基于协同表达纯化的协同显著性物体检测

朱子悦　张钊　林铮　孙鑫　程明明

摘要—协同显著性物体检测（Co-SOD）旨在发现一组相关图像中的共同物体。挖掘协同表示对于定位协同显著性物体至关重要。不幸的是，目前的协同显著性物体检测方法并没有充分注意到与协同显著性物体无关的信息被包含在协同表示中。这些协同表示中不相关的信息会干扰对于协同显著性物体的定位。在本文中，我们提出了一种用于搜索无噪声的协同表示的协同表示纯化（CoRP）方法。我们搜索了几个可能属于协同显著区域的像素级别的嵌入。这些嵌入组成了我们的协同表示并且指导我们进行预测。为了获得更加纯净的协同表示，我们利用预测结果迭代地减少我们协同表示中不相关的嵌入。在三个数据集上的实验表明，我们的 CoRP 在基准数据集上实现了最先进的性能。我们的源代码位于https://github.com/ZZY816/CoRP.

Index Terms—**协同显著性检测，显著性物体检测**

1　介绍

人类感知系统 [1] 可以毫不费力地发现最显著的区域。协同显著性物体检测（Co-SOD）旨在从一组相关的图像中发现共同的显著物体。同时，Co-SOD 还需要处理在训练过程中没有学习到的未知的物体种类。这样的能力可以用作许多真实世界应用的预处理步骤，例如，视频协同定位 [2], [3]，语义分割 [4]，图像质量评估 [5] 以及弱监督学习 [6]。Co-SOD 任务的难点在于在杂乱的现实世界环境中发现协同显著的物体。如图1所示，在多个不相关的显著物体中自动地发现并分割出协同显著的物体“香蕉”是一项挑战。

为了区分出协同显著的物体，大多数最先进的（SOTA）方法通过特征聚合 [7], [8]、聚类 [9], [10]、主成分分析 [11], [12]、全局池化 [13], [14], [15] 等方法直接地估计协同表示来捕获协同显著性物体的共享特征。这些方法的协同表示是从图像的所有区域 [9], [11], [13]，或者是预先预测的显著区域 [15], [16] 中总结得来的。尽管它们在许多场景下取得了令人满意的表现，但他们往往忽略了关于不相关的显著物体的噪声信息。

使用有噪声的协同表示会导致对于协同显著性物体不正确的定位，从而限制 Co-SOD 模型的性能，尤其是在复杂的现实场景中。为了克服这个瓶颈，我们尝试减少协同表示中的不相关信息。与目前通过总结图像中所有区域 [8], [9], [11], [13] 或显著性区域 [15], [16] 来直接获得协同表示的方法不同，我们提出了一种迭代过程来搜索仅属于协同显著区域的可信位置作为我们的指导协同显著性物体的完全分割的协同表示。

- 朱子悦、张钊、林铮和程明明在中国天津的南开大学计算机科学学院媒体计算实验室从事研究工作。
- 张钊在商汤研究院工作。
- 孙鑫在腾讯优图实验室工作。
- * 表示同等贡献。
- 程明明是通讯作者。(cmm@nankai.edu.cn).

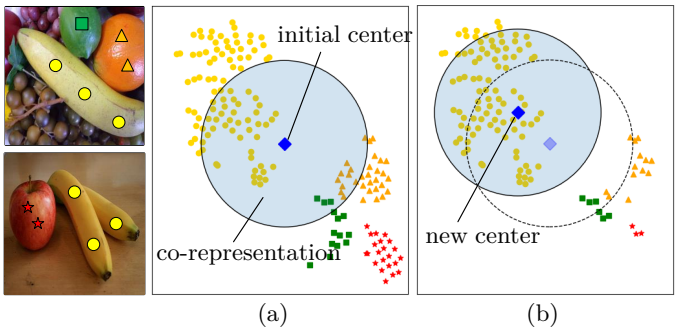

图 1. 嵌入的 t-SNE [17] 可视化结果。(a) “•” **表示协同显著性物体“香蕉”的嵌入。我们观察到非常靠近中心（蓝色圆圈区域内）的嵌入很有可能属于协同显著物体。我们利用它们作为我们的协同表示来定位协同显著的物体。**(b) **当通过我们的初始预测过滤掉许多无关物体的嵌入时，我们可以获得一个更少受到无关嵌入干扰的新中心。新中心有助于搜索更加纯净的协同表示，从而实现更准确的预测。**

具体来说，我们首先提出纯净协同表示搜索 (PCS) 来找到有信心属于协同显著区域的嵌入作为我们的协同表示。如图1所示，在所有显著性物体的像素嵌入中，协同显著性物体的嵌入占据主要地位，因为协同显著物体在图像组中具有重复性。当通过总结显著区域的所有嵌入来获得中心时，我们发现更接近中心的嵌入更有可能是协同显著的嵌入。基于这样的观察，我们没有直接使用不完美的中心来检测协同显著性物体 [12], [15]，而是将中心视为索引与协同表示有高相关性的嵌入的一个代理。与从所有显著区域总结的代理相比，我们的由有信心的协同显著嵌入组成的协同表示更少地受到无关噪声的干扰。

考虑从 PCS 得到的索引的协同表示仍然包含不相关的嵌入，我们提出了循环代理纯化 (RPP)，使用预测的协同显著性图来迭代地纯化协同表示。在获得协同显著性图的预测之后，我们使用该预测来过滤掉更多的噪声并获得一个新的代理。新的代理帮助 PCS 搜索具有更少噪声的协同表示，以获得更准确的预测。我们反复执行上述过程，以纯化我们的协同

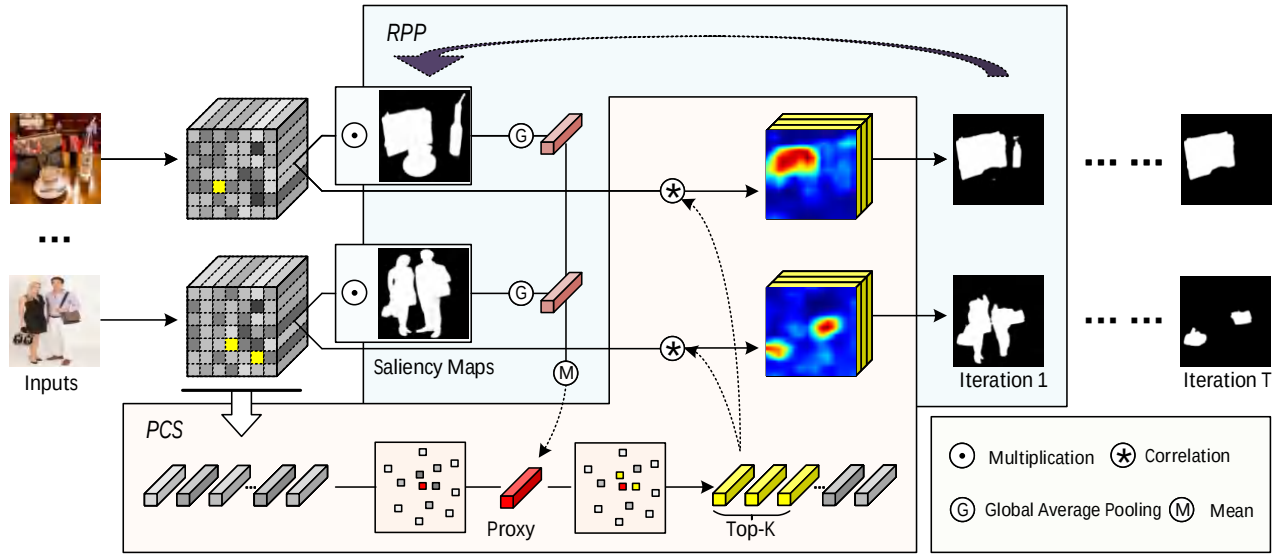

图 2. CoRP 的总体框架。 PCS" 和 "RPP" 表示提出的纯净协同表示搜索 (第3.2节) 和循环代理纯化 (第3.3节)。如上所示，当接收一组图像时，相应的显著性图首先被主干网络共享的显著性物体检测 (SOD) 头预测。在通过显著性图过滤背景噪声后生成了一个协同表示代理。在代理的帮助下，PCS 搜索纯净的协同表示，纯净的协同表示指导了协同显著性预测。RPP 对协同显著性图作出反馈而计算新代理，新代理有助于搜索更纯净的协同表示。随着 PCS 和 RPP 的协作，预测中的噪声被迭代地去除。简洁起见，我们没有绘制我们的编码器-解码器结构和与 Co-SOD 网络共享主干网络参数的 SOD 头。

表示。在 PCS 和 RPP 的交替工作下，我们的协同表示中的不相关嵌入逐渐减少。也就是说，迭代过程使我们的表示越来越纯净。我们在后面的章节中将我们的方法简称为 CoRP (协同表示纯化)。总之，我们的主要贡献如下。

- 我们提出了两种纯化策略：(i) PCS 用于挖掘无噪声的协同表示以及 (ii) RPP 基于先前的协同显著性图来迭代地减少噪声。
- CoRP 在具有挑战性的数据集 CoCA [13], CoSOD3k [18], 以及 CoSal2015 [19] 上实现了 SOTA 的性能。

2 相关工作

利用底层一致性的 Co-SOD. 自 Jacobs 等人 et al. [20] 首次提出在一系列相关图像中搜索共同显著物体的任务以来，Co-SOD 已经被学术界探索了十余年。早期的方法 [2], [21] 专注于探索多幅图像中协同显著性物体的底层的一致性。一些方法 [7], [22] 通过收集不同图像中相似的像素或特征分布中常见的底层线索来提取协同显著性物体。其他的研究通过聚类 [23]、度量学习 [24]，或者高效流形排序 [25] 来探索一组图像中的共享线索。在搜索图像间的重复性之前，一些其他工作 [21], [26], [27] 利用图像的显著性图来过滤背景噪声，以更准确地提取协同显著物体。

