International Journal of Computer Vision (2021) 129:2622–2638
https://doi.org/10.1007/s11263-021-01490-8

Structure-Measure: A New Way to Evaluate Foreground Maps

Ming-Ming Cheng[1] · Deng-Ping Fan[1]

Received: 30 October 2020 / Accepted: 4 June 2021 / Published online: 27 June 2021

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 measures (e.g., area-under-the-curve, F1-measure, average precision, etc.) have been used to evaluate the similarity between a foreground map and a ground-truth map. The existing 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 (0.005 s per image), and easy to calculate measure known as **S-measure** (structural measure) to evaluate foreground maps. Our new measure simultaneously evaluates region-aware and object-aware structural similarity between a foreground map and a ground-truth map. We demonstrate superiority of our measure over existing ones using 4 meta-measures on 5 widely-used benchmark datasets. Furthermore, we conduct a behavioral judgment study over a new database. Data from 45 subjects shows that on average they preferred the saliency maps chosen by our measure over the saliency maps chosen by the state-of-the-art measures. Our experimental results offer new insights into foreground map evaluation where current measures fail to truly examine the strengths and weaknesses of models. Code: https://github.com/DengPingFan/S-measure.

Keywords S-measure · Structure measure · Foreground maps · Evaluation · Salient object detection

1 Introduction

The evaluation of a predicted foreground map (FM) against a ground-truth (GT) annotation map is crucial in evaluating and comparing various computer vision algorithm for applications such as **AR COPY & PASTE** (Qin et al., 2021), object detection and recognition (Borji et al., 2013a, 2015; Kanan & Cottrell, 2010; Rutishauser et al., 2004; Islam et al., 2018), video summarization (Ghosh et al., 2012), video compression (Guo & Zhang, 2010; Itti, 2004), image segmentation (Yu et al., 2018), content-based image retrieval (Li et al., 2013a; Liu & Fan, 2013; Liu et al., 2015), visual tracking, photo synthesis (Chen et al., 2009), image-text matching (Zhuge et al., 2021b), image collection browsing (Chen et al., 2009), etc. As a specific example, here we focus on salient object detection (Borji et al., 2015; Borji & Itti, 2013; Borji, 2015; Bylinskii et al., 2015; Zhang et al., 2017, 2018a, 2018b, 2018c; Wang et al., 2018; Chen & Li, 2018; Gorji & Clark, 2018; Li et al., 2018; Zeng et al., 2018; Liu et al., 2018; Tiantian et al., 2018), although the proposed measure is general and can be used for other purposes. It is worth noting that the salient object does not necessarily need to be the foreground object (Feng et al., 2016; Borji et al., 2013b).[1]

The GT map is often the average of several manual annotations. Thus, it can be binary or non-binary. Similarly, the predicted foreground maps are either binary or non-binary. As a result, evaluation measures can be classified into two types:

1. **Binary map evaluation**: Common measures here include F_β-measure (Arbelaez et al., 2011; Cheng et al., 2015; Liu et al., 2011) and PASCAL's VOC segmentation measure (Everingham et al., 2010).

Communicated by O. Veksler.

A preliminary version of this work appeared at ICCV2017 (Fan et al., 2017).

✉ Ming-Ming Cheng
cmm@nankai.edu.cn

Deng-Ping Fan
dengpfan@gmail.com

[1] College of Computer Science, Nankai University, Tianjin, China

[1] https://clipdrop.co/.

2. **Non-binary map evaluation**: Three traditional and popular measures here include area under the curve (AUC), precision-recall (PR) curve, and average precision (AP) (Everingham et al., 2010). A newly released measure known as weighted F$_\beta$-measure (Fbw) (Margolin et al., 2014) has been proposed to remedy flaws of AUC, PR and AP.

It is often desired to have the foreground map contain the entire structure of an 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 map better captures the dog than the red-border map. In the latter case, shape of the 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 (SOTA) saliency detection algorithms to obtain 10 saliency maps (Fig. 2; 1st row) and then fed these maps to the SalCut[2] (Cheng et al., 2015) algorithm to generate corresponding binary maps (2nd row). Finally, we used the proposed **S-measure** to rank these maps. A lower score 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 entire structure of the object. In these ten binary maps (2nd row), there are six maps with score 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 (Xie et al., 2013), HKU-IS (Li & Yu, 2015), PASCAL-S (Li et al., 2014), and SOD (Martin et al., 2001)] 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 stringent than the F_β-measure on object structure.

To remedy the problem of existing measures (i.e., low sensitivity to entire object structure), we present a structure-sensitive similarity measure based on two observations:

1. **Region** perspectives: Although it is difficult to describe the 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).
2. **Object** perspectives: In high-quality foreground maps, the foreground regions contrast sharply with the background regions. These regions usually have approximately uniform contrast distributions.

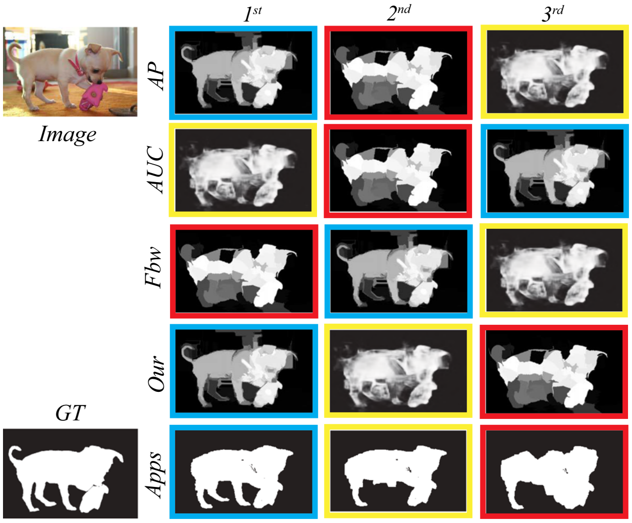


Fig. 1 Inaccuracy of existing evaluation measures. We compare the ranking of saliency maps generated by three state-of-the-art (SOTA) salient object detection algorithms: DISC (Chen et al., 2016), MDF (Li & Yu, 2015), and MC (Zhao et al., 2015). According to the segmentation application's ranking (last row; see details in Apps-Sec. 5), 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 three rows) fail to rank the maps correctly. Our new measure (4th row) ranks the three maps in the right order (Color figure online)

Consequently, the proposed structure measure consists of 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 object structure 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 (Wang et al., 2004). The object-aware similarity measure tries to compare global distributions of foreground and background regions in the foreground map and the GT map.

Our measure is compared against various existing measures including AP, AUC, PASCAL, Fbw, F$_\beta$-measure on several widely-used salient object detection benchmarks including ASD (Achanta et al., 2009), SOD (Martin et al., 2001), ECSSD (Xie et al., 2013), PASCAL-S (Li et al., 2014), and HKU-IS (Li & Yu, 2015). Extensive empirical investigations show that Structure-measure not only provides a reliable evaluation but also achieves significantly improved performance than current measures.

This work is an extension of our previous ICCV2017 version (Fan et al., 2017). The major differences between these two versions include: (1) We extend the preliminary version to binary foreground map evaluation and provide a unified evaluation applicable to both binary and non-binary fore-

[2] https://github.com/MCG-NKU/SalBenchmark/blob/master/Code/CmLib/Saliency/CmSalCut.cpp.

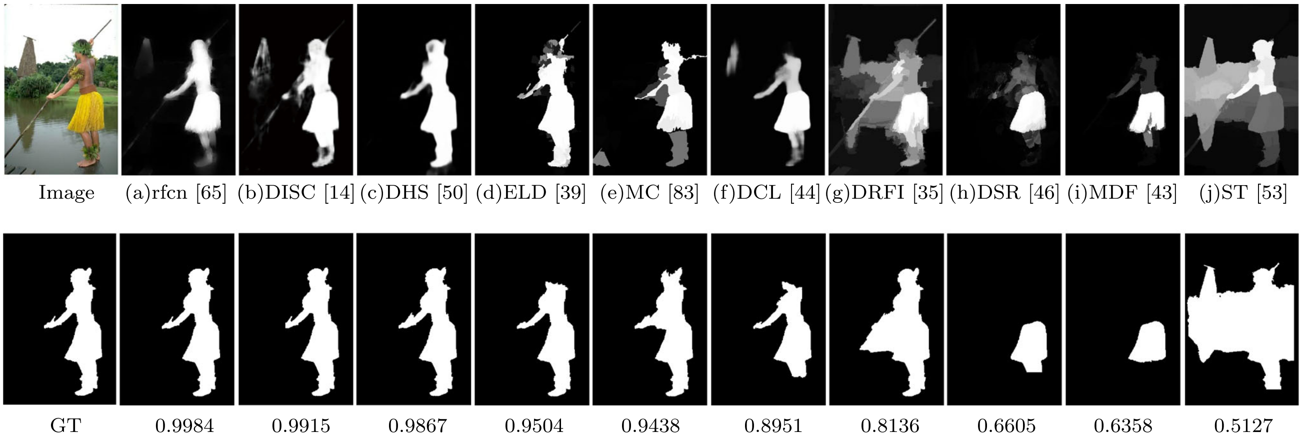


Fig. 2 S-measure score ($\lambda = 0.5$, $K = 4$) for the outputs of SalCut (Cheng et al., 2015) algorithm (2nd row) when fed with inputs of 10 saliency maps (1st row). The ranking results clearly indicate that our measure is good at capturing object structures and can provide a reliable ranking

