GCoNet+: 一种更强的组合作式协同显著物体检测器

郑鹏, 付华柱, 范登平†, 范琦, 秦杰, 戴宇荣, 郑智强, Luc Van Gool

Abstract—本文提出了一种新颖的端到端组协同学习网络*GCoNet+*，可以极高效率检测自然场景中的协同显著对象。本文提出的*GCoNet+*在CoSOD任务上达到了新的SoTA水准。该模型主要通过两种关键指标以挖掘共识信息：1）组内紧凑型：通过使用本文提出的组亲和模块（GAM），捕捉它们固有的共享属性，从而更好地构建协同显著对象间的一致性；2）组间分离性：通过引入本文的组对比模块（GCM）对不一致的共识信息进行筛选，从而有效抑制了噪声物体带来的负面影响。为了进一步提升准确性，本文设计了一系列简单但有效的如下结构：i）循环式辅助分类模块（RACM）以促进模型在语义等级的学习；ii）置信增强模块（CEM）以辅助模型提升最终预测结果图的质量；iii）基于群的对称三元组损失函数（GST loss）以引导模型学习更具区分度的特征。本文在三个富有挑战性的基准上（CoCA、CoSOD3k和CoSal2015）进行了更多的实验，实验得到的结果验证了本文的*GCoNet+*超越了现有的12种最新的先进模型。代码和权重已公开于：https://github.com/ZhengPeng7/GCoNet_plus。

Index Terms—协同显著性，CoSOD，组协同学习，深度学习。

1 引言

协同显著物体检测（**CoSOD**）旨在检测一组相关图片中的共同的显著物体。和标准的显著物体检测（SOD）任务相比，CoSOD更具有挑战且需要在区分不同图像中出现的共同物体的同时鉴别干扰物。为此，组内紧凑性和组间分离性作为两种重要线索，应当被同时学习。随着最新CoSOD方法的准确性和效率的提升，CoSOD不仅仅是被用于其他视觉任务的预处理模块 [3]–[7]，并用于许多实际应用中 [1], [8], [9]。

现有工作尝试在单个图像组内利用语义连接 [10]–[12]或不同的共享线索 [13]–[15]，从而得到图片中的一致性以处理CoSOD任务。在 [16], [17]中所提出的模型共同优化了一个统一的网络，用以生成显著图和协同显著信息。尽管这些方法带来了性能上的提升，但其中大部分仅仅依赖单个组内的一致性特征表达 [17]–[22]，这可能导致了以下限制。首先，同一个组的图像仅能提供正面关系而不是兼具不同物体之间的正面和负面关系。仅用单个组内的正样本训练模型可能导致过拟合与在边缘图像样本的模棱两可的结果。此外，一个组里通常仅拥有数量不多的图像（大多数的CoSOD数据集每组常常有20到40张图像）。因此，由单个组内学到的信息常常不足以构成一个具有区分度的表达。最终，单个的图像组们可能不容易挖掘语义线索，而这对于在复杂现实场景里测试中，区分噪声物体非常重要。由于现实场景图像上下文的复杂性，为挖掘共同信息设计一个模块非常有必要。在此之外，当用交叉熵（BCE）损失监督时，生成的结果图的像素值往往更接近0.5而非0或1。受困于不确定性，这些结果图很难直接应用到现实应用之中。

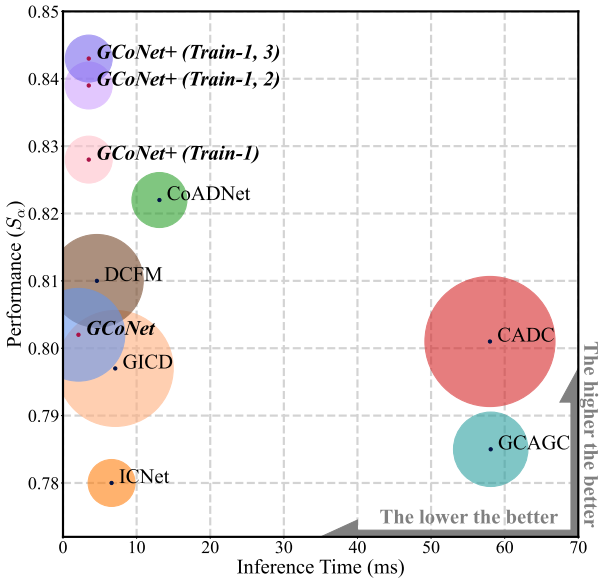

图 1. 七种现有的代表性**CoSOD**方法以及本文方法在**CoSOD3k**数据集上的比较. 本文对现有代表性深度学习CoSOD方法和本文方法从速度（横坐标）与准确度（纵坐标）两方面进行了比较。泡泡的大与小意味着模型的大与小。本文的*GCoNet+*从效果和效率两方面胜过了现有的模型。其中，“Train-1，2和3”分别代表了DUTS_class、COCO-9k和COCO-SEG数据集（更多相关细节见于表. 3）。所有模型都基于一块A100，batch size为2进行测试。本文的推理速度基准可见于：https://github.com/ZhengPeng7/CoSOD_fps_collection。

为了克服以上的限制，本文提出了一个新的组协同学习网络（GCoNet）以建立起相同组内的语义共识和不同组间的区分度。本文的GCoNet包含了三个基本模块：组亲和模块（GAM）、组对比模块（GCM）和辅助分类模块（ACM），同时共同引导本文的GCoNet更好地学习组间分离性和组内紧凑型。具体而言，GAM让模型能够学习同一个图像组内的共识特征，而GCM能够区分不同组之间的目标属

- 郑鹏于芬兰*Aalto University*，南京航空航天大学计算机学院和阿联酋*MBZUAI*从事研究工作。
- 付华柱于新加坡科学、技术和研究局（*A*STAR*）的高性能计算研究所从事研究工作。
- 范登平于瑞士*CVL-ETHZ*从事研究工作。
- 范琦于香港科技大学从事研究工作。
- 秦杰于中国南京航空航天大学计算机学院从事研究工作。
- 戴玉荣于香港科技大学从事研究工作。
- 郑智强于香港科技大学从事研究工作。
- *Luc Van Gool*于瑞士*CVL-ETHZ*和比利时*KU Leuven*从事研究工作。

- 这项工作的初步版本已经出现在*CVPR 2021 [1]*。

- † 本工作主要部分是郑鹏于*IIAI*做实习生时完成，由范登平指导。
- 郑鹏和付华柱同等贡献。通讯作者：范登平（*denfan@ethz.ch*）。

- 本文为*TPAMI2023 [2]*的中文翻译版。由郑鹏翻译，范登平校稿。

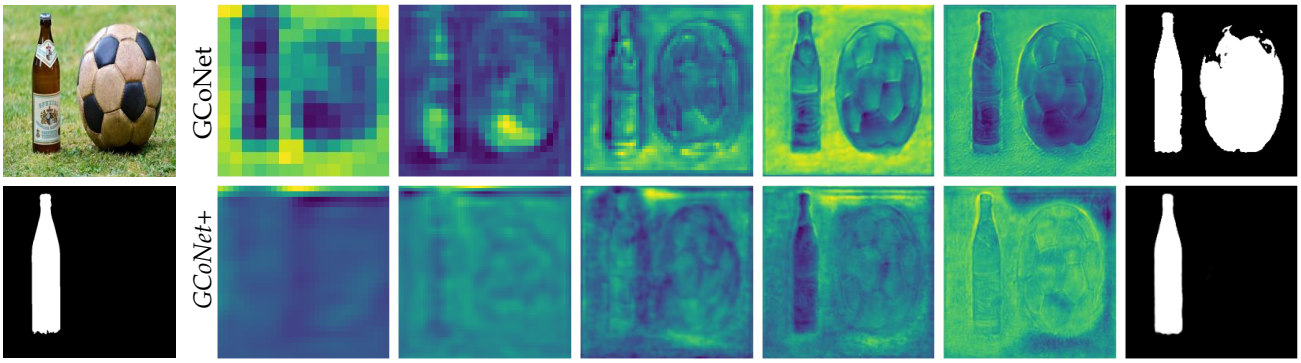

图 2. 特征图的可视化结果。（a）源图像和真值图。（b-f）为本文的GCoNet [1]和*GCoNet+*（Train-1，表. 3）的解码器在其不同阶段生成的特征图，由高层级往低层级采集。（b）列特征图的分辨率是最低的。如（b）所示，本文的*GCoNet+*提供了一个更全局的响应且并没有在非常早期的阶段就做出具体的预测，过早阶段的具体预测往往使得后续的结果图缺少更精确的信息。（g）协同显著图的预测结果。和GCoNet相比，*GCoNet+*能包含一个更全局的对显著物体和它周边事物的响应。

性，因此让网络在现有的丰富的SOD数据集上能够训练[1]。为了学得一个更好的嵌入空间，本文在每个图像上使用ACM以从一个全局语义层级上提升特征表达。

本文改进了原版的GCoNet，并对相应的贡献点提供了一个更精确的解释，即：一个用于CoSOD的更简洁的网络、三种额外的模块用以提升学习共识和差异的能力以及对于现有训练集缺点的讨论与对应的解决策略。总而言之，本文极大地拓展了GCoNet，主要不同如下所示：

- 新方法。本文提出了三种新模块以提升本文*GCoNet+*的性能和鲁棒性，即：置信增强模块（**CEM**）、基于群的对称三元组（**GST**）损失函数和循环式辅助分类模块（**RACM**）来解决本文GCoNet现有的缺点。1）置信增强模块（**CEM**）：为了使得输出结果图包含更少的不确定性，本文在置信增强模块中应用了可微二值化和一个混合显著性损失，给结果图带来了更高的质量并进一步提升了整体效果。2）基于群的对称三元组（**GST**）损失函数：本文的是最早之一将度量学习应用于CoSOD深度学习模型的工作，以一种度量学习的方式，使其在不同组习得的特征更具有区分度。3）循环式辅助分类模块（**RACM**）：为了更好地表达辅助分类特征，本文将原版的辅助分类模块拓展为循环式结构，使得它更精确地关注目标物体的像素区域。此外，本文将GCoNet [1]本身改进为一个更轻量且强大的网络以作为本文的基准模型。这三个模块和新的基准网络有机地结合在一起，以本文的实验所示图. 1，其在现有数据集和实际应用上都取得了很好的效果。
- 实验。尽管CoSOD领域的发展十分迅速，该领域有三个常用训练数据集，即：DUTS_class、COCO-9k和COCO-SEG，然而对于这个任务并没有一个选择训练集的标准。和训练集选取不一致的现有工作不同，本文用这三个数据集的所有排列组合进行了更全面的实验以便和先前工作进行公平的比较。凭借本文新提出的模块的组合，正如在以上新方法段落所指出的那样，使用相同的训练集和原版的GCoNet [1]相比，本文的新方法在所有的测试集上获得了$E_\xi^{\max}$ [23]和S_α [24]~3.2%的相对提升，达到了所有CoSOD测试集上的迄今为止的SoTA性能[2]。
- 新见解。基于所得的实验结果，本文揭示了现有CoSOD训练集的潜在问题，并分析了如何在未来对它们做进一步的提升。

