结构度量：一种新的评估前景图方法

范登平[1]　程明明[1]*　刘云[1]　李涛[1]　Ali Borji[2]
[1] 南开大学，媒体计算实验室　[2] 美国中佛罗里达大学
http://dpfan.net/smeasure/

Abstract

前景图的度量对于物体分割算法的发展有着重要的作用，特别是在显著物体检测领域，其目的是在场景中精确地检测和分割出最显著的物体。而当前几个广泛应用的评测非二进制显著性映射图*(SM)*和标准映射图*(GT)*之间的相似性的指标，比如曲线下面积*(AUC)*，平均精度*(AP)*和最近提出的F_β^ω*(Fbw)*，都是基于像素级误差的，并且通常忽略了结构相似性评测。然而，行为视觉研究显示，人类视觉系统对场景中的结构非常敏感。本文提出了一种新颖、高效且易于计算的指标*(**Structure-measure**)*来度量非二进制的前景映射图。本文的新指标在*SM*和*GT*图之间同时度量面向区域和面向物体的结构相似性。我们在*5*个基准数据集上采用*5*个元度量来证明新提出的指标优于现有的指标。

1. 引言

预测的前景映射图和标准手工标注映射图之间的度量对于衡量和比较计算机视觉应用中的各种算法如：对象检测[6, 25, 8, 41]、显著性检测[5, 20, 43]、图像分割[42]、基于内容的图像检索[12, 15, 22]、语义分割[21, 46, 47] 以及图像收集浏览[10, 30, 14] 有着重要的意义。尽管本文提出的指标可以用于其他目的，作为一个典型的例子，我们将集中于显著物体检测模型[6, 7, 4]的评估上。有必要指出，显著物体不一定是前景物体 [18]。

标准映射图（GT）通常是二进制的（我们假设如此）。前景映射可以是非二进制或二进制的。因此，评估指标可分为两类。第一类是二进制映射图评估，常见的指标有F_β-measure [2, 13, 34]和PASCAL VOC分割指标 [16]。第二类是非二进制映射图评估。包括AUC和AP [16]两种传统的指标以及最新公布的Fbw [37]指标，Fbw指标弥补了AP和AUC指标的缺点（详见Sec. 2）。几乎所有显著物体检测模型的结果都是非二进制的映射图。因此，本文专注于非二进制映射图的评估。

通常我们希望前景映射图应该包含完整的物体结构。因此，我们寄希望于评估指标能告诉我们哪个模型能够得到更加完整的物体。例如，Fig. 1(第一行)，

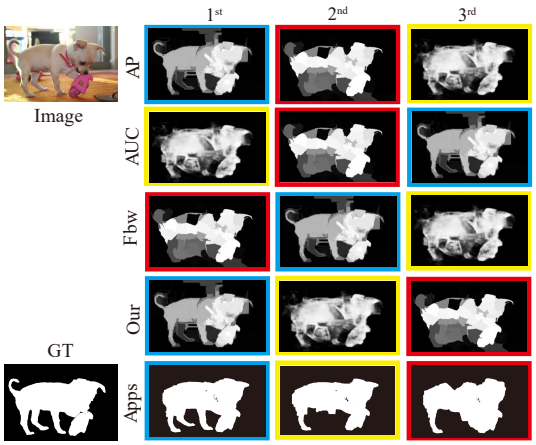


Figure 1. **当前的指标评价不准确.** 我们比较3种最先进的显著物体检测算法产生的显着性映射图的排序：DISC[11]、MDF[28]和MC[49]。根据应用程序的排序（Apps-Sec. 5;最后1行），蓝色边框映射图排在第1位，其次是黄色和红色边框映射图。就GT而言，蓝色边框的映射图最准确地捕获了狗的结构。黄色边框的映射图看起来较模糊，虽然狗的整体轮廓仍然存在。红边映射图则几乎完全破坏了狗的结构。令人惊讶的是，所有基于像素误差（前3 行）的指标都无法正确排序这些映射图，唯独我们新的指标（第4行）以正确的顺序排列了这三张映射图。

*本文为ICCV2017论文 [17]的中文翻译版.

