Enhanced-alignment Measure for Binary Foreground Map Evaluation
评估二值前景图的增强匹配指标

范登平[1], 龚成[1], 曹洋[1], 任博[1], 程明明[1] and Ali Borji[2]
[1] 计算机学院, 南开大学，[2] 美国中佛罗里达大学
http://dpfan.net/e-measure/

Abstract

现有的二进制前景图指标以像素或结构方式处理各种类型的错误。这些方法独立地考虑像素级匹配或图像级信息，而认知视觉研究表明人类视觉对场景中的全局信息和局部细节高度敏感。在本文中，我们详细介绍了当前的二值前景图评估方法，并提出了一种新颖有效的**E-measure** (增强匹配指标)[1]。我们的方法将局部像素与图像级平均值相结合，共同捕获图像级统计量和局部像素匹配信息。通过5个元度量来证明我们的方法在4个流行数据集优于当前评估方法，元度量包括为应用程序的排序比较，降低对通用图的偏好，随机高斯噪声，真值图切换以及认为判别。我们发现我们的指标在几乎所有的元度量上都有很大的提升。例如，在应用程序排名方面，与其他流行的方法相比，我们的指标提高了9.08% 到19.65%。

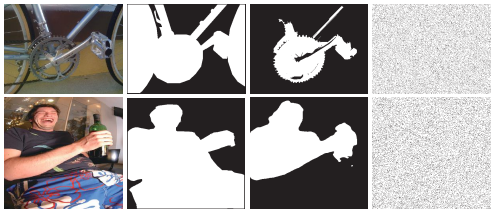
1 Introduction

请看Fig. 1。您将看到二值前景分割模型和随机高斯噪声图的输出。虽然估计的映射图（FM）与真值图（GT）非常接近，但是迄今为止，最常用的测量方法(*e.g.* , IOU [11], F1, 和JI [15])以及最近提出的包括Fbw [27]和VQ [34]在内的方法都认为噪声图比估计的映射图更好。这是本文要解决的问题(见实验部分Sec. 4)。为了解决这个问题，我们提出了比现有方法更好的新指标。

二值前景估计图与人标记二值真值图之间的比较在各种计算机视觉任务中比较常见，如图像分割 [32]，目标检测，识别 [16; 33]， [3]，和 [4; 5; 6; 7]，这对于说明哪个模型更好至关重要。

二值前景图与人标记的二值真值图之间的比较在各种计算机视觉任务中是普遍存在的，例如，图像检索

[1]本文为IJCAI2018论文 [13]的中文翻译版



(a) 图像 (b) 真值图 (c) 前景图 (d) 噪声图

Figure 1: 当前评估指标的不准确性。一个评价指标方法应该赋予最先进算法生成的前景图（c）比随机高斯噪声图（d）更高的分数。然而，目前常见的指标包括 [11], F1/JI [15] , Fbw [27], CM [30], 和VQ [34]都更偏向于噪声图。只有我们的方法正确地将（c）排在（d）之前。

[23], 图像分割 [32], 对象检测和识别 [16; 33], 前景提取 [3], 以及显著对象检测 [14; 4]。这样的比较对于判别那个模型更好至关重要。

三个广泛使用的比较前景图和真值图的指标是F_β measure [1], Jaccard Index (JI) measure [15], 以及intersection over union (IOU) [11]. 过去已经提出了基于F_β-measures [10; 27; 34] 和其他方法(例如, [30; 35; 29]) 的各种指标。但是，所有这些评估都采用了像素相似性的方法，通常忽略了结构相似性。最近，范等人 [12]，提出了一种结构度量方法，达到了最好的性能。然而，该方法是为非二值图像评估而设计的，而且一些函数（比如均匀分布项）并不适用于二值前景图的情况。

