基于梯度诱导的协同显著性检测（中译版）

张钊[1], 金闻达[2], 徐君[1]，程明明[1]

[1] 计算机学院, 南开大学

[2] 智能与计算学院, 天津大学

zzhang@mail.nankai.edu.cn; cmm@nankai.edu.cn

说明. 本文是论文《Gradient-Induced Co-Saliency Detection》的中文翻译版。获取原文及相关资源请访问 https://mmcheng.net/gicd/

摘要. 协同显著性检测（Co-SOD）致力于在一组相关的图像中分割出共同的显著前景。本文中，受人类行为的启发，我们提出了一种基于梯度诱导的协同显著性检测（GICD）方法。首先，我们在嵌入空间中抽象出一组图像的共识表示，接着，我们比较单个图片的表示与共识表示差异，利用反馈的梯度信息来诱导模型更加关注更具有鉴别性的特征。另外，由于 Co-SOD 任务的训练数据匮乏，我们提出了一种拼图训练策略，可以使 Co-SOD 网络在常规显著性物体检测数据集上进行训练，而不需要多余的像素级标注。为了评价 Co-SOD 模型在不同前景中发现协同目标的能力，我们构建了一个有挑战性的评测数据集 *CoCA*，其中每一张图像在包含协同显著目标之外还至少一个无关的前景干扰物体。实验表明我们的基于梯度诱导的协同显著性检测（GICD）方法取得了领先的性能。

关键词: 协同显著性检测，新数据集（*CoCA*），梯度诱导，拼图训练

1 介绍

协同显著性检测（Co-SOD）是一种通过探索多张相关图片之间的内在联系来发现图像共同的显著目标的任务。由于协同显著物体和图像背景的复杂多变，这是一个有挑战性的计算机视觉任务。作为一个可以理解多张图像联系的工具，协同显著性检测模型被广泛应用于很多视觉任务的预处理部分，比如弱监督语义分割 [43, 46]、监控摄像 [15, 29] 和视频分析 [18, 19] 等。

过去的研究从不同方面对协同显著性检测问题进行探索 [6, 20, 22]。在早期阶段，研究者利用手工设计的特征来探索一组相关图像的一致性，例如 SIFT [5, 18]，颜色与纹理 [13, 23] 或者多线索融合 [4] 等等。但是，这些浅

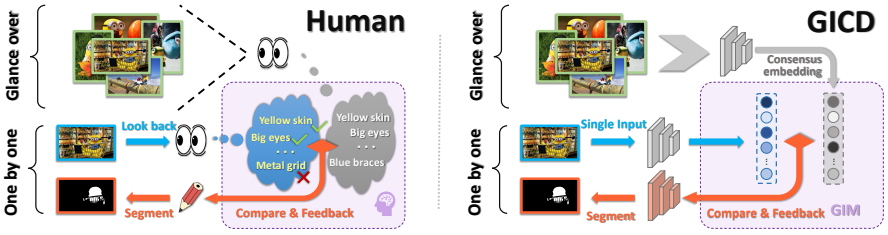

图 1. 受人类行为启发而生的 GICD，其中 GIM 是梯度诱导模块。

层特征的鉴别能力并不足以在现实场景中的分割协同显著的物体。最近，研究者利用基于学习的方法探索一组图像中的语义联系，达到了令人振奋效果。这些方法有深度学习 [22, 42]，自步学习 [17, 51]，度量学习 [16] 或者图学习 [20, 54] 等等。然而这些方法受制于特征内在的差异性，这种差异性是由变换的视角，外观，共同显著物体的位置等因素导致的。如何更好地利用相关图像的联系值得更深入地探究。

人类是如何从一组图像中分割出协同显著性目标呢？大体上，人类会先浏览这一组图像，通过“常识”总结出协同显著目标的共有属性 [32]，然后通过这些属性在每张图像中来寻找相关的目标，这个过程如图1所示。受人类行为的启发，我们设计了一种端到端的网络，其过程对应着上述两个阶段。为了像人类一样获得协同显著目标的共有属性，我们首先利用一个训练好的嵌入网络计算了在高维空间中多张相关图像的共识表示。当获取了共识表示后，我们提出了一个梯度诱导模块（GIM）来模仿人类通过比较特定场景与共识描述进而反馈匹配信息的行为。

在 GIM 中，首先单张图像与一组图像的共识表示之间的相似性可以被度量。由于不同的高层次的卷积核可以感知不同的语义信息 [35, 55]，我们可以找到那些和共识表示相关性强的卷积核，强化它们来检测协同显著目标。为此，通过局部的反向传播，我们计算了相似度相对于顶部卷积层的梯度作为反馈信息。高的梯度值意味着相应的卷积核对相似度具有正向的影响，因此，通过给这些卷积核赋予更高的权重，模型会被诱导关注与协同显著物体更相关特征。另外，为了更好地在自上而下解码器的每一层级保持协同显著目标特征的鉴别性，我们提出了一个注意保持模块（ARM）来连接我们模型中相应的编码-解码块。我们称这个具有 GIM 和 ARM 的双阶段框架为基于梯度诱导的协同显著性检测（GICD）网络。在基准数据集上的实验展示了我们的 GICD 方法相对于过去 Co-SOD 方法的优势。

由于专门的标注用于训练 Co-SOD 模型，现有的 Co-SOD 网络 [22, 42, 47] 常在语义分割数据集上进行训练，例如 Microsoft COCO [26]。然而，在语义分割数据集中的物体未必是显著的。在这篇论文中，我们提出了图中拼图策略来扩展现有显著目标检测（SOD）数据集用于训练 Co-SOD 模型，同时不需要进行额外的像素级标注。

除此之外，为了更好地评价 Co-SOD 在多个前景（包含无关物体）中找到协同显著目标的能力，用于评估模型的数据集中的图片应该至少包含一个除协同显著目标外的不相关显著前景物体。在图3中可以看出，这一重要的特性被当前的 Co-SOD 评测数据集 [2, 11, 44, 50] 所忽视。为了缓解这个问题，我们精心构建了一个更富有挑战的数据集，名为 *C*ommon *C*ategory *A*ggregation (*CoCA*)。