最近，多种基于深度学习的方法 [19], [28], [29] 如雨后春笋般涌现，并显著优于以前的传统方法。这些方法主要集中于学习图像组的协同表示，并使用它作为发现协同显著性物体的限制信号。

利用深度一致性的 Co-SOD. Wei 等人 [8] 将一组图像特征连接起来作为协同表示，并将其合并回每个单独的图像特征以进行自动预测。CSMG [11] 对图像特征应用改进的主成分分析法来提取用于生成先验掩膜的协同表示。Zha 等人 [9] 通过 SVM 分类器将一组特征压缩为一个向量，并将向量视为协同表示。GCAGC [10] 通过聚类生成协同注意力图并将聚类中心作为协同表示。GICD [13] 将一组图像的每个样本输入到预训练的分类器并将所有输出汇总作为协同表示。GCoNet [14] 使用组平均池化生成的协同表示来探索图像组内的紧密性与组间的可分离性。

利用显著性区域细化的 Co-SOD. 一些方法利用了显著性预测来提高它们的性能。CoAD-Net [16] 设计了一种按组进行的信道混洗来打破输入数量的固定限制。被显著性图掩蔽的混洗特征被连接在一起以获得协同表示。ICNet [15] 加权平均了每一个由预先预测的显著性图掩蔽的特征作为协同表示，然后将其与每个样本的每个位置比较余弦相似度。CoEGNet [12] 使用一个用于生成协同注意力图的协同注意力投影算法来提取协同表示。

现存的方法主要通过聚类 [9], [10], [23]，主成分分析 [11], [12]，度量学习 [24]，流形排序 [25] 或全局池化 [13], [14], [15] 直接地发现一致性表示来提取协同表示。即使对显著性区域进行细化，这些方法仍然难以避免与无关的显著性物体相关的有噪声的协同表示。

显著性物体检测. 显著性物体检测 (SOD) 旨在发现单个图像中吸引人类视觉注意力的区域 [30]。传统的 SOD 方法 [31],

图 3. 我们的纯净协同表示中的嵌入的位置以及通过我们的纯净协同表示转换的特征 A^t 的信道的可视化结果。在第一行中，图片中的每个红点 "•" 表示前 K 个空间位置中的一个，它们相应的表示组成了纯净协同表示。每个热力图表示转换后特征的一个信道 (见第 3.2节)。它是根据目标图像的深度表示和上述红点所代表的表示计算的相关图。

[32], [33], [34], [35] 主要依靠手工制作的特征来利用底层的线索。受益于神经网络在分割任务中的发展，许多最近的 SOD 方法 [36], [37], [38], [39], [40], [41] 设计了新颖的网络架构并进行了像素级预测。

SOD 和 Co-SOD 都是二进制分割任务，其真实值都是二进制掩码。SOD 为 Co-SOD 研究提供了几个方面的益处。Co-SOD 方法 [15], [16] 可以适用单一显著性图来过滤背景噪声从而更好地定位协同显著物体。同时，精确的单一显著性图可以提高 Co-SOD 方法 [12] 分割物体细节的能力。此外，在 GICD [13] 提出拼图策略后，大量 SOD 训练数据可以被用于训练 Co-SOD 模型。

3 提出的方法

3.1 概述

给定一组图像 $\mathcal{I} = \{\mathrm{I}_n\}_{n=1}^N$，协同显著性物体检测旨在发现它们由协同显著性图 $\mathcal{M} = \{\mathrm{M}_n\}_{n=1}^N$ 表示的共同的显著物体。我们的 CoRP 基于两个关键的互补纯化过程：纯净协同表示搜索 (PCS) 以及循环代理纯化 (RPP)。我们的 PCS 负责估计带有较少的无关信息的协同表示并定位协同显著性物体，而 RPP 利用预测结果来进一步纯化我们的我们的协同表示。

如图2所示，我们的 PCS 从提取 $\mathcal{I}$ 中每一个样本的 ℓ_2 标准化后的深度表示 $\mathcal{F} = \{\mathrm{F}_n \in \mathbb{R}^{D\times H\times W}\}_{n=1}^N$ 开始。对于图像组中的每一个 $N \times H \times W$ 的空间位置，都有一个 D 维的特征向量，它要么对应于目标的共同显著性物体，要么对应于不相关的前景和背景。类似于 [12], [15]，我们使用主干网络共享的 SOD 头预测出的显著性图作为掩膜来选择与潜在共同对象区域相对应的特征向量，这些特征向量被求平均以获得一个协同表示代理 $\boldsymbol{p} \in \mathbb{R}^D$。虽然该协同表示代理包含了关于协同显著性物体的丰富信息，但平均的表示 $\boldsymbol{p}$ 很容易包含不相关的前景信息。因此，我们尝试从图像特征组中搜索一些最有信心的像素级嵌入用于表示共同的物体。更具体来说，$\mathcal{F}$ 中的 K 个嵌入 $\{\boldsymbol{c}_k \in \mathbb{R}^D\}_{k=1}^K$ 根据它们到协同表示代理 $\boldsymbol{p}$ 的距离来选择，这 K 个嵌入作为我们的协同表示 $\mathrm{C} \in \mathbb{R}^{K\times D}$。

为了获取更纯净的协同表示 C，RPP 利用我们预测的协同显著性图来纯化我们的协同表示。具体而言，RPP 对现有的预测 $\mathcal{M}$ 作出反馈而计算出更准确的协同表示代理 $\boldsymbol{p}$。使用新的协同表示代理，PCS 可以探索出更纯净的协同表示，最终产生更好的预测。我们的 CoRP 逐渐消除了背景和外来前景噪声的干扰。我们在第3.2节和第3.3节中详细介绍了 PCS 和 RPP。

3.2 纯净协同表示搜索 (PCS)

假设我们已经获得了通过 RPP 获得了第 t 次迭代的协同表示代理 $\boldsymbol{p}^t \in \mathbb{R}^D$ (见第3.3节)。这里，协同表示代理 $\boldsymbol{p}^t$ 是一个由协同显著性物体主导的语义嵌入。然后，深度表示组 $\mathcal{F} = \{\mathrm{F}_n \in \mathbb{R}^{D\times H\times W}\}_{n=1}^N$ 包含 NHW 个像素嵌入 $\{\mathcal{F}^{(l)} \in \mathbb{R}^D\}_{l=1}^{NHW}$。PCS 旨在从像素嵌入中找到与受协同显著性主导的协同表示代理 $\boldsymbol{p}^t$ 最接近的前 k 个嵌入 $\{\boldsymbol{c}_k^t \in \mathbb{R}^D\}_{k=1}^K$，具体而言，我们通过以下公式计算每个像素嵌入 $\mathcal{F}^{(l)} \in \mathbb{R}^D$ 与 $\boldsymbol{p}^t$ 之间的相关分数。

$$\mathrm{Score}^{t(l)} = \boldsymbol{p}^t \mathcal{F}^{(l)\top}. \tag{1}$$

按降序排序得分后，我们记录具有最高相关性得分的前 k 个嵌入的空间位置 Index^t。

$$\mathrm{Index}^t = \arg\operatorname*{topk}_l \left(\mathrm{Score}^{t(l)}\right) \in \mathbb{R}^K. \tag{2}$$

根据位置 Index^t，我们从 $\mathcal{F}$ 中收集相应的前 k 个嵌入作为我们的协同表示。

$$\mathrm{C}^t = \mathrm{gather}\left(\mathcal{F}, \mathrm{Index}^t\right) \in \mathbb{R}^{K\times D}. \tag{3}$$

一旦获得协同表示 C^t，我们通过协同表示代理 $\boldsymbol{p}^t$ 初步过滤噪声，然后利用协同表示 C^t 将 $\mathcal{F}$ 中的每一个特征 F_n 转换为一组相关性图 $\mathrm{A}_n \in \mathbb{R}^{K\times H\times W}$，具体公式如下：

$$\mathbb{A}_n^t = \mathrm{C}^t\left(\left(\boldsymbol{p}^t\mathbb{F}_n\right)\odot\mathbb{F}_n\right) \in \mathbb{R}^{K\times HW}, \tag{4}$$

其中 $\mathbb{F}_n \in \mathbb{R}^{D\times HW}$ 是由 $\mathrm{F}_n \in \mathbb{R}^{D\times H\times W}$ 变形得到的。我们也将 $\mathbb{A}_n^t \in \mathbb{R}^{K\times HW}$ 变形为 $\mathrm{A}_n^t \in \mathbb{R}^{K\times H\times W}$，然后我们可以对于图像组得到 $\mathcal{A}^t = \{\mathrm{A}_n^t\}^N$。最终，我们解码 $\mathcal{A}^t$ 来预测协同显著性图 $\mathcal{M}^t$。

我们进一步详细地解释变换后的特征 $\mathcal{A}$。$\mathcal{A}^t$ 中的每一个 $\mathrm{A}_n^t \in \mathbb{R}^{K\times H\times W}$ 可以被视为用我们的协同表示的 K 个嵌入 $\{\boldsymbol{c}_k^t \in \mathbb{R}^D\}_{k=1}^K$ 计算的 K 个相关图。在图3中，我们可视化了一些相关图及其相应的嵌入的位置（用红点标注）。根据下面的三点观察，这种变换特征 A_n^t 带来了三个优点。