ground maps. Our work offers new insights into foreground maps evaluation where current measures fail to examine the strengths and weaknesses of models fully. (2) We provide a set of new experiment to validate the efficiency, robustness, and extensibility of our proposed measure. These extension focus on non-binary maps' evaluation. Besides, we also give more details about our application-ranking framework in Appendices. (3) We build the several representative online Benchmark and model zoo of saliency detection, which integrates various publicly available saliency datasets with uniform input/output formats (i.e., JPEG for image; PNG for GT). (4) We also provide Python[3] and Matlab[4] version code for existing benchmarking work which benefits many related tasks and our computer vision community. (5) We have made a lot of efforts to improve the presentations and organizations of our paper. First, several new figures are added or re-produced to better illustrate the meta-measure and key results of this work. Second, we have added several new sections to describe more details about the flaws of current measures, provide more theoretical details about our S-measure, and more theoretical details include region-aware (Sect. 4.1) and object-aware (Sect. 4.2) structure similarity evaluation. The new content will allow readers to better understand our approach.

2 Current Evaluation Measures

Foreground maps can be generated by various algorithms (e.g., for saliency detection or object segmentation). Saliency detection algorithms often generate non-binary maps, whereas object segmentation algorithms usually generate binary maps. As a result, the foreground maps can be divided as *non-binary* maps with values in the range [0, 1] or *binary* maps with values either 0 or 1. Each map value denotes the probability of a specific pixel belonging to the foreground (Peng et al., 2014; Margolin et al., 2014).

2.1 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:

$$Recall = TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{TN + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

The Precision and Recall are combined to compute the traditional F_β-measure.

$$\begin{aligned} F_\beta &= \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall} \\ &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP} \end{aligned} \tag{4}$$

where β is a parameter to balance the accuracy and the recall (typically $\beta = 1$ leading to harmonic mean). Another commonly used binary map evaluation metric is the PASCAL measure:

$$PASCAL = \frac{TP}{TP + FN + FP} \tag{5}$$

[3] https://github.com/DengPingFan/eval-co-sod.

[4] https://dpfan.net/One-KeyEvaluation/.

2.2 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 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 GT to get a set of TPR [see Eq. (1)] & FPR [see Eq. (2)] values. These values are plotted in a 2D plot (a.k.a ROC curve), from which the AUC distills the area under the curve.

The AP measure is computed in a similar way. One can get a Precision [see Eq. (3)] & Recall [see Eq. (1)] curve by plotting Precision $p(r)$ as a function of Recall r. AP measure (Everingham et al., 2010) 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 (Margolin et al., 2014) has offered an intuitive generalization of the F_β measure. 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.

$$F_\beta^\omega = \frac{(1+\beta^2) Precision^\omega \times Recall^\omega}{\beta^2 \cdot Precision^\omega + Recall^\omega} \tag{6}$$

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 Current Measures are Pixel-Wise Based

Traditional measures (AP, AUC, PASCAL, F_β and Fbw) rely on two types of basic errors: FN, FP. Since these basic errors are calculated in a pixel-wise manner (see the Fig. 3), they cannot fully capture the structural information of foreground maps. However, foreground maps with fine structural details are often desired in several applications (e.g., image retrieval, object detection and segmentation). Therefore, evaluation measures sensitive to foreground structures are favored. Unfortunately, the aforementioned measures fail to meet this expectation.

A contrived example is shown in Fig. 4 which contains two different types of foreground maps. In FM1, a black square falls inside the digit while in the FM2 it touches the boundary. In our opinion, FM1 is less favored than FM2 since it destroys the foreground map more drastically. However, the current pixel-level MAE (mean absolute error) measure favors FM1 over FM2. This seems to contradict our common sense.

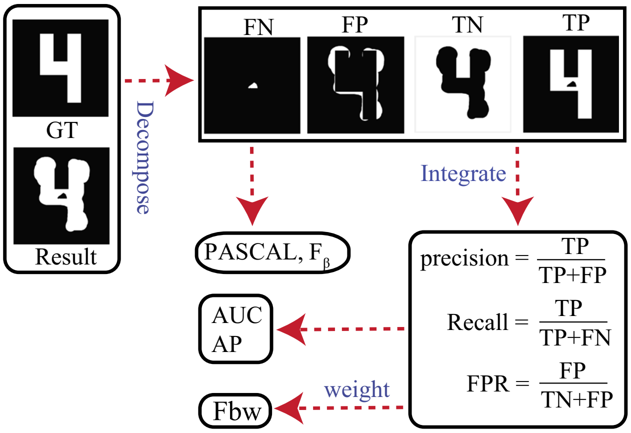


Fig. 3 The current evaluation measures' framework. The AP, AUC and Fbw evaluation measures are computed in a similar way. They are all calculated in a pixel-wise manner and ignore the structural similarities

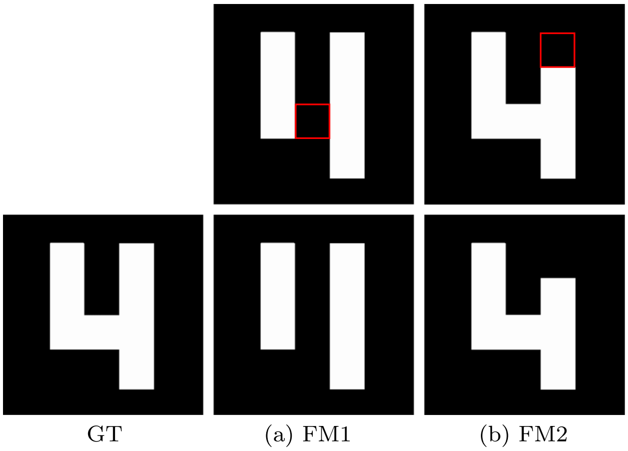


Fig. 4 A pixel-wise based evaluation example. In FM1, a black square falls inside the digit while in the FM2 it touches the boundary (1st row). They are two binary maps (2nd row) with the same TP, TN, FP and FN values. Visually, FM2 is favored over FM1 since FM1 destroys the foreground map's structure more drastically such that the the digit is hard to recognize. Current evaluation mean absolute error (MAE) measure is calculated in a pixel-wise manner and treat pixels independently. Hence, it ignores the structure of the foreground maps, thus it favor FM1 over FM2

A more realistic example is shown in Fig. 5. 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 (3rd row, the output of the SalCut algorithm fed with saliency maps; ranked by our measure, i.e., the 2nd row), the blue-border map offers a complete shape of the pyramid. In practice, this situation is very common. Thus, if the evaluation measure cannot capture the structural object information, it will not

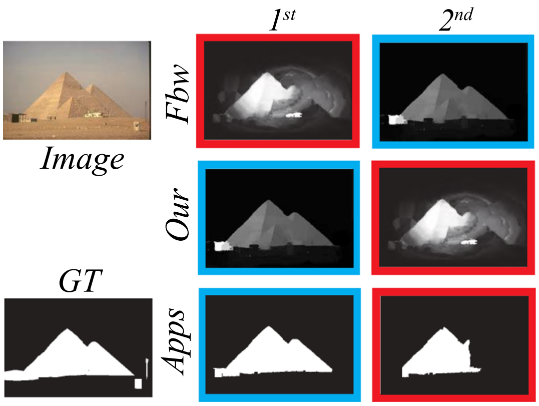


Fig. 5 A pixel-wise based evaluation example. Two foreground maps are generated by two saliency detection algorithms DSR (Li et al., 2013b), and ST (Liu et al., 2014). According to the application's ranking and our user-study (Apps-Sec. 5; last row), the blue-border map does the best, followed by the red-border map. Since Fbw measure does not account for the structural similarity, it results in the complete reverse ranking. Our measure (2nd row) correctly ranks the blue-border map as higher (Color figure online)

be able provide reliable information for model selection in applications.

4 Proposed Measure

In this section, we introduce our new measure to evaluate foreground maps. In the image quality assessment (IQA) field, a measure known as structural similarity measure (SSIM) (Wang et al., 2004) has been widely used to capture the structural similarity of the original image and a test image.

Let $x = \{x_i, i = 1, 2, \ldots, N\}$ and $y = \{y_i, i = 1, 2, \ldots, N\}$ be the FM and GT pixel values, respectively. The $\bar{x}$, $\bar{y}$, σ_x, σ_y are the mean and standard deviations of x and y, respectively. σ_{xy} is the covariance between the two. The SSIM is formulated as the product of three comparison terms including *luminance*, *contrast*, and *structure*:

$$ssim = \frac{2\bar{x}\bar{y} + C_1}{(\bar{x})^2 + (\bar{y})^2 + C1} \cdot \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \cdot \frac{\sigma_{xy} + C3}{\sigma_x\sigma_y + C_3} \tag{7}$$

where the constants C_1, C_2, and C_3 are set to very small values to avoid instability when denominator (e.g., $\bar{x}^2 + \bar{y}^2$) is very close to zero in each component.