综上，本文主要的贡献如下：

- **我们提出了一个基于梯度诱导的协同显著性目标检测 (GICD) 网络。**具体地，我们提出了一个梯度诱导模块（GIM）使得模型在训练过程中更加注意对协同显著目标更具有鉴别性的特征。我们同时提出了一个注意保持模块（ARM）来保持自上而下解码的过程中的注意力。
- **我们提出了一个拼图训练策略在通用的 SOD 数据集（如 DUTS）上训练 Co-SOD 模型**，从而缓解缺少 Co-SOD 训练数据的问题。
- **我们构建了一个有挑战性的 *CoCA* 数据集，同时仔细地为它赋予了标签**，提供了更加符合现实场景来更好地评价当前的 Co-SOD 方法。
- 在 *CoSal2015* [50] 数据集和我们的 *CoCA* 数据集上的实验展示了我们的 GICD 方法优于过去的 Co-SOD 方法。同时，充分的消融实验也验证了我们贡献的有效性。

2 相关工作

2.1 协同显著性目标检测 (Co-SOD)

与传统的显著性目标检测不同 (SOD) [8, 12, 14]，协同显著性目标检测致力于在一组相关图像中自动分割出共同的显著性物体。早期的 Co-SOD 方法假设多张图像中的协同显著目标具有浅层一致性 [51]。例如，Li 等人 [23] 引入了一种协同多层图来探索颜色和纹理属性。Fu 等人 [13] 探索对比、空间和对应线索并用聚类的方式来增强图像全局约束。Cao 等人 [4] 通过自适应加权方法整合了多种协同显著性线索。Tsai 等人 [38] 通过解决图的能量最小化问题提取了协同显著目标。

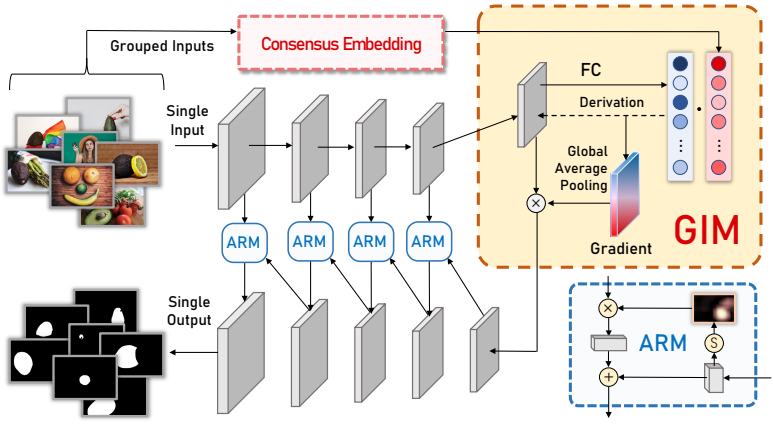

图 2. 基于梯度诱导的协同显著性目标检测（GICD）方法流程图。 GIM 代表梯度诱导模块，同时 **ARM** 代表注意保持模块。“•”，“⊗”，“⊕”，和 “Ⓢ” 分别代表内积，点积，点加，以及 sigmoid 函数。

最近，很多基于深度学习的方法被提出来为 Co-SOD 任务探索深层特征 [17, 50, 52]。这些方法可被分为两类，其一是从传统浅层一致性的方法向深度学习方法的自然拓展，它通过探索深层次相似性来增强多张图像中相似的候选区域。例如，Zhang 等人 [50] 通过深层卷积神经网络同时探究了组间的差异性和组内的一致性。Hsu 等人 [17] 提出了一种无监督的方法，通过图优化来最大化多样前景间的相似性同时最小化前景和背景间的相似性。Jiang 等人 [20] 利用图卷积网络构建内部和外部图来探索超像素级的相似性。Zhang 等人 [52] 论文中提出了一种掩码引导的网络预测粗略 Co-SOD 结果，接着通过多标签平滑处理来细化分割结果。第二类深度方法基于联合特征提取。他们通常提取出一组图像的共同特征，接着将共同特征与单幅图像的特征进行融合。例如，Wei 等人 [42] 通过分组学习方法学习了每五个图像的联合特征，然后将联合特征拼接在单幅图像提取的特征上。Li 等人 [22] 基于这个思路使用序列模型来处理不同长度输入。Wang 等人 [39] 和 Zha 等人 [47] 论文中学习了一组图像的类别向量，从而与单个图像特征在多个级别上的每个空间位置相拼接。

2.2 Co-SOD 数据集

当前的 Co-SOD 数据集主要有 *MSRC* [44]，*iCoseg* [2]，*CoSal2015* [50]，和 *CoSOD3k* [11]。在图3中，我们展示了这些数据集的一些图片样例以及

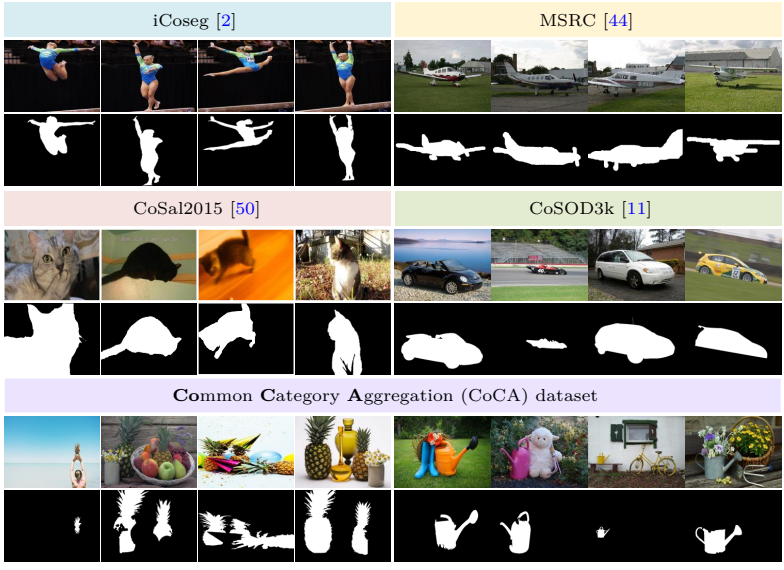

图 3. 当前常用的评测数据集和我们新提出的 *CoCA* 数据集的示例。在 *CoCA* 数据集中，除了协同显著目标，每一张图像上包含至少一个多余显著物体，使得这个数据集能更好地评价在多个前景中发现协同显著目标的能力。