1) **所找到的稀疏位置落在协同显著性区域中**。它意味着我们的协同表示 C^t 由从属于协同显著物体的语义嵌入中提取的嵌入 $\{\boldsymbol{c}_k^t\}_{k=1}^K$ 组成。这就是我们的协同表示比目前聚合所有位置信息的方法得到的协同表示更纯净的原因。

2) **不同的嵌入向量 $\boldsymbol{c}_k^t$ 聚焦于协同显著性对象的不同区域**。例如图3中的第一个示例（人），从人的头部、胸部和腿部提取的向量为目标人的相应区域提供了更多的激活。类似地，在第二组（人与海豚）中，海豚的向量和人的向量各自激活了两种物体的检测。总而言之，向量 $\boldsymbol{c}_k^t$ 学习协同显著物体的信息。对于每个向量，它所学习的信息与他被提取的位置相关。这种多样性有助于 CoRP 预测更加全面的目标物体区域图。

3) **语义特征 F_n 转换为一个拼接的相关性图** A_n^t，这个相关性图仅由组内表示的关联性产生。这一转换使我们的 CoRP 专注于发现组内的关联性，而不是适应训练集的语义类别。所有这些优点使我们的模型获得了更好的预测。我们将在消融实验中进一步证实这一点。

3.3 循环代理纯化 (RPP)

RPP 基于上一次迭代中预测的协同显著性图 $\mathcal{M}^{t-1} = \{\mathrm{M}_n^{t-1}\}_{n=1}^N$ 来计算协同表示代理 $\boldsymbol{p}^t \in \mathbb{R}^D$。$\mathcal{M}^0$ 由我们的 SOD 头初始化，该头与我们的 Co-SOD 网络共享主干网络特征。SOD 头的详细信息将在第4.1节中解释。当协同显著物体在每一个样本中重复出现时，由 $\mathcal{M}^{t-1}$ 掩蔽的语义表示会由协同显著性物体的嵌入主导。在这种情况下，我们直接对 $\mathcal{F}$ 中被 $\mathcal{M}^{t-1}$ 突出显示的空间位置进行平均，将其作为协同表示代理 $\boldsymbol{p}^t$。

$$\boldsymbol{p}^t = \sum_{n=1}^{N}\frac{\mathrm{GAP}(\mathrm{M}_n^{t-1}\odot\mathrm{F}_n)}{N} \in \mathbb{R}^D, \tag{5}$$

其中 $\odot$ 表示向量按元素相乘，并且全局平均池化 (GAP) 可以被公式化为 $\mathrm{GAP}(\mathrm{F}) = \frac{1}{HW}\sum_i\sum_j \mathrm{F}$，其中 $i = 1, ..., W$，$j = 1, ..., H$。$\boldsymbol{p}^t$ 再通过欧几里得距离标准化。

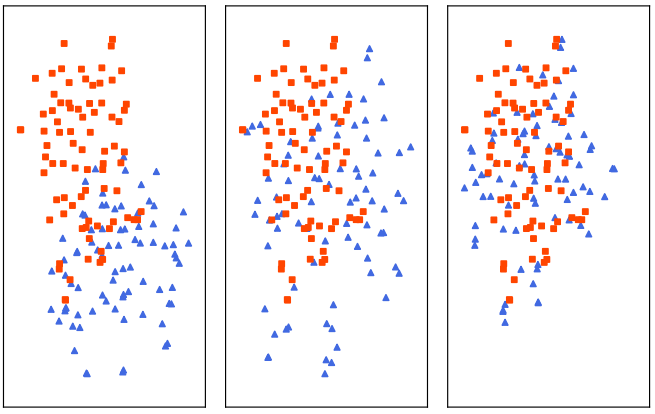

图 4. 在不同迭代轮数中获得的协同表示代理和相应的真实值得来的代理的分布。我们通过 t-SNE [17] 算法可视化 CoCA [13] 中的所有 80 个组。在这里，蓝色三角形 "▲" 是指由协同显著性图掩蔽后计算平均得到的协同表示代理，橙色正方形 "■" 表示由真实值掩蔽后计算平均得到的纯净协同表示代理。结果表明迭代过程使得协同表示代理逐渐接近真实值产生的代理。

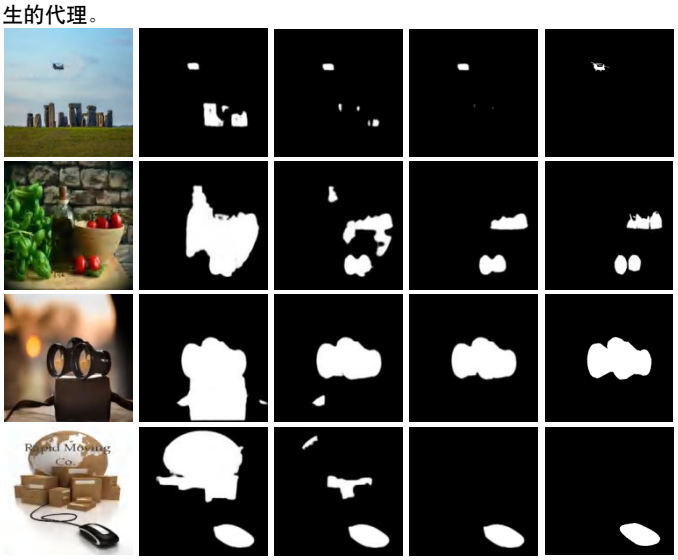

图 5. 不同迭代轮数时的预测结果。我们可以看到，预测结果逐渐地接近真实值标签。这要归功于迭代中逐渐纯净的协同表示。注意到具有较少前景噪声的预测有助于挖掘更纯净的协同表示。

注意，与 $\mathcal{M}^{t-1}$ 相比，$\boldsymbol{p}^t$ 有助于 PCS 预测更加无噪声的 $\mathcal{M}^t$。反过来，$\mathcal{M}^t$ 让 RPP 生成与 $\boldsymbol{p}^t$ 相比噪声更少的 $\boldsymbol{p}^{t+1}$。通过使用纯净协同表示搜索、循环代理纯化，我们的 CoRP 迭代类似于：

$$\begin{cases}\boldsymbol{p}^t &= \mathrm{RPP}\left(\mathcal{M}^{t-1}, \mathcal{F}\right)\\ \mathcal{M}^t &= \mathrm{PCS}\left(\boldsymbol{p}^t, \mathcal{F}\right)\end{cases}, t = \{1, 2, 3, ..., T\}. \tag{6}$$

此外，我们只使用我们模型的编码器一次，编码器不参与后续的迭代过程。换句话说，我们的 RPP 使 PCS 和解码器在没有编码器的情况下反复地工作。在这种情况下，迭代过程不会增加太多的计算负担。我们在算法1中详细介绍了迭代过程。

为了更好地理解迭代中发生了什么，我们分析了三个重要因素的变化。

Algorithm 1: 基于协同表示纯化的协同显著性物体检测

输入: 一组图像 $\mathcal{I} = \{\mathrm{I}_n\}_{n=1}^{N}$
输出: 协同显著性图 $\mathcal{M}^T = \{\mathrm{M}_n^T\}_{n=1}^{N}$
初始化: 提取深度表示 $\mathcal{F} = \{\mathrm{F}_n\}_{n=1}^{N}$; 使用我们主干网络共享的 SOD 头初始化 $\mathcal{M}^0$
for $t \leftarrow 1$ to T do
  RPP
    基于 $\mathcal{F}$ 和 $\mathcal{M}^{t-1}$ 生成协同表示代理 $\boldsymbol{p}^t$(Eq. 5)
  end
  PCS
    通过计算 $\mathcal{F}$ 中每一个空间位置与 pt 的分数获得 Score^t(Eq. 1)
    基于 Score^t 从 $\mathcal{F}$ 中搜索前 k 个最纯净的嵌入作为协同表示 (Eq. 2, 3)
    使用协同表示 C^t 将每一个 F_n 转换为一组相关性图 A_n^t (Eq. 4), 然后得到 $\mathcal{A}^t = \{\mathrm{A}_n^t\}^N$
    通过解码 $\mathcal{A}^t$ 来预测 $\mathcal{M}^t = \{\mathrm{M}_n^t\}_{n=1}^{N}$
  end
end

1) **协同表示代理**逐渐地接近通过正确答案标注得到的值。在图4中，我们利用 t-SNE 可视化了 CoCA [13] 的全部 80 组数据的协同表示代理。在每次迭代中，我们将经协同显著性图掩蔽后平均得到的协同显著代理表示为“蓝色三角形”，而将经真实值标注掩蔽后平均得到的代理表示为“橙色正方形”。随着协同显著性图的预测精度逐渐提高，两种代理之间的距离逐渐缩小。由于橙色节点表示完全由属于协同显著性物体的嵌入生成的无噪声代理，距离的减少意味着我们的 RPP 逐渐减少了我们代理中分散模型注意力的信息。
2) **协同表示**由越来越多的来自协同显著区域的表示组成。在表1中，我们计算嵌入的空间位置落在协同显著性物体上的比例。在三个基准数据集上，平均比例在每次迭代后都会增加。在经过三次迭代后，协同表示中超过三分之二的嵌入都位于协同显著性物体上。比例的增长意味着协同表示包含更多的协同显著信息和更少的分散注意力的噪音。由于代理中分散注意力信息的减少，协同表示随着迭代过程逐渐地更加纯净。
3) **协同显著性图**逐渐消除背景和不相关的前景物体。我们将预测结果展示在图5中。在第一次迭代中，背景干扰被抑制，但仍然包含大量的前景噪声，这些噪声在由更纯净的协同表示引导的进一步迭代中逐渐消失。