1. 现有60k的SOD图像公开可得，大约比现有的CoSOD数据集大十倍多。这意味着CoSOD任务上训练数据不充足的问题可在所提出的框架中一定程度地减轻了。

2. CoSOD模型排行榜：https://paperswithcode.com/task/co-saliency-detection

2 相关工作。

2.1 显著目标检测

在传统显著目标检测（SOD）方法中，手工特征在目标检测上扮演着重要的角色 [25]–[28]。在深度学习发展的早期，特征通常从图像块、目标候选框 [29]–[31]或超像素 [32]–[35]中提取。尽管这些方法已经取得一些进展，但其提取目标区域和自身特征的过程非常耗时。随着全卷积网络 [36]在分割任务上的成功，最近的SOD研究主要关注那些能进行逐像素预测的模型。更多细节和总结可以在最近的综述 [9], [37], [38]中找到，其中 [9]提供了最全面的基准和对性能、鲁棒性以及泛化性的分析。在这些综述中，对大量具有挑战性的SOD数据集进行了分析，并提出了对SOD开放问题的建设性讨论和未来研究方向的建议。在 [39]中，SOD方法网络的架构被分为5种，即：单流、多流、侧融合、U型结构和多分支。在这些架构中，U型结构是被使用的最为广泛的，尤其是FPN [40]和U-Net [41]的基本结构。通过聚合这些U型网络的不同阶段的特征，多阶段监督被用于早期阶段，以使得输出特征更鲁棒和稳定 [1], [17], [42]。在 [43]–[46]中，注意力机制和相关模块被用于提升网络的性能。此外，外部信息被当作额外的引导，在其训练过程中引入，例如边缘 [42]和轮廓 [47]。

在二值分割任务中（如：显著目标检测 [1], [43], [47]，光学字符识别 [48]–[50]），真值是目标物体的二值掩码图。然而，由于使用像素级损失的缘故（即：MSE损失、BCE损失），预测所得结果图并非完全二值化。因此，在实际应用中，有着过多不确定性的结果图不适合让程序作出决定 [51]。为了解决这一问题，一些最新的方法也被提出用以提升二值结果图的质量。在 [52]中，专门的模块被设计用以增强物体的整体性。在 [47]中，混合损失函数被用于让模型关注除像素损失外的其他属性。

2.2 图像共分割

图像共分割是一项基础而活跃的计算机视觉任务，旨在分割出一组图像中的共同物体。该项任务已被广泛适用于许多相关领域，例如：协同显著物体检测 [1], [16], [18], [53]、小样本学习 [54], [55]、语义分割 [56], [57]等。许多现有的共分割方法使用了孪生网络来挖掘输入图像对的共同特征 [54], [58]。基于图像对内的比较，Chang 等人 [59]和Rother 等人

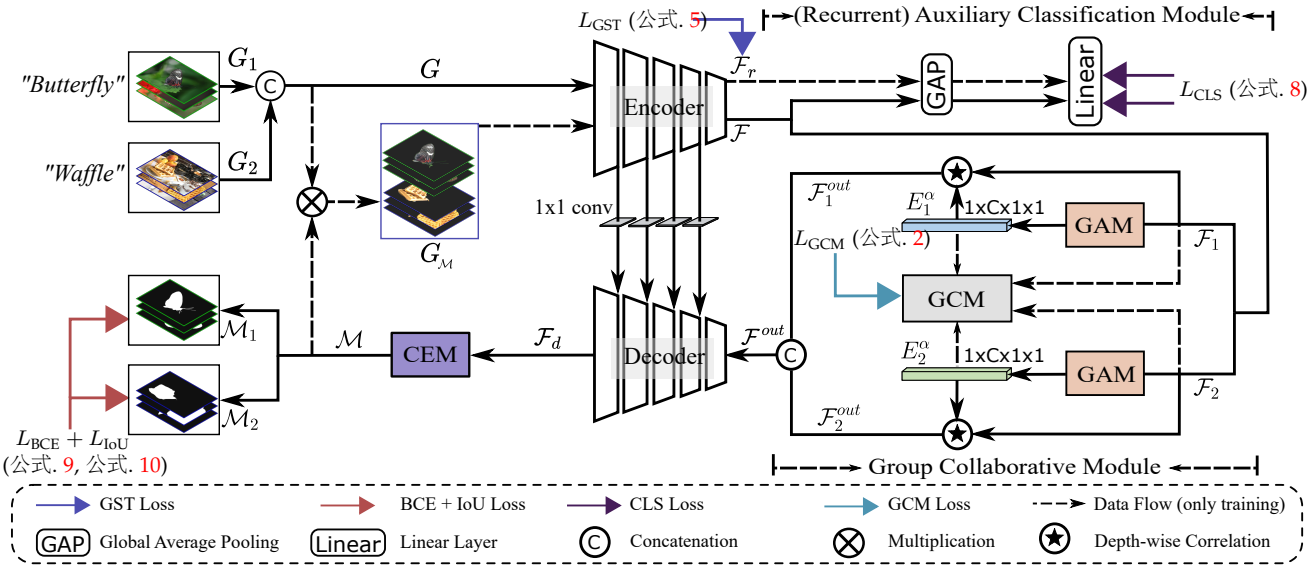

图 3. 本文提出的组协同学习网络加强版（***GCoNet+***）。输入图像从两组获取并送入编码器。然后使用本文的GCoM（GCoM）模块，其中，通过组亲和模块（GAM）进行组内协同学习，通过组对比模块（GCM）进行组间协同学习。源图像和经过掩码的RoI被送入编码器，以进行一次辅助分类使得不同类别的特征之间更具差异性。解码器的输出通过置信增强模块（CEM）让最终结果更二值化且易用。进一步地，本文通过一个三元损失函数来衡量两组图像RoI的间距，使组间特征的间距更大并降低组内特征的间距。

[60]分别使用显著性和颜色直方图来对视觉特征进行更精确的比较。随着深度学习方法的发展，共分割模型逐渐使用隐式的语义特征来发现共同物体。从模型的角度而言，Wei 等人 [18]和Fan 等人 [16]将共注意力嵌入他们的网络来生成组共识，Chen 等人 [61]运用了通道注意力来进行更好的物体共分割，同时LSTM [62]被Zhang 等人 [63]和Li 等人 [64]用来交互两图之间的信息从而增强组表达。从训练策略的方面而言，Wang 等人 [65]探索了以弱监督的方式，利用显著性引导对结果图进行的迭代式细化。Hsu 等人 [66]运用图像内的物体差异和图像间图与背景分离的特性，以一种无监督的方式实现图像共分割。

2.3 协同显著物体检测

SOD任务 [47], [67]–[69]旨在从单张图片分割出显著物体，而CoSOD则旨在找到一组语义相关的图像中的共同显著物体。之前的CoSOD方法主要通过挖掘组内线索来分割协同显著对象。如，早期的CoSOD方法过去常常基于手工特征去探索组内相关图片的相关性。通过从每张图片进行分割得到的计算性碎片（例如：超像素 [70]），这些方法建立起了相关性模型，并通过诸如聚类引导或转换对齐 [71]等排序机制，来挖掘共同区域。同时，度量学习 [10], [11]、直方图和对比度统计 [25]也被用来为后续的计算构建更好的语义属性。

在深度学习时代，许多端到端的深度CoSOD模型被陆续提出。文献 [11], [18]中的作者尝试通过学习单个组内的共识来发掘共同物体。随着上游深度学习方法的发展，现有方法 [1], [17], [19], [72], [73]利用强大的CNN模型（例如：ResNet [74]、VGGNet [75]和Inception [76]）或Transformer模型（例如：ViT [77]和PVT [78], [79]）来构建他们自己的模型，并达到了SoTA的水准。大多数现有方法利用全监督策略设计他们的模型，而基于弱监督的方法（例如：GWSCoSal [80]、FASS [81]、SP-MIL [10]、CODW [8]和GONet [12]）同样能达到可以接受的效果。

2.4 图像内和图像间一致性学习

随着深度学习的快速发展，深度模型通过如：图卷积网络（GCN） [82]–[84]、共注意力 [16]、共聚类 [85]、循环单元 [86]、相关性技术 [21]、自学习方法 [10]和质量度量 [87]等方法，在探究图像内与图像间的一致性上达到了极佳的性能。

共注意力是最为广泛使用的模块之一，用于实现图像内一致性学习。自它在 [88]中被提出后，共注意力模块常常被用于相似图像的分割。许多进一步的工作 [89]–[91]，如在像素对比、关系型数据和图网络等方面进行的深入探索，也被用于提高共注意力模块的性能和效果。这些工作展现出了强大的效果，并为相关领域的研究带来了可观的推动。

此外，图像内和图像间一致性还在其他研究领域崭露头角，例如目标检测 [92], [93]、语义分割 [94]和显著目标检测 [95]，尤其是用于建立物体间的相关性，从而以弱监督学习的方式获取不同类别的语义特征。

在过往的CoSOD方法中，组内一致性已被深入研究 [1], [8], [17], [18]。相反，组间信息则被相对忽视，然而它对于引导模型更好地习得差异性和通用特征起着十分重要的作用。在文献 [17]中，一种拼图训练策略被用于隐式地引入别组图像以辅助组训练。而在文献 [8]中，多组图像被送入模型以习得图像内的对比。然而这些模型仍主要关注组内信息，并未给学习组间信息显式地设计更高级的方法。本文方法和现有的模型在探索组间关系上颇具差异。本文尝试显式地、组层级地、更精准地学习语义层面的差异性特征。

3 方法

本文提出了一种用于CoSOD任务的*GCoNet+*，其架构概览在章节. 3.1中进行了介绍。之后，本文将依次介绍所提出的基础模块：组亲和模块（GAM）、组对比模块（GCM）、置信增强模块（CEM）、基于群的对称三元组（GST）损失函数和循环式辅助分类模块（RACM）。

3.1 概述

*GCoNet+*的基础框架基于GCoNet [1]，它也是CoSOD中最新的SoTA方法之一。不同于现有的CoSOD模型 [16], [17], [19], [21]，仅仅利用单个类别组内的共同信息，*GCoNet+*以一种孪生方式同时利用了不同组的组内和组外关系。

*GCoNet+*的流程图如图. 3所示。首先，本文的模型同时纳入两组源图像G_1, G_2作为输入。得到拼接起来©的两个图像

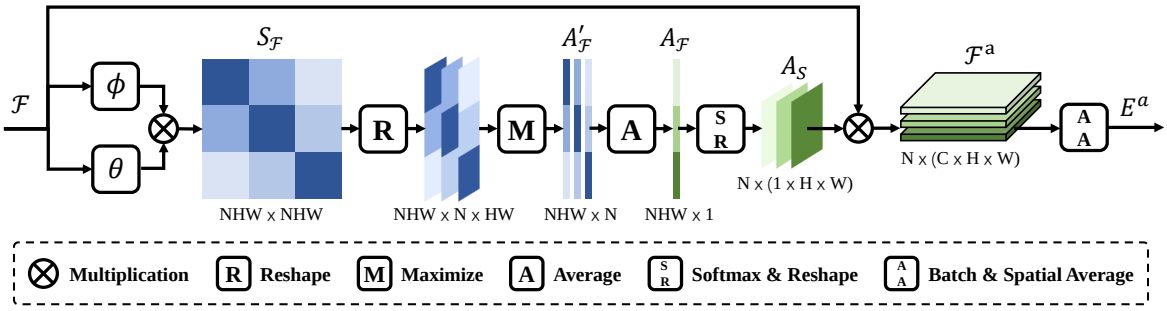

图 4. 组亲和模块。本文首先整合组内所有图像，利用相似性注意力来获取输入特征的注意力图。紧接着，将这些结果图与输入特征相乘，以生成这个组的共识。这个共识特征会被用于协调原始特征图，将其送入GCM模块进行组间协同学习。

组后，本文的编码器提取出了特征图$\mathcal{F}$，之后在将它送入循环式辅助分类模块（ACM）用于分类。在本文的组交互式模块中对$\mathcal{F}$进行进一步的处理。在组交互式模块中，$\mathcal{F}$依据它们的类别分为两部分，分别代表两个组的特征，即：$\mathcal{F}_1 = \{F_{1,n}\}_{n=1}^{N}, \mathcal{F}_2 = \{F_{2,n}\}_{n=1}^{N} \in \mathbb{R}^{N\times C\times H\times W}$，其中，C代表通道数，$H \times W$代表空间大小，而$N$表示组的大小。这两个特征被分开送入组亲和模块（GAM）以便将所有的单张图像特征结合来提取共识特征$E_1^a \in \mathbb{R}^{1\times C\times 1\times 1}$。同时，一个组对比模块（GCM）被应用以获得更具差异性的目标属性表达。组交互式模块的输出特征$\mathcal{F}_1^{out}, \mathcal{F}_2^{out}$被拼接并送入本文的解码器。解码器和编码器之间通过1x1卷积层连接。之后，置信增强模块（CEM）获得解码器的预测$\mathcal{F}_d$来细化并提供最终的协同显著性结果图$\mathcal{M}_1, \mathcal{M}_2$。最后，网络的输出和源图像$G$相乘以剔除不相关区域。本文的基于群的对称三元组（GST）损失函数作用于掩码后的图像$G_{\mathcal{M}}$并以度量学习的方式监督*GCoNet+*。此外，掩码的图像之后被再次送入编码器以获取掩码的编码特征$\mathcal{F}_r$。和$\mathcal{F}$不同，$\mathcal{F}_r$仅包含预测区域的特征，从而拥有一个更为精确的语义表达来输入循环式辅助分类模块（RACM）以获取分类损失。