我们的 *CoCA* 数据集的样例。*MSRC* [44] 主要用于在图像中识别物体。在 [13,50] 中，他们从 *MSRC-v1* 中的七组数据中选取了 233 张图像来评估检测的准确性。*iCoseg* [2] 包含不变场景的 38 组图像中的 643 张。在上述数据集中，协同显著目标大多是在相似场景下的同一个物体。*CoSal2015* [50] 和 *CoSOD3k* [11] 是两个相对规模较大的数据集，分别包含 2015 张和 3316 张图像，在这两个数据集中，有些属于同一类的目标物体在外貌上差异较大，这使这两个数据集更具有挑战性。然而上述这些数据级都不是特别适用于评价 Co-SOD 算法，因为它们的大多数图像都只有一个显著目标。以数据集 *iCoseg* 中的图3为例，虽然这个运动员在这组图像中是协同显著的，但是这些场景由于没有其它显著前景的干扰，完全可以直接用普通的 SOD 算法处理。虽然这一尴尬的局面已经在 *CoSal2015* 和 *CoSOD3k* 数据集里的部分图像中避免了，它在大多数情况下仍不容乐观。（作者注：在这些数据集上，SOD 的模型甚至可以取得比 Co-SOD 模型更好的效果。）由于在多张图片中探索协同显著目标是现实应用中 Co-SOD 方法的目标 [49]，为了更好地评估 Co-SOD 方法的这一能力，我们建立了一个有挑战性的 *CoCA* 数据

集，其中每张图像至少包含一个除协同显著前景外的无关的显著物体（作为干扰项）。

3 提出的方法

图2 展示了基于梯度诱导的协同显著性检测（GICD）网络的结构。我们的模型主要基于广泛使用的特征金字塔网络 (FPN) [25]。对于 Co-SOD 任务，我们结合了两个我们提出的模块：梯度诱导模块（GIM）和注意力保持模块（ARM）。GICD 通过两个阶段检测协同显著目标：首先它接收一组图像作为输入，接着用学习好的嵌入网络在高维空间探索这组图像的共识表示。这个共识表示描述了这组图像中协同显著目标的共有模式。接着，GICD 回过头来分割每一张图像的显著目标。在这个阶段，为了把模型的注意力诱导至协同显著区域，我们利用了 GIM 模块通过比较嵌入空间中单个目标和共识表示来增强与协同显著目标相关的特征。为了保持自上而下过程中解码的注意力，我们使用了 ARM 模块来连接编码-解码对。我们使用拼图策略来训练 GICD 网络，Co-SOD 模型可以在不需要额外像素级标注的条件下在 SOD 数据集上进行训练。

3.1 学习共识表示

给定一组图像 $\mathcal{I} = \{I_n\}_{n=1}^N$，为了定位每张图像中的协同显著目标，我们首先应该通过先验知识了解协同显著目标具有哪些模式。为此，我们提出用预先训练好的嵌入网络学习一组图像 $\mathcal{I}$ 中协同显著目标的共识表示。深度分类器可被自然地用于表征学习 [33]，其中语义属性的先验知识可以从 ImageNet [7] 上预训练好的参数上迁移而来。在这种情形下，我们使用一种预训练好的分类网络 $\mathcal{F}(\cdot)$，比如 VGG-16，在移除 softmax 层后作为我们的嵌入网络。它首先提取每一张图片 I_n 的代表 $e_n = \mathcal{F}(I_n) \in \mathbb{R}^d$，其中 d 是最后的全连接层的维数。接着一组图像的共识表示 $e^\dagger$ 可被计算为 $e^\dagger = \texttt{Softmax}\left(\sum_{n=1}^N e_n\right)$，用来描述这组图像的共同属性。

3.2 梯度诱导模块（GIM）

在获取了一组图像 $\mathcal{I}$ 的共识表示 $e^\dagger$ 后，对每一张图像，我们致力于找到对共识表示匹配有鉴别性的特。征如论文 [35, 55] 所述，深层卷积层自然地处理语义特定的空间信息。我们标记这个卷积神经网络 $\mathcal{F}(\cdot)$ 的 5 个卷积

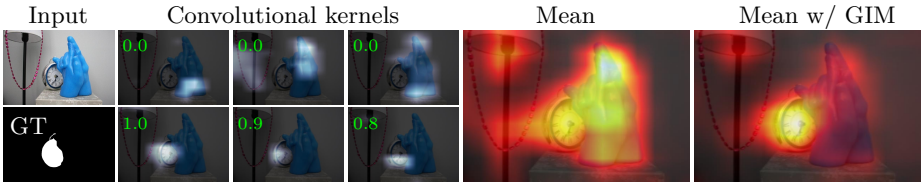

图 4. GIM 的深层梯度诱导可视化. 在左边的 6 个小图中，上方三张图像的卷积核对目标物体不敏感，然而下方的卷积核与目标物体相关。它们相应的基于梯度的重要性权重在图片左上角的绿色数字标出。F_n^5 以及梯度诱导的 $\tilde{F}_n^5$，两者均值用橘红色热力图展示出来。

特征块为 $\{F^1, F^2, \ldots, F^5\}$。在图4中，我们展示了最后的卷积层 F^5 的特征图。输入的图像（第一列）包含了一块怀表和一条蓝色围巾，不同卷积核聚焦于图像的不同区域（第二到第四列）。如果给密切关注协同显著目标的核赋予更高的重要性，模型会在解码时趋向于分割协同显著目标（怀表）。如论文 [35] 中表明的，神经网络中特征的鉴别性可以通过优化目标获得的梯度来衡量。因此，我们提出了一个梯度诱导模块（GIM），然后通过反馈的梯度信息来增强具有鉴别性的特征。因为我们的编码器和共识嵌入网络共享参数，它也能把每张图像嵌入与共识表示 $e^\dagger$ 相同的空间中。对于第 n 张图片中提取出的表征 e_n，e_n 和它的共识表示 $e^\dagger$ 的相似度 c_n 可以用内积来定义，*i.e.*，$c_n = e_n^\top e^\dagger$。接着我们计算局部反向传播最后一个卷积层 $F^5 \in \mathbb{R}^{w\times h\times c}$ 的正梯度 G_n 来选择 F_n^5 与协同显著目标相关的特征。具体地，

$$G_n = \mathrm{ReLU}\left(\frac{\partial c_n}{\partial F_n^5}\right) \in \mathbb{R}^{w\times h\times c}. \tag{1}$$