相反，我们在这里提出了一种新的方法，叫做**E-measure** (增强匹配指标)。它由一个同时考虑了像素和图像级别属性的表达式组成。我们证明了该方法是评估二值前景图的一种有效和高效的方式。为了说明这个概念，在Fig. 2中展示了一个例子。在三个用颜色（蓝色，红色和黄色）框标记的前景图与真值图的评估中，与3个最新的方法Fbw [27]，VQ [34]和CM [30]比较时，只有我们的方法同时考虑了结构信息与全局形状，因此能够正确地对三个前景图排序。我们通过考虑图像级统计（前

编号	指标	年份	出版方	优点	缺点
1	**IOU/F1/JI** [15]	1901	BSVSN	易于计算	图像级别的统计信息缺失
2	**CM** [30]	2010	CVPRW	同时考虑了区域和轮廓	对噪声敏感
3	**Fbw** [27]	2014	CVPR	为不同的错误分配权重	对错误发生的位置敏感, 计算复杂
4	**VQ** [34]	2015	TIP	通过心理学函数对错误进行权衡	是一个主观评价指标
5	**S-measure** [12]	2017	ICCV	考虑了结构相似性	聚焦于非二值图的特点

Table 1: 当前评估方法的总结.

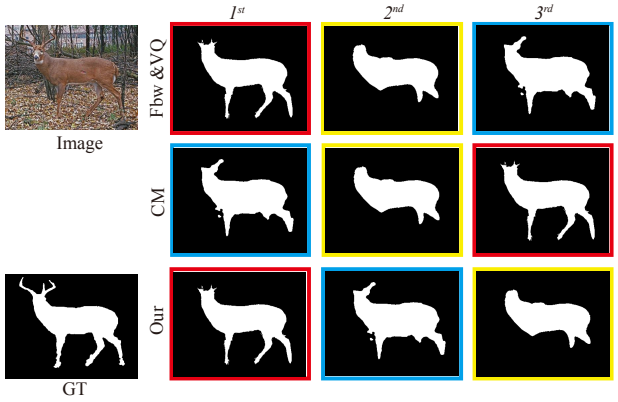


Figure 2: 证明我们的方法的有效性。由3个最先进的显著对象检测模型生成的二值前景图（阈值后）的排名，生成模型包括DCL,rfcn和DHS。3种不同类型的常用的度量方法（CM,Fbw和VQ）都无法正确排列前景图。但是，我们的方法给出了正确的顺序。

景图的平均值）和像素级匹配来实现这一点。我们的方法（将在第3节中详细描述）可以正确地对预测的分割图进行排序。本文的主要贡献如下：

- 我们提出了一个简单的方法仅仅通过一个紧凑项就能够同时捕捉图像级别统计信息和像素级别匹配信息。我们通过实验证明，在4 个流行数据集上使用5个元度量方法，我们的测量显着优于传统测量IOU, F1/JI, CM和最近提出的S-measure，VQ及Fbw。
- 为了评估这些度量，我们还提出了一个新的元度量（最新方法检测的前景图vs 噪声图），并建立一个新的数据集。该数据集包含了555个由人眼进行客观排序的二值前景图。我们使用这个数据集来检查当前度量与人类判断之间的排序一致性。

2 Related Work

Tab. 1中可以找到关于二值前景图评估常用评估方法的总结。接下来，我们解释这些方法并讨论它们的优缺点。

F_β measure [1; 9; 25] 是一个常用的方法，它同时考

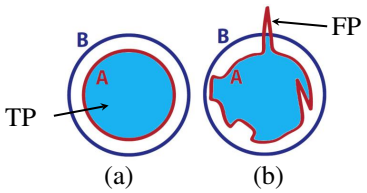


Figure 3: 基于区域的评估方法的局限性。蓝色圆圈表示GT，红色曲线表示FM。根据IOU [11]，F1/JI [15]的方法，与GT圆（蓝色圆形曲线）相比，（b）中的边界几乎与（a）中的边界一样好，即使它有很多尖峰，晃动和形状的差异 [30]。