In Eq. (7), 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 (Wang et al., 2004) associates 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 Eq. (7).

To build salient object detection or object segmentation algorithms, researchers are often more concerned 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 Wang et al. (2004), which aims to capture "object-part" structure information without any special concern regarding how complete is the foreground. The object-aware structural similarity, on the the other hand, is designed to mainly capture the structure information of "object-holistic" which focus on the complete object.

4.1 Region-Aware Structural Similarity Measure

Here we explain how to measure region-aware similarity. The region-aware similarity is designed to assess the object-part structure similarity in FM against the GT map. We first divide each of the FM 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 as in Lazebnik et al. (2006). The total number of blocks is denoted as K. An example is shown in Fig. 6. The region-aware similarity $ssim(k)$ of each block is computed independently using Eq. (7). We assign a different weight (w_k) to each block proportional to the GT foreground region that this block covers. Thus, the region-aware structural similarity measure can be formulated as:

$$S_r = \sum_{k=1}^{K} w_k \times ssim(k). \tag{8}$$

Our investigation shows that the proposed S_r can well describe the object-part similarity between a FM and a GT map. We also tried to directly use SSIM to assess the similarity between FM and GT at the image level or in the sliding

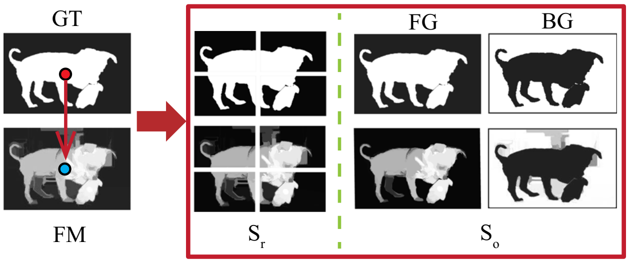


Fig. 6 The framework of our S-measure

window fashion as mentioned in Wang et al. (2004). These approaches fail to capture region structure similarities.

4.2 Object-Aware Structural Similarity Measure

Dividing the foreground map into blocks helps evaluate the object-part structural similarity. However, the region-aware measure (S_r) can not well account for the object similarity in a holistic way. 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 FM 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 characteristics.

4.2.1 Sharp Foreground–Background Contrast

Our first observation is that the foreground region of the GT map usually 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 FM and the foreground region of GT. Let x_{FG} and y_{FG} represent the probability values of foreground region of FM 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:

$$O_{FG} = \frac{2\bar{x}_{FG}\bar{y}_{FG}}{(\bar{x}_{FG})^2 + (\bar{y}_{FG})^2}. \tag{9}$$

Equation (9) has several appealing 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 $\bar{x}_{FG} = \bar{y}_{FG}$, then $O_{FG} = 1$, and
- The closer the two maps, the closer the O_{FG} to 1 (the most important property).

These properties make Eq. (9) suitable for our purpose.

4.2.2 Uniform Distribution

Our second observation is that the foreground and background regions of the GT maps usually have uniform distributions. So, it is important to assign a higher score to a FM with the object being uniformly highlighted (i.e., similar values across the entire object; see the Fig. 5). If the variability of the foreground values in the FM is high, then the distribution will not be uniform.

In probability theory and statistics, the coefficient of variation defined as the ratio of the standard deviation to the mean ($\sigma_x/\bar{x}$) is a standard measure of dispersion of a probability distribution. Here, we use it to represent the dispersion of the FM. In other words, the coefficient of variation is used to compute the dissimilarity between FM and GT distributions. According to Eq. (9), the total dissimilarity between FM and GT at object-level can be written as:

$$D_{FG} = \frac{(\bar{x}_{FG})^2 + (\bar{y}_{FG})^2}{2\bar{x}_{FG}\bar{y}_{FG}} + \lambda \times \frac{\sigma_{x_{FG}}}{\bar{x}_{FG}}, \tag{10}$$

where λ is a constant to balance the two terms. Then, the similarity between FM and GT at object level can be formulated as:

$$\begin{aligned} S'_{FG} &= \frac{1}{D_{FG}} \\ &= \frac{2\bar{x}_{FG}\bar{y}_{FG}}{(\bar{x}_{FG})^2 + (\bar{y}_{FG})^2 + 2\lambda \times \bar{y}_{FG} \times \sigma_{x_{FG}}}. \end{aligned} \tag{11}$$

Since in practice the mean probability of the GT foreground is exactly 1 ($\bar{y}_{FG} = 1$), the similarity between FM and GT in object level can be rewritten as:

$$S_{FG} = \frac{2\bar{x}_{FG}}{(\bar{x}_{FG})^2 + 1 + 2\lambda \times \sigma_{x_{FG}}}. \tag{12}$$

To compute background comparison S_{BG}, we regard the background as the complementary component of foreground by subtracting the FM and GT maps from 1 (change 1 to the maximum value of GT when GT is a non-binary map) as shown in Fig. 6. Then, S_{BG} can be similarly defined as:

$$S_{BG} = \frac{2\bar{x}_{BG}}{(\bar{x}_{BG})^2 + 1 + 2\lambda \times \sigma_{x_{BG}}}. \tag{13}$$

Let μ be the ratio of the foreground area in GT to the image area ($width \times height$). The final object-aware structural similarity measure can then be written as:

$$S_o = \mu \times S_{FG} + (1 - \mu) \times S_{BG}. \tag{14}$$

4.3 Structure Measure

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

$$S = \alpha \times S_o + (1 - \alpha) \times S_r, \tag{15}$$

where $\alpha \in [0, 1]$. We set $\alpha = 0.5$ in our implementation to assign equal contribution to both region similarity and object similarity. Using this measure to evaluate the three foreground maps in Fig. 1, we can correctly rank the maps consistent with the application rank and human rank.

5 Experiments

In order to assess the quality of our new measure, we utilized 4 meta-measures proposed by Margolin et al. (2014) and 1 meta-measure (human judgments) proposed by us. These meta-measures are used to assess the quality of evaluation measures (Pont-Tuset & Marques, 2013). To conduct fair comparisons, the 4 meta-measures are computed on the ASD (a.k.a ASD1000) dataset (Achanta et al., 2009). The non-binary foreground maps (5000 maps in total) were generated using five saliency detection algorithms including CA (Goferman et al., 2012), CB (Jiang et al., 2011), RC (Cheng et al., 2015), PCA (Margolin et al., 2013), and SVO (Chang et al., 2011) [binary maps are achieved by feeding non-binary maps to the SalCut (Cheng et al., 2015)].

Setting and Runtime We assign $\lambda = 0.5$ and $K = 4$ in all experiments as Fan et al. (2017). We also test our measure on the ASD1000 dataset using a single CPU machine. The average run time for a single image is 0.0053 s.

5.1 Meta-Measure 1: Application Ranking

The first meta-measure examines the ranking correlation of the evaluation measure to that of an application that uses foreground maps (Margolin et al., 2014). We assume that the GT map is the optimal input for the application (the top path in Fig. 7). Then, given a foreground map, we compare the application's output (the bottom path in Fig. 7) to that of the GT output. The closer the saliency map is to the GT, the closer its application output should be to the GT output. We compare the ranking result by each binary and non-binary evaluation measure: AP, AUC, Fbw, PASCAL, F_β and ours, to the ranking result by the application.

The work in Margolin et al. (2014) has examined three applications: object detection, segmentation and image retrieval. Here, we use the SalCut (Cheng et al., 2015) method (for non-binary) and image retrieval (for binary) application (see "Appendix Sect. 7") to compute this meta-measure.[5]

We utilize the 1-Spearman's ρ measure (Best & Roberts, 1975) to evaluate the ranking accuracy of the measures, where a lower values indicates better ranking consistency. The score of 0 indicates that the evaluation measure ranked the saliency maps identically to that of the application. The score of 2 indicates that the evaluation measure ranked the foreground maps in a complete reverse order. Comparison between different measures (AP, AUC, Fbw, Ours) is shown in Fig. 8a, which indicates that our structure measure produces the best ranking consistency among other alternative methods. According to the example shown in Fig. 1, all of the current non-binary measures fail to rank the foreground maps correctly. Our measure correctly ranks these maps. In the case of binary maps, S-measure also offers a 5.35%, 3.74%, 2.19%, improvement over the PASCAL, F_β, and Fbw measure, which score 0.897, 0.882, and 0.868, respectively, compared to 0.849 by our measure.

