[28] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *IEEE CVPR*, pages 478–487, 2016.

[29] L. Li, S. Jiang, Z.-J. Zha, Z. Wu, and Q. Huang. Partial-duplicate image retrieval via saliency-guided visual matching. *MultiMedia, IEEE*, 20(3):13–23, 2013.

[30] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *IEEE ICCV*, pages 2976–2983, 2013.

[31] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *IEEE CVPR*, pages 280–287, 2014.

[32] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *IEEE CVPR*, pages 678–686, 2016.

[33] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011.

[34] Z. Liu, W. Zou, and O. Le Meur. Saliency tree: A novel saliency detection framework. *IEEE TIP*, 23(5):1937–1952, 2014.

[35] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *IEEE CVPR*, pages 1139–1146, 2013.

[36] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *IEEE CVPR*, pages 248–255, 2014.

[37] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE ICCV*, 2001.

[38] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern recognition*, 26(9):1277–1294, 1993.

[39] J. Pont-Tuset and F. Marques. Measures and meta-measures for the supervised evaluation of image segmentation. In *IEEE CVPR*, pages 2131–2138, 2013.

[40] W. Qi, M.-M. Cheng, A. Borji, H. Lu, and L.-F. Bai. Saliencyrank: Two-stage manifold ranking for salient object detection. *Computational Visual Media*, 1(4):309–320, 2015.

[41] C. Qin, G. Zhang, Y. Zhou, W. Tao, and Z. Cao. Integration of the saliency-based seed extraction and random walks for image segmentation. *Neurocomputing*, 129:378–391, 2014.

[42] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 123(2):251–268, 2017.

[43] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841, 2016.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[45] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, 2017.

[46] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 2016.

[47] Y. Xie, H. Lu, and M.-H. Yang. Bayesian saliency via low and mid level cues. *IEEE TIP*, 22(5):1689–1698, 2013.

[48] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *IEEE CVPR*, pages 1265–1274, 2015.