4 实验

4.1 实现细节

模型细节. 我们利用预训练的 VGG-16 [42] 作为我们的主干网络并构建一个编码器-解码器结构。在 CoRP 中，初始显著性图 $\mathcal{M}^0$ 由我们主干网络共享的 SOD 头预测。在通过 $\mathcal{M}^0$ 获取到第一个协同表示代理 $\boldsymbol{p}^1$ 后，我们在编码器的最后四个输出上使用我们的 PCS。PCS 旨在定位协同显著性物体并生成四个协同显著特征。我们的解码器与 ICNet [15] 的解码器相同，解码四个协同显著特征和剩余的浅层特征，以预测协同显著性图 $\mathcal{M}^1$。我们的 RPP 对 $\mathcal{M}^1$ 作出反馈以生成新的协同表示代理，并重复刚才提到的过程来预测 $\mathcal{M}^2$。迭代过程产生 $\{\mathcal{M}^1, \mathcal{M}^2, \mathcal{M}^3, ..., \mathcal{M}^T\}$，我们在实验中设置 $T = 3$。原因将在第4.4节中解释。

生成初始显著性图 $\mathcal{M}^0$ 的 SOD 头与我们的 Co-SOD 网络共享主干网络权重。SOD 头的解码器是独立设计的，它直接合并编码器的输出并预测显著性图。SOD 头给我们的 CoRP 增加了 2.9MB 的模型参数。

表 1. 得到协同表示需要的嵌入的稀疏位置落在协同显著性物体上的比例。通过迭代过程，提取协同表示的位置更多位于协同显著性物体的对应特征位置；因此，我们的协同表示变得越来越纯净。

	CoCA [13]	CoSal2015 [19]	CoSOD3k [18]
$T = 1$	60.3%	93.2%	82.4%
$T = 2$	65.3%	94.6%	84.8%
$T = 3$	66.8%	94.9%	85.1%
$T = 4$	67.2%	95.0%	85.2%
$T = 5$	67.3%	95.0%	85.3%
$T = 6$	67.3%	95.0%	85.3%

训练细节. 借鉴 [15], 我们用于训练协同显著性模型（CoRP）的数据集是 DUTS 数据集 [54] 和 COCO 数据集 [55] 的子集，它包含 9213 幅图像。Co-SOD 网络使用 COCO [55] 的子集训练，而主干网络共享的 SOD 头则通过 DUTS 数据集 [54] 训练。Adam 优化器的初始学习率为 1×10^{-5}，权重衰减为 1×10^{-4}。我们一共对模型训练了 70 轮，在 Nvidia Titan X 上大约需要 4.5 个小时。在每个训练迭代中，我们从 COCO [55] 子集图像组中随机选择 10 个样本作为一个批次来训练 CoSOD 网络，并从 DUTS [54] 中随机选择 8 个样本来训练 SOD 头。在测试阶段，每个批次由 CoSOD 组内的所有图像组成，不需要来自 SOD 数据集的图像。在训练和测试阶段，图像都被调整为 224×224，并且我们在训练阶段随机水平翻转图像来进行数据增强。最后，预测结果被调整回原来大小以进行评估。我们的 CoRP 在 Pytorch [56] 和 Jittor [57] 中实现。当迭代次数设置为 3（T=3）时，PyTorch 版本的代码在 Nvidia Titan X GPU 上以 45.3FPS 的速度运行，总参数大小为 80.1MB。

在训练阶段，我们使用真实值作为掩膜来生成无噪声的协同表示代理。通过一个完全无噪声的代理，我们的 PCS 可以准确地定位协同显著物体，这种训练策略使我们的 Co-SOD 网络完全依赖 PCS 来定位物体。这样，我们的 RPP 就没有在训练阶段被使用。在测试阶段，第一个协同表示代理由我

表 2. 我们的 CoRP 与其他方法在 CoCA [13], CoSOD3k [18], 和 CoSal2015 [19] 数据集上对平均绝对误差（MAE）、最大 F 度量 [43]（$F_{\max}$）、S 度量 [44]（S_α）和平均 E 度量 [45]（E_ξ）进行定量比较。DUTS-Class [13], COCO-9k [8], COCO-SEG [9], 和 MSRA-B [46] 是 Co-SOD 中广泛使用的训练数据集，我们将它们分别表示为 Train-1、2、3 和 4。“↑” 标志意味着对应的数值越高，模型性能越好。带下划线的数值表示 VGG16 作为主干网络下的第二优秀的结果。特别的是，我们的 CoRP 显示的是第三次迭代（$T = 3$）的结果。我们提供了我们的模型使用三个骨干网络时的性能：VGG16 [42]、ResNet50 [47] 和 PVTv1-medium [48]。CoRP$^{\mathrm{NAS}}$ 表示，我们通过网络结构搜索 [49]. 在 PCS 中找到最佳超参数 K。

		GateNet ECCV20 [50]	GCPA AAAI20 [51]	CBCD TIP13 [23]	CSMG CVPR19 [11]	GCAGC CVPR20 [10]	GICD ECCV20 [13]	ICNet NIPS20 [15]	CoEG PAMI21 [12]	DeepACG CVPR21 [52]	GCoNet CVPR21 [14]	CADC ICCV21 [53]	CoRP 2021 VGG16	CoRP 2021 VGG16	CoRP$^{\mathrm{NAS}}$ 2021 VGG16	CoRP 2021 Res-50	CoRP 2021 PVT
	训练集	-	-	-	Train-4	Train-3	Train-1	Train-1,2	Train-1	Train-3	Train-1	Train-1,2	Train-2	Train-1,2	Train-1,2	Train-1,2	Train-1,2
CoSal2015	MAE↓	0.097	0.082	0.233	0.130	0.085	0.071	0.058	0.078	0.064	0.068	0.064	0.060	0.049	0.049	0.046	0.044
	$F_{\max}$ ↑	0.772	0.830	0.547	0.787	0.832	0.844	0.859	0.836	0.842	0.847	0.862	0.864	0.885	0.888	0.893	0.895
	S_α ↑	0.811	0.850	0.550	0.776	0.823	0.844	0.855	0.838	0.854	0.845	0.866	0.859	0.875	0.877	0.879	0.884
	E_ξ ↑	0.820	0.864	0.516	0.763	0.814	0.883	0.896	0.868	-	0.884	-	0.896	0.913	0.915	0.919	0.920
CoCA	MAE↓	0.173	0.188	0.180	0.114	0.111	0.126	0.148	0.106	0.102	0.105	0.132	0.101	0.121	0.110	0.104	0.093
	$F_{\max}$ ↑	0.398	0.435	0.313	0.499	0.517	0.513	0.514	0.493	0.552	0.544	0.548	0.564	0.551	0.575	0.607	0.619
	S_α ↑	0.600	0.612	0.523	0.627	0.666	0.658	0.657	0.612	0.688	0.673	0.681	0.699	0.686	0.703	0.719	0.732
	E_ξ ↑	0.609	0.612	0.535	0.606	0.668	0.701	0.686	0.679	-	0.739	-	0.750	0.715	0.741	0.745	0.773
CoSOD3k	MAE↓	0.112	0.104	0.228	0.157	0.100	0.079	0.097	0.084	0.089	0.071	0.076	0.067	0.075	0.072	0.057	0.057
	$F_{\max}$ ↑	0.697	0.746	0.468	0.730	0.779	0.770	0.766	0.758	0.756	0.777	0.759	0.794	0.798	0.801	0.828	0.835
	S_α ↑	0.763	0.795	0.529	0.727	0.798	0.797	0.798	0.778	0.792	0.802	0.801	0.820	0.820	0.825	0.842	0.850
	E_ξ ↑	0.772	0.813	0.509	0.675	0.791	0.845	0.843	0.817	-	0.857	-	0.864	0.862	0.866	0.887	0.891

们的 SOD 头生成的显著性图初始化。同时，RPP 迭代地将预测的协同显著性图送入 PCS，这提供了更精确的物体定位并最终生成更精确的协同显著性图。RPP 的迭代次数可以在测试阶段自由设置，我们将在第4.4节讨论这一点。

损失函数. 我们利用 IoU 损失 [58] 训练 CoRP。在 [13], [15] 中，IoU 损失已被证明对 Co-SOD 任务有效。其具体公式如下：

$$\mathrm{L}(\mathcal{M},\mathcal{G}) = 1 - \sum_{n=1}^{N} \frac{\sum_{w=1}^{W}\sum_{h=1}^{H}\min\{\mathrm{M}_n^{w,h},\mathrm{G}_n^{w,h}\}}{\sum_{w=1}^{W}\sum_{h=1}^{H}\max\{\mathrm{M}_n^{w,h},\mathrm{G}_n^{w,h}\}}. \quad (7)$$

$\mathcal{M} = \{\mathrm{M}_n\}_{n=1}^{N}$ 和 $\mathcal{G} = \{\mathrm{G}_n\}_{n=1}^{N}$ 分别表示由我们模型生成的协同显著性图和真实值。$\mathrm{M}_n^{w,h}$ 表示一组预测中第 n 个协同显著性图 M_n 的一个像素位置，在相应真实值组中的 $\mathrm{G}_n^{w,h}$ 也是同样如此。

为了监督我们的 SOD 头部，我们使用 $\mathcal{S} = \{\mathrm{S}_m\}_{m=1}^{M}$ 和 $\mathcal{T} = \{\mathrm{T}_m\}_{m=1}^{M}$ 来表示预测的显著性图和相应的真实值。我们同样将 IoU 损失 $\mathrm{L}(\mathcal{S},\mathcal{T})$ 用于 SOD 头。我们同时训练 Co-SOD 网络和 SOD 头的最终损失函数如下：

$$\mathrm{L}_{all} = \alpha\mathrm{L}(\mathcal{M},\mathcal{G}) + \beta\mathrm{L}(\mathcal{S},\mathcal{T}), \quad (8)$$