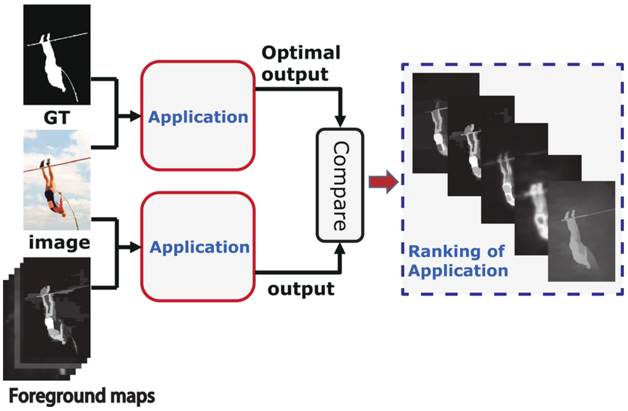
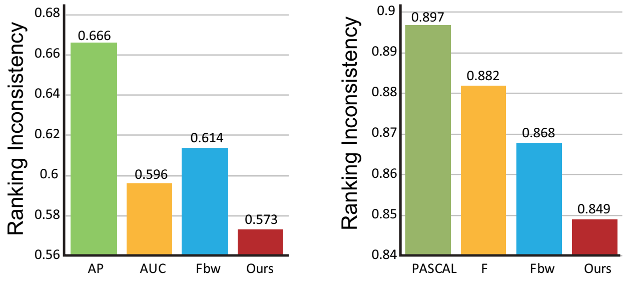


Fig. 7 Meta-measure 1: application ranking. To rank foreground maps according to an application, we compare the output obtained when using the GT, to the output when using the FM. The more similar a FM is to the GT map, the closer its application's output should be to the GT output



(a) Non-binary **(b)** Binary

Fig. 8 Meta-measure 1—results The ranking correlation of an evaluation measure to that given by the SalCut segmentation (Non-binary evaluation) and image retrieval (Binary evaluation) application. We used 1-spearman's rho as the results presentation. The lower the score, the better an evaluation measure is in term of predicting the preference of the application. Our measure achieves a better performance over other evaluation measure

[5] We follow the same experimental settings with Fbw (Margolin et al., 2014) for a fair comparison. Note that Fbw only provides the retrieval application, we can not achieve the other two application details.

5.2 Meta-Measure 2: SOTA Versus Generic

Our second meta-measure is that a measure should prefer the output achieved by a SOTA algorithm over generic baseline maps (e.g., centered Gaussian map, see Fig. 9d) that discard the image content. A good evaluation measure should rank the SM generated by a SOTA 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 SOTA models [CA (Goferman et al., 2012), CB (Jiang et al., 2011), RC (Cheng et al., 2015), PCA (Margolin et al., 2013), SVO (Chang et al., 2011)]. The mean score provides an indication of model robustness. Results are shown in Fig. 10. The lower the value here, the better the measure is. Over 1000 images, our measure has only 11 errors (i.e., generic winning over the s.t.a) for non-binary maps. Meanwhile, the AP and AUC measures are very poor and make significantly more mistakes. Our measure also offers a large improvement over the PASCAL, F_β, and Fbw.

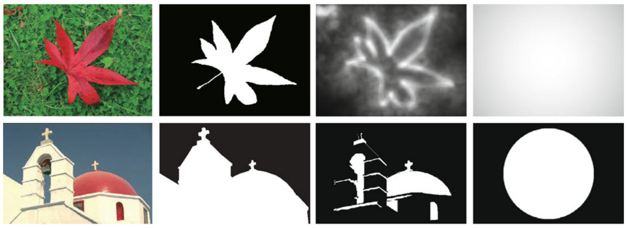
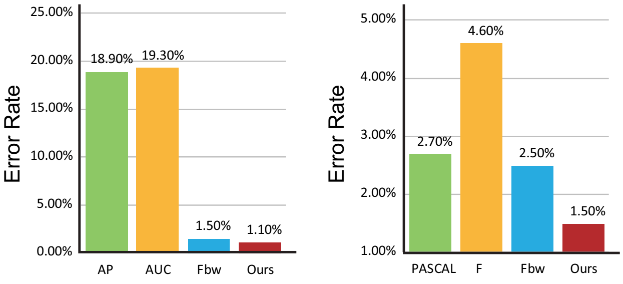

(a) Image **(b)** GT **(c)** FM **(d)** Generic

Fig. 9 Meta-measure 2: state-of-the-art versus generic. Given the input image **a** and the corresponding GT in **b**, an evaluation measure should give the FM generated by the SOTA method (**c**) a higher score than the generic map (**d**) that does not consider the content of the image. Unfortunately, all of the current evaluation measures give the map in **d** a higher score than **c**. Only our measure correctly ranks the SOTA result higher



(a) Non-binary **(b)** Binary

Fig. 10 Meta-Measure 2—results. The percentage of times that an evaluation measure ranked a generic map (non-binary circle or binary centered gaussian map) higher than the FM generated by the SOTA model. The lower the score, the better the evaluation measure is. Our measure achieves the best performance

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. (2014), 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 deemed 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 increased a saliency map's score when an incorrect GT map was used (see Fig. 11).

The Fig. 12 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 measure captures the object structural similarity between a FM and a GT map. Our measure will assign a lower value to the "good" FM when using a random selected GT since the object structure has changed in the random GT.

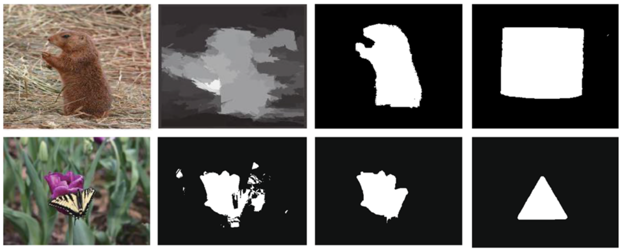
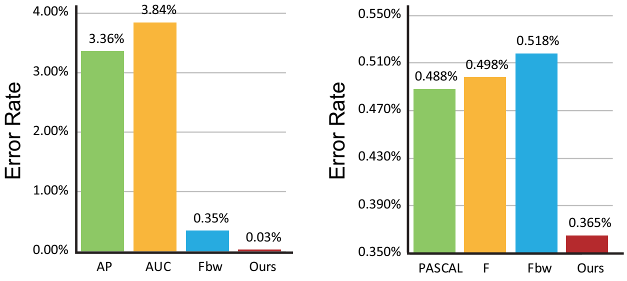

(a) Image **(b)** FM **(c)** GT **(d)** Switched GT

Fig. 11 Meta-measure 3: ground-truth switch. The score of a FM generated from **a** should decrease when using a wrong Switched GT as the reference. However, both AUC and AP gave the map in **b** a higher score when using **d** instead of **c** as the reference GT map. Using our measure, the score of **b** appropriately decreased when switching to random ground-truth (**d**)



(a) Non-binary **(b)** Binary

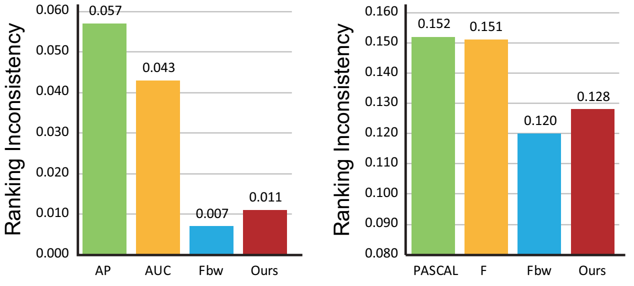
Fig. 12 Meta-measure 3—results. The percentage of times (tested on 1000 ASD dataset) that an evaluation measure assigned a higher score when using an incorrect GT map. The lower the score, the better the measure is. Our measure achieves significant improvement over other measures in both non-binary and binary maps

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. 13. While the two ground truth maps in (b) & (c) are slightly different, a good measure should not switch the ranking between the two foreground maps (d) & (e), when using (b) or (c) as the reference.

We use the 1-Spearman's ρ measure to examine the ranking correlation before and after annotation errors were introduced. The lower the score, the more robust an evaluation measure is to annotation errors (Margolin et al., 2014). Results are shown in Fig. 14. Our measure outperforms both the AP and the AUC but is not the best. Inspecting this reason, we realized that it is not always the case that the lower the score, the better an evaluation measure is. It 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 the structure change more carefully. Major structural changes often correspond to continuous large regions in the difference map between ground truth and its morphologically changed version. We used the sum of corroded version of the difference map as a measure of major structure change and to sort all the ground truth images.