GT 0.9984 0.9915 0.9867 0.9504 0.9438 0.8951 0.8136 0.6605 0.6358 0.5127
(a)rfcn[44] (b)DISC[11] (c)DHS[33] (d)ELD[27](e)MC[49] (f)DCL[29] (g)DRFI[24] (h)DSR[31] (i)MDF[28] (j)ST[35]

Figure 2. 10个显著性检测算法(第1行)作为SalCut [13]输入，输出结果(第2 行)用Structure-measure($\lambda = 0.25, K = 4$)度量。

蓝色边框映射图比红色边框映射更好地捕捉到小狗。对于后者而言，狗的形状已经严重退化到难以从其分割后的映射图中猜测物体的类别了。意想不到的是，目前的所有评价指标都无法正确地对这些映射图进行排序(就从结构的保留的角度)。

我们使用10个最先进的显著性检测模型来获得10个显著性映射图(Fig. 2; 第一行)，然后将这些映射图作为SalCut [13]算法的输入以生成相应的二进制映射图(第2行)。最后，使用我们的结构度量指标对这些映射图进行排名。

指标测量的值越小表示相应的人物的整体结构受到破坏更严重(e-j列)。实验结果清晰的表明我们的指标强调物体的整体结构。在这10张二进制映射图中(第2行)，有6张映射图的结构度量值低于0.95，那么比例就是60%。用相同的阈值(0.95)，发现四个主流的显著性数据集中物体被破坏的比例(例如, ECSSD[48], HKU-IS[28], PASCAL-S[32], and SOD[38])分别为66.80%，67.30%，81.82%和83.03%。而使用F_β measure 来度量这些二进制映射图，这些比例分别为63.76%，65.43%，78.32%和82.67%。这意味着我们的指标比F_β指标在物体结构的度量上更加严格。

为了解决现有方法的问题（即对全局物体结构的敏感度低），我们根据以下两个观察提出了一种结构相似性指标(**Structure-measure**)[1]：

- **区域角度**：虽然很难描述前景映射图的物体结构，但是我们注意到，一个物体的整体结构可以通过组合物体-部分（区域）的结构来很好地表达。
- **物体角度**: 在高质量的显著性映射图中，其前景与背景部分形成了强烈的对比，并且前景和背景部分通常近似于均匀分布。

本文提出的相似性指标由面向区域和面向物体的结构相似性度量两个部分组成。**面向区域**的结构相似性度量试图通过组合所有物体-部分的结构信息来捕捉整体的结构信息。在图像质量评估(IQA)领域中，区域的结构相似性的研究已有很好的研究成果。**面向物体**的结构相似性度量试图比较显著性映射图(SM)和标准映射图(GT)中前景和背景区域的全局分布。

我们采用5个元度量（其中1个由我们引入）在5个公开的基准数据集上进行实验，结果表明我们的指标比其他指标更有效。接下来的一节，我们将回顾一些流行的评价指标。

2. 现有的评价指标

显著性检测模型通常产生非二进制映射图。传统的评价指标通常将这些非二进制映射图转换为多个二进制映射图。

二进制显著图评估: 为了评估二进制显著图，需要从预测的混淆矩阵中计算出四个值：True Positives (TP), True Negatives (TN), False Positives (FP)和False Negatives (FN)。然后将这些值用于计算三个比值：正确率或召回率(TPR)、假阳性率(FPR)以及精度(Precision)。将精度(Precision)和召回率(TRP)结合起来就可以计算出传统的F_β-measure:

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

[1]源代码以及该指标在整个数据集上的评测结果可以在项目主页中找到：http://dpfan.net/smeasure/

非二进制映射图的评估： AUC和AP是两个普遍认可的评价指标。算法产生的结果为非二进制的映射图需要经过3个步骤，来评估模型预测图(非二进制映射图)和人工标注图(GT)之间的一致性。首先，将多个阈值应用于非二进制映射图以获得多个二进制映射图。其次，将这些二进制映射图与二进制的GT映射图进行比较，得到一系列TPR和FPR值。最后，将这些值绘制成2D图，AUC指标就是计算这个曲线下的面积。

AP指标的计算方法类似。它通过绘制精度$p(r)$作为召回率r的函数来获得精度和召回率曲线。AP指标[16]是x轴从$r = 0$到$r = 1$的均匀间隔点的$p(r)$的平均值。

最近，一个名为Fbw[37]的指标对F_β-measure做了直观的概括。它被定义为：

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

Fbw的作者确定了AP和AUC评价指标不准确的三个原因。为了解决这些缺陷：1) 他们将四个基本量TP，TN，FP和FN扩展到非二进制值，2) 根据错误发生的不同位置和以及错误的邻域信息分配不同的权重(w)。虽然Fbw改进了其它指标的缺点，但是有时它也不能给前景映射图一个正确的排序结果(参见Fig. 1的第3 行)。在下一节中，我们将分析为什么当前的指标不能正确地排序这些映射图。

3. 当前指标的局限性

传统的指标(AP、AUC和Fbw)使用四类基本度量(FN、TN、FP和TP)来计算精度(Precision)、召回率(Recall)和FPR。由于这些基本度量都是以逐像素的方式计算的，所得到的基本度量(FN，TN，FP和TP) 不能完全捕捉到预测映射图的结构信息。而在很多应用中通常都需要预测的映射图具有精细的结构细节。因此，评估指标对前景映射图中的结构敏感是有好处的。不幸的是，上述指标(AP，AUC和Fbw)未能达到预期。