3.2 组亲和模块（GAM）

大多数现实生活例子中，相同类别的物体通常具有更高的外观和特征相似性，这是许多计算机视觉任务的基础。例如，自监督视频跟踪方法 [96]–[99]经常基于相邻两帧的逐像素相关性生成目标物体的分割结果图。因此，基于该动机，本文通过计算组内全局相似性图来为CoSOD任务提供这种特性。

对于任何图像的$\{F_{1,n}, F_{1,m}\} \in \mathcal{F}_1$特征[3]，本文以内积的形式计算了它们的逐像素相关性：

$$S_{(n,m)} = \theta(F_n)^T \phi(F_m), \tag{1}$$

其中，θ, ϕ表示线性嵌入函数（$3\times 3\times 512$的卷积层）。相似性图$S_{(n,m)} \in \mathbb{R}^{HW\times HW}$有效地捕捉了所给图像对$(n,m)$中的协同显著对象的共同特征。接着，本文可以通过发现每个F_n的像素对于F_m的最大值，生成F_n的相似性图$A_{n\leftarrow m} \in \mathbb{R}^{HW\times 1}$，从而减轻了图中噪声相关性值的负面影响。

与之相似地，本文将一个图像对的局部相似性的用法拓展应用到整个图像组的全局相似性。具体来说，本文使用 公式. 1计算所有图像特征$\mathcal{F}$的相似性图$S_{\mathcal{F}} \in \mathbb{R}^{NHW\times NHW}$。接着，本文从$S_{\mathcal{F}}$中找到每张图像的最大值$A'_{\mathcal{F}} \in \mathbb{R}^{NHW\times N}$，并取$N$张图像的所有最大值的均值，生成全局相似性注意力结果图$A_{\mathcal{F}} \in \mathbb{R}^{NHW\times 1}$。以这种方式，相似性注意力结果图被全局优化在所有图像上，偶然共同出现的偏置也因此被减弱。接着，本文使用了一个softmax操作来将$A_{\mathcal{F}}$归一化并调整形状以生成注意力结果图$A_S \in \mathbb{R}^{N\times(1\times H\times W)}$。有了注意力结果图$A_S$，本文将它与原特征$\mathcal{F}$相乘以获得注意力特征结果图$\mathcal{F}^a \in \mathbb{R}^{N\times C\times H\times W}$。最终，如图. 4所示，整个组的注意力特征结果图$\mathcal{F}^a$经过batch维度的和空间维度的平均池化，被用于生成注意力共识E^a。

GAM关注捕捉相同组内、共同出现的、显著物体的相同之处，并借此提升共识表达的组内紧凑性。这样的组内紧凑性缓解了由同时出现的噪声造成的干扰并鼓励模型聚焦共同显著区域。这能让协同显著对象的共享属性更好地被捕捉，促成更好的共识表达。通过逐深度相关性 [100], [101]，我们将注意力共识E^a与原特征图$\mathcal{F}$相乘，以达到高效的信息交互。不同组别生成的特征图$\mathcal{F}^{out}$之后被拼接并送入解码器。在置信增强模块（CEM）后，得到所有图像的最终共同显著性结果图。

3.3 组对比模块（GCM）

当前，大多数CoSOD方法倾向于关注共识的组内紧凑性。尽管如此，组间分离性仍然对区分干扰物体起着非常重要的作用，尤其是当处理超过一个显著物体的复杂图像时。为此，本文提出了一个简单却有效的模块，即：GCM，通过学习来编码组间分离性。

有了GAM，本文能够得到两组图像的注意力共识$\{E_1^a, E_2^a\}$。接着，本文对相应的特征$\{\mathcal{F}_1, \mathcal{F}_2\}$与注意力共识进行叉乘（$\cdot$），得到组内协同特征：$\mathcal{F}_1^1 = \mathcal{F}_1 \cdot E_1^a$和$\mathcal{F}_2^2 = \mathcal{F}_2 \cdot E_2^a$。相对应地，组间相乘很好地处理了不同组的特征和共识，即：$\mathcal{F}_1^2 = \mathcal{F}_1 \cdot E_2^a$与$\mathcal{F}_2^1 = \mathcal{F}_2 \cdot E_1^a$，以对组间交互进行表达。组内特征$\mathcal{F}^+ = \{\mathcal{F}_1^1, \mathcal{F}_2^2\}$被用于预测共显著图，组间特征$\mathcal{F}^- = \{\mathcal{F}_1^2, \mathcal{F}_2^1\}$则赋给共识组分离性。

具体来说，本文将组间和组内特征$\{\mathcal{F}^+, \mathcal{F}^-\}$送入一个含单个上采样层的小卷积网络，获得显著图$\{\mathcal{M}^+, \mathcal{M}^-\}$[4]，对两者使用不同的监督信号。如图. 5所示，本文使用真值图对$\mathcal{F}^+$进行监督，而用全0图对$\mathcal{F}^-$进行监督。损失函数为：

$$L_{\text{GCM}} = \frac{1}{N}\sum_{n}^{N} L_{\text{FL}}(<\mathcal{M}_n^+, \mathcal{M}_n^->, <\mathcal{G}_n, \mathcal{G}_n^0>), \tag{2}$$

其中，L_{FL}表示focal loss [40]，$\mathcal{G}_n$表示真值，$\mathcal{G}_n^0$表示全0图，而$<\cdot>$表示拼接操作。

最后，GCM使得共识具备高度组间分离性，并使得网络在复杂环境中能够更好地区分干扰物。而且，这个模块并没有在做推理时引入额外的计算，可以完全被丢弃。

3. 章节. 3.2中对$\mathcal{F}_1$的所有分析都可被用于$\mathcal{F}_2$。本文省略了组上标使标注更简洁，即：本文使用F_n来表示$F_{1,n}$。

4. $\mathcal{M}^+ = \{\mathcal{M}_1^+, \mathcal{M}_2^+\}$且$\mathcal{M}^- = \{\mathcal{M}_1^-, \mathcal{M}_2^-\}$.

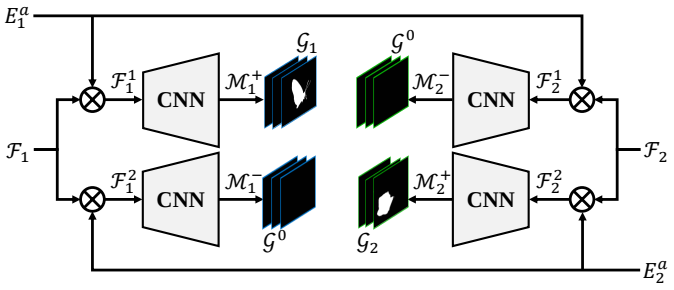

图 5. 组对比模块。两组的原始特征图和共识都被送入GCM。预测输出被可获得的真值标签监督。若真值标签不可得，则它被全0图监督。

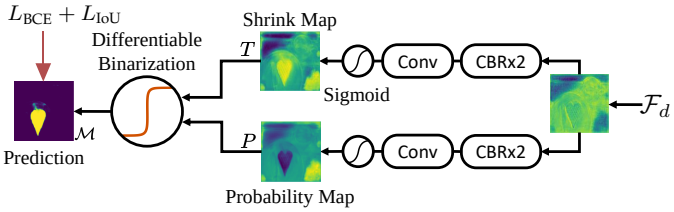

图 6. 置信增强模块。在解码器之后，本文调节了CEM来为预测的显著图带来更高的质量和二值化。CBR意味着紧跟着BN层与ReLU激活函数的一个卷积层。

3.4 置信增强模块（CEM）

SOD任务中，由于网络通常在结尾带着一个Sigmoid函数，预测所得显著图的像素值介于0到1之间。尽管真值非0即1，但所预测的显著图的像素值可能趋近0.5，这也意味着预测中携带更多的不确定性与噪声。在非常难的样本中，拥有更多的不确定性与噪声的结果往往会在某些经典指标上获得更高的分数 [23]，例如：Fbw [102]、IoU [103]、MAE等，然而在实践中表现糟糕，这背离了最终目标。

为了处理预测结果中的不确定值，本文从损失函数和网络结构两个方面进行了研究。从损失函数的角度来看，本文建立了比较实验来验证不同的损失函数会给网络引入不同的优化方向。更具体来说，IoU损失引导输出成为几乎全0或1，但在现有指标上有更低的准确率，即：S-measure [24]、E-measure [23]。相反，BCE损失引导网络去输出更不确定的像素值，可能在上述指标上达到更高的性能。正如意料中的结果图（图. 7）所示，尽管IoU损失给最终结果图带来了更高的置信，但优化过于粗糙，表现在显著图的整体性上。因此，BCE损失仍然是训练中的必要项。为了提升显著图在实际应用中的需要的图像质量，本文以一种混合的像素损失监督，试图平衡BCE和IoU损失。

从模型架构的角度来说，本文将置信增强模块（CEM）运用在末尾处图. 3。在之前的SOD方法中，Sigmoid函数经常被用于压缩0到1的输出值。然而，正如 [48]中所说，Sigmoid激活函数不够陡峭，产出的值不够二值化。为了解决这个问题，如图. 6所示，解码器的输出特征$\mathcal{F}_d$被送入CEM。首先，特征$\mathcal{F}_d$经过两个含一对3x3卷积的平行分支，在卷积之后紧跟着BN层、卷积激活函数、一个1x1卷基层和一个Sigmoid激活函数。在这之后，概率图P和阈值图T被生成并被送入可微二值化函数以获取最终预测结果。根据 [48]，最终的共显著图$\mathcal{M}$可被表示为：

$$\mathcal{M}_{i,j} = \frac{1}{1+e^{-k(P_{i,j}-T_{i,j})}}, \tag{3}$$

其中，k是控制阶跃函数陡峭程度的因子。在本文的实现中，k的默认值为300。当训练中损失得到NaN时，它会被替换为50用于当前的传播。

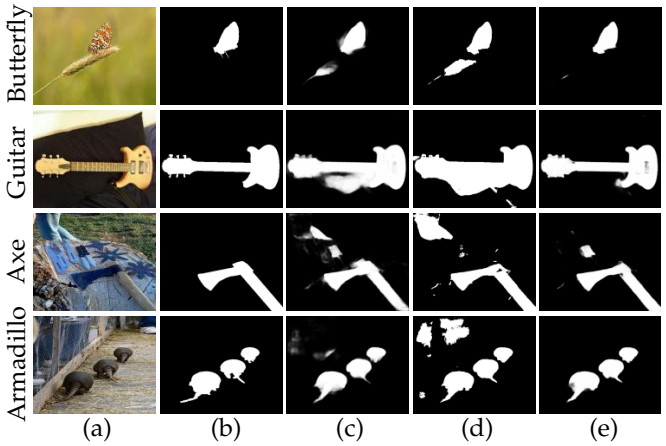

图 7. 本文的***GCoNet+***在使用不同损失函数时产生的预测结果。（a）源图像。（b）真值。（c）*GCoNet+*仅经BCE损失训练的结果。（d）*GCoNet+*仅经IoU损失训练的结果。（e）*GCoNet+*经BCE和IoU的平衡损失训练的结果。这里所有的结果都由在DUTS_class数据集上训练所得模型生成。