在这个局部反向传播过程中，正梯度 G_n 反应了相应位置对最终相似度得分的敏感度；也就是说，以更大的梯度增加激活值将使得特定的表示 e_n 与共识表示 $e^\dagger$ 更加相似。因此，对检测一个特定目标来说，一个卷积核的重要性可用它的特征梯度的均值来度量。具体地，通道重要性可以通过全局平均池（GAP）来计算，公式为：$w_n = \texttt{GAP}(G_n) = \frac{1}{wh}\sum_i\sum_j G_n$，其中 $i = 1, ..., w$ 并且 $j = 1, ..., h$。当获取了权重，我们可以通过给每一个卷积核分配权值 $\tilde{F}_n^5 = F_n^5 \otimes w_n$（其中 $\otimes$ 代表点乘），诱导出深层特征 F_n^5。像图4中所展示的，我们可视化了 F_n^5 和 $\tilde{F}_n^5$ 的均值热力图，如果没有我们的 GIM 模块，卷积核同时关注怀表和手套。我们可以发现和协同显著类别更相关的卷积核（以绿色数字标注）以及网络的注意力在梯度诱导后已经转移到协同显著目标上。

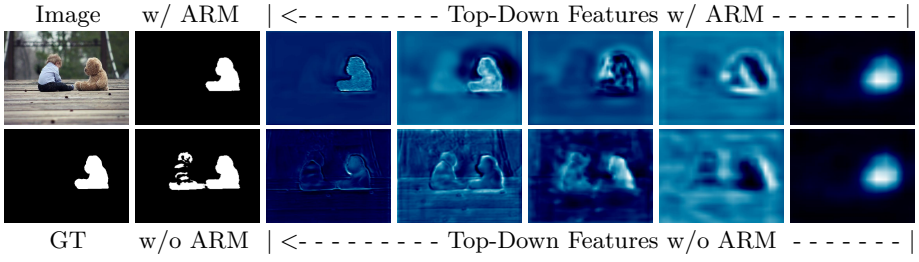

图 5. ARM 的注意保持可视化。第一行展示了模型中有 (w/)ARM 情形下的多层级的中间特征，相应地，第二行显示了没有 (w/o)ARM 情形下。具备 ARM 的预测（第二列上面）比不具备 ARM 的预测（第二列下面）更准确，因为我们的 ARM 更注意协同显著区域。

3.3 注意保持模块（ARM）

在 GIM 中，深层特征已经被梯度所诱导。然而，自上而下的解码器建立于自下而上的框架，以致被诱导的深层特征在传到浅层的过程中会逐渐稀释。为此，我们提出了一个注意保持模块（ARM）来连接 GICD 网络中的对应编译-解码对。像图2中所展示的，对于每一个 ARM，用于跳层连接的编码器被深层预测所指引。通过自上而下的迭代提示，网络会专注于细化协同显著区域而不被其它不相关目标干扰。我们将 $\tilde{F}_n^5$ 的通道平均值作为第一个低分辨率导向图 S_n^5，并且把 $\tilde{F}_n^5$ 减少为包含 64 个通道的特征 P_n^5。ARM 的解码过程如下所示：

$$\begin{cases} \tilde{F}_n^i = (S_n^{i+1}) \uparrow \odot F_n^i \\ P_n^i = \mathcal{E}^i \left((P_n^{i+1}) \uparrow + \mathcal{R}^i \left(\tilde{F}_n^i \right) \right), \ i \in \{4,3,2,1\}, \\ S_n^i = \mathcal{D}^i \left(P_n^i \right), \end{cases} \tag{2}$$

其中，$(\cdot)\uparrow$ 是上采样。$\mathcal{R}^i(\cdot)$ 包含两个卷积层，同时亚索被增强的特征 $\tilde{F}_n^i$ 至 64 个通道。$\mathcal{E}^i(\cdot)$ 是相应两个卷积层，在解码器中包含 64 个卷积核。$\mathcal{D}^i(\cdot)$ 被应用于深度监督，用两个卷积层和一个 sigmoid 层输出预测。最后的 S_n^1 是 GICD 的最终输出。

为了验证我们 ARM 的有效性，在图5中，我们展示了使用 ARM（第一行）和不使用 ARM（第二行）的解码器在不同层级中的中间特征。我们观察到，通过 GIM 两张图像都成功定位了协同显著目标（泰迪熊），然而没有使用 ARM 的 GICD 在上采样的过程中逐渐被干扰并且产生不准确的检测结果。这些结果表明我们的 ARM 有效地将注意力保持在了协同显著目标上。

3.4 拼图训练策略

策略： 一个在 Co-SOD 任务中的重要问题是当前 SOD 数据集，例如 DUTS [40] 和 MSRA-B [28] 不适用于训练 CO-SOD 网络。原因有两方面：1）它们没有类别信息，所以不可能分组训练模型；2）大多数数据集中的图像仅仅包含一个显著前景。这样的网络很难促使网络在多个前景中区分协同显著目标。最近的 Co-SOD 方法 [22, 39, 42] 在语义分割数据集 [26] 上训练的。这也遇到了两个难题：1）和 SOD 数据集相比，语义分割数据集往往粗糙，所以训练的网络细节修复的能力不理想，从而满足不了下游任务对精度要求。2）语义分割数据集中的目标未必是显著的。为了缓解这些问题，我们设计了一个拼图策略来将传统 SOD 数据集转化为适用于 Co-SOD 模型的训练数据。第一步：由于 SOD 数据集原本没有类别信息，我们使用了一个分类器 [30] 来为它们分类，并组合成不同的组别。第二步：像图6中所示，我们将不同种类中的样例进行拼接来形成一个新拼图。这一步保证了输入图像包含了除协同显著前景外的额外的前景目标。通过上述步骤，现有的 SOD 数据集可以无缝地用于训练 Co-SOD 网络，而无需额外的像素级标注。