Fig. 3(a)中展示了一个典型的例子，其中包含两种不同类型的前景映射图。在SM1中，一个黑色的方块落在数字的内部，而SM2中黑色方块则触及边界。在我们看来，SM2比SM1更受青睐，因为SM1更严重地破坏了前景映射图，但是，在目前的评价指标中两者结果一样。这似乎与我们的常识相矛盾。

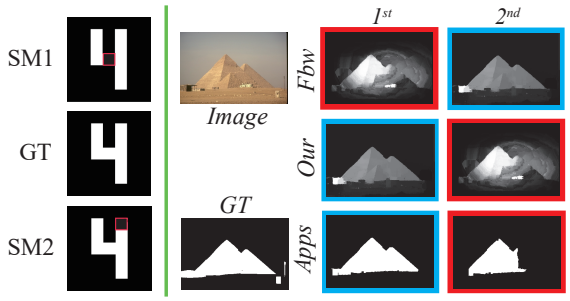


Figure 3. **结构相似性度量**. 在子图(a)中，两个不同的前景图得到相同的FN、TN、FP和TP分数。在子图(b)中，两个映射图由两个显著性模型DSR[31]和ST[35]产生。根据应用程序的排序和我们的用户研究(Apps-Sec. 5;最后一行)，蓝色边框的映射图是最好的，其次是红色边框的映射图。但是由于Fbw指标没有考虑结构相似性，所以得到的排序结果与应用程序排序结果不同。我们的指标(第2行)正确地将蓝色的映射图排名靠前。

Fig. 3 (b)展示了一个更实际的例子。蓝色边框映射图比红边映射图能更好地捕捉到金字塔，因为红色边框的映射图模糊不清，主要突出了金字塔的顶部而忽略了其余部分。从应用的角度(第3行，SalCut输出的显著性映射图;第2行，根据我们的指标排序)，蓝色的边框映射图提供了金字塔的完整形状。因此，如果评价指标不能捕捉到物体的结构信息，它就不能为应用场景中的模型选择上提供一个可靠的信息。

4. 本文的指标

在本节中，我们将介绍我们的新指标来评价前景映射图。在图像质量评价（IQA）领域中，结构相似性指标(SSIM)[45]被广泛用于衡量原始图像和测试图像之间的结构相似性。

设$x = \{x_i|i = 1, 2, \cdots, N\}$和$y = \{y_i|i = 1, 2, \cdots, N\}$分别是SM和GT的像素值。$\bar{x}$, $\bar{y}$, σ_x, σ_y分别是x和y的均值和标准差。σ_{xy}是它们的协方差。从而SSIM即可表示为三个部分的乘积：亮度比较，对比度比较和结构比较。

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

在Equ. (3)中，前两项分别表示亮度比较和对比度比较。两者越接近(例如, $\bar{x}$ and $\bar{y}$, or σ_x and σ_y)，则他

们的比值越趋近于1(即亮度或对比度)。图像中的物体结构与明亮度（受照明和反射率影响）无关。因此，图像结构比较公式的设计应该与亮度和对比度无关。SSIM [45]将两个单元向量$(x-\bar{x})/\sigma_x$ 和$(y-\bar{y})/\sigma_y$结合起来表示两个图像的结构。由于这两个向量之间的相关性等价于x 和y之间的相关系数，所以图像结构比较公式可由Equ. (3)中的第三项表示。

在显著性物体检测领域，研究人员更关心前景物体结构。因此，我们提出的结构度量指标同时考虑了面向区域和面向物体之间的结构相似性度量。面向区域的结构相似性度量和[45]相似，其目的是捕获物体-部分的结构信息，并没有考虑到整个前景部分。面向物体的结构相似性度量的设计主要是为了捕获完整的前景物体的结构信息。

4.1. 面向区域的结构相似性度量

在本节中，我们研究如何度量面向区域的相似性。面向区域相似性的度量旨在评价物体-部分与GT 映射图之间的结构相似性。我们先找到GT的重心，然后沿着该中心点采用水平和垂直分割线将SM和GT 映射图分成4块。与文献 [26]一样，我们接着将每个块递归地分割，最后分块的总数为K。Fig. 4中展示了一个简单的例子。使用Equ. (3)独立地计算每个块的区域相似度$ssim(k)$。用每个块所包含的GT前景区域面积的比例为每个块分配不同的权重(w_k)。因此，面向区域的结构相似性度量可以表达为

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

根据我们的研究表明，我们提出的S_r可以很好地描述SM和GT映射图之间的物体-部分的相似性。我们还试图在整个图像（部分块）级别中采用ssim[45]中提到的滑动窗口方式去度量SM 和GT 之间的相似性，然而这一方式并不能捕获面向区域的结构相似性。