3.5 基于群的对称三元组（GST）损失

近年来，已经有一些从度量学习角度解决CoSOD任务的方法被提出 [10], [11]。然而，大多现有的基于度量学习的CoSOD方法使用超像素 [104]来提取碎片作为度量的单元。这些方法的绝大多数并不常是端到端的并且效率较低。此外，现有的工作一般引入类别标签来帮助模型学习含高级语义的更具代表性的特征。具体来说，在 [17]中，Zhang 等人将DUTS数据集 [105]通过主显著物体的绝对类别标签，分为不同组。然而，绝对类别标签可能不能在现实场景中获取。相反，通常仅有两个组的相对标签（它们是否属于相同组）。在2015年，Schroff 等人提出triplet loss [106]来辅助人脸识别，通过拉近正样本和推开负样本来习得更有差异性的不同身份的特征。Triplet loss的在人脸识别 [106]、视觉跟踪 [107]、行人重识别 [108]等任务上取得了成功。本文改变了原始的triplet loss成为GST loss来学习不同组图像的更具差异性的特征。这能够提升不同类别物体的共识特征的独特性与差异性。

要注意的是，本文的GST loss仅仅在训练过程中激活。具体而言，它作用于$\mathcal{F}_r$这一由编码器从$G_{\mathcal{M}}$（预测显著性结果图$\mathcal{M}$和源图像G的相乘结果）中提取出的输出特征（见图. 3）。仅有目标物体的像素被用于度量。以图. 8中的G_1为例，骨干网络Φ_θ从原始图像经由$\mathcal{M}_1$掩码提取语义表达$\mathcal{F}_r^1$。相同组的特征互相被看作正样本，而其余组的特征作为负样本。如图. 8所示，本文的GST loss以一种对称的结构被计算得到。最终，triplet loss被应用在$(\mathcal{F}_r^{1A}, \mathcal{F}_r^{1B}, \mathcal{F}_r^{2A})$和$(\mathcal{F}_r^{1B}, \mathcal{F}_r^{2B}, \mathcal{F}_r^{2A})$，其中特征间的距离由欧氏距离计算。具体而言，$L_{\text{Tri}}(\mathcal{F}_r^{1A}, \mathcal{F}_r^{1B}, \mathcal{F}_r^{2A})$可被表示如下：

$$||\mathcal{F}_r^{1A} - \mathcal{F}_r^{1B}||_2 - ||\mathcal{F}_r^{1B} - \mathcal{F}_r^{2A}||_2 + \alpha, \tag{4}$$

其中，α表示margin，一个在正负样本对之间强制设置的超参 [106]。$||\cdot||_2$表示输入的二范数。因为GST loss的对称性，$L_{\text{Tri}}(\mathcal{F}_r^{1B}, \mathcal{F}_r^{2B}, \mathcal{F}_r^{2A})$也使用相同的方式进行度量。

最终的GST loss是一对L_{Tri}的结合，G_1和G_2从用预测图掩盖的图像中交替充当正样本：

$$L_{\text{GST}} = L_{\text{Tri}}(\mathcal{F}_r^{1A}, \mathcal{F}_r^{1B}, \mathcal{F}_r^{2A}) + L_{\text{Tri}}(\mathcal{F}_r^{1B}, \mathcal{F}_r^{2B}, \mathcal{F}_r^{2A}), \tag{5}$$

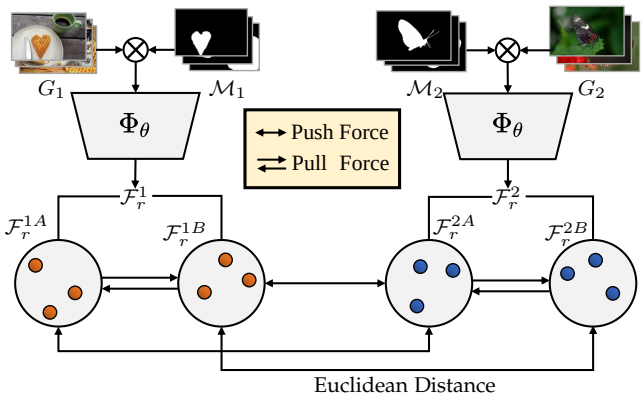

图 8. 基于群的对称三元组 **Loss**。在GST loss中，每个组被均分为两个副组。来自同一个组的副组们相互之间彼此拉近并和其他组的特征相互推远。Φ_θ表示骨干网络（见图. 3）。

3.6 循环式辅助分类模块（**RACM**）

现有的工作往往使用相同组内的图像训练模型以提取共同信息。具体来说，某一个batch内的图像仅仅有属于同一个类别的物体的真值图，因而仅组内共同特征能够被学习到。然而，由于在习得的特征上没有约束，不同类别的共同特征可能彼此之间过于靠近，进而难以区分。

在文献 [1]中，辅助分类模块（ACM）利用高阶语义表达来获得更具差异性的特征用于共识学习。具体而言，一个包含了全局平均池化的类别预测器和一个全连接层被用在骨干网络之后。相同类别物体的特征通过类别等级的监督被聚类到一起。尽管ACM在GCoNet [1]中起到很好的作用，它仍有一些不足：骨干网络中的特征不够稳定且可能是正确物体之外的一些东西。结果而言，ACM可能给出一个错误的优化方向。同时，它隐式地运行并很难被监测。

本文提出使用Recurrent ACM（RACM）来克服上述问题。RACM的流程几乎和原始ACM的相同。对应地，RACM将模型的输出作为掩码来获取目标物体的像素而不是ACM中所使用的整个图像的像素。被掩码的图像会再被送入编码器和类别预测器。在排除其他干扰区域之后，本文的RACM仅关注于感兴趣区域。当本文的*GCoNet+*和真值图相距甚远时，RACM能够给予一个增强的惩罚用来帮助加速训练的收敛。结合源图像和真值图来构建损失，RACM能够让模型习得更具差异性的特征，从而分别获取更好的组间分离性与组内紧凑性。分类损失函数的表示如下：

$$\hat{Y}_{\text{ACM}} = \varphi(\Phi_\theta(G)), \tag{6}$$

$$\hat{Y}_{\text{RACM}} = \varphi(\Phi_\theta(G \otimes \mathcal{M})), \tag{7}$$

$$L_{\text{CLS}} = L_{\text{CE}}(\hat{Y}_{\text{RACM}}, Y_{\text{CLS}}) + L_{\text{CE}}(\hat{Y}_{\text{ACM}}, Y_{\text{CLS}}), \tag{8}$$

其中，φ和Φ_θ分别表示类别预测器（GAP和一个线性层）和编码器。L_{CE}是交叉熵损失，Y_{CLS}是真值类别标签，而$\hat{Y}_{\text{ACM}}$和$\hat{Y}_{\text{RACM}}$分别是ACM和RACM预测的类别标签。

3.7 目标函数

目标函数是一个显著图损失（BCE损失和IoU损失的结合）、GCM损失、本文的GST损失和分类损失的加权和。BCE损失和IoU损失如下所示：

$$L_{\text{BCE}} = -\sum[Y log(\hat{Y}), (1-Y) log(1-\hat{Y})], \tag{9}$$

$$L_{\text{IoU}} = 1 - \frac{1}{N}\sum\frac{Y \cap \hat{Y}}{Y \cup \hat{Y}}, \tag{10}$$

其中，Y是真值而$\hat{Y}$是预测结果。有了GCM损失（公式. 2）、GST损失（公式. 5）和分类损失（公式. 8），本文最终的目标函数为：

$$L = \lambda_1 L_{\text{BCE}} + \lambda_2 L_{\text{IoU}} + \lambda_3 L_{\text{GCM}} + \lambda_4 L_{\text{GST}} + \lambda_5 L_{\text{CLS}}, \tag{11}$$

其中，$\lambda_1, \lambda_2, \lambda_3, \lambda_4$和$\lambda_5$分别被设置为30、0.5、250、3和3来保持所有的损失在训练初期处于同一个数量级。

4 实验

这一部分分别提供了本文基础实验与拓展实验的方针与细节，即：数据集、设置、评测协议和训练与测试的分析。

4.1 数据集

训练集。本文遵循了GICD [17]使用DUTS_class作为本文的训练集以设计实验。在Zhang 等人移除噪声样本之后，整个DUTS_class数据集被分为291组，包含共计8250张图像。DUTS_class数据集是在本文对照试验中唯一的训练集。现如今，仍缺少一个完全被公认的训练集。为了做出和现有最新方法 [18], [19], [21], [109], [110]的公平对比，本文使用广泛采用的COCO-9k [18]（COCO [111]的一个子集，拥有65组的9213张图像）和COCO-SEG [109]（同样也是COCO的一个子集，包含200k张图像）用于训练本文的*GCoNet+*作为补充实验。

测试集。为了得到一个对本文*GCoNet+*的全面评估，本文在三个广泛使用的CoSOD数据集进行测试，即：CoCA [17]、CoSOD3k [16]和CoSal2015 [8]。在这三个数据集中，CoCA是最具挑战的。它在背景、遮挡、光照和周边物体等方面更具多样性和复杂性。遵循最新的基准 [16]，由于iCoseg [112]和MSRC [113]中的大多数图像中仅有一个显著物体，本文没有评测这两个数据集。在有更多显著物体的图像上对CoSOD方法的评估更具有说服力，这也更契合现实应用。

4.2 评测协议

遵循GCoNet [1]，本文运用了S-measure [24]、maximum F-measure [114]、maximum E-measure [23]和mean absolute error（MAE）在本文的实验中评估性能。评估工具箱可参考 https://github.com/zzhanghub/eval-co-sod。

S-measure [24]是一种对显著图和它的真值图进行结构化相似度度量的方法。使用S_α进行评测能够不通过二值化就快速地得到结果。S-measure的计算如下：

$$S_\alpha = \alpha \times S_o + (1-\alpha) \times S_r, \tag{12}$$

其中，S_o和S_r表示物体感知和区域感知的结构化相似度，α遵循 [24]中的建议，默认被设置为0.5。

F-measure [114]被设计用于以准确率和召回率的加权调和平均值进行评估。显著图输出被以不同的阈值进行二值化来获取一组二值显著图预测。预测所得的显著图和真值图被计算以得到准确率和召回率。整个数据集上，在所有阈值中得到的最好的F-measure分数被定义为F_β^{max}。F-measure可被如下计算：

$$F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 (Precision + Recall)}, \tag{13}$$

其中，β^2遵循 [37]被设为0.3，来强调准确性多于召回率。

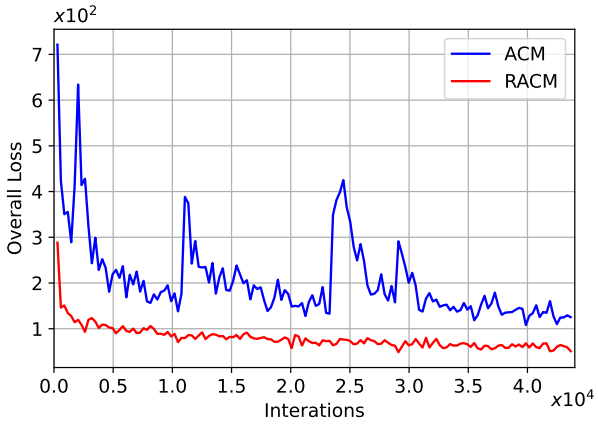

图 9. 学习率比较。本文记录了基线在整个训练中的整体损失（见于章节. 4.4），分别是有额外的RACM和只有原始ACM的。其中DUTS_class被用作训练集。

E-measure [23]作为一个感知型指标被设计出来，用以从局部和全局的视角评估预测图和真值图之间的相似度。E-measure被定义为：

$$E_\xi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_\xi(x, y), \tag{14}$$