其中我们设置 $\alpha = 0.8$ and $\beta = 0.2$。

4.2 评估数据集和评价指标

数据集. 我们在三个具有挑战性的数据集上评估了我们的 CoRP：CoSal2015 [19], CoSOD3k [18], 以及 CoCA [13]。这三个数据集被广泛用于评估 Co-SOD 方法。Cosal2015 [19] 包含 50 个类别的 2015 幅图像，其中每一组都有一个或多个挑战性的问题，例如复杂的环境、物体外观的变化、遮挡和背景杂乱。CoSOD3k [18] 是迄今为止最大的 CoSOD 评估数据集，它涵盖 13 个父类，包含 160 个组，共 3316 幅图像。每个组都包含了多样的现实场景和不同的物体外观，并涵盖了 Co-SOD 中的主要挑战。CoCA [13] 数据集由 80 个类别共 1295 张图像组成。与 Cosal2015 和 CoSOD3k 相比，CoCA 具有两个特殊的特征。首先，除了协同显著性物体外，每个图像都包含了外部显著性物体。同时，CoCA [13] 数据集中的图像组对我们的 CoRP 来说是完全不可见的类别，因为 CoCA 与我们的训练集之间没有相同的类别。由于这两方面的关键原因，CoCA 上的结果可以更好地反映每个方法在类无关的 Co-SOD 任务上的性能。

评价指标. 借鉴 GICD [13]，我们报告了四个广泛使用的评价指标，即平均绝对误差（MAE）、最大 F 度量（$F_{\max}$）[43]、S 度量（S_α）[44] 以及平均 E 度量（E_ξ）。评估代码位于https://github.com/zzhanghub/eval-co-sod。

4.3 与现有方法的比较

为了说明 CoRP 的性能，我们将我们的方法与九种 SOTA Co-SOD 方法进行了比较，包括 CBCD [23], CSMG [11], GCAGC [10], GICD [13], CoEG-Net [12], ICNet [15], DeepACG [52], GCoNet [14], 以及 CADC [53]。同时，借鉴 [13], [15], [16]，我们还报告了两种 SOTA-SOD 方法的结果作为基线，即 GateNet [50] 和 GCPA [51]。

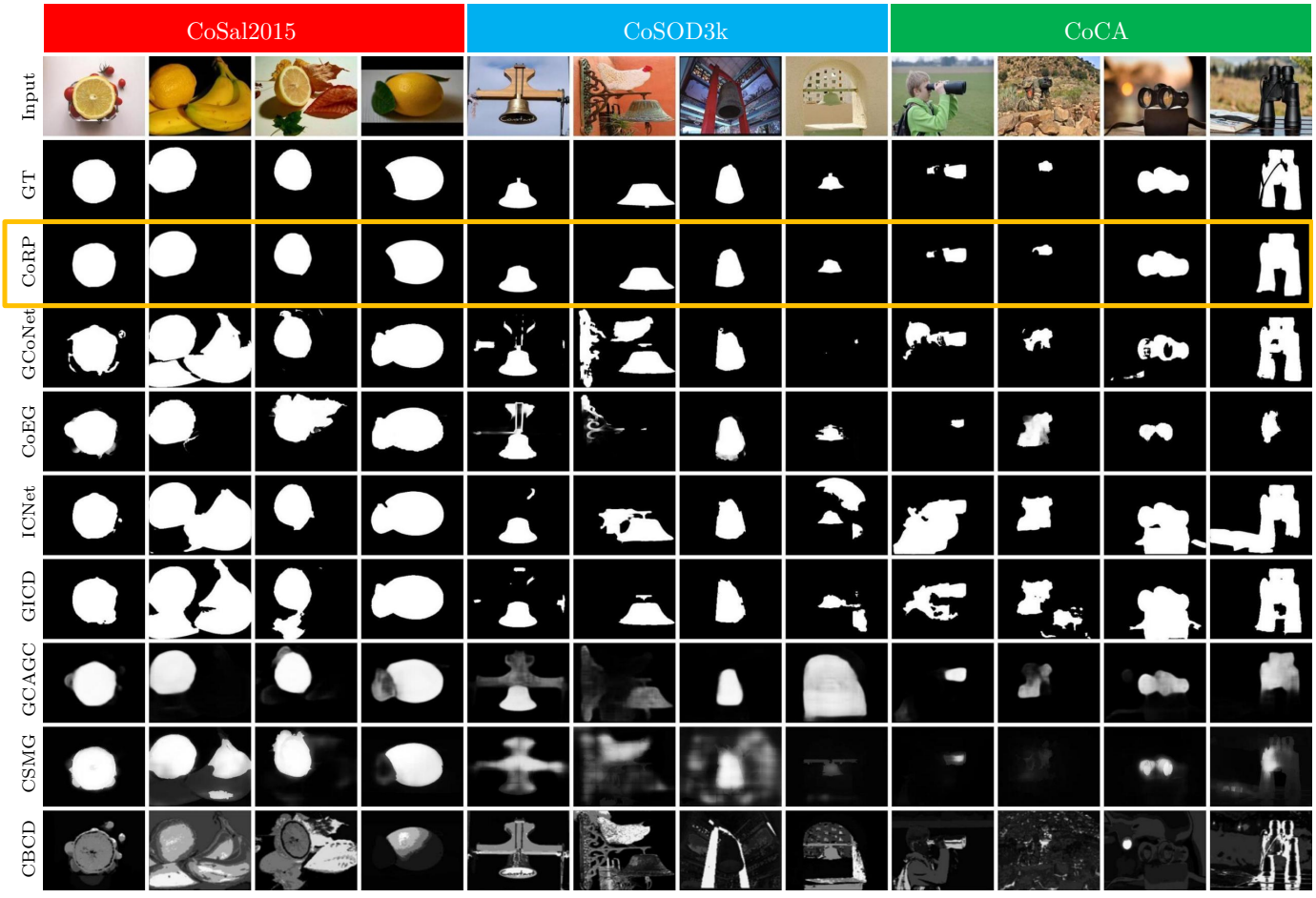

图 6. 我们的 CoRP 与六种 SOTA 方法的可视化比较。我们展示了属于三个基准数据集 (CoSal2015 [19], CoSOD3k [18], 以及 CoCA [13]) 的三个类别（柠檬、钟铃和双筒望远镜）的预测。我们用橙色框突出显示 CoRP 的结果。

定量评估. 在表2中,我们将我们的方法 CoRP 与其他 SOTA Co-SOD 和 SOD 方法的定量结果进行了比较。我们可以看到，我们的 CoRP 在三个基准数据集上实现了 SOTA 性能。在 CoCA [13] 数据集上，在最大 F 度量和 S 度量方面，我们的方法的性能在很大程度上优于其他方法。与 CoCA [13] 数据集上第二优秀的方法相比，我们的 CoRP 方法在 S 度量方面领先 1.5%，在最大 F 度量方面领先 2.3%。由于在 CoCA 的每个图像中都存在协同显著性物体外的外部显著性物体，因此我们的优越性能表明，我们的方法具有更好的定位协同显著性物体的能力。同时，CoCA [13] 不包含出现在我们的训练数据集 COCO [55] 中的任何图像类别。我们在这个数据集上的领先性能表明，我们的方法对类无关任务中的未知类是鲁棒的。注意到，SOD 方法尤其是 GCPANet [51] 的性能，与 CoSOD3k [18] 和 CoSal2015 [19] 数据集上的一些 Co-SOD 方法相当甚至更好，这是因为这两个数据集中的大部分图像仅包含一个显著对象。然而，我们的方法在 S 度量方面也超过了 CoSal2015 [19] 和 CoSOD3k [18] 上第二优秀的 Co-SOD 方法 1.1% 和 2.3%。我们在三个基准数据集上的表现充分证明了我们方法的有效性。此外，在 MAE 方面，我们的方法并不优于所有其他方法。这可能是因为我们的方法分割对象细节的能力不足。

定性结果. 在图6中，我们将我们的方法的预测与其他方法在三个基准数据集上进行了比较。我们的方法成功地将柠檬与图像的其他部分分离，而其他一些方法无法将香蕉与柠檬区分开来。尽管钟铃的图像有着巨大变化的复杂背景，我们的方法仍然定位了钟并将其与背景准确地分离。同时，望远镜上的手很小但是我们的方法可以将望远镜与手分开。

4.4 消融实验

在表3中，我们验证了我们的纯净协同表示搜索（PCS）和循环代理纯化（RPP）的有效性。加入 SOD 头增加了训练数据集和 2.9MB 的模型参数。为了更好地比较，我们同样提供了具有 SOD 头的基线模型的性能。如果没有 PCS，我们的方法将失去检测协同显著性物体的能力，因此在消融实验中，我们使用像 ICNet [15] 和 GCoNet [14] 那样计算代理和特征之间的余弦相似度而产生的相关性图来证明我们的 PCS 比直接使用代理作为协同代表更好。

PCS 的有效性. PCS 被设计用于搜索属于协同显著性物体的多个嵌入作为协同表示，从而精确地探索协同显著信息。与

表 3. **提出的 CoRP 在 CoCA 和 CoSOD3k 数据集上进行的消融实验。**“SOD”是指我们的 SOD 头，它与我们的 Co-SOD 网络共享主干网络权重。“PCS”和“RPP”是提出的纯净协同表示搜索和循环代理纯化。“baseline”是指我们的没有“SOD”、“PCS”和“RPP”的 Co-SOD 编码器-解码器架构。“baseline + SOD + proxy”表示直接使用代理作为协同代表。