虑$recall = \frac{TP}{TP+FN}$ & $precision = \frac{TP}{TP+FP}$：

$$F_\beta = \frac{(1+\beta^2)precision \cdot recall}{\beta^2 \cdot precision + recall}, \quad (1)$$

其中β是一个参数来权衡$recall$和$precision$。真正类（TP），真负类（TN），假正类（FP）和假负类（FN）是4个基本量。设置$\beta = 1$即为经典的F1度量：另一个广泛使用的基于F1的测量是Jaccard指数（JI [15]，也称为IOU测量：

$$JI = IOU = \frac{TP}{TP + FN + FP}. \quad (2)$$

F1和IOU的关系为：$JI = \frac{F1}{2-F1}$。shi等人 [34]提出了另一种主观的对象分割评估方法。他们的衡量标准基本上也是基于F1指标。Margolin 等人 [27] 提出了一种称为加权F_β （Fbw）的复杂测量方法：

$$F_\beta^\omega = \frac{(1+\beta^2)Precision^\omega \cdot Recall^\omega}{\beta^2 \cdot Precision^\omega + Recall^\omega}. \quad (3)$$

它为不同位置的错误分配不同的权重。

上述所有方法都与F_β密切相关。他们通过独立地考虑每个像素位置来估计，但是忽略了重要的图像级别的信息，这导致了在识别不同形状(Fig. 3), noise (Fig. 1)，噪声(Fig. 1)和结构错误(Fig. 2) 方面的糟糕表现。

Movahedi等人 [30]提出了轮廓映射（CM）测量。然而，这种测量对噪声敏感（见Fig. 1），导致性能不佳，特别是使用后面所述的元度量3 （Sec. 4.3和Tab. 2）。最近提出的称为S-measure[12]的测量方法，侧重于非二值前景图（FM）评估。它考虑了分割图上$2 * 2$ 网格上的区域级结构相似性和对象级属性（例如，均匀度和对比度）。但是，这些属性不适用于二值映射。

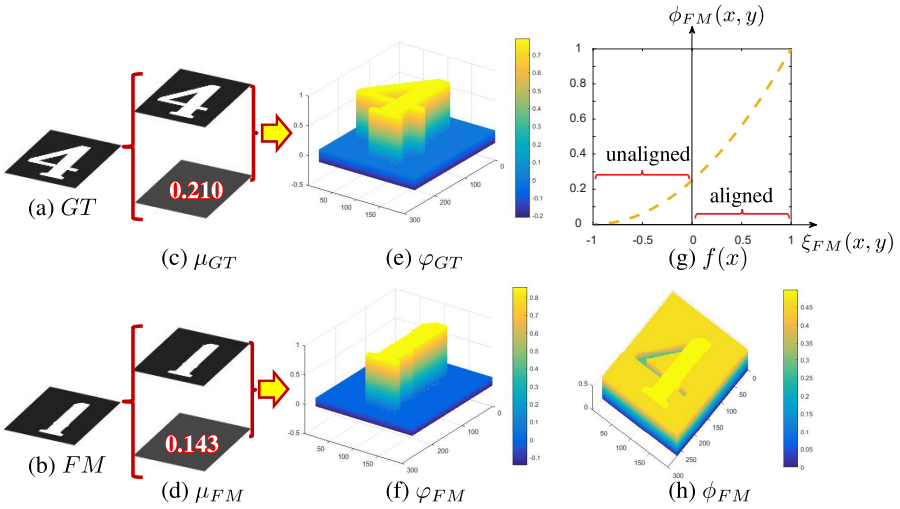


Figure 4: 本文E-measure的框架图. (a) 真值图GT。(b) 评估前景图FM。(c)和(d) GT和FM的均值。(e) 和(f) 是由公式 (Eq. 4)计算的偏差矩阵。(g) 是非线性映射函数。(h) 根据公式 (Eq. **??**)计算的对齐矩阵。'align'和'unaligned'分别表示$GT(x;y) = FM(x;y)$和$GT(x;y) \neq FM(x;y)$时的位置

