# How to Evaluate Foreground Maps?

Ran Margolin
Technion
Haifa, Israel
margolin@tx.technion.ac.il

Lihi Zelnik-Manor
Technion
Haifa, Israel
lihi@ee.technion.ac.il

Ayellet Tal
Technion
Haifa, Israel
ayellet@ee.technion.ac.il

## Abstract

*The output of many algorithms in computer-vision is either non-binary maps or binary maps (e.g., salient object detection and object segmentation). Several measures have been suggested to evaluate the accuracy of these foreground maps. In this paper, we show that the most commonly-used measures for evaluating both non-binary maps and binary maps do not always provide a reliable evaluation. This includes the Area-Under-the-Curve measure, the Average-Precision measure, the $F_\beta$-measure, and the evaluation measure of the PASCAL VOC segmentation challenge. We start by identifying three causes of inaccurate evaluation. We then propose a new measure that amends these flaws. An appealing property of our measure is being an intuitive generalization of the $F_\beta$-measure. Finally we propose four meta-measures to compare the adequacy of evaluation measures. We show via experiments that our novel measure is preferable.*

## 1. Introduction

The comparison of a foreground map against a binary ground-truth is common in various computer-vision problems, e.g., salient object detection [10], object segmentation [11], and foreground-extraction [6]. These comparisons are crucial in assessing the quality of an algorithm.

The foreground maps are either binary or non-binary. The common measures for evaluating a binary map are PASCAL's VOC (Visual Object Classes) segmentation measure (referred to henceforth as PASCAL) [3, 11, 14] and $F_\beta$-measure [2, 4, 10, 17, 23]. Many algorithms that output a non-binary map still compare against a binary ground-truth [1, 9, 10, 12, 13, 19, 21, 24]. They do this via two steps. First, multiple thresholds are applied to it, to obtain multiple binary maps. Then, these binary maps are compared to the ground-truth. Common methods to integrate this into a single measure are Area-Under-the-Curve (AUC) [7, 12, 15] and Average-Precision (AP) [10, 23].

We expect these measures to provide an indicator as to which algorithm offers the best quality of detection. What happens if this is not the case? Then a better algorithm may receive a lower score than a lesser one. For instance, in Figure 1 the cyan map better captures the teddy-bear than the red map, which offers a fuzzy detection that mostly captures the red flower. Yet, the commonly-used AUC incorrectly prefers the red over the cyan.

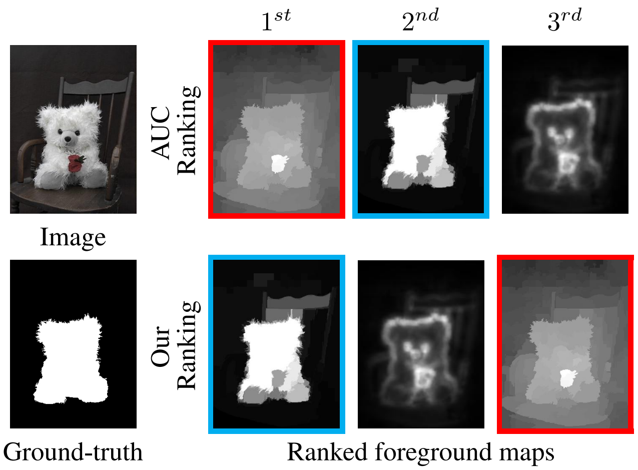


Figure 1. **Inaccuracy of current evaluation measures.** We compare the ranking of foreground maps that are the output of three state-of-the-art salient-object detection algorithms [9, 12, 13]. According to the most common measure, AUC, the red map offers the best result despite its fuzziness, which captures much of the background. Conversely, our proposed measure correctly ranks first the cyan map, which most accurately captures the foreground object, according to the ground-truth. Our ranking is supported by four meta-measures.

Our first contribution is identifying three assumptions in commonly-used measures (AUC, AP, PASCAL and $F_\beta$-measure), which lead to inaccurate evaluations (Section 3). For instance, typically the locations of the errors in the map are ignored, while they are highly important.

Next, we proceed to amend each of these flaws and to suggest a novel measure that evaluates foreground maps at an increased accuracy (Section 4). Two appealing properties of our measure are: i) being a generalization of the

F$_\beta$-measure, and, ii) providing a unified evaluation to both binary and non-binary maps.

Our third contribution is proposing four *meta-measures* to analyze the performance of evaluation measures (Section 5). Much like a measure is used to evaluate an algorithm, a meta-measure is used to evaluate a measure [22]. For instance, one of our meta-measures verifies that the ranking of foreground maps by an evaluation measure agrees with the preferences of applications that use these foreground maps (e.g. image retrieval, object detection and segmentation). Using these meta-measures we compare the evaluation measures, and show that our measure outperforms all others.