ID	Combination	CoCA [13]				CoSal2015 [19]				CoSOD3k [18]			
		$F_{\text{avg}}\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\text{avg}}\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\text{avg}}\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
1	baseline	0.381	0.397	0.560	0.563	0.652	0.662	0.699	0.727	0.576	0.590	0.654	0.687
2	baseline + SOD	0.436	0.443	0.608	0.640	0.760	0.772	0.800	0.830	0.688	0.697	0.753	0.795
3	baseline + SOD + proxy	0.381	0.399	0.551	0.531	0.798	0.825	0.836	0.866	0.675	0.697	0.744	0.766
4	baseline + SOD + PCS	0.474	0.488	0.640	0.658	0.860	0.874	0.869	0.906	0.753	0.764	0.800	0.839
CoRP	baseline + SOD + PCS + RPP	0.541	0.551	0.686	0.715	0.872	0.885	0.875	0.913	0.788	0.798	0.820	0.862

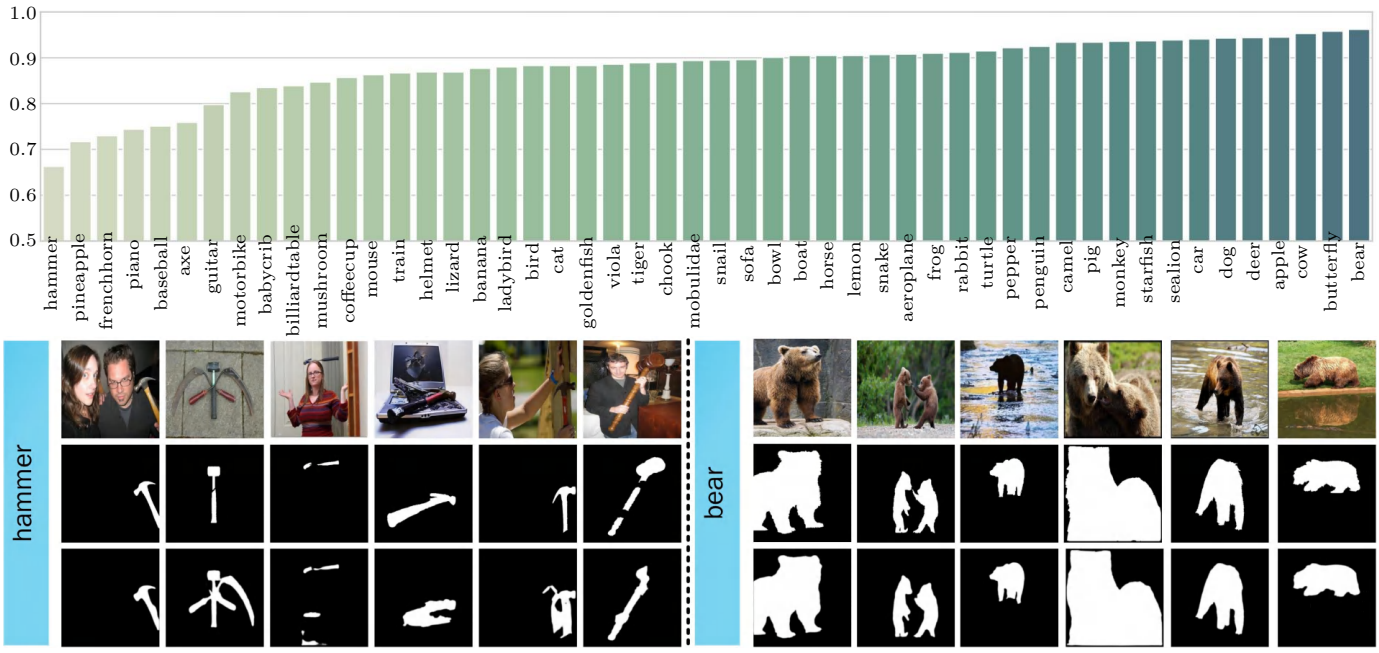

图 7. **我们的模型在 CoSal2015 不同类别上的依据 S 度量评估的表现。**我们展示的预测结果中最差组为“锤子”，最佳组为“熊”。从上到下的这三行对应于输入图像、真实值和我们的预测。

表 4. **在 CoSal2015 和 CoSOD3k 数据集上 PCS 中搜索的稀疏位置数量的影响。**K 表示协同表示中嵌入的数量。K^* 是我们手动选择的使用值。K^{NAS} 是由网络结构搜索找到的值。

	CoCA [13]			CoSal2015 [19]			CoSOD3k [18]		
	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
$K=16$	0.491	0.635	0.650	0.865	0.864	0.900	0.758	0.791	0.827
$K=24$	0.542	0.678	0.708	0.880	0.871	0.909	0.788	0.815	0.856
$K^*=32$	0.551	0.686	0.715	0.885	0.875	0.913	0.798	0.820	0.862
$K=48$	0.556	0.689	0.724	0.889	0.873	0.913	0.793	0.815	0.860
$K=56$	0.528	0.668	0.692	0.877	0.872	0.907	0.780	0.809	0.847
$K=64$	0.504	0.650	0.668	0.866	0.862	0.900	0.767	0.799	0.837
$K^{\text{NAS}}=45$	0.575	0.703	0.741	0.888	0.877	0.915	0.801	0.825	0.866

其他方法相比，我们的 PCS 生成具有较少干扰噪声的协同表示。通过纯净的协同表示，我们的方法可以精确地定位协同显著性物体，并最终将它们与图像的其他部分分离。表3表示，与“baseline+SOD”相比，我们的 PCS 在 CoCA [13]，CoSOD3k [18] 和 CoSal2015 [19] 数据集上的 S 度量方面分别提高了 3.2%、4.7% 和 6.9%。在我们的方法的参数数量不大于“baseline+SOD“的情况下，该结果证明了我们的 PCS 的有效性。此外，协同表示代理可以直接被用作协同表示来探索协同显著物体，但是与“baseline+SOD+proxy”相比，我们的纯净协同表示搜索在 CoCA [13]，CoSOD3k [18] 和 CoSal2015 [19] 数据集上的 S 度量方面相应地提高了 8.9%、5.6% 和 3.3%。该改进得益于第3.2节中分析的优势。

CoRP 的协同表示由 K 个嵌入组成，其中大部分属于协同显著物体。如果 K 太大，我们的协同表示将包含更多的噪声，如果 K 太小，协同显著信息的多样性将会降低。在表4中，我们设计了一个实验，其中超参数 tab:K 被设置为不同的值，以便我们的 CoRP 找到合适的 tab:K。根据在 CoSal2015，

表 5. 在 CoCA 和 CoSOD3k 数据集上，CoRP 随着迭代次数变化的性能结果。T 表示迭代次数。

		CoCA [13]			CoSal2015 [19]			CoSOD3k [18]		
	FPS	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
$T=1$	62.5	0.488	0.640	0.658	0.874	0.869	0.906	0.764	0.800	0.839
$T=2$	56.6	0.532	0.672	0.699	0.883	0.873	0.901	0.790	0.815	0.857
$T=3$	45.3	0.551	0.686	0.715	0.885	0.875	0.913	0.798	0.820	0.862
$T=4$	39.5	0.561	0.691	0.722	0.887	0.875	0.914	0.800	0.821	0.863
$T=5$	34.8	0.565	0.693	0.725	0.887	0.875	0.914	0.800	0.821	0.864
$T=6$	30.1	0.567	0.693	0.727	0.887	0.875	0.914	0.800	0.821	0.864

表 6. 我们的方法在设置批次大小不同时的结果。n_{train} 和 n_{test} 分别表示训练和测试输入数量大小。“all”表示类别的所有图像。

		CoCA [13]			CoSal2015 [19]			CoSOD3k [18]		
n_{train}	n_{test}	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
5	5	0.540	0.674	0.702	0.878	0.872	0.906	0.791	0.810	0.854
5	10	0.549	0.678	0.707	0.881	0.873	0.910	0.795	0.815	0.856
5	15	0.549	0.681	0.712	0.882	0.874	0.912	0.794	0.819	0.860
10	5	0.541	0.676	0.704	0.879	0.871	0.907	0.792	0.813	0.853
10	10	0.548	0.683	0.710	0.883	0.874	0.914	0.797	0.818	0.860
10	15	0.550	0.686	0.713	0.885	0.875	0.912	0.798	0.819	0.859
10	all	0.551	0.686	0.715	0.885	0.875	0.913	0.798	0.820	0.862

CoCA 和 CoSOD3k 数据集上具有不同 K 时的性能，我们在所有其他实验中将 K 设置为 32。

此外，我们使用网络结构搜索（NAS）来找到最佳的 K。具体来说，我们采用 NAS [49] 方法来实现这个搜索过程。表4的最后一行显示，我们使用 NAS 为三个基准 Co-SOD 数据集找到了更好的 K。与随机设置 K 相比，NAS 搜索的 K 对 CoCA 数据集带来了明显的改进。

RPP 的有效性. 我们使用 RPP 来纯化用于搜索纯净协同表示的协同表示代理。许多可视化结果和统计数据证明了迭代过程的有效性。在图4中，基于迭代过程，协同表示代理的分布更加接近了真实值分布。由于有更好的协同表示代理，在表1中，每次迭代后协同表示会由更高百分比的属于协同显著性区域的嵌入组成。图5直观地说明了我们的方法逐渐消除了错误预测。表3和表2显示 RPP 提高了我们的性能，并使我们的 CoRP 能够达到 SOTA 方法。