损失函数： 考虑到协同显著性检测的最重要目标是正确发现共同前景的位置。我们采用了交并比损失（IoU） [24, 34] 训练 GICD 网络，特别的，

$$\mathcal{L}(S,G)=1-\frac{\sum_{c} S(c)\,G(c)}{\sum_{c}\left[S(c)+G(c)-S(c)\,G(c)\right]}, \tag{3}$$

其中 S 是预测值，G 代表真实值，c 代表每一张图像中的像素位置。我们整个模型的损失函数可以表示为：

$$L_{total}=\sum_{n=1}^{N}\sum_{i=1}^{4}\mathcal{L}\left(S_n^i,G_n\right). \tag{4}$$

4 提出的 *CoCA* 数据集

指导原则： 我们基于四条规则建立我们的 *CoCA* 数据集。**G1**: 每张图像至少包含一个除协同显著目标外的多余前景。**G2**: 在每张图像组中，协同显著目标的外貌最好有区别。**G3**: 数据集需要与常用训练集包含的类别错开，以

图 6. 拼图训练策略示例。一个猫的图片和其它类别的图片组成了多张用于训练的拼图。

探索模型处理未知类别的能力。规则 **G1** 反应了模型是否能检测协同显著目标，而不是仅仅分割出前景和背景。规则 **G2** 可以评价模型是否对组内差异鲁棒。规则 **G3** 确保了可以评估模型从未知类别中检测出协同显著目标的能力。

建立过程： 有了上述规则，我们从 pixabay[3]上收集图像。我们把它们分为 80 类，包含日常室内和室外的场景。值得注意的是这些种类与微软 COCO [26] 数据集完全错开，而这个 COCO 数据集合常被用于 Co-SOD 模型 [22,39,42] 的训练。更重要地，通过手动筛选，我们数据集中的图像包含至少一个协同显著目标之外的显著目标。我们提供了四个标注级别：类级别，边界框级别，物体级别和实例级别。高质量物体级别标注可用于这篇论文中的协同显著性检测任务。不同级别的标注对应不同的任务，比如协同定位 [21,37]，小样本目标分割 [48,53]，以及实例协同分割 [36]。

数据集统计： 我们的 *CoCA* 数据集包含 80 个种类，共计 1295 张图像。如图3中所展示的那样，这些图像由于遮挡，背景杂乱，多余目标干扰等方面具有挑战性。每个类别中的图像数量不同，从 8 到 40 不等。数量上的差异有助于评价模型处理不同图像数据集大小的能力。一张图像中的协同显著实例的个数也是不一样的。我们的数据集中有 336 张图像有两个以上的协同显著实例。这些实例数量上的不同可以帮助评估模型对多目标场景的有效性。

[3] https://pixabay.com

5 实验

5.1 实现细节

我们使用我们提出的拼图策略在 *DUTS* [40] 训练集上进行 GICD 网络训练。这些图片被分到了 291 组，其中包含 8250 张图片（去除了一些样本过少和前景过于杂乱的类别）。每一个例子都会和其它例子相结合形成三个拼图作为补充。如图6中所示，候选训练数据变为原来四倍。在每一轮训练中，我们从每组中选了至多 20 个训练样本。使用 Adam 优化器进行训练，设置初始化学习速率为 0.0001，$\beta_1 = 0.9$, 同时 $\beta_2 = 0.99$。在第五十轮迭代时将学习速率降低为原来的十分之一。最终我们总共训练 GICD 网络 100 轮。为了将图像数据输入我们的 FPN 框架（VGG 网络），我们在训练和测试阶段将它们的大小调整为 224×224，同时输出图像调整为原始大小以用来评估模型。我们的 GICD 网络在 [31] 上实现，它以 ~ 55 FPS 的速度在英伟达 GeForce RTX 2080Ti 上运行。

5.2 评测数据集和评测指标

数据集：我们使用了两个有挑战的数据集来评价不同方法的表现。第一个数据集是 *CoSal2015* [50]。在一些图像组中，比如棒球组，由于多余显著目标的干扰这组图像在实验中具有挑战性。第二个数据集是我们的 *CoCA*，其中大多数图像拥有至少一个除协同显著目标之外的不相干显著目标。
指标：我们使用了 [17, 49, 52] 中提到的五个常用指标：平均 F-measure (F_{avg}) [1]，最大 F-measure ($F_{\max}$) [3]，PR 曲线，S-measure (S_α) [9]，以及平均 E-measure (E_ξ) [10].

5.3 与前沿的方法的对比实验

对比方法： 我们将我们的 GICD 方法与七个先进方法作比较，其中包含四个 Co-SOD 方法：RCAN [22]，CSMG [52]，GW [41] 和 CBCD [13]，以及三个 SOD 方法：BASNet (ResNet-34) [34]，PoolNet (ResNet-50) [27] 和 SCRN (ResNet-50) [45].