## 2. Current Evaluation Measures

We discern between *binary* maps, which consist of values of either 0 or 1, and *non-binary* maps, which consist of values in the range $[0, 1]$. These values represent the probability that a pixel belongs to the foreground [8].

**Evaluation of binary maps:** All common measures for evaluating a binary map are based on a subset of the following four basic quantities: true-positive ($TP$), true-negative ($TN$), false-positive ($FP$) and false-negative ($FN$). These quantities are used to assess different qualities of the binary map. The most common qualities are: *Hit-rate* & *False-alarm*, and *Precision* & *Recall*:

$$\text{Hit-rate} = \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{False alarm} = \frac{FP}{TN + FP} \tag{2}$$

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

These qualities are typically combined into a single score. One common score is the F$_\beta$-measure:

$$\begin{aligned} \text{F}_\beta\text{-measure} &= (1+\beta^2)\frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \\ &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP} \end{aligned} \tag{4}$$

where $\beta$ is a parameter that controls the preference between complete-detection and over-detection (typically $\beta = 1$). A second common score is the PASCAL measure:

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

**Evaluation of non-binary maps:** Non-binary maps are compared against a binary ground-truth as well. The two most common evaluation measures are AUC and AP. Both measures are computed by first thresholding the non-binary map into multiple binary maps. In the case of the AUC, the binary maps are then compared against the ground-truth map using the Hit-rate & False-alarm measures. Each of the comparisons is marked on a Hit-rate & False-alarm graph. A curve is then interpolated between the marked points. The final AUC score is the area under the curve.

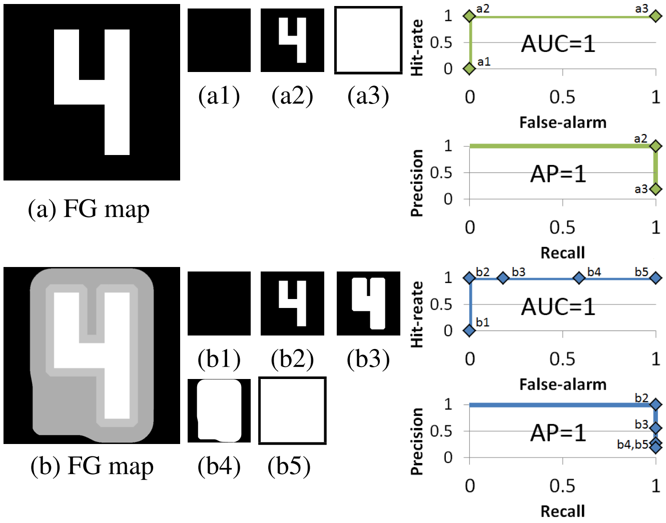


Figure 2. **Interpolation flaw.** Foreground map (a), which is identical to the ground-truth, is better than foreground map (b). (a1-a3), which are the only possible binary maps thresholded from (a), are used to generate the green curves. (b1-b5), which are the only possible binary maps thresholded from (b), are used to generate the blue curves. The curves of (a) and (b) are identical. Therefore, both AUC and AP cannot discern between (a) and (b), and rank them both as perfect.

The AP score is computed in a similar fashion. A curve is interpolated from the Precision and Recall values of the binary maps. The interpolated precision value at each recall level, $r$, is computed as the maximum precision measured at higher recall levels [11]: $p(r) = \max_{\tilde{r}:\tilde{r}\geq r} p(\tilde{r})$. The AP score is computed by averaging the precision values at evenly spaced recall levels.

## 3. Limitations of Current Measures

While the current evaluation measures often perform well, they posses several limitations that hinder their performance. In what follows, we present three assumptions that are the cause for these limitations. We begin by discussing an assumption of AUC & AP (non-binary) and then present two additional assumptions that apply to all four measures (non-binary and binary).

**Interpolation flaw:** Both AUC and AP assume that the interpolated curve (between binary maps) is a valid tool for evaluating non-binary maps. Figure 2 demonstrates why this assumption is inaccurate. (a) and (b) present two foreground maps to be evaluated. (a) is identical to the ground-truth, so it should be scored as much better than (b). Sur-

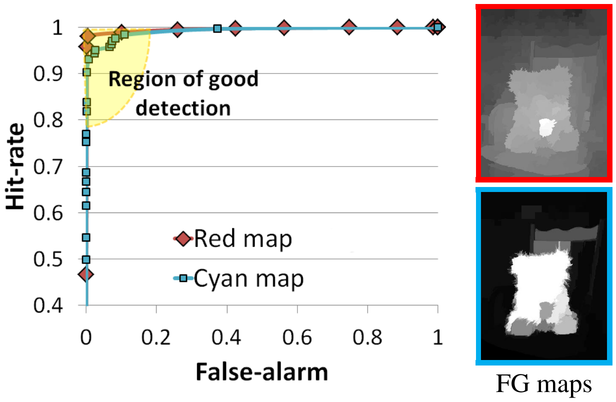


Figure 3. **Interpolation flaw.** These AUC curves are generated for the cyan and the red foreground maps. Since the score relies solely on the interpolated curve and not on the location of the points used to create it, it incorrectly ranks the red map as better.