3 The Proposed Measure

在本节中，我们将解释我们评估二值前景图的新方法的细节。我们方法的一个重要优势是其简单性，因为它包含一个同时捕获**全局统计信息**和**局部像素匹配**信息的紧凑项。因此，我们的方法比现有的最好的方法表现更好。Fig. 4中展示了我们的算法过程。

3.1 Motivation

尽管二值映射评价之前已经取得一些成功，但最近的评价方法如S-measure在二值映射上仍然表现不佳。通常的情况是，这些评价方法评价二值通用映射图的分数高于最先进的（SOTA）模型的输出结果（参见Sec. 4.2）。背后的原因在于，在二值图中，S-measure强调亮度比较，对比度比较和离差率。然而，虽然为非二值映射计算这些项是有意义的，其值为实数，范围为[0,1] 并且将值视为表示前景拥有像素的概率，但在二值映射中这样的属性是没有明确定义且不太有效。因此，使用连续假设可能导致对二值映射的错误评估。

认知视觉研究表明，人类视觉系统对场景中的结构（例如全局信息，局部细节）高度敏感。因此，在评估FM和GT之间的相似性时，同时考虑局部信息和全局信息是很重要的。

基于上述观察，我们设计了一种适合二值映射评估的新方法。我们的方法在二值图上有明确的定义，并且在一个单元中组合了局部像素值和图像级平均值，这有助于联合捕获图像级统计和局部像素匹配信息。实验（Sec. 4.2）表明，我们的测量比二值图上的其他指标表现更好。

3.2 Alignment Term

为了设计一个同时捕捉全局统计量和局部像素匹配信息的紧凑项，我们定义一个*偏差矩阵* φ 作为输入二值映射I 的每个像素值与它的全局平均值μ_I之间的距离：

$$\varphi_I = I - \mu_I \cdot \mathbb{A}, \tag{4}$$

其中，$\mathbb{A}$ 是一个矩阵，其中所有元素值都是1，它的尺寸与输入I相同。我们分别为二值真值图GT和二值前景图FM计算偏差矩阵φ_{GT}和φ_{FM}。$I \in \{GT, FM\}$。通过从信号中去除平均强度，可以将偏差矩阵视为信号中心。它可以消除由于内在变化或大的数值差异引起的误差。

我们的偏差矩阵与亮度对比度有很强的关系[38]。因此，我们将φ_{GT}和φ_{FM}之间的相关性（Hadamard乘积）视为量化偏差矩阵相似性的简单有效的量度。因此，我们将**对齐矩阵**ξ定义如下：

$$\xi_{FM} = \frac{2\varphi_{GT} \circ \varphi_{FM}}{\varphi_{GT} \circ \varphi_{GT} + \varphi_{FM} \circ \varphi_{FM}}, \tag{5}$$

$\circ$表示Hadamard乘积，对齐矩阵ξ_{FM}具有如下属性:$\xi_{FM}(x,y) \geq 0$仅当φ_{GT}和φ_{FM}符号相同，即两个输入在(x,y)的位置处对齐。对齐矩阵的元素值考虑了全局统计信息，即全局的均值。这些属性使公式(Eq. 5)符合我们的目标。

3.3 Enhanced Alignment Term

$\xi_{FM}(x,y)$的绝对值取决于μ_{FM}和μ_{GT}的相似性。当两幅图高度相似时，μ_{FM}和μ_{GT}之间的进一步相似性可能会增加对齐位置的正值，并减少未对齐位置的负值。将$\xi_{FM}(x,y)$每个位置的值叠加后得到的总数并总是提升从而不符合我们的期望（期望提升）。因此，我们需要一个映射函数来抑制负值（$\xi_{FM}(x,y) \leq 0$）区域的减少（这意味着具有较小的微分值）并且增强正值的增加（$\xi_{FM}(x,y) \geq 0$）区域。