定量评测： 在 表1中,我们展示了我们的 GICD 和其它前沿方法在 *CoSal2015* 和我们的 *CoCA* 两个数据集上的定量评测结果，可以看出，我们的 GICD 表

	Metric	CBCD [13]	GW [41]	CSMG [52]	RCAN [22]	BASNet [34]	PoolNet [27]	SCRN [45]	GICD Ours
CoSal2015	$F_{\text{avg}}\uparrow$	0.378	0.639	0.721	0.670	**0.778**	0.768	0.755	**0.835**
	$F_{\max}\uparrow$	0.547	0.706	0.787	0.764	**0.791**	0.785	0.783	**0.844**
	$S_\alpha\uparrow$	0.550	0.744	0.776	0.779	0.822	**0.823**	0.817	**0.844**
	$E_\xi\uparrow$	0.516	0.727	0.763	0.742	**0.841**	0.836	0.822	**0.883**
CoCA	$F_{\text{avg}}\uparrow$	0.230	0.358	0.390	0.360	**0.398**	0.394	0.394	**0.504**
	$F_{\max}\uparrow$	0.313	0.408	**0.499**	0.422	0.408	0.404	0.413	**0.513**
	$S_\alpha\uparrow$	0.523	0.602	**0.627**	0.616	0.592	0.602	0.612	**0.658**
	$E_\xi\uparrow$	0.535	0.615	0.606	0.614	0.600	0.616	**0.625**	**0.701**

表 1. 量化对比. 我们 GICD 和其它方法在 *CoSal2015* [50] 和 *CoCA* 数据集上的平均 F-measure [1] (F_{avg})，最大 F-measure [3] ($F_{\max}$)，S-measure [9] (S_α) 以及平均 E-measure。“↑” 表示数值越大，模型表现越好。

现更好。结果表现出了一些有趣的现象，在 *CoSal2015* 数据集中，SOD 方法超越了除 GICD 外的大多数 Co-SOD 方法。原因是 *CoSal2015* 数据集中的大部分图像只有一个显著目标，这样可以用 SOD 算法解决，Co-SOD 算法的优势不能在这个数据集上得以体现，所以这些细节导向的 SOD 方法轻易地超越了 Co-SOD 方法的表现。然而，在我们新提出的 *CoCA* 数据集上，这种现象不再明显，因为一张图像上的显著目标包很多非协同显著目标，这就是为什么我们的 *CoCA* 数据集更适用于评价 Co-SOD 算法。然而，值得注意的是我们的 CICD 仍然在 *CoSal2015* 数据集上超越了 SOD 方法。与最好的 Co-SOD 方法相比较，GICD 在平均 F-measure 上带来了 11.4% 的提升，比比 SOD 方法也有 5.7% 的提升。以 S-measure 为衡量标准，CICD 方法比最好的 Co-SOD 方法提升了 3.1%，比 SOD 方法提升了 4.6%。从图7中可以看出，我们的方法在 PR 曲线上和 F-measure 曲线上也超越了其它方法。曲线的趋势表明我们的方法受阈值影响较小，这可以避免在现实应用中难以选择阈值的问题。

定性对比： 在图8中，我们展示了各种方法生成的显著图用于定性的视觉比较。我们展示的这些例子是具有挑战性的，因为输入的每张图像中的显著目标不仅包含了协同显著目标，还包含了无关前景的干扰，这一点也在 SOD 算法的预测结果中体现出来，其中 SOD 算法分割了许多不相干的区域。从总体结果上来看，我们的 GICD 对于图像预测即使在边缘上也有很高的置信度，然而大多数其它算法出现了大量的不确定区域。来看具体的例子，棒

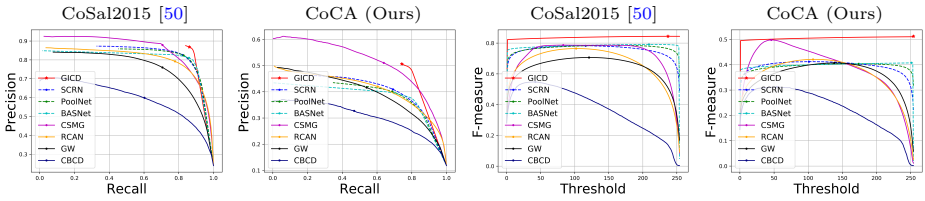

图 7. PR 曲线和 F-meausre 曲线衡量我们的 GICD 方法和七个其它先进方法在 *CoSal2015* 和 *CoCA* 数据集上的表现。在每一个 PR 曲线上的点表示用于计算最大 F-meansure 的点。

球图片数据集是 *CoSal2015* [50] 中最具挑战性的子集，因为棒球在图像间的尺寸变化很大并且受其它显著目标的影响。结果表明，我们的方法成功处理了小目标和遮挡的问题。在 *CoCA* 中，靴子类面对背景颜色的干扰，草莓类有多个分割目标。在这些例子中，GICD 都准确地定位了目标物体。

5.4 消融实验

为了探索梯度诱导模块（GIM）、注意保持模块（ARM）、拼图训练策略（JT）这三者的贡献和工作机理，我们评估了这三者的所有可能组合。请注意，这三者相互依赖，不推荐单独使用。像表2中所展示的那样，“A” 是没有 JT、GIM 和 ARM 的基础模型，由于它不考虑图像间的联系，实际上它是一个 SOD 模型。

GIM 的有效性： GIM 是我们 GICD 的核心模块。有了 GIM，变种 “C”，“E”，“G”，和我们的 GICD 可被看作 Co-SOD 网络。然而，没有 GIM，变种 “A”，“B” “D” 和 “F” 实际上 SOD 模型。通过直接应用 GIM，变种 “C” 将注意力转移到深层特征中的协同显著目标，其表现相对于基础模型有了一定的提升。然而，在这种情形下，训练集会因为没有 JT 而不适用于协同显著性检测，同时在解码过程中注意力会因没有 ARM 而被干扰。这些因素限制了模型的表现，通过引入 JT 或者 ARM（变种 “E”，“G” 以及我们的 GICD），GIM 的效果得到了进一步增强。

ARM 的有效性： ARM 对于自上而下解码过程中保持深层预测具有重要作用。像变种 “D” 中所示，单独使用 ARM 不能提升 Co-SOD 的表现。原因在于，缺少 GIM 的诱导，深层预测实际上是显著目标而不是协同显著目标。当在变种 “G” 中结合了 GIM 时，虽然在不当的数据上进行训练，该变