其中，ϕ_ξ表示增强的对齐矩阵。和F-measure相似，本文也采用了max E-measure（E_ξ^{max}）作为本文的评测指标。

MAE ϵ是一种简单的像素级评测指标，无需二值化，衡量预测结果图和真值图之间的绝对之差。它被定义为：

$$\epsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\hat{Y}(x, y) - \mathrm{GT}(x, y)|. \tag{15}$$

4.3 实现细节

基于GCoNet [1]，本文使用带BN层 [115]的VGG-16作为骨干网络。本文从两个不同组中随机挑选N个样本作为训练的一个batch。

$$N = \min(\#groupA, \#groupB, 32), \tag{16}$$

其中，N表示训练的batch size，而#表示在对应组中的图像数量。由于一些组内图像的数量很小，本文选取32和随机选取的两组中图像数量的更小数。要注意的是，N对于训练和测试可以是不同的。在测试中，本文遵循之前的工作 [1], [17], [19], [110], [116], [117]将所给组的图像数作为batch size N。

为了阐明本文提出的网络，本文提供了新提出模块中的超参。陡阶函数在置信增强模块（CEM）的反向传播中会产生一些NaN值。因此，本文将可微二值化中的k设为一个基础值300和一个保守值50。当NaN在某一个步骤中产生，50会被用于替换它，这使得本文的实验中损失不会变为NaN。在基于群的对称三元组（GST）损失中，边界值被设为1.0。

将图像调整为256x256以进行训练和测试。将输出图调整为原始大小以进行评估。三种数据增广策略被应用到本文的训练过程中，即：水平翻转、颜色增强和旋转。本文的*GCoNet+*使用Adam优化器训练320轮。初始学习率被设为3e-4，$\beta_1 = 0.9$和$\beta_2 = 0.99$。整个训练过程花费20小时。所有的实验基于PyTorch [118]和一块Tesla V100 GPU实现。

4.4 对照实验

我们研究了应用在本文*GCoNet+*中的每个拓展模块的效果（即：RACM、CEM和GST），并探讨了为什么它们能够在本文的框架中，既帮助学习好的共识特征，又帮助学习具有差异性的特征。每个模块定性结果可见于图. 11。更多的对照实验和实验设定可以参考本文的会议版本 [1]。

基线。本文遵循GCoNet以孪生方式设计本文的*GCoNet+*。请注意，GCoNet沿用了GICD [17]的架构，没有对GICD中每个组件的有效性进行广泛的实验，包括多头监督、损失函数、特征归一化等。尽管这些组件给网络本身带来了额外的参数和复杂性，实验结果仍然不能证明其有效性。本文没有将这些组件视为理所当然，而是对每个组件进行了广泛的实验。首先，本文尝试像原始FPN [40]那样，仅用一个1x1卷积层替换横向连接中的多个卷积块。其次，本文尝试去除解码器上显着图的多阶段监督。第三，本文尝试在除1x1卷积层之外的每个卷积层后面添加批量归一化。最后，正如本文的实验所示，BCE损失为本文的实验带来了更高的准确性，而IoU损失带来了更多二值化的最终显着图和更快的收敛效果。为了更好地结合这两种损失，本文将初始BCE和IoU损失控制在具有不同权重的相同数量水平上，并将它们相加。

这些改动可以被概括为三个部分，即：网络结构简化、归一化层和混合损失。遵循奥卡姆剃刀原则，本文尝试去除现有工作中的所有没有足够有效实验证据的不确定模块。和基线相比，这些改动在简易程度和准确性两方面大大提升了本文的*GCoNet+*（表. 1中的ID:1。）如表. 1所示，将它们全部结合，在CoSOD3k和CoSal2015上分别带来2.6%和2.8%的E-measure相对提升。这还在最具挑战性的CoCA测试集上获得了2.5%的E-measure相对提升。

RACM的效果。ACM引导模型去学习更具有差异性的特征以分辨不同类别的物体。与原始ACM相比，它表现得更精确并加速了本文*GCoNet+*的收敛（见图. 9）。正如表. 2所见，RACM轻微提升了基线在CoCA和CoSOD3k中的大部分指标上的效果。图. 10中的激活图展示了本文的*GCoNet+*在许多样例中给出一个更高的准确率并引导模型更精准地关注目标。图. 2展示了*GCoNet+*中解码器的每个阶段的特征图。正如结果所示，*GCoNet+*相比GCoNet，能够更准确地区分不同类别的物体。

CEM的效果。在所有的CoSOD方法中，IoU和BCE损失往往被用作训练损失。然而，在大多数的方法中，仅有一个损失被用于训练时的监督。BCE引导像素级别的监督，而IoU从区域的角度引导训练。尽管现有的许多方法取得了优异的性能 [1], [19], [84], [116], [117]，分别单独使用BCE和IoU还是会导致一些问题。具体来说，有了IoU损失从区域级别监督模型，预测的显著图常常粗糙而不能很好地处理小细节。BCE能够引导模型关注细节。同时，由它监督所生成的显著图往往包含许多不确定性，使得预测结果直接投入实际应用更具挑战性。

在此，本文使用CEM来同时预测更准确和更二值化的结果图，使其更贴近现实世界的应用。正如图. 11和表. 2所示，CEM能使预测结果图在准确率和可视化结果上更好。

GST损失的效果。共识特征在CoSOD任务中，对检测共同物体，扮演着重要的角色。然而，一些类别的共识特征可能过于靠近彼此。为此，本文需要使共识特征之间保持差异性，使它们远离其他特征。本文引入了GST损失来使习得的不同类别的特征对于彼此更具差异性。正如表. 2和图. 11中的实验所示，GST损失成功地从全局和RoI级别区分开了不同特征，并进一步提升了模型的竞争力。

4.5 竞争方法

由于不都是所有的CoSOD方法都开源了，本文仅仅比

表 1
对***GCoNet+***框架整体改动的定量对照实验。本文对*GCoNet+*的整体改动的效果进行了对照实验，其中包括网络简化（Net-Sim）、批量归一化（BN）和混合损失（HL）。

ID	模块			CoCA [17]				CoSOD3k [16]				CoSal2015 [8]			
	Net-Sim	BN	HL	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$
1				0.760	0.673	0.544	0.105	0.860	0.802	0.777	0.071	0.888	0.845	0.847	0.068
2	✓			0.752	0.676	0.538	**0.100**	0.872	0.815	0.796	0.063	0.895	0.853	0.858	0.063
3	✓	✓		0.747	0.683	0.556	0.110	**0.884**	0.824	0.806	**0.062**	0.912	0.868	0.874	**0.051**
4	✓		✓	0.774	**0.691**	**0.562**	0.106	0.879	**0.831**	0.806	0.065	0.901	0.867	0.865	0.062
5	✓	✓	✓	**0.779**	0.681	0.558	0.119	0.882	0.828	**0.807**	0.068	**0.913**	**0.875**	**0.877**	0.055

表 2
在本文***GCoNet+***中所提出的模块的定量实验。本文对*GCoNet+*在提出的模块上进行对照实验，包括RACM（循环式辅助分类模块）、CEM（置信增强模块）、GST（基于群的对称三元组）和它们的组合。

ID	模块			CoCA [17]				CoSOD3k [16]				CoSal2015 [8]			
	RACM	CEM	GST	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$
1				0.779	0.681	0.558	0.119	0.882	0.828	0.807	0.068	0.913	0.875	**0.877**	0.055
2	✓			0.780	0.684	0.570	0.120	**0.884**	0.829	0.809	**0.067**	0.912	0.873	0.875	0.056
3	✓	✓		0.779	0.686	0.565	0.117	0.881	0.829	0.805	0.068	0.913	0.872	0.873	0.057
4	✓		✓	0.780	0.683	0.559	0.118	0.882	**0.831**	**0.810**	0.068	0.914	**0.876**	0.876	0.055
5	✓	✓	✓	**0.786**	**0.691**	**0.574**	**0.113**	0.881	0.828	0.807	0.068	**0.917**	0.875	0.876	**0.054**

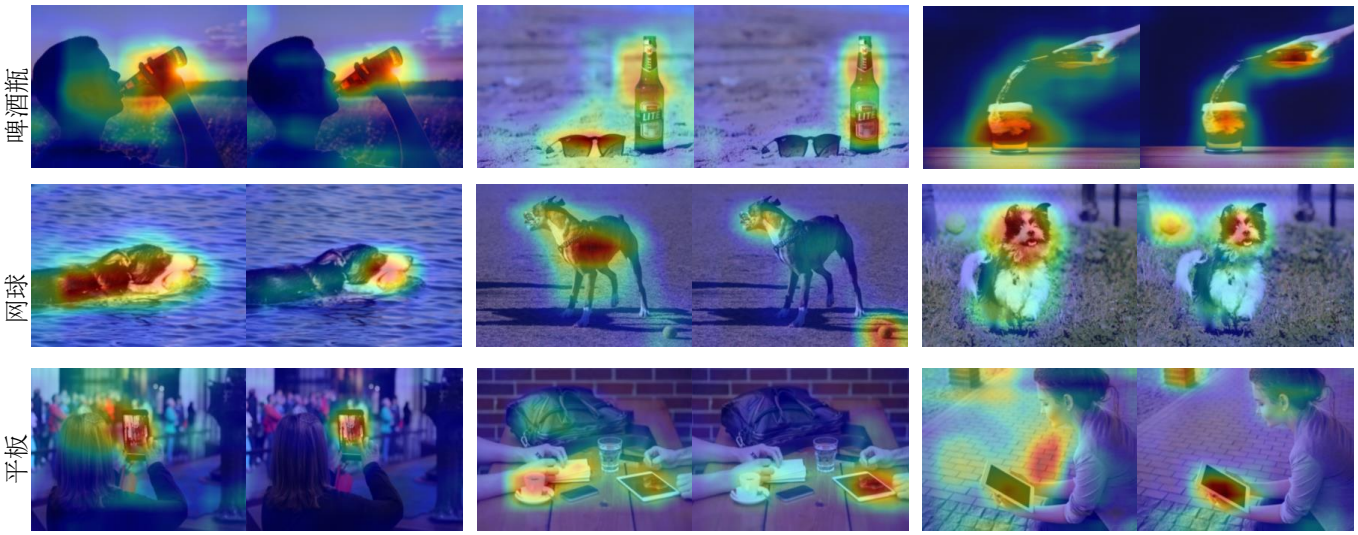

图 10. 在分类分支上得到的类激活图 **[119]**。对于每一对相邻格子，左半边为使用ACM从原版GCoNet [1]得到的激活图；右半边结果为本文*GCoNet+*通过额外的RACM生成的激活图。如列（a）所示，本文的*GCoNet+*在关注到周围充满干扰物的目标物体时更精准。在列（b）中，本文的*GCoNet+*在关注正确类别物体上展现出更好的性能，尽管它受周围物体的干扰。在最后的列（c）中，一些复杂样本让两个模型都误判了。尽管错误的注意力被置于错误类别的物体上，本文的*GCoNet+*仍然能够将大部分注意力放在正确的物体上，并将它们看作图像的整体部分。此处提供的分类激活图的*GCoNet+*仅在DUTS_class上训练。

较本文的GCoNet和*GCoNet+*、一个具有代表性的传统算法CBCS [14]以及11 个基于深度学习的CoSOD模型，包括了全部最新的模型，即：GWD [18]、RCAN [86]、CSMG [120]、GCAGC [84]、GICD [17]、ICNet [21]、CoADNet [19]、CoEGNet [16]、DeepACG [117]、CADC [110]、UFO [53]和DCFM [116]。因为最新的CoSOD相比于单SOD模型在性能上具有明显优势，所以本文并不列出单SOD的方法。关于现有方法的更完整的列表可见文献 [16]。