4.2. 面向物体的结构相似性度量

将显著性映射图划分成块可以帮助评价物体-部分的结构相似性。然而，这一面向区域的度量(S_r)并不能很好地表达全局的结构相似性。对于显著性物体检测这一类高级视觉任务来说，物体级别的相似性度量至关重要。为了实现这一目标，我们提出了一种新的方法将前景和背景分开度量。由于GT映射图通常有强烈的前景-背景对比度和均匀分布这两个重要特性。因此，预测的SM也被期望具有这样的特性。这有利于我们更加容易地区分前景与背景。针对这两个特征，我们设计了面向物体的结构相似性度量。

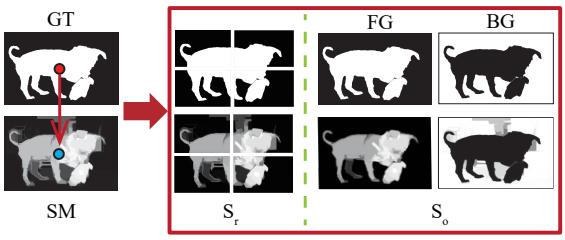


Figure 4. **Structure-measure的框架.**

强烈的前景-背景对比. GT的前景区域与背景区域形成强烈的对比。我们采用与SSIM的亮度比较相似的公式来度量SM的前景区域和GT的前景区域之间接近的平均概率。令x_{FG}和y_{FG}分别表示SM和GT的前景区域的概率值。$\bar{x}_{FG}$和$\bar{y}_{FG}$分别表示x_{FG}和y_{FG}的均值。前景比较可以表示为，

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

Equ. (5) 有几个令人满意的性质:

- 交换$\bar{x}_{FG}$ 和$\bar{y}_{FG}$的值。O_{FG}的结果不变。
- O_{FG}的范围是[0,1]。
- 当且仅当$\bar{x}_{FG} = \bar{y}_{FG}$时，我们会得到$O_{FG} = 1$。
- 但是最重要的性质是，当两个映射图越相似，则O_{FG}越接近为1。

这些性质使Equ. (5) 适用于我们的目的。

均匀的显著性分布. GT的前景和背景区域通常是均匀分布的。因此，给予显著性物体被均匀检测（即，整个物体有相似的显著值）出来的显著图一个更高的评测值是非常重要的。如果SM中前景部分的显著值变化很大，那么它的分布就不均匀。

在概率理论和统计学中，标准差与平均数的比值$(\sigma_x/\bar{x})$称为变异系数，是一个标准的衡量概率分布离散程度的统计量。在这里，我们用它来度量SM的离散程度。换句话说，可以使用变异系数来计算SM和GT之间的不相似度。根据Equ. (5)，SM和GT之间物体级别的总体不相似性可以写成：

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

其中λ是平衡这两个部分的常数。由于GT前景部分的平均概率在实际中恰好为1，所以SM和GT之间物体级别的相似性可以表示为

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

为了计算背景的相似性O_{BG}，我们将背景视为前景的补集，可以用1减去SM和GT映射图，如Fig. 4所示。那么，O_{BG} 可以类似地定义为

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

设μ为GT中的前景区域与图像区域(宽*高)的比值。最后，面向物体的结构相似性度量被定义为，

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

4.3. 本文新的结构度量

有了面向区域和面向物体的结构相似性定义，最后本文的结构度量就可以表示为

$$S = \alpha * S_o + (1 - \alpha) * S_r \tag{10}$$

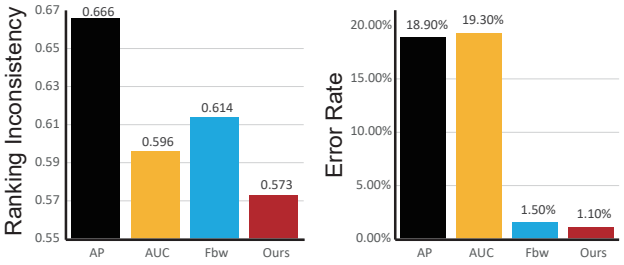
其中，$\alpha \in [0, 1]$。本文实验中设置$\alpha = 0.5$。使用这个指标去度量Fig. 1中的三个SM映射图，就能够得到与应用排序结果一致的正确排序。

5. 实验

为了评测我们的指标的性能，我们采用了Margolin等人[37]提出的4个元度量以及本文提出的1个元度量。这些元度量是用于衡量评价指标的性能[40]。为了公平比较，所有元度量都是在ASD(a.k.a ASD1000) 数据集[1]上计算的。非二进制前景映射图(总共为5000 张)由5个显著性检测模型包括CA[19], CB[23], RC[13], PCA[36], 和SVO[9]得到的。我们在所有实验中设置$\lambda = 0.5$和$K = 4$。在单个线程CPU（4 GHz）上计算一张图像的结构度量，我们的Matlab版本代码平均需要5.3 ms。