(a) Non-binary **(b)** Binary

Fig. 14 Meta-measure 4—results. The ranking correlation of an evaluation measure under small manual annotation inaccuracies. We use the 1-Spearman's Rho measure to present the results. The lower the score, the better

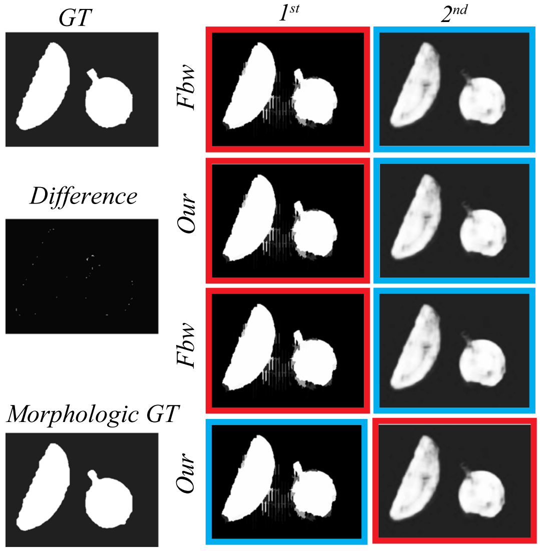
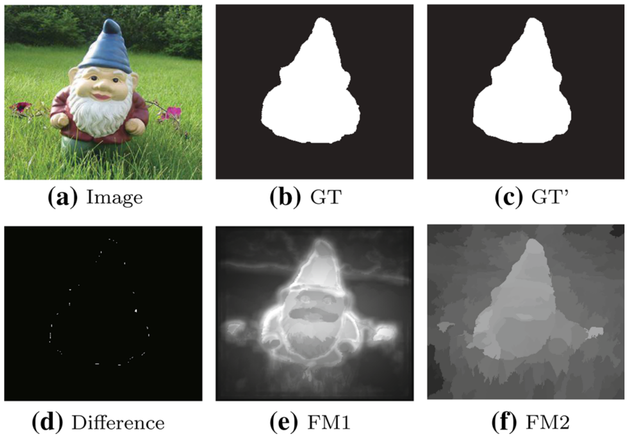


Fig. 15 Structural unchanged case. Both of ours and the Fbw measure are not sensitive to inaccuracies (structural unchanged) in manual annotations of the GT boundaries



(a) Image **(b)** GT **(c)** GT'

(d) Difference **(e)** FM1 **(f)** FM2

Fig. 13 Meta-measure 4: annotation errors. An evaluation measure should not be sensitive to slight changes in the manual annotation of the GT boundaries. While GT (**b**) & GT' (**c**) are almost identical, some measures switched the ranking order of the two foreground maps (**e, f**), depending on the different (**d**) GT used. Our measure consistently ranked **e** higher than **f**. Best viewed in color (Color figure online)

Among top 10% least changed images, our measure and Fbw have the same MM4 scores (both of them are 0). When the topology of ground truth map does not change, our measure and Fbw preserve the original ranking. This can be seen in the example in Fig. 15. While ground truth maps (GT and Morphologic GT) differ slightly, both Fbw and our measure preserve the ranking order of the two saliency maps, depending on the used ground truth map.

For top 10% most changed images, we asked 3 users to judge whether the ground truth images have major structure changes. 95 out of 100 ground truth images were considered to have major structure changes (e.g., small bar, thins legs, slender foot and minute lines in each group; see Fig. 16), for which we believe that keeping the same ranks is not good. Figure 17 is a typical example. 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. A good measure should rank the blue-border SM higher than red-border SM. So the rank-

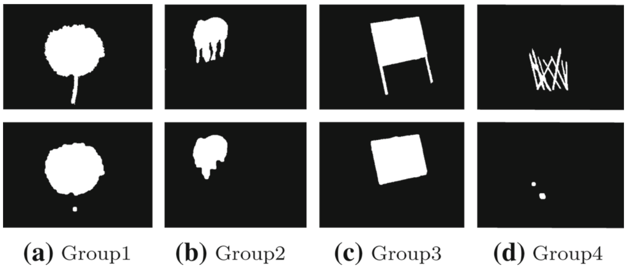


Fig. 16 Structural changed examples. The first row shoes the GT maps. The second row shows morphologically changed versions. We observe significant structural changes

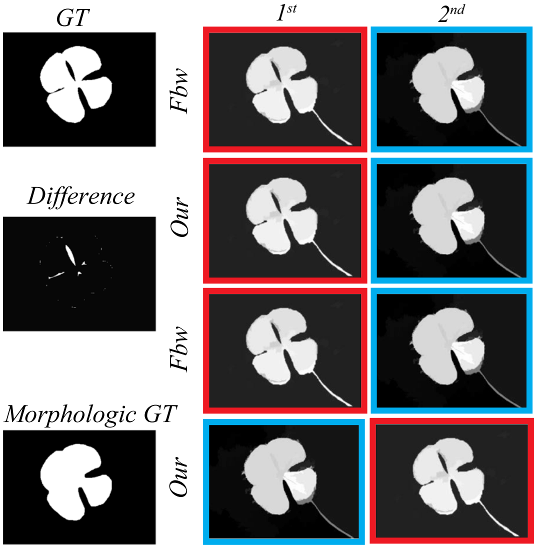


Fig. 17 Structural changed case. The ranking of an evaluation measure should be sensitive to the structural changes. Surprisingly, the current best measure (Fbw) does not account for structural changes. Using our measure, we rank the maps correctly. Best viewed on screen

ing 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 comparison.

5.5 Further Comparison

The results in Figs. 8, 10, and 12 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 use 10 SOTA algorithms for salient object detection to perform experiments on another 4 widely-used benchmark datasets.

Foreground Maps Collection We used 10 SOTA algorithms including 3 traditional models [i.e., ST (Liu et al., 2014), DRFI (Jiang et al., 2013), and DSR (Li et al., 2013b)] and 7 deep learning based models [DCL (Li & Yu, 2016), rfcn (Wang et al., 2016), MC (Zhao et al., 2015), MDF (Li & Yu, 2015), DISC (Chen et al., 2016), DHS (Liu & Han, 2016), and ELD (Lee et al., 2016)] to generate the binary and non-binary foreground maps. Binary maps are obtained by thresholding the non-binary maps using image dependent adaptive thresholding method in Achanta et al. (2009).

Benchmark Datasets The 4 widely-used datasets include PASCAL-S (Li et al., 2014), ECSSD (Xie et al., 2013), HKU-IS (Li & Yu, 2015), and SOD (Martin et al., 2001). PASCAL-S contains 850 challenging images, which have multiple objects in high background clutter. ECSSD contains 1000 semantically meaningful but structurally complex images. HKU-IS is another large dataset that contains 4445 large scale images. Most of the images in this dataset contain more than one salient object with low contrast. Finally, we also evaluate our measure over the SOD dataset, which is a subset of the BSDS dataset. It contains a relatively small number of images (300), but with multiple complex objects.

Results Non-binary and binary maps' quantitative comparison results are shown in Table 1. Our measure performs the best according to the first meta-measure for both binary and non-binary maps evaluation. This indicates that our measure is more useful for applications than other measures.

For the evaluation results (binary and non-binary) in MM2, our measure performs better than the existing four measures (AP, AUC, F, PASCAL) with a large margin. The results on two easier datasets (ECSSD and HKU-IS) show that our measure and Fbw perform on par for both binary and non-binary maps.

According to meta-measure 3, our measure reduces the non-binary error rate by 67.62%, 44.05%, 17.81%, 69.23% on PASCAL, ECSSD, SOD and HKU-IS, respectively compared to the second ranked measure. For binary maps, our measure also reduces the error rate by 62.86%, 52.38%, 10.96%, 61.54% on PASCAL, ECSSD, SOD and HKU-IS, respectively compared to the second ranked measure. This indicates that our measure has higher capacity to capture the structural similarity between FM and GT maps.

Overall, our measure wins in the majority of cases indicating that it 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 highly agree with the human ranking. It is argued that "a human being is the best judge to evaluate the output of any segmentation algorithm" (Pal & Pal, 1993). However, sub-

Table 1 Non-binary (N-binary) & Binary maps' quantitative comparison with current measures on 3 meta-measures

Type	Measure	PASCAL-S (Li et al., 2014)			ECSSD (Xie et al., 2013)			SOD (Martin et al., 2001)			HKU-IS (Li & Yu, 2015)		
		MM1	MM2 (%)	MM3 (%)	MM1	MM2 (%)	MM3 (%)	MM1	MM2 (%)	MM3 (%)	MM1	MM2 (%)	MM3 (%)
N-binary	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**
Binary	F	0.757	21.65	1.62	0.619	12.5	1.53	0.760	25.3	1.81	0.593	4.54	0.41
	PASCAL	0.905	19.41	1.59	0.787	11.0	1.54	0.911	24.0	1.81	0.786	3.62	0.41
	Fbw	0.802	13.76	1.72	0.675	7.50	1.54	0.814	19.0	**1.42**	0.665	**2.16**	0.58
	Ours	**0.655**	**12.50**	**1.11**	**0.523**	**7.30**	**1.29**	**0.707**	**15.0**	1.56	**0.483**	2.27	**0.32**

The best result is highlighted in bold
MM meta-measure

jective 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 in the object segmentation domain that meets these requirements. Here, we focus on the non-binary maps to collect such a database.

Stimuli The source foreground maps are sampled from three large scale datasets: PASCAL-S, ECSSD, and HKU-IS. As mentioned above, we use 10 SOTA saliency models to generate the maps for each dataset. Therefore, we have 10 foreground 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 $|n - 1|$. Figures 18, 19 and 20a, b, c show the histogram of rank distances between the two measures. 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 from the three datasets. The top panel in Fig. 21b shows one example trial with the best map according to our measure on the left, and the best map according to the Fbw on the far right. The user is asked to choose the map 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. 21b), the two maps are too similar making it difficult for the used 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 (see Fig. 21a) 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; from 10 different majors such as history, medicine and finance; 25 males and 20 females).