定量结果。 表. 3展示了本文*GCoNet+*和之前SoTA方法在各种数据集上的定量结果。可以发现本文的*GCoNet+*在所有的指标上均优于其他方法，尤其是在CoCA和CoSOD3k数据集上。和其他两个数据集相比，CoCA是难度最高的数据集，这是由于需要发现共同物体、单张图像中物体数量更多以及更多样的背景等因素造成的。本文的*GCoNet+*展示出分割上更强的能力，这得益于从显著性检测和共识学习两方面得到改进的特征。CoSOD3k有相似的属性，且本文的*GCoNet+*在这个数据集上保持着它最强的性能。CoSal2015是最简单的数据集，由于它大多数的图像仅仅包含一个显著物体，这使得它很容易被单SOD方法处理。尽管它的难度较低且缺乏共显著性，本文的*GCoNet+*仍然以一个相对更小的提升，击败了其他方法。此外，如表. 4所示，本文的*GCoNet+*拥有更少的参数，这使得其推理速度相较于大多数现有方法更快。

定性结果。 图. 12展示了不同方法生成的显著性结果图用于定性比较。啤酒瓶图像组包含很多类别的多个显著物体，在此*GCoNet+*能够精确检测出共显著物体，而其他方法不行。在拐杖图像组中，目标是细长的枝杆，但本文的*GCoNet+*仍然能够以高准确率将其分割出来，而其他方法甚至不能做正确的分割。本文使用网球图像组来比较模型检测小目标的能力，在分类和准确度两个方面，本文

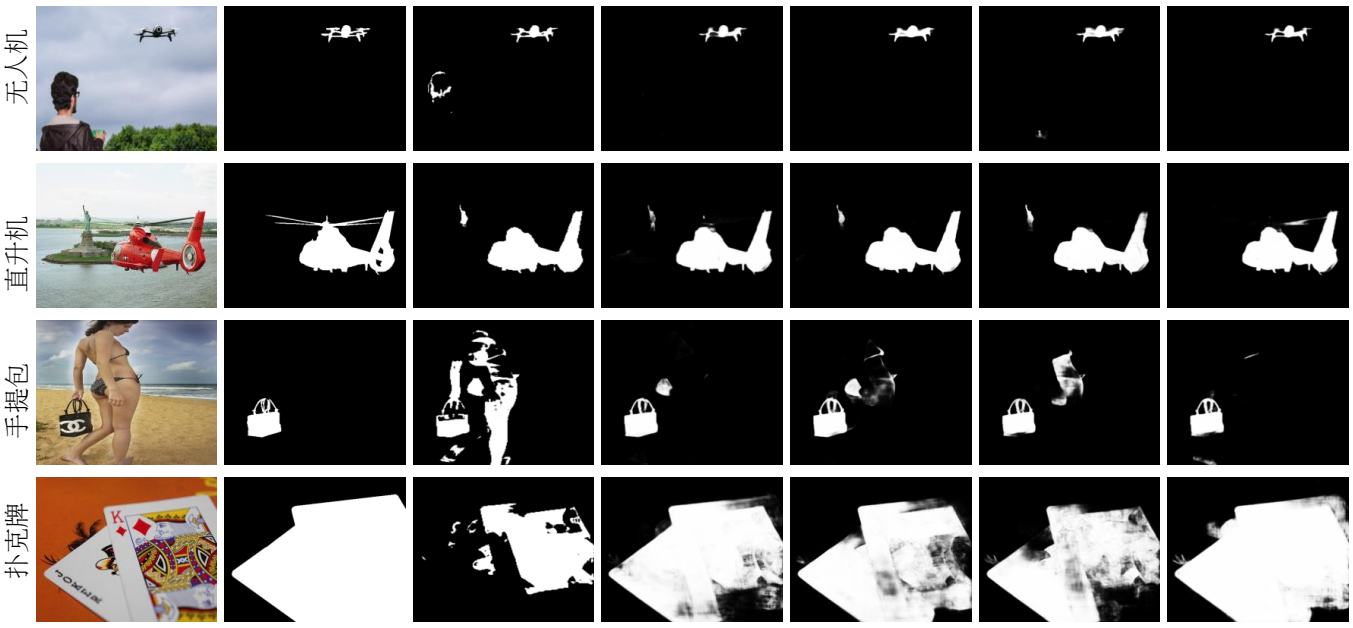

图 11. 对于***GCoNet+***的不同模块和它们的组合的定性对照实验。（a）源图像；（b）真值；（c）GCoNet [1]；（d）本文的新基线；（e）基线+RACM；（f）基线+RACM+CEM；（g）基线+RACM+CEM+GST (*GCoNet+*的最终版)。为了和GCoNet保持一致，此处的预测结果图均由在DUTS_class上训练的*GCoNet+*生成。

表 3
本文***GCoNet+***和其他方法的定量比较。"↑"和（"↓"）分别代表更高（更低）是更好的。方法的开源代码或论文源被放在链接中。由于有多个数据集用于CoSOD任务的训练，本文将对应方法所用到的所有的训练集都列出来，即：Train-1，2和3分别代表DUTS_class [17]、COCO-9k [18]和COCO-SEG [109]。

方法	出版刊物& 年份	训练集	CoCA [17]				CoSOD3k [16]				CoSal2015 [8]			
			$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$
CBCS [14]	TIP 2013	-	0.641	0.523	0.313	0.180	0.637	0.528	0.466	0.228	0.656	0.544	0.532	0.233
GWD [18]	IJCAI 2017	Train-2	0.701	0.602	0.408	0.166	0.777	0.716	0.649	0.147	0.802	0.744	0.706	0.148
RCAN [86]	IJCAI 2019	Train-2	0.702	0.616	0.422	0.160	0.808	0.744	0.688	0.130	0.842	0.779	0.764	0.126
GCAGC [84]	CVPR 2020	Train-3	0.754	0.669	0.523	0.111	0.816	0.785	0.740	0.100	0.866	0.817	0.813	0.085
GICD [17]	ECCV 2020	Train-1	0.715	0.658	0.513	0.126	0.848	0.797	0.770	0.079	0.887	0.844	0.844	0.071
ICNet [21]	NeurIPS 2020	Train-2	0.698	0.651	0.506	0.148	0.832	0.780	0.743	0.097	0.900	0.856	0.855	**0.058**
CoADNet [19]	NeurIPS 2020	Train-1, 3	-	-	-	-	0.878	0.824	0.791	0.076	0.914	0.861	0.858	0.064
CoEGNet [16]	TPAMI 2021	Train-1	0.717	0.612	0.493	0.106	0.837	0.778	0.758	0.084	0.884	0.838	0.836	0.078
DeepACG [117]	CVPR 2021	Train-3	0.771	0.688	0.552	0.102	0.838	0.792	0.756	0.089	0.892	0.854	0.842	0.064
CADC [110]	ICCV 2021	Train-1, 2	0.744	0.681	0.548	0.132	0.840	0.801	0.859	0.096	0.906	0.866	0.862	0.064
DCFM [116]	CVPR 2022	Train-2	0.783	0.710	0.598	0.085	0.874	0.810	0.805	0.067	0.892	0.838	0.856	0.067
UFO [53]	ArXiv 2022	Train-3	0.782	0.697	0.571	0.095	0.874	0.819	0.797	0.073	0.906	0.860	0.865	0.064
GCoNet (Ours)	CVPR 2021	Train-1	0.760	0.673	0.544	0.105	0.860	0.802	0.777	0.071	0.887	0.845	0.847	0.068
***GCoNet+* (Ours)**	TPAMI 2023	Train-1	0.786	0.691	0.574	0.113	0.881	0.828	0.807	0.068	0.917	0.875	0.876	0.054
***GCoNet+* (Ours)**	TPAMI 2023	Train-2	0.798	0.717	0.605	0.098	0.877	0.819	0.796	0.075	0.902	0.853	0.857	0.073
***GCoNet+* (Ours)**	Submission	Train-3	0.787	0.712	0.602	0.100	0.875	0.820	0.793	0.075	0.899	0.853	0.852	0.071
***GCoNet+* (Ours)**	TPAMI 2023	Train-1, 2	0.808	0.734	0.626	0.088	0.894	0.839	0.822	0.065	0.919	0.876	0.880	0.058
***GCoNet+* (Ours)**	TPAMI 2023	Train-1, 3	**0.814**	**0.738**	**0.637**	**0.081**	**0.901**	**0.843**	**0.834**	**0.062**	**0.924**	**0.881**	**0.891**	**0.056**

表 4
不同方法的运行时间比较。所有方法在推理时的batch size都被设为2。

方法	推理时间(ms)	参数量(MB)
CoEGNet [16]	2300	412.3
GICD [17]	18.2	1060.7
ICNet [21]	12.5	**70.3**
CoADNet [19]	70.0	113.2
GCAGC [84]	103	280.7
CADC [110]	54	1498.7
DCFM [116]	3.73	542.9
GCoNet [1] (**Ours**)	**3.3**	541.7
***GCoNet+* (Ours)**	4.0	**70.3**

的*GCoNet+*都优于其他方法。在番茄组中有许多番茄作为共显著目标，他们应该被同时检测出来。本文方法能够以一个高质量的显著图找到所有的番茄，而其他方法会漏掉其中的一些显著番茄或分割出别的非常接近的物体。在上述这些例子中，本文的*GCoNet+*更好地发现组内共识信息并区分组间信息。

4.6 对现有**CoSOD**训练集的讨论

尽管CoSOD领域已有很多好工作，但仍然缺少一个标准训练集。DUTS_class、COCO-9k和COCO-SEG是常用的三个训练集但都有局限，例如：不正确的真值图和过少的目标。

DUTS_class。由于DUTS_class仅关注检测显著目标，会有不同类别的显著物体出现在同一张图像中。正如图. 13所

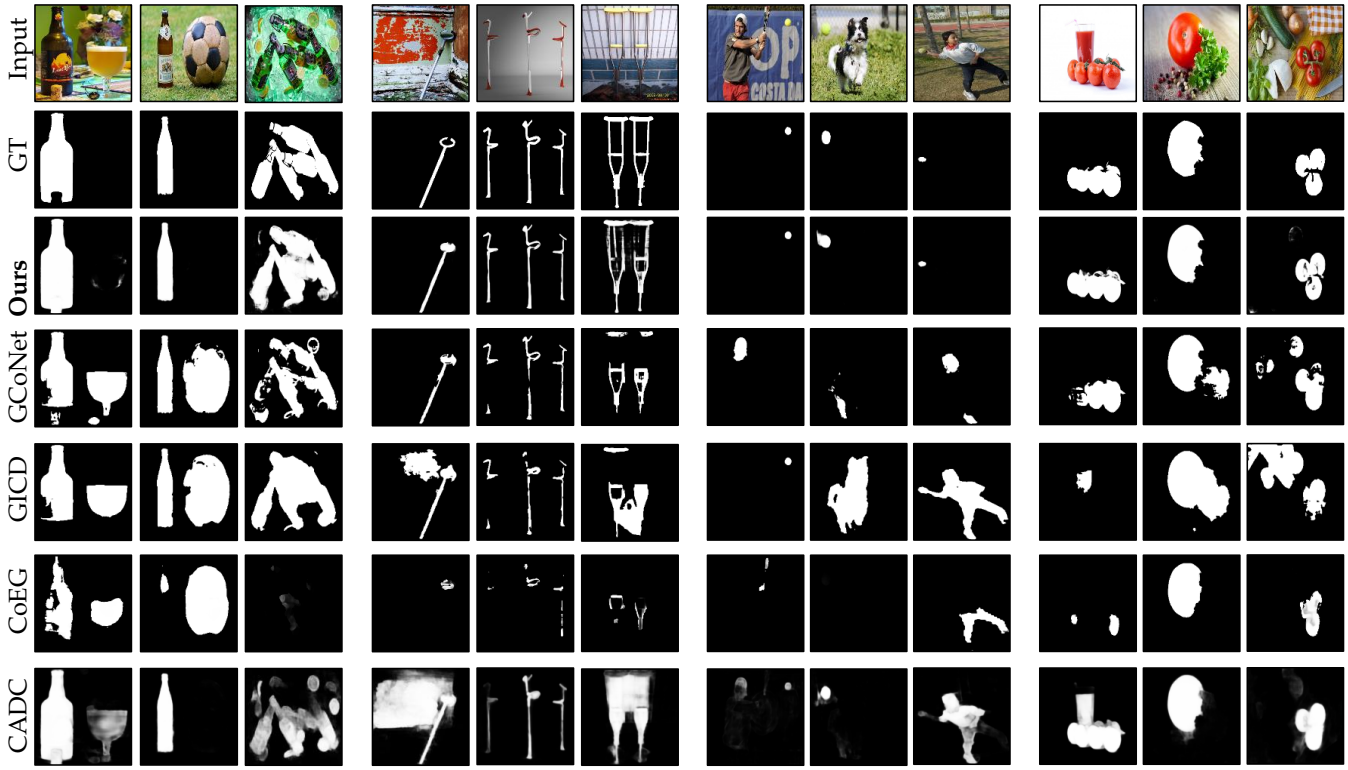

图 12. 定性比较本文的***GCoNet+***和其他方法。“GT”表示真值。**Ours**行的预测结果由经DUTS_class训练的*GCoNet+*生成。