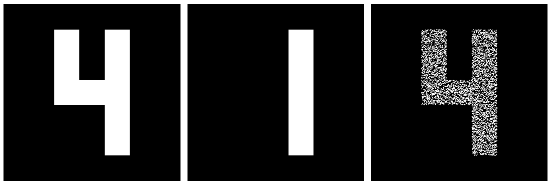
prisingly, both maps obtain a perfect score by both AUC and AP.

To understand why this happens, note that for Figure 2(a), there are only three possible unique binary maps that can be extracted (by setting thresholds) and plotted on the graph. Differently, for Figure 2(b), there are five possible unique binary maps that can be extracted and plotted. In both cases, however, the resulting interpolated curves are identical. Since both AUC and AP rely solely on the interpolated curve, ignoring the distribution of points along the curves, they deem (b) as perfect as (a).

A more realistic example is presented in Figure 3, which presents the AUC curves of the cyan and the red maps of Figure 1. These maps are the results of state-of-the-art saliency detection algorithms [9, 13]. Intuitively, the cyan map is better than the red, since it is much less fuzzy. Furthermore, when using these maps as priors in three different applications (image retrieval, object detection and segmentation – Section 5), the cyan map produced better results. However, both AUC and AP rank the red map as better. This is since they ignore the location of the points in the graph. Both are blind to the fact that many of the binary maps, obtained from the cyan map, have both high Hit-rate and low False-alarm (the region of good detection; see Figure 3). It is important to note that the difference in point distribution along the curves between the cyan and red curves, would not change regardless of the chosen thresholding intervals.

The interpolation flaw applies solely to the measures of non-binary maps. We next describe two more flaws that apply to the evaluation of both binary and non-binary maps.

**Dependency flaw:** Current measures assume that the pixels are independent of each other. Figure 4 demonstrates why this assumption may be wrong. Both Figures 4(a) and 4(b) have identical $TP, TN, FP$ and $FN$ values. Hence, they get the exact same score by all current evaluation measures. However, the false-negatives in Figure 4(a) are concentrated, thus a whole piece of the foreground is not detected at all. Conversely, in Figure 4(b) the false-negatives are sparsely scattered among the true-positives, hence, the entire object is sampled. For most applications, the maps in Figures 4(a) and 4(b) are not of the same quality and should not receive the same score.



Ground-truth (a) FG map (b) FG map

Figure 4. **Dependency flaw.** (a-b) are two binary maps with the same $TP, TN, FP$ and $FN$ values. Current measures consider each pixel as independent. Hence, they ignore the fact that the false-negatives in (b) are sparsely spread within true-positive detections, thus offering a good sampling of the foreground region.

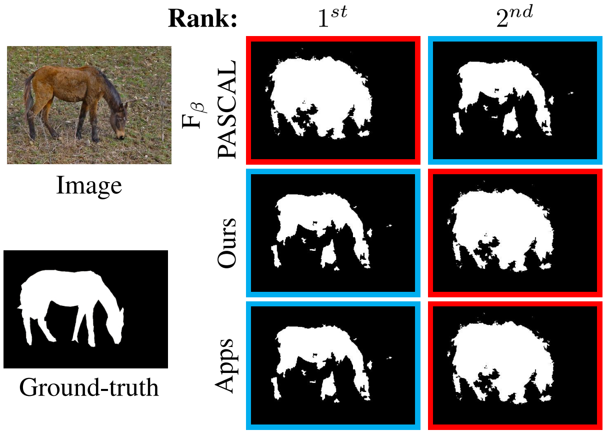


Figure 5. **Dependency flaw.** Based on three applications (“Apps” – Section 5), the detection offered by the cyan map is superior to that of the red. However, both $F_\beta$-measure and PASCAL rank the red map as higher. By incorporating pixel dependency, our measure correctly ranks the cyan map as higher.

Figure 5 illustrates another case of the dependency flaw, this time on a real-world example. The cyan map contains false-negatives that are mostly in regions of true-positive detections, thus offering a good sampling of the foreground region. Conversely, while the red map has more true-positive detections, it also has numerous false-positive detections. When using these maps as priors in three different applications (“Apps” – Section 5) the cyan map produced the best results. Yet, both PASCAL and the $F_\beta$-measure rank the red map higher than the cyan map.