5.1. 元度量1: 应用排序

评价指标应与使用SM作为输入的应用程序的偏好一致。我们假设GT图是最适合应用程序的。给定一个SM，我们将应用程序的输出与GT的输出进行比较。



(a) 元度量1　　(b) 元度量2

Figure 5. **元度量1&2-结果.**

SM与GT映射图越相似，其应用程序的输出与GT的输出就越接近。

为了量化排序的准确性，我们使用SalCut [13]作为应用程序来执行此元度量。采用1-斯皮尔曼系数ρ的度量[3]来度量评价指标的排序准确性，其中较小的值表示更好的排序一致性。不同指标之间的比较显示在Fig. 5(a)中，这表明我们的结构指标在其他可选的指标中具有最佳的排序一致性。

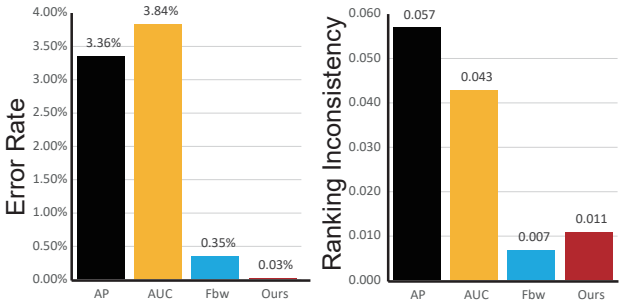
5.2. 元度量2: 最新水平vs. 标准

第二个元度量为一个指标应该优先选择那些采用最先进算法得到的结果而不是那些没有考虑图像内容的通用的基准映射图（例如，中心高斯映射图）。即，一个良好的评价指标应该把由最先进的模型生成的SM排在通用映射图的前面。

我们统计基准映射图得分高于由五个最先进模型(CA[19], CB[23], RC[13], PCA[36], SVO[9])生成的映射图的平均分数的次数，平均分数象征着模型鲁棒性。结果显示在Fig. 5 (b)中。值越小越好。在1000张图像测试中，我们的指标仅仅只有11个错误(即一般映射图胜过最先进算法得到的映射图的次数)。同样的测试中，AP和AUC指标表现非常差，产生了大量的错误。

5.3. 元度量3: 标准映射图替换

第三个元度量规定，当替换成错误的GT映射图时，“好”的SM不应该获得更高的分数。在Margolin [37]等人的论文中，当SM评分大于0.5（用原始GT作为参考）时，SM被认为是“好”。使用该阈值(0.5)，在5000张映射图中有41.8%被认为是“好”的。为了公平比较，我们和Margolin等人一样选择相同百分比的“好”映射图进行试验。在1000张图像中的每一张图像，我们都进行100次的随机GT 映射图替换。然后，我们统计当使



(a) 元度量3 (b) 元度量4

Figure 6. **元度量3&4-结果.**

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

Figure 7. **元度量4: 标注错误.** (a) GT映射图, (b) 另外一个GT映射图, (c) (a)和(b)之间的差异图, (d) 显著性映射图1, (e) 显著性映射图2.

用不正确的GT时，显著性图的分数指标增加的次数百分比。

Fig. 6 (a)结果显示。分数越小，表示指标匹配正确GT映射图的能力越强。我们的指标比排名第二的指标好上10倍。这归功于我们的指标捕获了SM和GT映射图之间的物体结构相似性。由于物体结构在随机GT中发生变化，所以使用随机选择的GT时，我们的度量值将为“好”的SM赋予较低的分数。

5.4. 元度量4: 标注错误

第四个元度量规定，评价指标不应对GT边界手工标注时的轻微错误或者不准确性敏感。为了执行这个元度量，我们通过使用[37]提到的形态操作对GT映射图做微小的改动。Fig. 7中显示了一个示例。(a) & (b)中的两个GT映射图几乎相同，所以在使用(a)或(b)作为参考标准时，指标不应该改变两个显著图之间的顺序。

我们使用1-斯皮尔曼系数来度量引入标注错误前后的排序相关性。分数越低，评价指标对注释错误的鲁棒性就越好 [37]。结果显示在Fig. 6 (b)中。我们的指标优于AP和AUC，但不是最好的。检查这一现象，我们意识到并不总是得分越低评价指标越好。原因是有时“轻微”不准确的手动注释可能会改变GT的结构，从而可能改变排序。我们仔细检查了结构变化的影响。当GT映射图与其形态学变化版本之间的差异图具有连续大块区域时常常引起主要结构的变化。我们试图将