图 13. **DUTS_class**数据集、**COCO-9k**和**COCO-SEG**数据集中有问题的真值图。如给出的例子所示，DUTS_class数据集中可能存在不同类别的显著物体同时出现在同一张图像中，而它们的区域被错误地标注为真值。在COCO-9k和COCO-SEG中，许多物体并非显著的，这可能会对训练CoSOD模型产生负面影响。

示，错误类别的物体仍然存在于真值图中，这可能会导致模型朝着错误方向优化。此外，对于某些图像，其中只有少量的目标物体可供分割，这可能会导致模型无法充分学习到共同物体的分割能力。

COCO-9k/COCO-SEG。 正如 [18], [109]中所提到的，COCO-9k和COCO-SEG都从COCO数据集 [111]里提取图片而构建的。然而，它们都没有考虑显著目标的因素。因此，有真值的物体可能并非显著。由于这原因，仅在COCO-9k或仅在COCO-SEG上训练的模型会在分割共同物体上表现很好，然而在分割显著物体上表现较差。

实验。 在这三个公开测试集和真实场景中，样例或复杂或简单，可以是包含各种物体或复杂背景的，亦或是简简单单地由一个主要物体占据一个全白背景的。为了以令人满意的结果处理这些样本，模型需要在共同物体分割和显著目标检测上都表现很好，这也是可以从COCO-9k [18]/COCO-SEG [109]和DUTS_class [17]分别习得的。正如章节. 4.1所提到的，CoCA [17]更关注复杂上下文中共同物体的分割，而CoSal2015 [8]在评测模型检测显著物体上扮演着更为重要的角色。本文使用这两个数据集从不同方面检验模型性能。

本文在DUTS_class和COCO-9k/COCO-SEG上分别分开和联合地训练*GCoNet+*。从图. 14中的CoSal2015 [8]的结果来看，经过DUTS_class [17]训练的模型展现出了更好的SOD性能，而在检测共同类别的物体上示弱。然而，在COCO-9k或COCO-SEG上的训练能让模型更好地习得分割共同类别物体的能力，却在检测简单的显著物体上表现得相对更弱。与在DUTS_class上训练的模型相比，仅在COCO-9k/COCO-SEG上训练的模型常不能很好地检测显著物体。

为了解决CoSOD的两个子任务，即共同物体分割和显著物体检测，本文需要在两个方向上优化*GCoNet+*。因此，本文设置了联合训练，将*GCoNet+*在DUTS_class [17]和COCO-9k/COCO-SEG [18], [109]上进行联合训练。在联合训练的设置下，上述两个方向同一个模型展现出了更鲁棒的性能。正如表. 3中的性能所示，联合训练的（即：Train-1，3）模型在所有三个测试集上获得了更好的结果。特别是与仅在DUTS_class上训练的模型相比，本文的*GCoNet+*在CoSal2015上表现相当，但在CoCA上提升显著。同时，与仅在COCO-9k上或仅在COCO-SEG上训练的模型相比，联合训练的模型在CoCA上展现出相似的性能，但在CoSal2015上提升明显。如图. 14所示，相同的现象也体现在预测结果图上。

4.7 失败案例

协同显著物体检测网络有两个子目标，可以从两方面划分网络的能力，即：找到共同物体和分割其中的显著物体。因此，本文选出了这两种典型的失败样例用于分析。如图. 15所示，当有许多不同类别的相似物体出现在同一张图像中或是目标物体很难从周边物体中分离出来时，本文的模型可能错判它们而产生不准确的预测结果。

具体来说，对于图. 15左侧的草莓，本文训练好的*GCoNet+*往往关注物体的纹理和颜色。麻点纹理可能被误

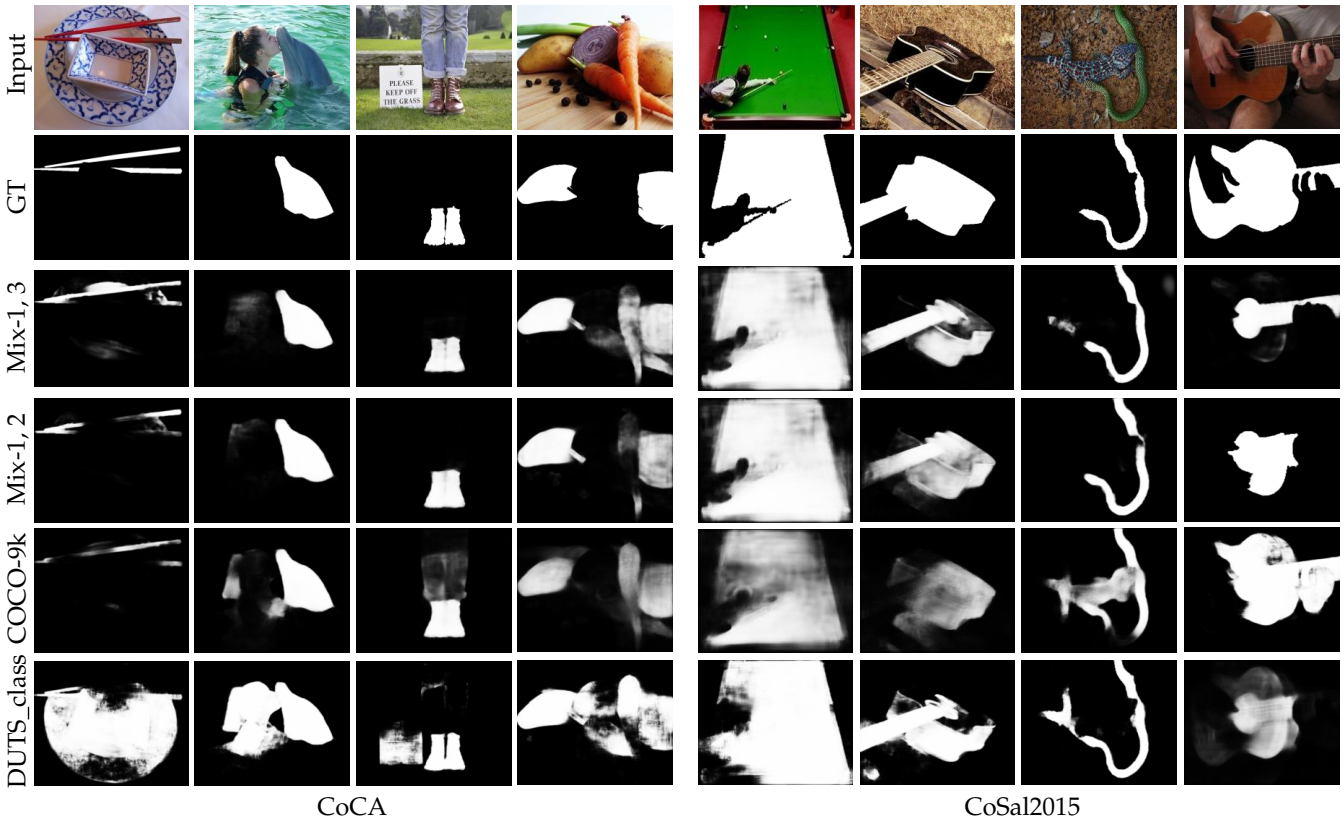

图 14. ***GCoNet+***在不同训练集上训练所产生的定性结果。本文采用了不同的数据集来构建实验以验证DUTS_class数据集 [105]和COCO-9k/COCO-SEG数据集 [18], [109]的不同的优化方向。“Mix-1, 2”表示DUTS_class和COCO-9k被用于训练。“Mix-1, 3”表示DUTS_class和COCO-SEG被用于训练。CoCA是最复杂的CoSOD测试集且需要花更多的注意力找到相同类别的物体。同时，CoSal2015是一个相对简单的数据集，在大多数样例中仅衡量了显著目标检测的能力。

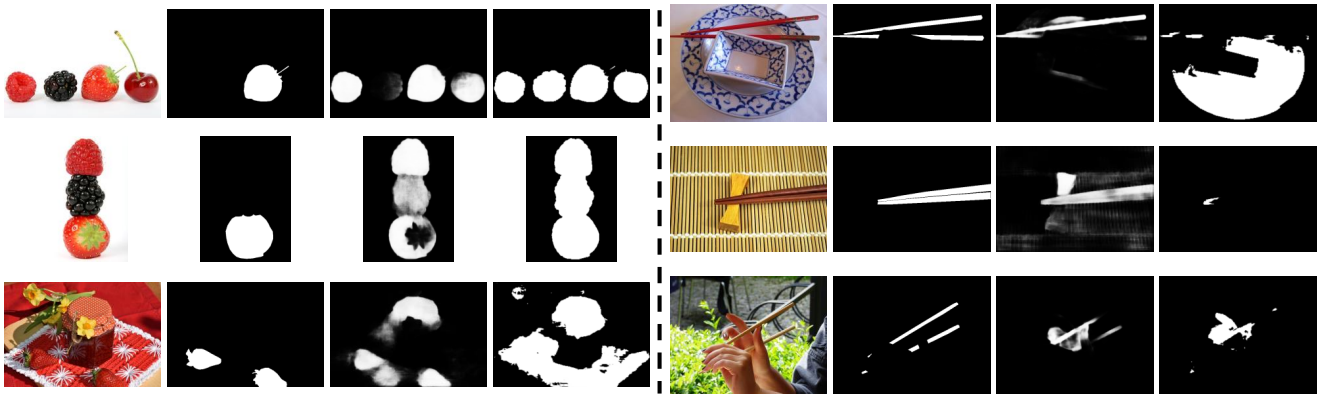

图 15. 本文的***GCoNet+***所生成结果中的失败样例。这里提供了本文的*GCoNet+*和GCoNet [1]中的典型失败样例。在左侧，对周边物体的错误分类导致了较差的CoSOD结果。在右侧，缺乏精细分割能力导致不精准的显著图预测结果。

认为是草莓纹理。因此，蔓越莓和樱桃被误认为是草莓，而蓝莓则可以被鉴别出来。对于图中右侧的筷子，*GCoNet+*更能找到目标对象，但仍无法解决困难的分割问题。尽管本文的*GCoNet+*仍然在这些非常困难的样例中面临问题，它依然能展现出很高的潜力且超越本文之前的GCoNet。

为了进一步提升模型在这些困难样例中的效果，一个更大、包含更多类别的训练集是当务之急。一个更大的类别树能够提供更强的区分不同类别物体的能力，而更多的分割样本则能增强通用分割能力，以准确分割复杂场景中的物体。正如章节. 4.6中所提到的，这可能会是CoSOD任务在未来的一个潜在突破点。

5 潜在应用

本文展示了利用提取出的共显著图生成高质量分割掩码的可能性，并且这种方法可以辅助下游相关图像处理的任务。

应用**1**：内容感知图像共分割。 共显著图已经被广泛应用于图像预处理任务中。以本文实现的无监督物体分割为例，本文首先通过关键词在互联网中检索一组图像。接着，本文的*GCoNet+*被用以生成共显著图。遵循 [25]，本文用GrabCut [121]得到最终的分割结果。这里选择自适应阈值 [122]来初始化GrabCut，用于显著图的二值化处理。如图. 16所示，本文方法能够很好地完成内容感知物体共分割任务，这能够对现有的电子商务应用在背景替换方面具有帮助作用。