为了实现这一目标，需要用到“凸函数”。我们也测试了其他形式的映射函数，如高阶多项式或三角函数，但发现二次型（$f(x) = \frac{1}{2}(1+x)^2$ 如图4 （g）所示）是一个简单的且有效的函数，在我们的实验中效果最好。在这里，我们用它来定义 **增强的对齐矩阵**ϕ如下：

$$\phi_{FM} = f(\xi_{FM}) \tag{6}$$

3.4 Enhanced Alignment Measure

使用增强的对齐矩阵 ϕ 来捕获二值映射的两个属性（像素级匹配和图像级统计），我们将最后的**E-measure**定义为：

$$Q_{FM} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi_{FM}(x,y), \tag{7}$$

其中h和w分别是映射图的高度和宽度。使用此方法来评估Fig. 1中的前景图（FM）和噪声，我们的指标的排序结果和应用程序的排序结果一致（见下文）。

4 Experiments

在本节中，我们在4个公共的显着物体检测数据集上比较了我们的**E-measure**方法与5个最先进的方法的二值前景图评估结果，见[12]。

元度量。 为了测试评估指标的质量，我们使用了元度量方法。其基本思想是定义一些关于结果质量的理想标准，并评估一个衡量标准如何满足这些标准 [31]。我们利用 [27; 31; 12] 中提出的4个元度量，以及我们在这里引入的一个新的元度量(Sec. 4.3)。Tab. 2中列出了所有的结果。

数据集和模型。 所使用的数据集包括PASCAL-S [22], ECSSD [39], HKU-IS [19]和SOD [28]。我们使用了包括3 个传统的（ST [26]，DRFI [36]和DSR [21]）和7个深学习的（DCL [20]，RFCN [37]，MC [40]，MDF [19]，DISC [8]，DHS [24]和ELD [17]）生成非二值图。为了进一步获得前景二值图，我们使用图像相关的自适应阈值方法去阈值化非二值图。阈值被确定为非二值图的平均值的两倍。

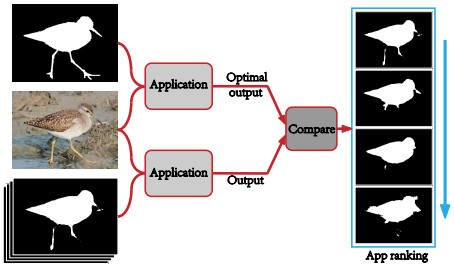


Figure 5: 应用排序。要根据应用程序对前景图进行排序，我们将使用GT时的输出与使用FM 前景图时的输出进行比较。FM前景图与GT越相似，其应用程序输出将越接近GT输出。

4.1 元度量1：应用排序

第一个元度量指定评估前景图的评估结果应该与应用程序的评估结果一致。Fig. 5 说明了应用程序的排名。假设在使用GT映射时应用程序的输出是最优输出。然后，我们向应用程序提供一系列估计的映射图，并获得从最相似到最不相似的程度排序的输出序列。我们将输出序列与最佳输出序列进行比较。与GT越相似，应用程序的输出顺序越接近GT输出顺序。

正如Margolin等人 [27]所声称的那样，包括图像检索，对象检测和分割在内的应用程序具有相似的结果。为了公平的比较，我们使用基于上下文的图像检索应用程序作为Margolin *et al.* 中提及的应用来执行这个元度量。在**应用实现**小节中提到了此应用程序的实现，其他应用程序的实现方式相似。

在这里，我们使用$\theta = 1 - \rho$ [2]度量来检查度量排名和应用排名之间的排名相关性。θ的值落在范围[0,2] 中。0的值意味着测量顺序和应用顺序是相同的。2则表示完全相反的顺序。