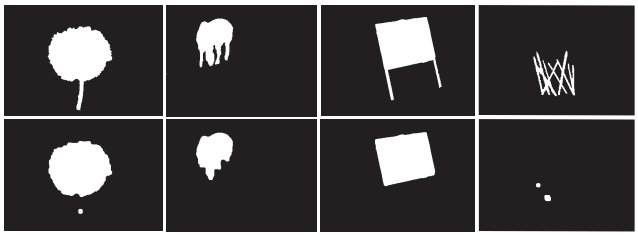
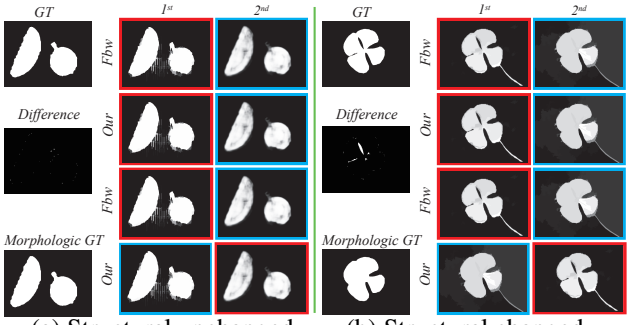


Figure 8. **结构改变示例.** 第1行表示GT映射图。第2行表示其相应的形态学操作后的结果。



(a) Structural unchanged (b) Structural changed

Figure 9. **Structural unchanged/changed.** (a)我们和Fbw的指标对GT边界手工标注的不准确性(结构不变)都不敏感。(b)评价指标的排序应该对结构发生变化敏感令人惊讶的是，目前的最佳指标(Fbw)无法适应结构变化。使用我们的指标，我们可以根据结构改变而正确的改变排序结果。最好在屏幕上观看。

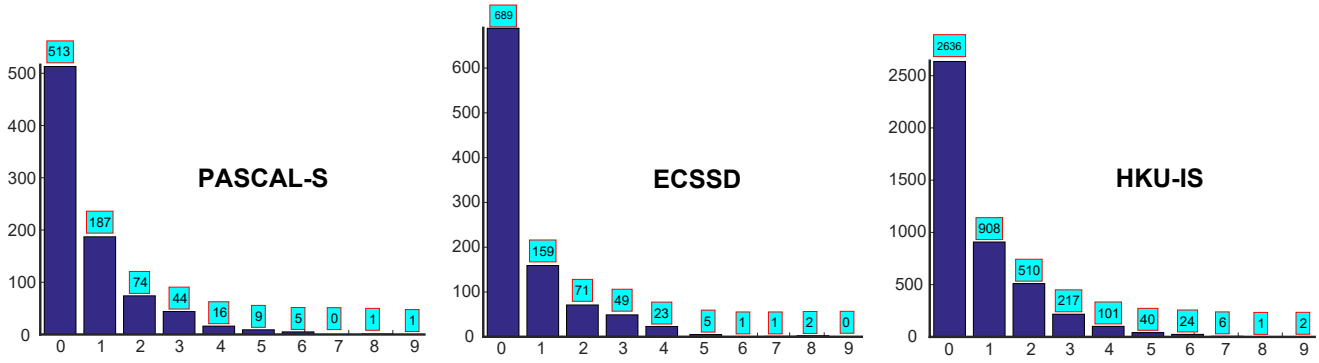
差异图腐蚀后再进行求和作为衡量结构变化的指标，然后根据这些求和结果来排序GT映射图。

在前10%求和结果变化最小的GT映射图中，我们的指标和Fbw具有相同的MM4分数(均为0)。因此，当GT的拓扑结构不变时，我们的指标和Fbw指标可以保持度量的映射图排序不变。从示例Fig. 9 (a)可见。虽然GT映射图与形态学操作的GT映射图略有不同，但是Fbw和我们的指标都能依据所用的GT而保持两个显著图排序不变。

在前10%求和结果变化最大的GT映射图中，我们邀请3位用户判断GT映射图是否具有重大的结构变化。100个GT映射图中有95个被认为具有重大的结构变化。(类似于Fig. 8，例如每组中的小棒，瘦腿，细长脚和细线)，因此，我们认为保持排序不变是不合理的。Fig. 9 (b)证明了这一观点。当我们使用GT映射图作为参考时，Fbw和我们的指标都能正确地排序这两个图。然而，当使用形态学操作后的GT映射图作为参考时，排序结果就不同了。显然，从视觉和结构上看，蓝色边框的SM比红色边框的SM更像形态学操作后的GT映射图。指标应该赋予蓝框的SM更高的分数。所以这两个