图 16. 应用**1**。由本文的*GCoNet+*生成的内容感知的物体共分割视觉结果（“直升机”）。

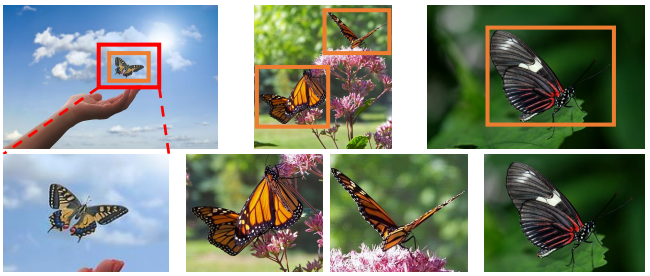

图 17. 应用**2**。基于*GCoNet+*生成的自动缩略图协同定位（“蝴蝶”）。

应用**2**：自动缩略图。本文的成对图像缩略图想法参考了 [71]，并有着相同的目标[5]。为了更好地在网站上分享图片，本文引入了基于CNN的摄影分类任务。如图. 17所示，*GCoNet+*生成的显著图处理成为橘色框。本文还能放大橘色框过得更大的红色框。最终，集合感知切片技术 [71]能够被应用以产生第二行所示结果。

6 结论

本研究提出了一种创新的组协同模型（*GCoNet+*）来处理CoSOD任务。基于实验结果，本文发现组级共识能够引入有效的语义信息，而辅助分类和度量学习能从组内紧凑性和组间分离性两方面来提升特征表达。定性和定量实验表明，*GCoNet+*的性能优于SoTA。本文还展示了*GCoNet+*的易迁移性和应用性，如在共检测与共分割等相关应用的效果。

References

[1] Q. Fan, D.-P. Fan, H. Fu, C.-K. Tang, L. Shao, and Y.-W. Tai, “Group collaborative learning for co-salient object detection,” in *Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 283–12 293.

[2] P. Zheng, H. Fu, D.-P. Fan, Q. Fan, J. Qin, Y.-W. Tai, C.-K. Tang, and L. Van Gool, “GCoNet+: A stronger group collaborative co-salient object detector,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2023.

[3] W. Wang and J. Shen, “Higher-order image co-segmentation,” *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1011–1021, 2016.

[4] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, “Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection,” in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8846–8855.

[5] 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.

[6] K. R. Jerripothula, J. Cai, and J. Yuan, “Efficient video object co-localization with co-saliency activated tracklets,” *IEEE Trans. Circ. Syst. Video Technol.*, vol. 29, no. 3, pp. 744–755, 2018.

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

[8] 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.

[9] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, “Salient object detection in the deep learning era: An in-depth survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2021.

[10] D. Zhang, D. Meng, and J. Han, “Co-saliency detection via a self-paced multiple-instance learning framework,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, 2016.

[11] J. Han, G. Cheng, Z. Li, and D. Zhang, “A unified metric learning-based framework for co-saliency detection,” *IEEE Trans. Circ. Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, 2018.

[12] 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. 502–518.

[13] 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.

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

[15] 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.

[16] 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.*, pp. 1–1, 2021.

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

[18] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, and F. Wu, “Group-wise deep co-saliency detection,” in *Int. Joint Conf. Artif. Intell.*, 2017, pp. 3041–3047.

[19] 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, pp. 6959–6970.

[20] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, and S. Kwong, “Global-and-local collaborative learning for co-salient object detection,” *IEEE Trans. Cybern.*, pp. 1–1, 2022.

[21] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, “Icnet: Intra-saliency correlation network for co-saliency detection,” in *Adv. Neural Inform. Process. Syst.*, 2020, pp. 18 749–18 759.

[22] L. Tang, B. Li, S. Kuang, M. Song, and S. Ding, “Re-thinking the relations in co-saliency detection,” *IEEE Trans. Circ. Syst. Video Technol.*, pp. 1–1, 2022.

[23] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in *Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.

[24] 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. 4558–4567.

[25] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S. Hu, “Global contrast based salient region detection,” in *Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 409–416.

[26] H. Jiang, Z. Yuan, M.-M. Cheng, Y. Gong, N. Zheng, and J. Wang, “Salient object detection: A discriminative regional feature integration approach,” *Int. J. Comput. Vis.*, vol. 123, pp. 251–268, 2013.

[27] 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.

[28] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Sorkine-Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in *Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 733–740.

[29] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, “Deep networks for saliency detection via local estimation and global search,” in *Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3183–3192.

[30] J. Zhang, S. Sclaroff, Z. L. Lin, X. Shen, B. L. Price, and R. Mech, “Unconstrained salient object detection via proposal subset optimization,” in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 5733–5742.

[31] J. Kim and V. Pavlovic, “A shape-based approach for salient object detection using deep learning,” in *Eur. Conf. Comput. Vis.*, 2016, pp. 455–470.

5. Jacobs 等人的工作 [71]限于图像对的例子。

[32] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5455–5463.

[33] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1265–1274.

[34] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 660–668.

[35] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, pp. 330–344, 2015.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.

[37] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, pp. 5706–5722, 2015.

[38] Ali Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, pp. 117–150, 2019.

[39] 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.*, pp. 1–1, 2022.

[40] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 936–944.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[42] J.-X. Zhao, J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 8778–8787.

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

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

[45] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3080–3089.

[46] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Eur. Conf. Comput. Vis.*, 2018, pp. 236–252.

[47] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7479–7489.

[48] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2022.

[49] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9806–9815.

[50] Y. Zhu and J. Du, "Textmountain: Accurate scene text detection via instance segmentation," *Pattern Recognit.*, vol. 110, p. 107336, 2021.

[51] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suàrez, M. Jagersand, and L. Shao, "Boundary-aware segmentation network for mobile and web applications," *arXiv preprint arXiv:2101.04704*, 2021.

[52] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[53] Y. Su, J. Deng, R. Sun, G. Lin, and Q. Wu, "A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection," *arXiv preprint arXiv:2203.04708*, 2022.

[54] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4165–4173.

[55] M. Siam, N. Doraiswamy, B. N. Oreshkin, H. Yao, and M. Jägersand, "Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings," in *Int. Joint Conf. Artif. Intell.*, 2020.

[56] T.-W. Ke, J.-J. Hwang, Y. Guo, X. Wang, and S. X. Yu, "Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers," in *Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 2571–2581.

[57] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 548–557.

[58] W. Li, O. Hosseini Jafari, and C. Rother, "Deep object co-segmentation," in *Asian Conf. Comput. Vis.*, 2018, pp. 638–653.

[59] 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 *Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2129–2136.

[60] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in *Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 993–1000.

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

[62] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[63] C. Zhang, G. Li, G. Lin, Q. Wu, and R. Yao, "Cyclesegnet: Object co-segmentation with cycle refinement and region correspondence," *IEEE Trans. Image Process.*, vol. 30, pp. 5652–5664, 2021.

[64] B. Li, Z. Sun, Q. Li, Y. Wu, and A. Hu, "Group-wise deep object co-segmentation with co-attention recurrent neural network," in *Int. Conf. Comput. Vis.*, 2019, pp. 8519–8528.

[65] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1354–1362.

[66] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Co-attention cnns for unsupervised object co-segmentation," in *Int. Joint Conf. Artif. Intell.*, 2018, pp. 748–756.

[67] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 678–686.

[68] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 478–487.

[69] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, 2016.

[70] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[71] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *ACM symposium on User interface software and technology*, 2010, pp. 219–228.

[72] L. Tang, "Cosformer: Detecting co-salient object with transformers," *arXiv preprint arXiv:2104.14729*, 2021.

[73] G. Ren, T. Dai, and T. Stathaki, "Adaptive intra-group aggregation for co-saliency detection," in *IEEE Int. Conf. Acoust. Speech SP*, 2022, pp. 2520–2524.

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

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

[76] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2818–2826.

[77] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021.

[78] 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. 548–558.

[79] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 1–10, 2022.

[80] X. Qian, Y. Zeng, W. Wang, and Q. Zhang, "Co-saliency detection guided by group weakly supervised learning," *IEEE Trans. Multimedia*, pp. 1–1, 2022.

[81] X. Zheng, Z. Zha, and L. Zhuang, "A feature-adaptive semi-supervised framework for co-saliency detection," in *ACM Int. Conf. Multimedia*, 2018, pp. 959–966.

[82] B. Jiang, X. Jiang, A. Zhou, J. Tang, and B. Luo, "A unified multiple graph learning and convolutional network model for co-saliency estimation," in *ACM Int. Conf. Multimedia*, 2019, pp. 1375–1382.

[83] B. Jiang, X. Jiang, J. Tang, B. Luo, and S. Huang, "Multiple graph convolutional networks for co-saliency detection," in *Int. Conf. Multimedia and Expo*, 2019, pp. 332–337.

[84] 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 *Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9047–9056.

[85] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3196–3209, 2017.

[86] B. Li, Z. Sun, L. Tang, Y. Sun, and J. Shi, "Detecting robust co-saliency with recurrent co-attention neural network," in *Int. Joint Conf. Artif. Intell.*, 2019, pp. 818–825.

[87] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2466–2477, 2018.

[88] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3623–3632.

[89] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Int. Conf. Comput. Vis.*, 2021, pp. 7303–7313.

[90] X. Lu, W. Wang, J. Shen, D. J. Crandall, and L. Van Gool, "Segmenting objects from relational visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7885–7897, 2021.

[91] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Int. Conf. Comput. Vis.*, 2019, pp. 9236–9245.

[92] X. Zhang, Y. Wei, and Y. Yang, "Inter-image communication for weakly supervised localization," in *Eur. Conf. Comput. Vis.*, 2020, pp. 271–287.

[93] J. Jeong, S. Lee, J. Kim, and N. Kwak, "Consistency-based semi-supervised learning for object detection," *Adv. Neural Inform. Process. Syst.*, vol. 32, 2019.

[94] T. Zhou, L. Li, X. Li, C.-M. Feng, J. Li, and L. Shao, "Group-wise learning for weakly supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 799–811, 2021.

[95] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 324–336, 2019.

[96] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Eur. Conf. Comput. Vis.*, 2018, pp. 391–408.

[97] Z. Lai and W. Xie, "Self-supervised video representation learning for correspondence flow," in *Brit. Mach. Vis. Conf.*, 2019, p. 299.

[98] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2566–2576.

[99] Z. Lai, E. Lu, and W. Xie, "Mast: A memory-augmented self-supervised tracker," in *Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6479–6488.

[100] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4282–4291.

[101] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4013–4022.

[102] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 248–255.

[103] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–308, 2009.

[104] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1874–1883.

[105] 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 *Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3796–3805.

[106] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 815–823.

[107] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Eur. Conf. Comput. Vis.*, 2018, pp. 459–474.

[108] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[109] C. Wang, Z. Zha, D. Liu, and H. Xie, "Robust deep co-saliency detection with group semantic," in *AAAI Conf. Art. Intell.*, 2019, pp. 8917–8924.

[110] 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.

[111] T.-Y. Lin, M. Maire, S. J. 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.

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

[113] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Int. Conf. Comput. Vis.*, vol. 2, 2005, pp. 1800–1807.

[114] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1597–1604.

[115] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[116] S. Yu, J. Xiao, B. Zhang, and E. G. Lim, "Democracy does matter: Comprehensive feature mining for co-salient object detection," in *Conf. Comput. Vis. Pattern Recog.*, 2022.

[117] K. Zhang, M. Dong, B. Liu, X.-T. Yuan, and Q. Liu, "Deepacg: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance," in *Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 13 698–13 707.

[118] 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.*, vol. 32, 2019.

[119] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2921–2929.

[120] K. Zhang, T. Li, B. Liu, and Q. Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in *Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3090–3099.

[121] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[122] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.