**Equal-importance flaw:** The last assumption made by all the current measures is that all erroneous detections have

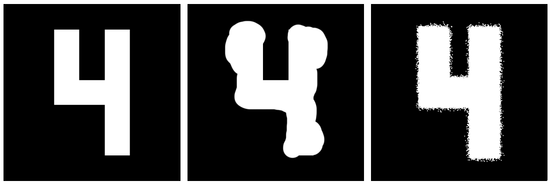


Figure 6. **Equal-importance.** (a-b) are two binary maps with the same number of false-positives. Current measures consider all the false-positives as equally important, deeming the errors in (b), which are near the foreground boundary, as equally harmful as the errors in (a). However, (b) is less damaging for many applications.

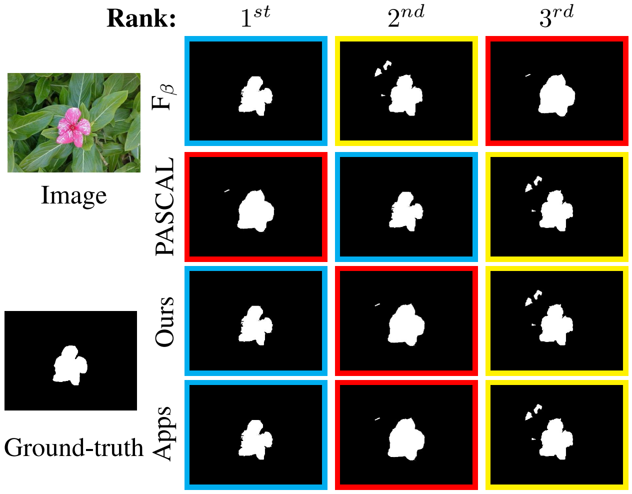


Figure 7. **Equal-importance flaw.** According to three applications ("Apps" – Section 5), the cyan map is the best, followed by the red and yellow maps. Due to the equal-importance assumption, both the $F_\beta$-measure and PASCAL result in different rankings.

equal importance. Figure 6 demonstrates why this may not be the case. Both foreground maps in Figures 6(a) and 6(b) have identical $TP, TN, FP$ and $FN$ values, and hence are scored as equal by all current evaluation measures. However, the false-positive detections in Figure 6(b) are located near the foreground boundary and hence are often less damaging for many applications.

This behavior is further presented on a real-world example in Figure 7. When using the cyan, red, and yellow maps as priors in three different applications ("Apps" – Section 5), the cyan map was ranked as best, the red map second, and the yellow map last. This is since the cyan map is the most accurate with the least number of false-positives. The red and the yellow maps have a similar number of false-positive detections, but the false-positive detections in the red map are near the foreground. This is ignored by the $F_\beta$-measure, which scores the yellow map as better than the red.

The PASCAL measure aims to resolve this by using a 5-pixel wide *don't care* band around the foreground. It thus correctly ranks the red higher than the yellow. However, due to the *don't care* band it also erroneously ranks the red higher than the cyan.

## 4. Our Solution

In this section we present a novel measure to evaluate both non-binary and binary maps. By amending the three assumptions presented in Section 3, our measure provides better evaluation than previous measures (Section 5).

### 4.1. Resolving the Interpolation Flaw

The source of the interpolation flaw is the thresholding of the non-binary maps. We propose a simple solution that avoids this stage by evaluating the non-binary map directly. The key idea is to extend the four basic quantities: $TP, TN, FP$ and $FN$, to deal with non-binary values. These quantities will later be the basis of our measure.

Let $G_{1\times N}$ denote the column-stack representation of the binary ground-truth, where $N$ is the number of pixels in the image. Let $D_{1\times N}$ denote the non-binary map to be evaluated against the ground-truth.

Previously, the four basic quantities were defined solely for the case of a binary map. Each pixel $i$ of the map $D$ was classified as either correct ($D(i) = G(i)$) or incorrect ($D(i) \neq G(i)$). To extend these definitions to the case of non-binary maps, we allow for pixels to be partially correct. Therefore, instead of summing the number of pixels that are correct or incorrect, we sum the partial correctness or incorrectness. Our four basic quantities are defined as follows:

$$\begin{aligned} TP' &= D \cdot G \\ TN' &= (1-D)\cdot(1-G) \\ FP' &= D \cdot (1-G) \\ FN' &= (1-D)\cdot G. \end{aligned} \tag{6}$$

By basing our measure on these quantities, we do not need to threshold the non-binary map into multiple binary maps, thus avoiding the interpolation flaw. Note that when $D$ is binary, these definitions are identical to the conventional ones.

### 4.2. Resolving the Dependency Flaw & the Equal-Importance Flaw

Recall that the remaining assumptions deal with detection **errors**. The first deals with the dependency between false-negatives, and the latter deals with the location of the false-positives. Our key idea is to attribute different importance to different errors.