我们在表5中研究了迭代次数对模型性能的影响。随着迭代过程的进展，我们的模型性能越来越好，但同时性能增长也在逐渐减少。从 T=1 到 T=2，RPP 的有效性非常显著，在 CoCA [13], CoSOD3k [18] 和 CoSal2015 [19] 的 S 度量方面，RPP 相应地提高了 3.2%、1.5% 和 0.4%。由于 CoCA [13] 中有许多分散注意力的前景对象，因此该改进充分揭示了纯化协同表示的进行过程。尽管第四次和第五次迭代在性能上仍有提高，但考虑到方法的性能和计算成本之间的平衡，在所有其他实验中，我们让 CoRP 在第三次迭代（T=3）中结束。

表 7. 我们的模型在不同 α 和 β 下的性能。

		CoCA [13]			CoSal2015 [19]			CoSOD3k [18]		
α	β	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$	$F_{\max}\uparrow$	$S_\alpha\uparrow$	$E_\xi\uparrow$
0.2	0.8	0.541	0.677	0.716	0.881	0.867	0.905	0.792	0.814	0.859
0.4	0.6	0.542	0.677	0.703	0.883	0.874	0.910	0.784	0.813	0.853
0.5	0.5	0.540	0.674	0.700	0.883	0.874	0.910	0.788	0.817	0.856
0.6	0.4	0.547	0.680	0.717	0.886	0.874	0.911	0.790	0.815	0.855
0.8	0.2	0.551	0.686	0.715	0.885	0.875	0.913	0.798	0.820	0.862

输入大小的影响. 在表6中，我们提供了我们的模型在训练和测试期间在一组样本中样本数量不同时的性能。增加训练时样本的数量可以让训练过程更加稳定。但由于 GPU 内存受限，我们的最终模型将数量设置为 10。同时，与训练阶段相比，测试中更大的输入量会带来更明显的改进。

不同图像组的影响. 在图7中，我们展示了不同类别图片的定性和定量的结果。我们发现我们的性能表现相对稳定。大多数类别的 S 度量值都超过 0.850。对于最好的“熊”组，分割结果非常接近真实值。然而，在极少数“困难”组上，性能可能会下降。对于“锤子”组来说，锤子很小，而且相对隐蔽。我们仍然成功定位了锤子，但分割的细节仍然不完美。

超参数 α 和 β 的影响. 在我们的损失函数中，α 和 β 分别控制了协同显著性图和显著性图的监督。为了探索它们对结果的影响，我们将 α 和 β 设置为不同的值，并在表7中展示相应的定性结果。我们可以看到当设置 $\alpha > \beta$ 时可以产生更好的结果，我们最终设置 $\alpha = 0.8, \beta = 0.2$。

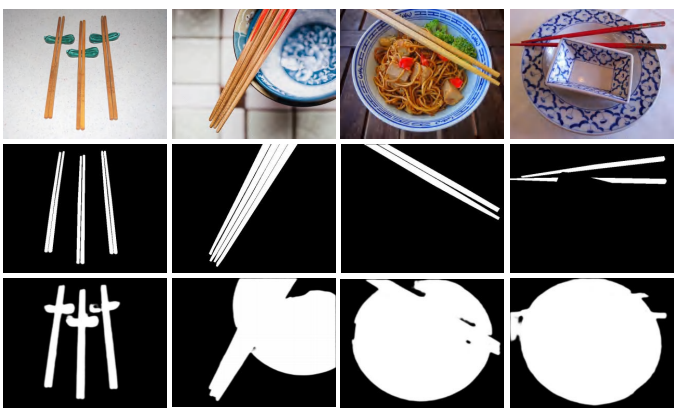

图 8. 我们方法的失败情况。从上到下是图像、真实值和我们的预测。筷子是这一组图像的协同显著性物体。

表 8. 我们的方法和其他协同分割方法在协同分割基准数据集 Internet 上的精确率（$\mathcal{P}$）和杰卡德系数（$\mathcal{J}$）的定量比较。

Internet Dataset	Airplane		Car		Horse		Average	
	$\mathcal{P}(\%)\uparrow$	$\mathcal{J}(\%)\uparrow$	$\mathcal{P}(\%)\uparrow$	$\mathcal{J}(\%)\uparrow$	$\mathcal{P}(\%)\uparrow$	$\mathcal{J}(\%)\uparrow$	$\mathcal{P}(\%)\uparrow$	$\mathcal{J}(\%)\uparrow$
Jerripothula et al. [59]	90.5	61.0	88.0	71.0	88.3	60.0	88.9	64.0
Li et al. [60]	94.1	65.4	93.9	82.8	92.4	69.4	93.5	72.5
Chen et al. [61]	94.1	65.0	94.0	82.0	92.2	63.0	93.4	70.0
Zhang et al. [62]	94.6	66.7	89.7	68.1	93.2	66.2	92.5	67.0
Ours	94.4	83.0	94.1	93.0	93.9	77.0	94.1	84.3

表 9. 我们的方法和其他协同分割方法在协同分割数据集 iCoseg 上的杰卡德系数（$\mathcal{J}$）的定量比较。

iCoseg Dataset	Average$\mathcal{J}(\%)\uparrow$	bear2	brownbear	cheetah	elephant	helicopter	hotballoon	panda1	panda2
Jerripothula et al. [59]	70.4	67.5	72.5	78.0	79.9	80.0	80.2	72.2	61.4
Li et al. [60]	84.2	88.3	92.0	68.8	84.6	79.0	91.7	82.6	86.7
Chen et al. [63]	86.0	88.3	91.5	71.3	84.4	76.5	94.0	91.8	90.3
Zhang et al. [62]	88.0	87.4	90.3	84.9	90.6	76.6	94.1	90.6	87.5
Ours	90.5	91.6	92.8	90.1	91.2	79.3	95.6	93.9	89.5

表 10. 我们的方法和其他方法在协同分割基准数据集 MSRC 上的精确率（$\mathcal{P}$）和杰卡德系数（$\mathcal{J}$）的定量比较。

MSRC Dataset	$\mathcal{P}(\%)\uparrow$	$\mathcal{J}(\%)\uparrow$
Mukherjee et al. [64]	84.0	67.0
Li et al. [60]	92.4	79.9
Chen et al. [63]	95.2	77.7
Zhang et al. [62]	94.3	79.4
Ours	96.0	83.1

4.5 失败的情况

对于一些极具挑战性的场景，我们的方法无法成功地分割协同显著性物体。以图8中的图像为例，筷子是这一组图像的协同显著性物体。然而，显著性物体碗出现在大多数图像中。同时，与碗相比，筷子是更小的物体，碗比筷子更加显著。在这种情况下，很难以消除碗带来的噪声并提取纯净的协同表示，因此预测是不准确的。

4.6 扩展到协同分割问题

为了验证 CoRP 的普遍有效性，我们将我们的方法扩展到图像协同分割领域。事实上，图像的协同分割类似于 Co-SOD。这两个任务都旨在分割一组相关图像中的共同物体。区别在于图像协同分割不需要协同物体是显著的。我们的方法旨在解决 Co-SOD 中的复杂场景问题，也同样在协同分割问题中非常有效。

我们同样将我们的 CoRP 与流行的协同分割方法在三个协同分割基准数据集（包括 MSRC 数据集 [65], iCoseg 数据集 [66], 以及 Internet 数据集 [67]）上进行了定性比较。我们使用两个广泛使用的指标：精确率和杰卡德系数，报告了协同分割模型的性能。

表8显示了在 Internet 数据集上，我们的方法在所有类别和所有指标上都优于其它方法。我们在杰卡德系数指标上的提升非常明显。与第二优秀的方法相比，我们的方法在整体数据集上带来了 17.3% 的提升。在精确率方面，我们的方法也带来了 0.6% 的提高。在表9中，我们在杰卡德系数方面将我们的方法与其他方法进行了比较。我们在所有类别上再一次优于其他方法。在整个 iCoseg 数据集上，与第二好的方法相比，我们的方法提高了 2.5%。此外，表10显示，在 MSRC 数据集上，我们的方法在精确率和杰卡德系数方面分别带来 1.7% 和 3.7% 的性能改进。在图9中，我们展示了两组（汽车和鸟类）的预测结果。结果表明，我们的方法对目标物体进行了高精度的分割。

5 结论

在本文中，我们观察到，当前的 Co-SOD 方法没有充分注意到对 Co-SOD 任务至关重要的协同表示中的噪声，因此它们的性能受到复杂前景物体的干扰。为了克服这一缺点，我们专注于消除协同表示中分散注意力的信息，并提出了一种有效的方法（CoRP）。我们的 CoRP 采用两种协作策略（PCS 和 RPP）迭代工作，旨在抑制协同表示中的噪声，以获得更准确的预测。简而言之，PCS 被设计用于在 RPP 提供的协同表示代理的帮助下，从属于协同显著性物体的稀疏位置搜索纯净的协同表示，然后我们的 PCS 将新预测的协同显著性图反馈到 RPP。反过来，RPP 再一次使用新的预测结果来计算带有更少分散注意力信息的协同表示代理。以这种方式，在迭代过程中，协同表示和预测结果被用来互相改进对方。大量

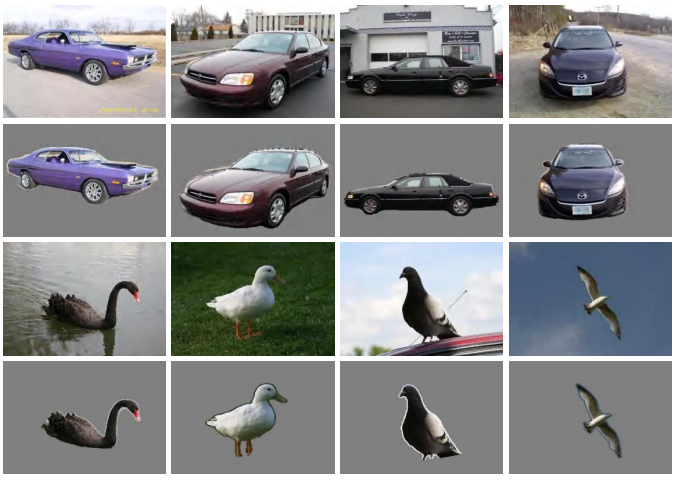

图 9. 我们的方法在协同分割数据集上得到的定性结果。

的可视化结果和实验分析证明了我们的贡献。我们的 CoRP 在三个具有挑战性的数据集上实现了 SOTA 结果。

致谢. 本研究得到了国家自然科学基金（62176130）和中央高校基本科研业务费（南开大学，NO.63223050）的资助。

参考文献

[1] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," IEEE Trans. Circuit Syst. Video Technol., vol. 29, no. 10, pp. 2941–2959, 2018.