Table 1. 当前指标在3个元度量上的定量比较. 最好的结果用蓝色 凸显出来。MM:元度量。

	PASCAL-S[32]			ECSSD[48]			SOD[38]			HKU-IS[28]		
	MM1	MM2(%)	MM3(%)	MM1	MM2(%)	MM3(%)	MM1	MM2(%)	MM3(%)	MM1	MM2(%)	MM3(%)
AP	0.452	12.1	5.50	0.449	9.70	3.32	0.504	9.67	7.69	0.518	3.76	1.25
AUC	0.449	15.8	8.21	0.436	12.1	4.18	0.547	14.0	8.27	0.519	7.02	2.12
Fbw	0.365	7.06	1.05	0.401	**3.00**	0.84	0.384	16.3	0.73	0.498	0.36	0.26
Ours	**0.320**	**4.59**	**0.34**	**0.312**	3.30	**0.47**	**0.349**	**9.67**	**0.60**	**0.424**	**0.34**	**0.08**



(a) (b) (c)

Figure 10. **Fbw与我们指标之间的排序距离.** (a)-(c)是展现了三个数据集上，Fbw和我们的**Structure-measure**之间的排名距离，y轴是图像的数量。x轴是排序距离。

映射图的排序应该改变。Fbw指标未能达到此目的，而我们的指标给出了正确的排序。

上述分析表明，这一元度量不是很可靠。因此在接下来的进一步比较的数据集中我们不把它列入其中。

5.5. 进一步比较

Fig. 5和Fig. 6 (a)中的结果表明，在数据集ASD1000上进行的3个元度量实验，我们的指标获得了最佳的性能。但是，一个好的评价指标应该能够在几乎所有数据集中表现良好。为了证明我们指标的鲁棒性，我们进一步在4个广泛使用的基准数据集上进行了实验。

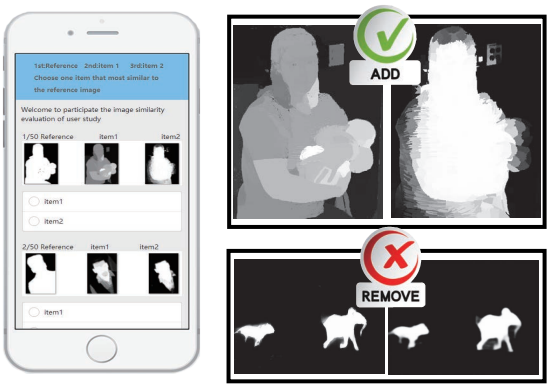
数据集. 使用的数据集包括PASCAL-S[32], ECSSD[48], HKU-IS[28]和SOD[38]。PASCAL-S包含850幅具有挑战性的图像，有多个物体并且背景杂乱。ECSSD包含1000幅具有语义信息但是结构复杂的图像。HKU-IS是另一个大型数据集，其中包含4445幅大尺度图像。该数据集中的大多数图像包含多个显著性物体并且对比度低。最后，我们还对SOD数据集进行了评估，该数据集是BSDS数据集的一个子集。它包含相对较少数量的图像(300)，但具有多个复杂物体。

显著性模型. 我们使用10个最先进的模型，包括3个传统模型(ST[35], DRFI[24], 和DSR[31])和7个基于深度学习的模型(DCL[29], rfcn[44], MC[49], MDF[28], DISC[11], DHS[33], 和ELD[27])来测试我们的指标。

结果. 结果显示在Tab. 1中。根据第1个元度量我们的指标结果最好。这表明在实际应用中我们的指标比其他指标更有效。根据元度量2，除了ECSSD数据集我们的指标排名第2以外，在其他数据集上我们的指标表现优于现有的指标。对于元度量3，我们的指标分别在PASCAL，ECSSD，SOD和HKU-IS中比排名第2的指标的错误率降低了67.62%, 44.05%, 17.81%, 69.23%。这表明我们的指标具有较高的表达SM和GT图之间结构相似性的能力。总而言之，我们的指标在大多数情况下胜出，这清楚地表明我们新的指标比其它指标具有更强的鲁棒性。

5.6. 元度量5: 人的判别

在这里，我们提出了一个新的元度量来评估前景评价指标。该元度量规定，一个评价指标对映射图的排序结果应该与人对映射图的排序结果一致。有学者认为[39]，“人类最适合衡量任何一个分割算法的输出”。然而，由于时间和货币成本，对数据集的所有图像进行主观评价是不切实际的。而且据我们所知，没有符合这些要求的视觉相似度评估数据集。



(a) (b)

Figure 11. 我们的用户调研平台.