在Tab. 2中，可以看到我们的方法相对于当前流行评估方法性能有重大提升。我们的测量分别在PASCAL-S [22]，ECSSD [39]，SOD [28]和HKU-IS [19]数据集比现有最先进方法提高了19.65%，9.08%，18.42%和9.64%。Fig. 2说明了我们的方法如何很好地预测这些应用程序的偏好

应用实现. 基于上下文的图像检索系统查找查询图像数据集 [18]中最相似的图像。相似性由诸如颜色直方图，颜色和边缘方向性描述符（CEDD）等各种特征确定。我们用LIRE [18]和CEDD来权衡二值前景图。

首先，为了忽略背景并获得前景特征，我们通过将图像与其GT图或FM图组合（Fig. 6 (a)-(d)）来生成组合的GT图像或FM图像。组合图像结果由此表示: $GT_{combine} = \{G_1, \cdots, G_n\}$ and $FM_{combine} = \{F_1, \cdots, F_n\}$ 其次，对于每个组合的图像，我们使用LIRE来检索到100个最相似组合图像的列表。这些图

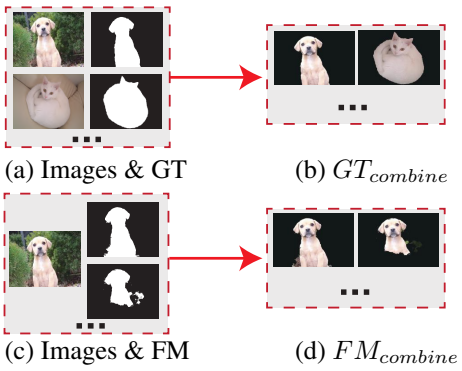


Figure 6: 组合图像与其前景映射图。

Figure 7: 元度量2: 最先进vs. 通用映射图。评价方法应该使得由最新模型生成的FM（c）比不考虑图像内容的通用映射图（d）得分高。

像列表是预先从最相似到最不相似进行排序的。GT输出$GTout_i = \{G_{1-i}, \cdots, G_{100-i}\}$是使用组合GT(比如，$G_i$)时返回的有序列表。有序分数列表是$GTscore_i = \{GTs_{1-i}, \cdots, GTs_{100-i}\}$。得分意味着相似的程度。同样地，对于FM我们得到$Fout_i = \{F_{1-i}, \cdots, F_{100-i}\}$和$Fscore_i = \{Fs_{1-i}, \cdots, Fs_{100-i}\}$。第三，让$I_i = \{GTout_i \cap Fout_i\}$。在$Fout_i$中查找$F_k$等于$G_i$。如果$F_k$存在，表明$G_i \in Fout_i$，我们可得到索引$k$以及相应的得分$Fs_{k-i}$。每个FM的分数$S_i$是：

$$S_i = \begin{cases} Fs_{k-i} + \frac{1}{k} + \frac{\|I_i\|}{100}, & G_i \in I_i \\ \frac{\|I_i\|}{100}, & otherwise \end{cases} \quad (8)$$

4.2 元度量2: 最先进vs. 通用映射图

第二个元度量是，评估方法应该为通过最新模型获得的映射图赋予更高的分数，而不是没有考虑内容的随意的映射图。在这里，我们使用一个中心圆作为通用映射图。可以在Fig. 7中看到一个例子。我们期望前景图（c）相对（d）会得到更高的分数。

我们统计了一张通用映射图的得分高于在Sec. 4中提及的10个最先进模型生成的映射图得到的平均得分的次数作为错误排序比率。正如 [27]中所建议的那样，在考虑到某些模型会产生一个糟糕的映射图结果的情况情况下，取平均得分是鲁棒的。如果10张映射图的评分都高于一个阈值，则视为“好的映射图”。基于此，我们选择了数据集中约80%的“好的映射图”来检验这个元度量。得分越低，评价方法表现的效果就越好。除了在PASCAL-S数据集外，我们的评价方法在ECSSD，SOD和HKU-IS数据集上都胜过当前的其它评价方法。