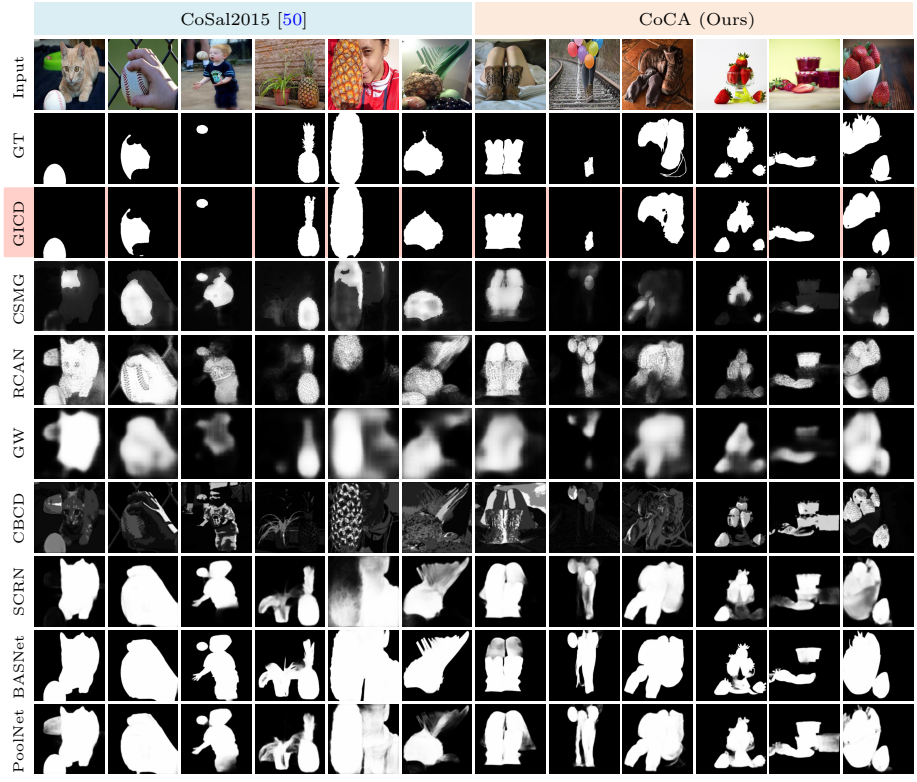

图 8. 我们 GICD 方法和其它 7 个高性能方法（4 个 Co-SOD 方法和 3 个 SOD 方法）在 *CoSal2015* [50] 和我们的 *CoCA* 数据集上的视觉对比。

种仍然强制地保持诱导信息，因此，变种 “G” 比变种 “C” 的表现要好很多。变种 “E” 具备 GIM 和 ARM 模块。像图5中所展示的，没有了 ARM，模型在复原目标细节的时候很容易被不相关前景干扰。因此，它的表现弱于我们的 GICD。

JT 的有效性： 拼图策略（JT）帮助把 SOD 数据集转化成 Co-SOD 数据集，可作为训练 Co-SOD 网络的有效策略。在表2中，没有 GIM，变种模型 “B” 和 “F” 是 SOD 模型，而不是 Co-SOD 模型。由于没有考虑到图像间的交互线索，一个 Co-SOD 数据集上训练的 SOD 模型无法发现组内的联系，同时生成的 JT 标签将在这个病态的场景中带来无意义的预测；因此，JT 在这些情形下不适用。当有了变种模型 “E” 中的 GIM 协助时，JT 提升了模型在 *CoCA* 数据集中的表现。相似地，这样的提升可在我们 GICD 和变种模型 “G” 的对比中体现出来。

Variant	Candidate			CoCA				CoSal2015 [50]			
	JT	GIM	ARM	$F_{\text{avg}}\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\text{avg}}\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
A				0.420	0.430	0.601	0.627	0.788	0.800	0.818	0.852
B	✓			0.424	0.430	0.602	0.655	0.750	0.759	0.782	0.821
C		✓		0.446	0.462	0.618	0.643	0.809	0.824	0.833	0.868
D			✓	0.429	0.437	0.607	0.628	0.800	0.809	0.829	0.860
E	✓	✓		0.470	0.478	0.631	0.689	0.795	0.803	0.808	0.850
F	✓		✓	0.436	0.442	0.612	0.654	0.762	0.770	0.795	0.832
G		✓	✓	0.471	0.480	0.636	0.667	0.826	0.835	0.845	0.879
GICD	✓	✓	✓	0.504	0.513	0.658	0.701	0.835	0.844	0.844	0.883

表 2. 我们的 GICD 方法在 *CoCA* 和 *CoSal2015* 数据集上的消融研究。备选项是拼图训练策略（JT）梯度诱导模块（GIM），以及注意保持模块（ARM）。注意到，没有 GIM 的变种 “A”、“B”、“D”、和 “F” 实际上是 SOD 模型而不是 Co-SOD 模型。实验反应了我们提出的三个贡献的相互作用机制。

综上所述，我们的三个贡献：GIM，ARM 和 JT 可以相互增益，以实现更好的协同显著性检测表现。

6 结论

在这篇论文中，受到人类在 Co-SOD 任务上行为的启发，我们提出了基于梯度诱导的协同显著性检测（GICD）方法；由于缺少 Co-SOD 训练数据，我们提出了一种的拼图训练策略，可以帮助我们在通用 SOD 数据集上训练 Co-SOD 模型；除此之外，我们构建了一个用于 Co-SOD 评估的 *CoCA* 数据集来推进后续在真实场景下 Co-SOD 方法的研究。我们正在项目主页 https://mmcheng.net/gicd/收集前沿 Co-SOD 方法在 *CoCA*，*CoSOD3k* 和 *CoSal2015* 的预测结果（显著图）。这些结果可以方便后续的研究进行对比实验。我们鼓励之前工作和后续新工作的作者通过邮箱 zzhang@mail.nankai.edu.cn 向我们提供这三个数据集上的检测结果，具体说明可以在项目主页找到。

致谢： 程明明是通讯作者，张钊和金闻达是共同一作。本研究得到了国家重大项目“新一代人工智能”（No. 2018AAA0100400），国家自然科学基金 (61922046)，天津市自然科学基金 (18ZXZNGX00110)，和南开大学中央高校基础研究经费 (63201169) 的资助。