Results Results (Fbw vs. our measure) are shown in Fig. 22. 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 procedure to conduct two additional user studies (AP vs. our measure, AUC vs. our measure). The results are 72.11% and 73.56%, respectively. This indicates that our measure correlates better with human judgments.

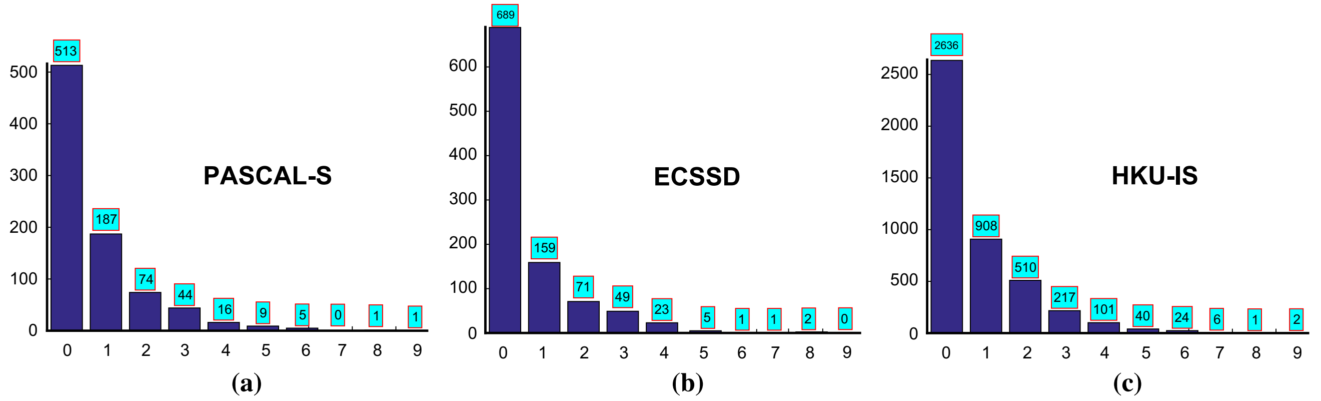


Fig. 18 The rank distance between Fbw and our measure. The **a**–**c** are the three datasets used to compute the rank distance between Fbw and our S-measure. The y axis of the plot is the number of the images. The x axis is the rank distance

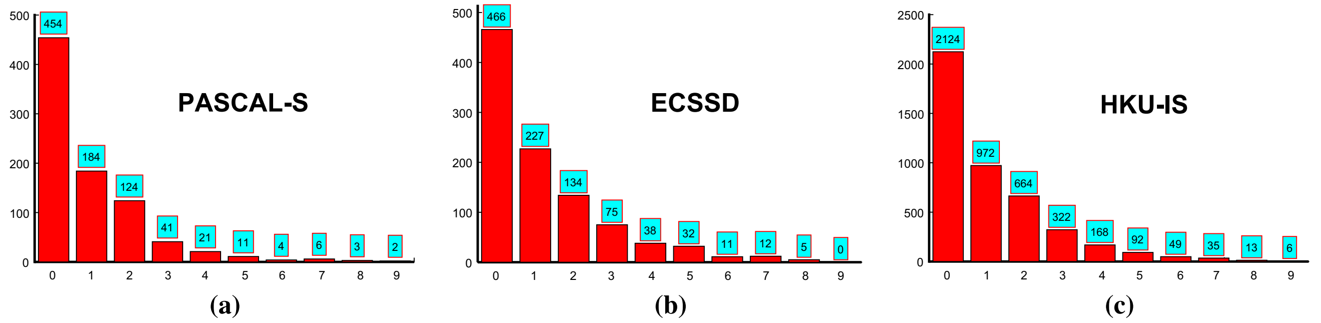


Fig. 19 The rank distance between AP and our measure. The **a**–**c** are the three datasets used to compute the rank distance between AP and our S-measure. The y axis of the plot is the number of the images. The x axis is the rank distance

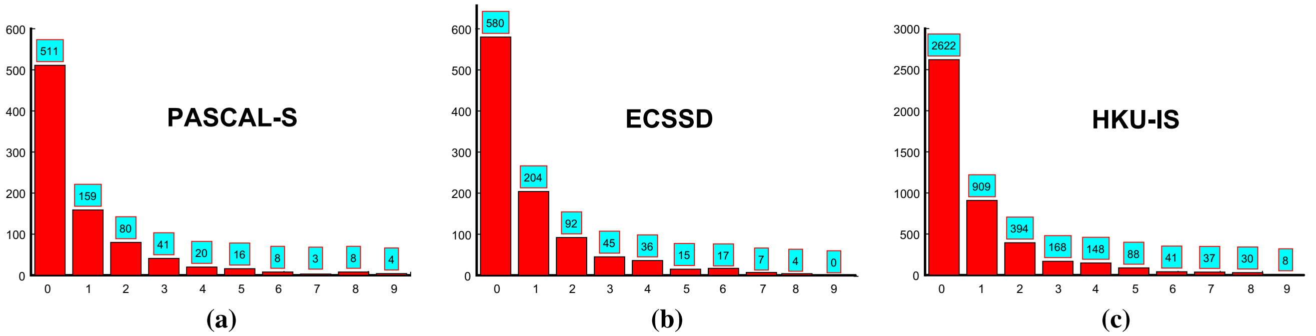
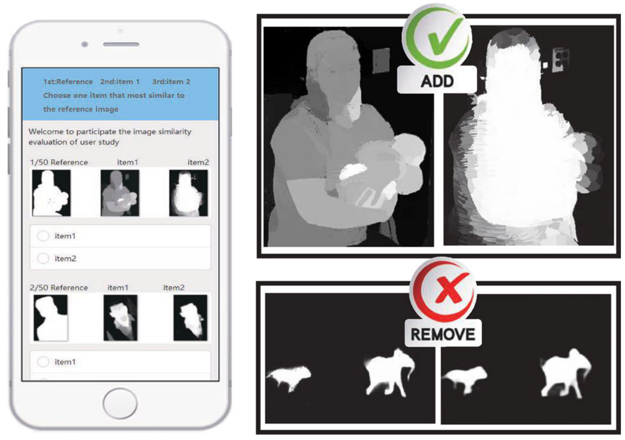


Fig. 20 The rank distance between AUC and our measure. The **a**–**c** are the three datasets used to compute the rank distance between AUC and our S-measure. The y axis of the plot is the number of the images. The x axis is the rank distance



1st:Reference 2nd:item 1 3rd:item 2
Choose one item that most similar to the reference image

Welcome to participate the image similarity evaluation of user study

1/50 Reference item1 item2

○ item1
○ item2

2/50 Reference item1 item2

○ item1

(a) the mobile phone app **(b)** the one example trial

Fig. 21 Our user study platform

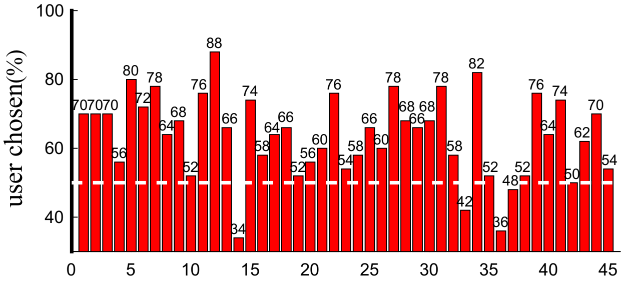


Fig. 22 Results of our user study (Fbw & S-measure). The x axis is the viewer id. The y axis shows the percentage of the trials in which a viewer preferred the map chosen by our measure

5.7 Saliency Model Comparison

Establishing that our S-measure offers a better way to evaluate foreground maps, here we compare 10 SOTA models on 4 datasets (PASCAL-S, ECSSD, HKU-IS, and SOD). Figure 23 shows the rank of the 10 models. According to our measure, top three models in order are dhsnet, DCL, and rfcn. Moreover, we also establish many representative online benchmarks (1. http://dpfan.net/socbenchmark; 2. http://dpfan.net/d3netbenchmark; 3. http://dpfan.net/cosod3k/; 4. http://dpfan.net/camouflage/) to compared our S-measure with other measures.

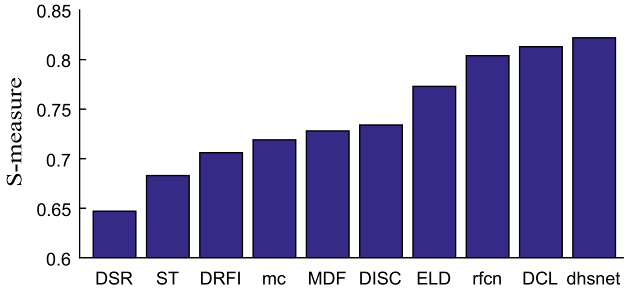


Fig. 23 Ranking of 10 saliency models using our new measure. The y axis shows the average score on each dataset [PASCAL-S (Li et al., 2014), ECSSD (Xie et al., 2013), HKU-IS (Li & Yu, 2015), SOD (Martin et al., 2001)]

6 Ablation Study

To investigate the contribution of each part in our S-measure, we further conduct the ablation study on a new human ranking dataset (Fan et al., 2021b). The FMDatabase[6] consists of 185 color images and 555 ranked maps. Similar to meta-measure1, we also utilize the 1-Spearma's ρ metric to evaluate the ranking performance of the measures.