We start by reformulating the basic quantities in terms of errors. Equation (6) is rewritten as functions of the errors in detection. Let $E_{1\times N}$ denote the absolute error of detection:

$$E = |G - D|. \tag{7}$$

Recalling that $G$ is binary, we can rewrite the quantities of Equation (6) as follows:

$$\begin{aligned} TP' &= (1-E)\cdot G \\ TN' &= (1-E)\cdot(1-G) \\ FP' &= E\cdot(1-G) \\ FN' &= E\cdot G. \end{aligned} \tag{8}$$

By writing the quantities in this form, one can see that indeed all errors are of equal importance, regardless of any dependency or location constraints.

We suggest applying a weighting function to the errors, such as to take into consideration both the dependency between pixels and the location of the errors. Our weight function consists of two components, a matrix, $\mathbb{A}_{N\times N}$, which captures the dependency between pixels and a vector $\mathbb{B}_{N\times 1}$, which represents the varying importance of the pixels. We incorporate this by weighting the error map:

$$E^w = \min(E, E\mathbb{A})\cdot\mathbb{B}, \tag{9}$$

where $\min$ is a per-element minimum function. By multiplying the error map, $E$, by the matrix $\mathbb{A}$ and then taking the minimum between the two, we never increase the error, only reduce it. $\mathbb{B}$ then re-weights the result according to the pixels' location. The basic quantities are now redefined as:

$$\begin{aligned} TP^w &= (1-E^w)\cdot G \\ TN^w &= (1-E^w)\cdot(1-G) \\ FP^w &= E^w\cdot(1-G) \\ FN^w &= E^w\cdot G, \end{aligned} \tag{10}$$

Note that when independency and equal-importance are assumed (i.e. $\mathbb{A} = I$ and $\mathbb{B} = 1$), these definitions are identical to the conventional ones.

In what follows we elaborate on the dependency matrix $\mathbb{A}$ and the importance vector $\mathbb{B}$ that were found to work well.

**Incorporating pixel dependency – $\mathbb{A}$.** The matrix $\mathbb{A}$ should capture the dependency between foreground pixels. We wish for the dependency between pixels to be based on their relative Euclidean distance. The smaller the distance, the greater the impact should be. We realize this via a Gaussian weight as follows:

$$\mathbb{A}(i,j) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{d(i,j)^2}{2\sigma^2}} & \forall i,j \quad G(i)=1, G(j)=1 \\ 1 & \forall i,j \quad G(i)=0, i=j \\ 0 & otherwise \end{cases} \tag{11}$$

where $d(i,j)$ is the Euclidean distance between pixel $i$ and pixel $j$. $\sigma^2$ controls the influence of pixels that are farther away. The larger $\sigma^2$ is, the greater the influence of distant pixels. To mostly rely on the close neighborhood of a pixel, we used $\sigma^2 = 5$ as the default value for all our results.

**Incorporating pixels of varying importance – $\mathbb{B}$.** The vector $\mathbb{B}$ should assign importance weights to false detections according to their distance from the foreground. Thus, we define $\mathbb{B}$ as:

$$\mathbb{B}(i) = \begin{cases} 1 & \forall i, G(i)=1 \\ 2 - e^{\alpha\cdot\Delta(i)} & otherwise \end{cases} \tag{12}$$

where $\Delta(i) = \min_{G(j)=1} d(i,j)$. The constant $\alpha$ determines the decay rate. A value of $\alpha = \frac{\ln(0.5)}{5}$ was found to perform well. This results in a minimal weight of ~1 given to false-positives ($FP$) that are adjacent to the foreground, and a weight of ~1.5 given to $FP$ that are located at a distance of 5 pixels. Note that the exponent was shifted to generate values in the range [1,2), for numerical convenience.

### 4.3. The New Measure – $F^w_\beta$-measure

Having dealt with all three flaws, we proceed to construct our evaluation measure. We follow the methodology of the F$_\beta$-measure, replacing the four quantities with our weighted quantities of Equation (10). We define *weighted Precision*, which is a measure of exactness, and *weighted Recall*, which is a measure of completeness:

$$\text{Precision}^w = \frac{TP^w}{TP^w + FP^w} \quad \text{Recall}^w = \frac{TP^w}{TP^w + FN^w}. \tag{13}$$

Finally, we define the *weighted $F^w_\beta$-measure* as:

$$F^w_\beta = \left(1+\beta^2\right)\frac{\text{Precision}^w\cdot\text{Recall}^w}{\beta^2\cdot\text{Precision}^w + \text{Recall}^w}. \tag{14}$$

Here, similarly to the F$_\beta$-measure, $\beta$ signifies the effectiveness of detection with respect to a user who attaches $\beta$ times as much importance to Recall$^w$ as to Precision$^w$.