参考文献

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604 (2009)
2. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: iCoseg: Interactive co-segmentation with intelligent scribble guidance. In: CVPR. pp. 3169–3176. IEEE (2010)
3. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE TIP **24**(12), 5706–5722 (2015)
4. Cao, X., Tao, Z., Zhang, B., Fu, H., Feng, W.: Self-adaptively weighted co-saliency detection via rank constraint. IEEE TIP **23**(9), 4175–4186 (2014)
5. Chang, K.Y., Liu, T.L., Lai, S.H.: From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In: CVPR 2011. pp. 2129–2136. IEEE (2011)
6. Chen, H.T.: Preattentive co-saliency detection. In: ICIP. pp. 1117–1120. IEEE (2010)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
8. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV. pp. 186–202 (2018)
9. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. pp. 4548–4557 (2017)
10. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: IJCAI. pp. 698–704 (2018)
11. Fan, D.P., Lin, Z., Ji, G.P., Zhang, D., Fu, H., Cheng, M.M.: Taking a deeper look at the co-salient object detection. In: CVPR (2020)
12. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: ECCV (2020)
13. Fu, H., Cao, X., Tu, Z.: Cluster-based co-saliency detection. IEEE TIP **22**(10), 3766–3778 (2013)
14. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: ECCV (2020)
15. Gao, Z., Xu, C., Zhang, H., Li, S., de Albuquerque, V.H.C.: Trustful internet of surveillance things based on deeply-represented visual co-saliency detection. IEEE Internet of Things Journal (2020)
16. Han, J., Cheng, G., Li, Z., Zhang, D.: A unified metric learning-based framework for co-saliency detection. IEEE TCSVT **28**(10), 2473–2483 (2017)

17. Hsu, K.J., Tsai, C.C., Lin, Y.Y., Qian, X., Chuang, Y.Y.: Unsupervised CNN-based co-saliency detection with graphical optimization. In: ECCV. pp. 485–501 (2018)
18. Jerripothula, K.R., Cai, J., Yuan, J.: CARS: Co-saliency activated tracklet selection for video co-localization. In: ECCV. pp. 187–202. Springer (2016)
19. Jerripothula, K.R., Cai, J., Yuan, J.: Efficient video object co-localization with co-saliency activated tracklets. IEEE TCSVT **29**(3), 744–755 (2018)
20. Jiang, B., Jiang, X., Zhou, A., Tang, J., Luo, B.: A unified multiple graph learning and convolutional network model for co-saliency estimation. In: ACM Multimedia. pp. 1375–1382 (2019)
21. Joulin, A., Tang, K., Fei-Fei, L.: Efficient image and video co-localization with frank-wolfe algorithm. In: ECCV. pp. 253–268. Springer (2014)
22. Li, B., Sun, Z., Tang, L., Sun, Y., Shi, J.: Detecting robust co-saliency with recurrent co-attention neural network. In: IJCAI. pp. 818–825 (2019)
23. Li, H., Ngan, K.N.: A co-saliency model of image pairs. IEEE TIP **20**(12), 3365–3375 (2011)
24. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: CVPR. pp. 577–585 (2018)
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
26. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
27. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR. pp. 3917–3926 (2019)
28. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE TPAMI **33**(2), 353–367 (2010)
29. Luo, Y., Jiang, M., Wong, Y., Zhao, Q.: Multi-camera saliency. IEEE TPAMI **37**(10), 2057–2070 (2015)
30. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: ECCV. pp. 181–196 (2018)
31. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS. pp. 8024–8035 (2019)
32. Plaut, D.C.: Graded modality-specific specialisation in semantics: A computational account of optic aphasia. Cognitive neuropsychology **19**(7), 603–639 (2002)

33. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: CVPR. pp. 5822–5830 (2018)
34. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR. pp. 7479–7489 (2019)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
36. Sun, H., Zhen, X., Zheng, Y., Yang, G., Yin, Y., Li, S.: Learning deep match kernels for image-set classification. In: CVPR. pp. 6240–6249 (2017)
37. Tang, K., Joulin, A., Li, L.J., Fei-Fei, L.: Co-localization in real-world images. In: CVPR. pp. 1464–1471 (2014)
38. Tsai, C.C., Li, W., Hsu, K.J., Qian, X., Lin, Y.Y.: Image co-saliency detection and co-segmentation via progressive joint optimization. IEEE TIP **28**(1), 56–71 (2018)
39. Wang, C., Zha, Z.J., Liu, D., Xie, H.: Robust deep co-saliency detection with group semantic. AAAI **33**, 8917–8924 (2019)
40. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
41. Wei, L., Zhao, S., Bourahla, O.E.F., Li, X., Wu, F.: Group-wise deep co-saliency detection. In: IJCAI. pp. 3041–3047 (2017)
42. Wei, L., Zhao, S., Bourahla, O.E.F., Li, X., Wu, F., Zhuang, Y.: Deep group-wise fully convolutional network for co-saliency detection with graph propagation. IEEE TIP (2019)
43. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI **39**(11), 2314–2320 (2016)
44. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: ICCV. pp. 1800–1807 (2005)
45. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: ICCV. pp. 7264–7273 (2019)
46. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. In: ICCV. pp. 7223–7233 (2019)
47. Zha, Z., Wang, C., Liu, D., Xie, H., Zhang, Y.: Robust deep co-saliency detection with group semantic and pyramid attention. IEEE TNNLS pp. 1–11 (2020)
48. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: CVPR. pp. 5217–5226 (2019)
49. Zhang, D., Fu, H., Han, J., Borji, A., Li, X.: A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. ACM TIST **9**(4), 1–31 (2018)

50. Zhang, D., Han, J., Li, C., Wang, J., Li, X.: Detection of co-salient objects by looking deep and wide. IJCV **120**(2), 215–232 (2016)
51. Zhang, D., Meng, D., Han, J.: Co-saliency detection via a self-paced multiple-instance learning framework. IEEE TPAMI **39**(5), 865–878 (2016)
52. Zhang, K., Li, T., Liu, B., Liu, Q.: Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In: CVPR. pp. 3095–3104 (2019)
53. Zhang, X., Wei, Y., Yang, Y., Huang, T.: SG-One: Similarity guidance network for one-shot semantic segmentation (2018)
54. Zheng, X., Zha, Z.J., Zhuang, L.: A feature-adaptive semi-supervised framework for co-saliency detection. In: ACM Multimedia. pp. 959–966 (2018)
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)