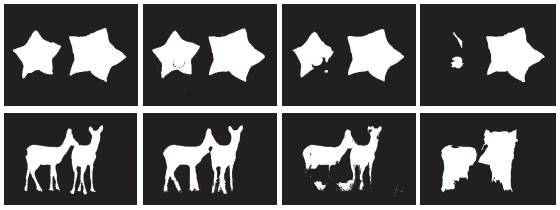


Figure 8: 元度量4: 人为排序. 示例图像来自我们新创建的FMDatabase。

4.3 元度量3: 最先进vs. 随机噪声

我们的第三个元度量所依据的性质是评估方法应该偏好于最先进模型生成的映射图的平均值而不是随机噪声图。

我们采用与元度量2相似的实验来进行元度量3的实验，但是这次我们使用高斯随机噪声图替代了Sec. 4.2中的通用映射图。由于考虑到局部像素匹配和全局统计，我们的方法达到了最好的性能。值得注意的是，如Sec. 4.2中所述。从某个最先进模型生成的FM可能出现的偶然错误的情况来说，因此最先进模型的结果考虑的是平均得分来提现鲁棒性。从而，来自最先进模型生成的前景图的平均得分应始终高于从噪声图评估的得分。只有我们的方法和S-measure 达到最低的错误排序率。

4.4 元度量4: 人为排序

第四个度量方法是考察评价指标与人为排序之间的相关性。据我们所知，在此之前没有人为排名的二值前景映射图数据集。为了创建这样一个数据集，我们在PASCAL-S [22]，SOD [28]，ECSSD [39]和HKU-IS [19]这4个数据集中随机选择根据元度量1中的应用程序排序过的显著图。接着，让10名受试者对这些映射图进行排序，并保留那些所有受试者赋予了一致性排序的显著图。我们将我们的数据集命名为**FMDatabase**[2]，其包含了185张图像。每张图片都有三张估计的映射图（共555 张图）。

为了定量评估人类排序和测量排序之间的相关性，我们同样使用θ measure（在元度量1中提到）来检验这个元度量。得分越低，说明评估指标与人的排序结果越一致。从结果可以看出，我们的指标胜过其他指标。Fig. 2 举例说明了我们的度量如何预测人类排序偏好的。

4.5 元度量5：真值图替换

第五个元度量表达的是，当我们使用错误的GT图

[2]FMDatabase: http://dpfan.net/e-measure/

Table 2: E-measure与当前评估方法在4个元度量上的定量比较。最好的结果用**blue.**高亮显示。MM: 元度量。下列差异的统计意义在$\alpha < .05$级别。

	PASCAL			ECSSD			SOD			HKU			
Measure	MM1	MM2	MM3	MM1	MM2	MM3	MM1	MM2	MM3	MM1	MM2	MM3	MM4
CM	0.610	49.78%	100.0%	0.504	34.62%	100.0%	0.723	29.89%	56.22%	0.613	25.26%	100.0%	1.492
VQ	0.339	17.97%	15.32%	0.294	7.445%	6.162%	0.335	9.143%	14.05%	0.331	3.067%	1.800%	0.161
IOU/F1/JI	0.307	9.426%	5.597%	0.272	4.097%	1.921%	0.342	4.571%	6.857%	0.303	0.900%	0.197%	0.124
Fbw	0.308	5.147%	4.265%	0.280	2.945%	1.152%	0.361	6.286%	5.714%	0.312	0.535%	0.083%	0.149
S-measure	0.315	**2.353%**	0.000%	0.279	1.152%	0.000%	0.374	1.714%	0.000%	0.312	0.141%	0.000%	0.140
Ours	**0.247**	3.093%	**0%**	**0.247**	**0.641%**	**0%**	**0.273**	**0.571%**	**0%**	**0.274**	**0.084%**	**0%**	**0.121**