## 5. Experimental Validation

One of the most difficult tasks in devising an evaluation measure, is proving its quality. Inspired by [22], we adopt the *meta-measure* (a measure that evaluates measures) methodology. Each meta-measure is based on an expected property of an evaluation measure, which it verifies. We employ four meta-measures, based on the following properties:

1. The ranking of an evaluation measure should agree with the preferences of an application that uses the map as input.

2. A measure should prefer a good result by an algorithm that considers the content of the image, over an arbitrary map [22].

3. The score of a map should decrease when using a wrong ground-truth map [22].

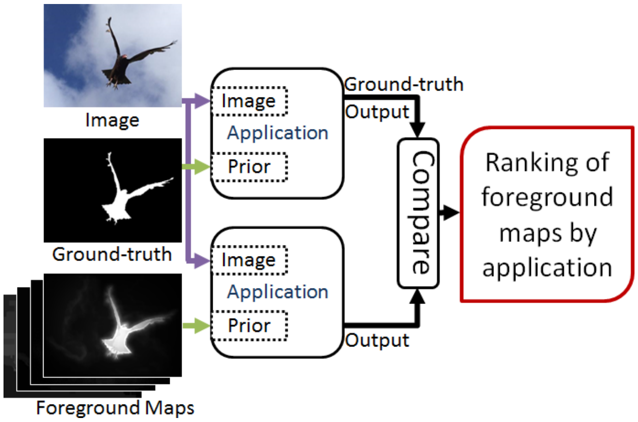


Figure 8. **Application Ranking:** To rank foreground maps according to an application, we compare the output achieved when using the ground-truth, to the output when using the foreground map. The closer the foreground is to the ground-truth, the closer its application output should be to the ground-truth output.

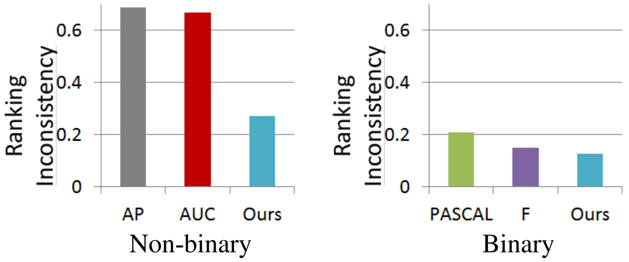


Figure 9. **Meta-measure 1 – results:** The ranking correlation of an evaluation measure to that given by the image retrieval application. The results presented are $1 - \rho$ ($\rho$ denoting Spearman's Rho measure). The lower the score, the better an evaluation measure is at predicting the preference of the application. Our measure offers improvement over the other measures.

4. The ranking of an evaluation measure should not be sensitive to inaccuracies in the manually marked boundaries in the ground-truth maps.

All of our meta-measures were examined on the ASD dataset [1], which consists of 1000 natural images with binary ground-truth maps (similar results were found on the SOD dataset [20]). Binary and non-binary foreground maps were generated for each image using five state-of-the-art algorithms for salient object detection [9, 10, 12, 13, 19] (binary maps are obtained by thresholding the non-binary maps).

### 5.1. Meta-Measure 1: Application Ranking

Our first meta-measure examines the ranking correlation of the evaluation measure to that of an application that uses foreground maps. We assume that the ground-truth map is the optimal prior for the application (upper path in Figure 8). Then, given a foreground map, we compare the application's output (lower path in Figure 8) to that of the ground-truth output. The more similar a foreground map is to the ground-truth map, the closer its application's output should be to the ground-truth output. The ranking of the foreground maps is determined by the similarity of their output to that obtained when using the ground-truth. Finally, the first meta-measure compares the ranking by each evaluation measure: AP, AUC, PASCAL, $F_\beta$-measure and ours, to the ranking by the application.

We examined three applications: image retrieval, object detection and segmentation. Similar results were found in all three applications. For lack of space, Appendix A discusses the realization of only one application: context-based image retrieval. The realization of the other application was performed similarly.

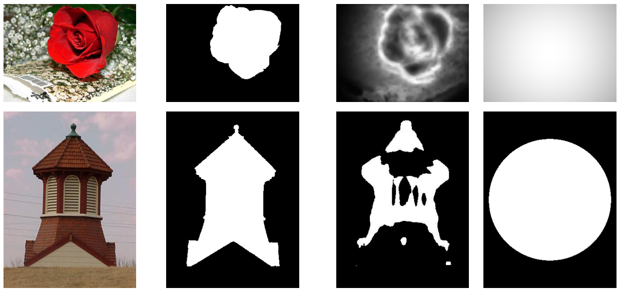
We performed this experiment using the results of five state-of-the-art algorithms [9, 10, 12, 13, 19]. The results on the 1000 images of the ASD dataset [1] are shown in Figure 9. 1−Spearman's Rho measure [5] was used to assess the ranking accuracy of the measures. A score of 0 is given to evaluation measures that ranked the detection algorithms identically to that of the application. A score of 2 is given to measures that ranked the foreground maps in a complete reversed order. In the case of non-binary maps, we can see a great improvement over the previously used AUC and AP measures. Some improvement is also achieved for binary maps, when compared to PASCAL and $F_\beta$-measure. Figures 5 and 7 illustrate several examples of how our measure better predicts the preference of these applications.