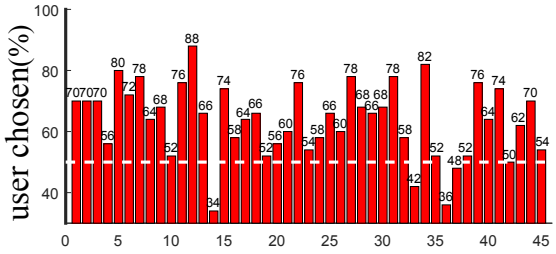


Figure 12. **我们的用户调查结果**. x轴是观察者id，y轴表示观察者选择了我们的指标所选映射图的实验次数百分比。

原始的显著性映射图收集. 原始的显著性映射图采集于三大数据集：PASCAL-S，ECSSD和HKU-IS。如上所述，我们使用10个最先进的显著性模型来为每个数据集生成相应的显著性映射图。因此，我们每个图像都有10个显著性映射图。我们使用Fbw和我们的指标来评估这10张图，分别根据每个指标选择出排序第一的映射图。如果两个指标选择相同的映射图，则其排序距离为0.如果某个指标将一张映射图排第一，而另一个指标把这张映射图排列在第n的位置，那么它们之间的排序距离就是$|n-1|$。Fig. 10 (a), (b) 和(c)展示了这两个指标的排名距离。蓝框表示每个排序距离下的图像数目。排序距离大于0的显著性图被选为我们做用户调研的候选图。

用户调研. 我们从三个数据集中随机选取了100对显著性映射图。在Fig. 11 (b)顶部面板显示了一个示例性试验，其中根据我们指标选择的最佳显著性图在左边，根据Fbw选择的最佳显著性图在最右边。用户被要求选择她认为与GT最相似的映射图。在这个例子中，这两个显著性图是显然是不同的，这使用户很容易做出决定。在另一个例子中(Fig. 11 (b)中的底部面板)，两个显著性映射图太相似了以至于难以选择一个与GT最相似的映射图。因此，我们要避免向受试者显示这种情况。最后，我们留下了包含50对实验映射图的集合。我们开发了一个手机应用程序来进行用户调研。我们收集了来自45位观察者的数据，这些观察者不知道实验目的。观察者的视力正常或被矫正(年龄分布为19-29岁;教育背景从本科到博士; 10类专业，如历史，医药和金融; 25名男性和20名女性)

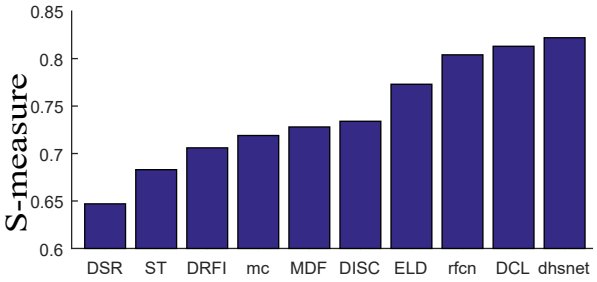


Figure 13. **用我们新的指标对10个显著性模型排序。** y轴表示在每个数据集上的平均得分。(PASCAL-S[32], ECSSD[48], HKU-IS[28], SOD[38]).

结果. 结果显示在Fig. 12中。观察者偏好于我们指标选择的显著性映射图的次数百分比(对所有观察者的结果取平均)为63.689%。我们用同样的方式做了另外两项用户调研实验(AP与我们的指标比较，AUC与我们的指标比较)。结果分别是72.11% 和77.133%。这意味着我们的指标选择更符合用户的选择。

5.7. 显著性模型比较

确定了我们的指标能够更好的度量显著性物体检测模型，这里我们在4个数据集(PASCAL-S, ECSSD, HKU-IS, 和SOD)上比较10个最先进的显著性模型。Fig. 13显示了这10个模型的排名。根据我们的指标，按顺序来说最好的模型是dhsnet,DCL和rfcn。更多模型示例图参见补充材料。

6. 讨论和结论

在本文中，我们分析了当前显著性评价指标是基于像素误差的，并指出它们忽略了结构相似性。然后，我们提出了一种称为**Structure-measure**的新的结构相似性指标，它同时评估显著性图和真值图之间面向区域和面向物体的结构相似性。我们的指标基于两个重要特征：1)强烈的前景-背景对比和2)均匀的显著性分布。更重要的一点是该指标计算简单有效。

在5个数据集上的实验结果表明，我们的指标优于当前的AP、AUC和Fbw指标。最后，我们在包含了100个显著性映射图和50个GT映射图的数据集上进行了用户调研。来自45个受试者的数据表明，相对于AP、AUC和Fbw指标选择的显著性映射图，用户更偏向我们的指标所选择的显著性映射图。总之，我们的指标为显著性物体检测的评估提供了新的思路，而目前的指标未能真正检验显著性模型的优缺点。我们希望显著性领域能在未来的模型评价和比较中考虑这一指标。

致谢 我们要感谢匿名评委为本文提供了宝贵的意见。本研究得到了NSFC(NO. 61572264, 61620106008)，华为创新研究计划，CAST YESS Program和IBM Global SUR award的支持。

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE ICCV*, pages 4548–4557, 2017.

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

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

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

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

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

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

[24] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature

integration approach. In *IEEE CVPR*, pages 2083–2090, 2013.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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