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

Figure 9: 元度量5：真值图替换。评价指标应该使用正确的真值图（c）作为参考，而不是使用随机切换的真值图（d），为“好”的映射图（b）给予更高的分数。

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

Figure 10: E-measure在MM2上的失败案例。由于忽略了语义信息，E-measure将（c）排名高于（d）。

时，“好的映射图”的分数应该减少。我们分析了4个流行的数据集（PASCAL-S [22]，SOD [28]，ECSSD [39]，HKU-IS [19]），发现一个映射图在评分（使用F1-measure）至少为0.8 时被认为“好”的。我们遵循Margolin等人 [27]来计算这个元度量。我们计算一个评估方法在使用错误的GT而获得较高分数的次数的百分比。我们发现，所有的评价方法都表现出色（4个数据集的平均结果为：VQ [34]为0.000925%，CM [30]0.001675%，IOU/JI/F1 [15] 0.0014%和我们的方法0.0523%）。我们的方法相对于其他方法有0.05%的差距。

5 结论和未来的工作

在本文中，我们分析了考虑像素，区域，边界和对象层次等不同层次的各种二值前景评价指标。它们被划分为要么考虑像素级错误要么是仅考虑图像级别的错误。为了解决这个缺点，本文提出了同时考虑这两种类型的错误的**E-measure** (增强匹配指标)。我们的方法是非常有效和高效的。我们采用5个元度量的方法在4个流行的数据集上，通过大量的实验证明了我们的方法比当前的方法更有效。最后，我们创建了一个新的数据集（740张映射图），它由185张真值图和555张经由人眼排序过的映射图组成，数据集用以检验评价指标与人类判断之间的相关性。

局限性. 与我们的指标相比，S-measure主要用于解决结构相似性问题。PASCAL数据集中的图像比其他3个数据集（ECSSD，SOD，HKU-IS）具有更多结构对象。因此，S-measure在PASCAL数据集略好于我们的方法。Fig. 10 给出了一个失败案例。

未来的工作. 在未来的工作中我们将研究基于E-measure提出新的分割模型的可能性。此外，由于度量指标由简单的可导函数组成，因此可以开发基于E-measure的新的损失函数。为了促进该领域的探索，我们的代码和数据集将在网上公开发布。

致谢

This research was supported by NSFC (NO. 61620106008, 61572264), Huawei Innovation Research Program, and Fundamental Research Funds for the Central Universities.

References

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

[2] 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.

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

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

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013.

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

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

[8] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li. Disc: Deep image saliency computing via progressive representation learning. *IEEE T Neur. Net. Lear.*, 27(6):1135–1149, 2016.

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

[10] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation? In *BMVC*, volume 27, page 2013. Citeseer, 2013.

[11] 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.

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

[13] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. pages 698–704, 2018.

[14] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 2018.

[15] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

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

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

[18] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM T Multim. Comput.*, 2(1):1–19, 2000.

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

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

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

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

[23] G. Liu and D. Fan. A model of visual attention for natural image retrieval. In *Information Science and Cloud Computing Companion (ISCC-C)*, pages 728–733. IEEE, 2013.

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

[25] 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.

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

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

[28] 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 *ICCV*, pages 416–423. IEEE, 2001.

[29] K. McGuinness and N. E. O' connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

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

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

[32] 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.

[33] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–37. IEEE, 2004.

[34] R. Shi, K. N. Ngan, S. Li, R. Paramesran, and H. Li. Visual quality evaluation of image object segmentation: Subjective assessment and objective measure. *IEEE TIP*, 24(12):5033–5045, 2015.

[35] P. Villegas and X. Marichal. Perceptually-weighted evaluation criteria for segmentation masks in video sequences. *IEEE TIP*, 13(8):1092–1103, 2004.

[36] 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.

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

[38] 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.

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

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