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

The property on which we base our second meta-measure is that an evaluation measure should prefer a result obtained by a state-of-the-art method over a map created without taking into account the content of the image. We use a centered Gaussian and centered circle as generic maps that do not consider the content of the image.

Two examples are provided in Figure 10, one non-binary and the other binary. We expect the evaluation measure to score the result obtained by the state-of-the-art algorithm in Figure 10(c) higher than the generic Gaussian or circle maps in Figure 10(d). Yet, currently used measures prefer the generic results. Conversely, our measure correctly ranks the state-of-the-art result higher.

We examined the number of times a generic map scored higher than the mean score obtained by the five state-of-the-art algorithms [9, 10, 12, 13, 19]. (The mean score provides robustness to cases in which a specific algorithm produces a poor result.) Figure 11 summarizes the results: the lower the score, the better the measure is. Our measure outperforms the current methods of both non-binary and binary measures. This is thanks to our consideration of the neighborhoods of detections and their location.



(a) Image (b) Ground-truth (c) Algorithm (d) Generic

Figure 10. **Meta-measure 2:** A measure should prefer the result of a state-of-the-art algorithm (c), over a result obtained without taking into account the content of the image (d). Surprisingly, all of the current measures prefer the generic result. Only our measure correctly ranked (c) higher than (d).

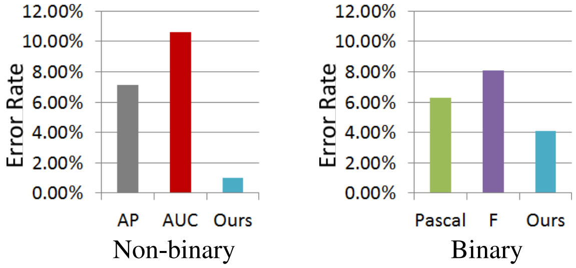


Figure 11. **Meta-measure 2 – results:** The percent of times (tested on 1000 images) that an evaluation measure scored a generic map (centered gaussian or circle) higher than the results of the state-of-the-art algorithms. The lower the score, the better. Our measure outperforms the other measures.

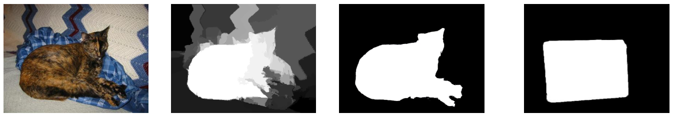
### 5.3. Meta-Measure 3: Ground-truth Switch

The third meta-measure assumes that a good result should not get a higher score when switching to a wrong ground-truth. A foreground map is considered as "good" when it scores at least 0.5 out of 1 (when compared to the original ground-truth map).

In Figure 12 we expect that evaluating the foreground map in (b) against the ground-truth in (c) would produce a higher score than when switching the ground-truth to (d). However, both AUC and AP score otherwise.

Quantitative results of the rate of incorrectly increasing a detection's score when using a wrong ground-truth map are reported in Figure 13. For each of the 1000 images, 100 random ground-truth switches were performed. The lower the score, the better a measure can correctly match between a good foreground map and its true ground-truth map. Since our method directly evaluates the non-binary maps, it outperforms both the AUC and AP measures.

In the case of binary maps, we found that all three methods performed well with respect to this meta-measure ($F_\beta$-measure with $0.02\%$, and Pascal & ours with $\sim 0\%$).



(a) Image (b) FG map (c) Ground-truth (d) Switched GT

Figure 12. **Meta-measure 3:** The score of a foreground map should decrease when using a wrong ground-truth map. Yet, both AUC and AP gave the map in (b) a higher score when using (d) instead of (c) as the reference ground-truth map. Using our measure, the score of (b) appropriately decreased when switching to (d).

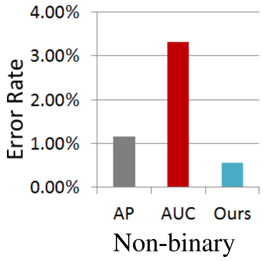


Figure 13. **Meta-measure 3 – results:** The percent of times (tested on 1000 images) that a measure increased a foreground map's score when an incorrect ground-truth map was used. The lower the score, the better. Our measure outperforms both AUC and AP.

### 5.4. Meta-measure 4: Annotation errors

Our fourth meta-measure is inspired by the PASCAL VOC Challenge [11], which uses a *don't care* band around the borders of the manually annotated data, to decrease the effect of slight annotation inaccuracies. We assert that the rankings of given foreground maps should not change much with small inaccuracies in the ground-truth maps.