As shown in Table 2, we observe that our S-measure outperforms other settings (i.e., object-aware, region-aware) on FMDatabase. It clearly shows that only region-level or object-level structural similarity cannot provide stable evaluation performance. Since the object-aware structure similarity mainly focusing on assessing the property of sharp foreground–background contrast. It more like a global evaluation. On the other hand, the region-aware links to the local evaluation which based on window-level statistics.

[6] http://dpfan.net/e-measure/.

Table 2 Ablation studies of human ranking (Using FMDatabase-IJCAI'18) in terms of 1-Spearman's measure

Settings	1-Spearman's measure
Object-aware	0.195
Region-aware	0.142
S-measure (Ours)	**0.140**
Fbw (Margolin et al., 2014)	0.149
SSIM (Wang et al., 2004)	0.223

The best (the lower the better) score is highlighted in bold

Compared with existing two classical metrics (i.e., Fbw and SSIM), we also found that our metric achieve the best results.

7 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 **S-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 ground-truth 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 Fwb.

All metrics are double-edged swords. Generally, it's hard to argue which measure is the best one. These measures are deeply coupled to the actual applications, e.g., some applications may favor the correctness of important regions while some may prefer the continuity in local structures. We observe a few failure cases where a prediction map without (or less) object structure will achieve a higher S-measure score. For example, as shown in Fig. 24, the proposed S-measure does not work well in this situation in which the object of the GT (b) without a clear structure. Consequently, to evaluate the foreground maps, we need to assess whether it performs well on multi metrics at the same time.

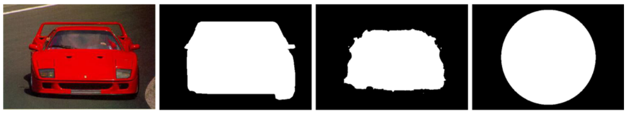

(a) Image (b) GT (c) rfcn (d) Generic

Fig. 24 Failure case. With the given image (**a**) and its corresponding GT in (**b**), our S-measure ranked (**d**) generic higher than **c** rfcn due to these prediction map without obvious structure

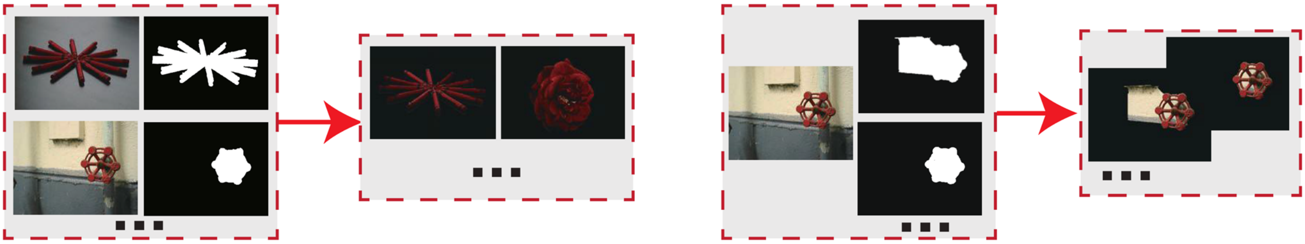
In summary, our measure offers new insights into foreground map evaluation where current measures fail to examine the strengths and weaknesses of models fully. For now, we have found that the saliency [e.g., RGB SOD (Zhuge et al., 2021a; Zhang et al., 2020a; Chen et al., 2020; Zhao et al., 2019), RGB-D SOD (Zhang et al., 2021; Zhao et al., 2020; Fan et al., 2021d; Fu et al., 2021; Zhou et al., 2021), RGB-T SOD (Zhang et al., 2019b), light field SOD (Zhang et al., 2019a; Piao et al., 2020; Jiang et al., 2020), VSOD (Ji et al., 2021), 360 SOD, SID (Li et al., 2017), Saliency Ranking (Amirul Islam et al., 2018), Co-SOD (Fan et al., 2021c, e; Zhang et al., 2020b), and HR SOD (Zeng et al., 2019)] community has begun to widely adopt this measure even in the camouflaged object detection (Fan et al., 2021a; Zhai et al., 2021; Mei et al., 2021) and medical image segmentation (Fan et al., 2020).

Acknowledgements This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, NSFC (61922046), and S&T innovation project from Chinese Ministry of Education.

Appendix

Image Retrieval Application We use the publicly available content based image retrieval system LIRE (Lew et al., 2000) as our application. Firstly, we generate a combined image as Fig. 25a–d. For each combined GT image (e.g., $GT_1, \ldots, GT_n$; n denotes the total number of images), we use LIRE to extract the CEDD feature and then search a list of 100 most similar images $GT_{lst-i} = \{G_{1-i}, \ldots, G_{100-i}\}$. LIRE also assigns the 100 images with score ($GT_{score-i} = \{G_{s1-i}, \ldots, G_{s100-i}\}$) which indicates the similarity value. Accordingly, for each FM_{comb} image we can obtain the 100 most similarity images FM_{lst-i} and corresponding score sets $FM_{score-i}$. Finally, let $Q_i = \{GT_{lst-i} \cap FM_{lst-i}\}$. We search FM_k which equals to G_i in the FM_{lst-i}. If FM_k exists, we record the index k and the corresponding score F_{sk-i}. The similarity S_i of each FM assigned by LIRE is:

$$S_i = \begin{cases} F_{sk-i} + \frac{1}{k} + \frac{\|Q_i\|}{100}, & G_i \in Q_i \\ \frac{\|Q_i\|}{100}, & otherwise \end{cases} \tag{16}$$



(a) Imgs&GT **(b)** GT_{comb} **(c)** Imgs&FM **(d)** FM_{comb}

Fig. 25 Integration of the image with its foreground map. **a**, **c** Are the images (Imgs) with its GT and FM. **b**, **d** Are the combined images

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE conference on computer vision and pattern recognition* (pp. 1597–1604) .

Amirul Islam, M., Kalash, M., & Bruce, N. D. (2018). Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *IEEE conference on computer vision and pattern recognition* (pp. 7142–7150).

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 898–916.

Best, D., & Roberts, D.: 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)

Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing*, *24*(2), 742–756.

Borji, A., Cheng, M. M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5706–5722.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

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

Borji, A., Sihite, D. N., & Itti, L. (2013b). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, *91*, 62–77.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2015). Mit saliency benchmark. http://saliency.mit.edu/results_mit300.html

Chang, K. Y., Liu, T. L., Chen, H. T., & Lai, S. H. (2011). Fusing generic objectness and visual saliency for salient object detection. In *International conference on computer vision* (pp. 914–921).

Chen, H., & Li, Y. F. (2018). Progressively complementarity-aware fusion network for RGB-D salient object detection. In *IEEE conference on computer vision and pattern recognition*.

Chen, T., Cheng, M. M., Tan, P., Shamir, A., & Hu, S. M. (2009). Sketch2photo: Internet image montage. *ACM Transactions on Graphics*, *28*(5), 124.

Chen, T., Lin, L., Liu, L., Luo, X., & Li, X. (2016). Disc: Deep image saliency computing via progressive representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(6), 1135–1149.

Chen, Z., Xu, Q., Cong, R., & Huang, Q. (2020). Global context-aware progressive aggregation network for salient object detection. In *AAAI conference on artificial intelligence*.

Cheng, M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. (2015). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(3), 569–582.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fan, D. P., Cheng, M. M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps. In *International conference on computer vision* (pp. 4548–4557).

Fan, D. P., Ji, G. P., Cheng, M. M., & Shao, L. (2021a). Concealed object detection. arXiv preprint arXiv:2102.10274

Fan, D. P., Ji, G. P., Qin, X., & Cheng, M. M. (2021b). Cognitive vision inspired object segmentation metric and loss function. *SSI*,. https://doi.org/10.1360/SSI-2020-0370.

Fan, D. -P., Ji, G. -P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020). Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 263–273). Springer.

Fan, D. -P., Li, T., Lin, Z., Ji, G. -P., Zhang, D., Cheng, M. -M., Fu, H., & Shen, J. (2021c). Re-thinking co-salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3060412

Fan, D. -P., Lin, Z., Zhang, Z., Zhu, M., & Cheng, M. -M. (2021d). Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(5), 2075–2089. https://doi.org/10.1109/TNNLS.2020.2996406

Fan, Q., Fan, D. P., Fu, H., Tang, C. K., Shao, L., & Tai, Y. W. (2021e). Group collaborative learning for co-salient object detection. In *IEEE conference on computer vision and pattern recognition*.