[2] K. R. Jerripothula, J. Cai, and J. Yuan, "Cats: Co-saliency activated tracklet selection for video co-localization," in Eur. Conf. Comput. Vis., 2016, pp. 187–202.

[3] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-wolfe algorithm," in Eur. Conf. Comput. Vis., 2014, pp. 253–268.

[4] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in Int. Conf. Comput. Vis., 2019, pp. 7223–7233.

[5] X. Wang, X. Liang, B. Yang, and F. W. Li, "No-reference synthetic image quality assessment with convolutional neural network and local image saliency," Computational Visual Media, vol. 5, no. 2, pp. 193–208, 2019.

[6] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 6024–6033.

[7] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," IEEE Trans. Image Process., vol. 28, no. 1, pp. 56–71, 2018.

[8] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, F. Wu, and Y. Zhuang, "Deep group-wise fully convolutional network for co-saliency detection with graph propagation," IEEE Trans. Image Process., vol. 28, no. 10, pp. 5052–5063, 2019.

[9] Z.-J. Zha, C. Wang, D. Liu, H. Xie, and Y. Zhang, "Robust deep co-saliency detection with group semantic and pyramid attention," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 7, pp. 2398–2408, 2020.

[10] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu, "Adaptive graph convolutional network with attention graph clustering for co-saliency detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 9050–9059.

[11] K. Zhang, T. Li, B. Liu, and Q. Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in IEEE Conf. Comput. Vis. Pattern Recog., 2019, pp. 3095–3104.

[12] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," IEEE Trans. Pattern Anal. Mach. Intell., 2021.

[13] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, "Gradient-induced co-saliency detection," in Eur. Conf. Comput. Vis. Springer, 2020, pp. 455–472.

[14] Q. Fan, D.-P. Fan, H. Fu, C. K. Tang, L. Shao, and Y.-W. Tai, "Group collaborative learning for co-salient object detection," IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[15] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, "Icnet: Intra-saliency correlation network for co-saliency detection," Adv. Neural Inform. Process. Syst., vol. 33, 2020.

[16] Q. Zhang, R. Cong, J. Hou, C. Li, and Y. Zhao, "CoADNet: Collaborative aggregation-and-distribution networks for co-salient object detection," in Adv. Neural Inform. Process. Syst., 2020.

[17] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." Journal of machine learning research, vol. 9, no. 11, 2008.

[18] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 2919–2929.

[19] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," Int. J. Comput. Vis., vol. 120, no. 2, pp. 215–232, 2016.

[20] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in Proceedings of the 23nd annual ACM symposium on User interface software and technology, 2010, pp. 219–228.

[21] H. Li and K. N. Ngan, "A co-saliency model of image pairs," IEEE Trans. Image Process., vol. 20, no. 12, pp. 3365–3375, 2011.

[22] H.-T. Chen, "Preattentive co-saliency detection," in IEEE Int. Conf. Image Process. IEEE, 2010, pp. 1117–1120.

[23] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," IEEE Trans. Image Process., vol. 22, no. 10, pp. 3766–3778, 2013.

[24] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," IEEE Trans. Circuit Syst. Video Technol., vol. 28, no. 10, pp. 2473–2483, 2017.

[25] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," IEEE Sign. Process. Letters, vol. 22, no. 5, pp. 588–592, 2014.

[26] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," IEEE Trans. Image Process., vol. 23, no. 9, pp. 4175–4186, 2014.

[27] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 2011, pp. 2129–2136.

[28] K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, X. Qian, and Y.-Y. Chuang, "Unsupervised cnn-based co-saliency detection with graphical optimization," in Eur. Conf. Comput. Vis., 2018, pp. 485–501.

[29] B. Li, Z. Sun, L. Tang, Y. Sun, and J. Shi, "Detecting robust co-saliency with recurrent co-attention neural network." in IJCAI, 2019, pp. 818–825.

[30] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," Computational Visual Media, vol. 5, no. 2, pp. 117–150, 2019.

[31] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, 1998.

[32] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, 2015.

[33] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," Int. J. Comput. Vis., vol. 123, no. 2, pp. 215–268, 2017.

[34] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in Int. Conf. Comput. Vis., 2013, pp. 2976–2983.

[35] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 2012, pp. 733–740.

[36] J. Wei, S. Wang, and Q. Huang, "F^3net: fusion, feedback and focus for salient object detection," in Association for the Advancement of Artificial Intelligence, 2020, pp. 12 321–12 328.

[37] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 3203–3212.

[38] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "Poolnet+: Exploring the potential of pooling for salient object detection," IEEE Trans. Pattern Anal. Mach. Intell., 2022.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2117–2125.

[40] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 714–722.

[41] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 3089–3098.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Int. Conf. Learn. Represent., 2015.

[43] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," IEEE Trans. Image Process., vol. 24, no. 12, pp. 5706–5722, 2015.

[44] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in Int. Conf. Comput. Vis., 2017, pp. 4548–4557.

[45] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in IJCAI, 7 2018, pp. 698–704.

[46] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 353–367, 2010.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 770–778.

[48] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Int. Conf. Comput. Vis., 2021, pp. 568–578.

[49] Z. Guo, X. Zhang, H. Mu, W. Heng, Z. Liu, Y. Wei, and J. Sun, "Single path one-shot neural architecture search with uniform sampling," 2020.

[50] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in Eur. Conf. Comput. Vis., 2020, pp. 35–51.

[51] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in Association for the Advancement of Artificial Intelligence, 2020, pp. 10 599–10 606. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6633

[52] K. Zhang, M. Dong, B. Liu, X.-T. Yuan, and Q. Liu, "Deepacg: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance," in IEEE Conf. Comput. Vis. Pattern Recog., 2021, pp. 13 703–13 712.

[53] N. Zhang, J. Han, N. Liu, and L. Shao, "Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection," in Int. Conf. Comput. Vis., 2021, pp. 4167–4176.

[54] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 136–145.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[56] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "PyTorch: An imperative style, high-performance deep learning library," in Adv. Neural Inform. Process. Syst., 2019, pp. 8024–8035.

[57] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, "Jittor: a novel deep learning framework with meta-operators and unified graph execution," Science China Information Sciences, vol. 63, no. 222103, pp. 1–21, 2020.

[58] H. Lin, X. Qi, and J. Jia, "Agss-vos: Attention guided single-shot video object segmentation," in Int. Conf. Comput. Vis., 2019, pp. 3949–3957.

[59] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," IEEE Trans. Multimedia, vol. 18, no. 9, pp. 1896–1909, 2016.

[60] W. Li, O. H. Jafari, and C. Rother, "Deep object co-segmentation," in Asia Conf. Comput. Vis. Springer, 2018, pp. 638–653.

[61] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3632–3647, 2020.

[62] K. Zhang, J. Chen, B. Liu, and Q. Liu, "Deep object co-segmentation via spatial-semantic network modulation," in Association for the Advancement of Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12 813–12 820.

[63] H. Chen, Y. Huang, and H. Nakayama, "Semantic aware attention based deep object co-segmentation," in Asia Conf. Comput. Vis. Springer, 2018, pp. 435–450.

[64] P. Mukherjee, B. Lall, and S. Lattupally, "Object coseg-

mentation using deep siamese network," arXiv preprint arXiv:1803.02555, 2018.

[65] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in Eur. Conf. Comput. Vis. Springer, 2006, pp. 1–15.

[66] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 2010, pp. 3169–3176.

[67] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 1939–1946.

朱子悦 目前是南开大学计算机学院的硕士生，师从程明明教授。他的研究兴趣包括深度学习和计算机视觉

张钊 目前是商汤集团有限公司的研究员。他在南开大学获得硕士学位，师从程明明教授，并在扬州大学获得学士学位。他的研究兴趣主要集中在图像处理、计算机视觉和深度学习。

林铮 目前是南开大学计算机学院的博士研究生，导师为程明明教授。他的研究兴趣包括深度学习、计算机图形学和计算机视觉。

孙鑫 目前是腾讯优图实验室的首席研究员。在此之前，他于 2016 年获得香港大学博士学位。他的研究兴趣包括图像处理、机器学习和计算机视觉。

程明明 于 2012 年获得清华大学博士学位。然后，他在牛津大学与 Philip Torr 教授做了两年的研究员。他现在是南开大学的教授，领导媒体计算实验室。他的研究兴趣包括计算机图形学、计算机视觉和图像处理。他获得了 ACM 中国新星奖、IBM Global SUR Award、CCF-Intel Young Faculty Riesearcher Program 等研究奖项。