To realize this meta-measure, we generate a slightly modified ground-truth map by applying morphological operations. Figure 14 illustrates an example. While the two ground-truth maps in (b) & (c) are almost identical, both the AUC & AP switched the ranking between the two foreground maps when using (b) or (c). Conversely, our measure consistently ranked the map in (d) higher than (e).

To offer a quantitative assessment of the change in ranking we used $1-$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. The results are shown in Figure 15. Our measure outperforms both the AP and the AUC. It also offers a slight improvement over the $F_\beta$-measure and PASCAL, which score $0.023$ and $0.025$ respectively, compared to $0.022$ scored by our measure.

## 6. Conclusion

In this paper, we analyzed the currently-used evaluation measures that compare a foreground map against a binary ground-truth. We showed that they suffer from three flawed assumptions: interpolation, dependency and equal-

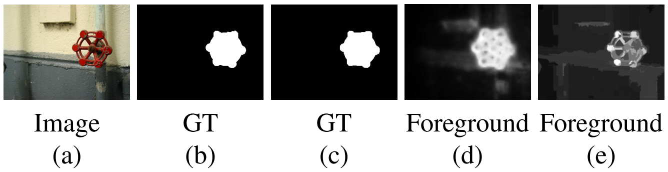


Figure 14. **Meta-Measure 4:** The ranking of an evaluation measure should not be sensitive to inaccuracies in the manually marked boundaries in the ground-truth maps. While ground-truth maps (b) & (c) differ slightly, both AUC and AP switched the ranking order of the two foreground maps (d) & (e), depending on the ground-truth used. Our measure consistently ranked (d) higher than (e). Best viewed on screen.

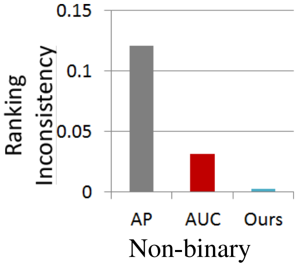


Figure 15. **Meta-measure 4 – results:** The ranking consistency of an evaluation measure under small annotation inaccuracies. The results presented are $1-\rho$ of Spearman's Rho measure. The lower the score, the better.

importance. We further suggested an evaluation measure that amends these assumptions. Our measure is based on two key ideas. The first is extending the basic quantities ($TP, TN, FP$ and $FN$) to non-binary values. The second is weighting errors according to their location and their neighborhood. Based on these, our measure can be defined as a weighted $F^w_\beta$-measure. An additional benefit of our measure is offering a unified solution to the evaluation of non-binary and binary maps. The advantages of our measure were shown via four different meta-measures, both qualitatively and quantitatively.

**Acknowledgments:** This research was funded (in part) by the Intel Collaborative Research Institute for Computational Intelligence (ICRI–CI), Minerva, the Ollendorff Foundation, and the Israel Science Foundation under Grant 1179/11.

## References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 1, 6

[2] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*, pages 1–8, June 2007. 1

[3] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, pages 3378–3385, 2012. 1

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5), 2011. 1

[5] D. J. Best and D. E. Roberts. Algorithm AS 89: The upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society*, 24(3):377–379, 1975. 6

[6] A. Blake, C. Rother, M. Brown, P. Pérez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, pages 428–441, 2004. 1

[7] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012. 1

[8] A. Borji, D. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, pages 414–429, 2012. 2

[9] K. Chang, T. Liu, H. Chen, and S. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, pages 914–921, 2011. 1, 3, 6

[10] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011. 1, 6

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010. 1, 2, 7

[12] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, 2010. 1, 6

[13] 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*, volume 3, page 7, 2012. 1, 3, 6

[14] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, pages 542–549, 2012. 1

[15] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, MIT, 2012. 1

[16] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2:1–19, 2006. 8

[17] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *PAMI*, pages 1–8, 2010. 1

[18] M. Lux. Content based image retrieval with LIRE. In *ACM International Conference on Multimedia*, 2011. 8

[19] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, pages 1139–1146, 2013. 1, 6

[20] V. Movahedi and J. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, pages 49–56, 2010. 6

[21] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 1

[22] J. Pont-Tuset and F. Marqués. Measures and meta-measures for the supervised evaluation of image segmentation. In *CVPR*, pages 2131–2138, 2013. 2, 5

[23] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*. 1

[24] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42, 2012. 1

## A. Meta-Measure 1: Application Realization

Content-based image retrieval finds for a given query image the most similar images in a dataset [16]. The similarity is determined by various features such as color-histograms, histograms of oriented gradients (HOG), and Gabor responses. We used LIRE [18], a publicly available image retrieval system with 12 different features, weighted according to the foreground maps. For each image we used LIRE to retrieve an ordered list of the 100 most similar images. The ground-truth output is the ordered list returned when using the ground-truth map. The comparison between the ground-truth output to that of a foreground map is performed using Spearman's Rho measure.