Feng, D., Barnes, N., You, S., & McCarthy, C. (2016). Local background enclosure for RGB-D salient object detection. In *IEEE conference on computer vision and pattern recognition* (pp. 2343–2350).

Fu, K., Fan, D. -P., Ji, G. -P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3073689

Ghosh, J., Lee, Y. J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *IEEE conference on computer vision and pattern recognition* (pp. 1346–1353).

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

Gorji, S., & Clark, J. (2018). Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *IEEE conference on computer vision and pattern recognition*.

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, *19*(1), 185–198.

Islam, M.A., Kalash, M., D. B. Bruce, N. (2018). Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *IEEE conference on computer vision and pattern recognition*.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304–1318.

Ji, Y., Zhang, H., Jie, Z., Ma, L., & Wu, Q. J. (2021). Casnet: A cross-attention siamese network for video salient object detection. *IEEE Transactions on Neural Networks and Learning Systems, 32*(6), 2676–2690. https://doi.org/10.1109/TNNLS.2020.3007534

Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., & Li, S. (2011). Automatic salient object segmentation based on context and shape prior. In *British machine vision conference* (p. 9).

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., & Li, S. (2013). Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2083–2090).

Jiang, Y., Zhou, T., Ji, G. P., Fu, K., Zhao, Q., & Fan, D. P. (2020). Light field salient object detection: A review and benchmark. arXiv preprint arXiv:2010.04968

Kanan, C., & Cottrell, G. (2010). Robust classification of objects, faces, and flowers using natural image statistics. In *IEEE conference on computer vision and pattern recognition* (pp. 2472–2479).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 2169–2178).

Lee, G., Tai, Y. W., & Kim, J. (2016). Deep saliency with encoded low level distance map and high level features. In *IEEE conference on computer vision and pattern recognition* (pp. 660–668).

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2000). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing*, *2*(1), 1–19.

Li, G., Xie, Y., Lin, L., & Yu, Y. (2017). Instance-level salient object segmentation. In *IEEE conference on computer vision and pattern recognition* (pp. 2386–2395).

Li, G., Xie, Y., Wei, T., & Lin, L. (2018). Flow guided recurrent neural encoder for video salient object detection. In *IEEE conference on computer vision and pattern recognition*.

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *IEEE conference on computer vision and pattern recognition* (pp. 5455–5463).

Li, G., & Yu, Y. (2016). Deep contrast learning for salient object detection. In *IEEE conference on computer vision and pattern recognition* (pp. 478–487).

Li, L., Jiang, S., Zha, Z., Wu, Z., & Huang, Q. (2013a). Partial-duplicate image retrieval via saliency-guided visually matching. *IEEE Transactions on Multimedia*, *20*(3), 13–23.

Li, X., Lu, H., Zhang, L., Ruan, X., & Yang, M. H. (2013b). Saliency detection via dense and sparse reconstruction. In *International conference on computer vision* (pp. 2976–2983).

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *IEEE conference on computer vision and pattern recognition* (pp. 280–287).

Liu, G., & Fan, D. (2013). A model of visual attention for natural image retrieval. In *International conference on information science and cloud computing companion* (pp. 728–733).

Liu, G. H., Yang, J. Y., & Li, Z. (2015). Content-based image retrieval using computational visual attention model. *Pattern Recognition*, *48*(8), 2554–2566.

Liu, N., & Han, J. (2016). Dhsnet: Deep hierarchical saliency network for salient object detection. In *IEEE conference on computer vision pattern recognition* (pp. 678–686).

Liu, N., Han, J., & Yang, M. H. (2018). Picanet: Learning pixel-wise contextual attention for saliency detection. In *IEEE conference on computer vision and pattern recognition*.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367.

Liu, Z., Zou, W., & Le Meur, O. (2014). Saliency tree: A novel saliency detection framework. *IEEE Transactions on Image Processing*, *23*(5), 1937–1952.

Margolin, R., Tal, A., & Zelnik-Manor, L. (2013). What makes a patch distinct? In *IEEE conference on computer vision and pattern recognition* (pp. 1139–1146).

Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps? In *IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International conference on computer vision* (Vol. 2, pp. 416–423).

Mei, H., Ji, G. P., Wei, Z., Yang, X., Wei, X., & Fan, D. P. (2021). Camouflaged object segmentation with distraction mining. In *IEEE conference on computer vision and pattern recognition*.

Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition*, *26*(9), 1277–1294.

Peng, H., Li, B., Xiong, W., Hu, W., & Ji, R. (2014). RGBD salient object detection: A benchmark and algorithms. In *European conference on computer vision* (pp. 92–109).

Piao, Y., Rong, Z., Zhang, M., & Lu, H. (2020). Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In *AAAI conference on artificial intelligence* (pp. 11865–11873).

Pont-Tuset, J., & Marques, F. (2013). Measures and meta-measures for the supervised evaluation of image segmentation. In *IEEE conference on computer vision and pattern recognition* (pp. 2131–2138).

Qin, X., Fan, D. P., Huang, C., Diagne, C., Zhang, Z., Sant'Anna, A. C., Suàrez, A., Jagersand, M., & Shao, L. (2021). Boundary-aware segmentation network for mobile and web applications. arXiv preprint arXiv:2101.04704

Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? In *IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. II–37).

Tiantian, W., Zhang, L., Lu, H., & Borji, A. (2018). Detect globally, refine locally: A novel approach to saliency detection. In *IEEE conference on computer vision and pattern recognition*.

Wang, L., Wang, L., Lu, H., Zhang, P., & Ruan, X. (2016). Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision* (pp. 825–841).

Wang, W., Shen, J., Dong, X., & Borji, A. (2018). Salient object detection driven by fixation prediction. In *IEEE conference on computer vision and pattern recognition*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Xie, Y., Lu, H., & Yang, M. H. (2013). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing*, *22*(5), 1689–1698.

Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E. K., & Yuille, A. L. (2018). Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *IEEE conference on computer vision and pattern recognition*.

Zeng, Y., Lu, H., Zhang, L., Feng, M., & Borji, A. (2018). Learning to promote saliency detectors. In *IEEE conference on computer vision and pattern recognition*.

Zeng, Y., Zhang, P., Zhang, J., Lin, Z., & Lu, H. (2019). Towards high-resolution salient object detection. In *International conference on computer vision* (pp. 7234–7243).

Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., & Fan, D. P. (2021). Mutual graph learning for camouflaged object detection. In *IEEE conference on computer vision and pattern recognition*.

Zhang, J., Fan, D. -P., Dai, Y., Anwar, S., Saleh, F., Aliakbarian, S., & Barnes, N. (2021). Uncertainty inspired RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3073564

Zhang, J., Xie, J., & Barnes, N. (2020a). Learning noise-aware encoder–decoder from noisy labels by alternating back-propagation for saliency detection. In *European conference on computer vision*.

Zhang, J., Zhang, T., Dai, Y., Harandi, M., & Hartley, R. (2018a). Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *IEEE conference on computer vision and pattern recognition*.

Zhang, L., Dai, J., Lu, H., He, Y., & Wang, G. (2018b). A bi-directional message passing model for salient object detection. In *IEEE conference on computer vision and pattern recognition*.

Zhang, M., Li, J., Wei, J., Piao, Y., & Lu, H. (2019a). Memory-oriented decoder for light field salient object detection. In *Advances in neural information processing systems* (pp. 898–908).

Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. In *International conference on computer vision*.

Zhang, Q., Cong, R., Hou, J., Li, C., & Zhao, Y. (2020b). Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection. In *Advances in neural information processing systems*.

Zhang, Q., Huang, N., Yao, L., Zhang, D., Shan, C., & Han, J. (2019b). RGB-T salient object detection via fusing multi-level CNN features. *IEEE Transactions on Image Processing*, *29*, 3321–3335.

Zhang, X., Wang, T., Qi, J., Lu, H., & Wang, G. (2018c). Progressive attention guided recurrent network for salient object detection. In *IEEE conference on computer vision and pattern recognition*.

Zhao, J. X., Liu, J. J., Fan, D. P., Cao, Y., Yang, J., & Cheng, M. M. (2019). EGNet: Edge guidance network for salient object detection. In *International conference on computer vision* (pp. 8779–8788).

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. In *IEEE conference on computer vision and pattern recognition* (pp. 1265–1274).

Zhao, X., Pang, Y., Zhang, L., Lu, H., & Zhang, L. (2020). Suppress and balance: A simple gated network for salient object detection. In *European conference on computer vision*.

Zhou, T., Fan, D. -P., Cheng, M. -M., Shen, J., & Shao, L. (2021). RGB-D salient object detection: A survey. *Computational Visual Media, 7*(1), 37–69.

Zhuge, M., Fan, D. P., Liu, N., Zhang, D., Xu, D., & Shao, L. (2021a). Salient object detection via integrity learning. arXiv:2101.07663

Zhuge, M., Gao, D., Fan, D. P., Jin, L., Chen, B., Zhou, H., Qiu, M., & Shao, L. (2021b). Kaleido-bert: Vision-language pre-training on fashion domain. In *IEEE conference on computer vision and pattern recognition.*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.