关于协同显著物体检测的思考

范登平，李腾鹏，林铮，季葛鹏，张鼎文，程明明，付华柱，沈建冰

摘要—本文对图像的协同显著物体检测 (CoSOD) 问题进行了全面的调研评估。协同显著物体检测是显著物体检测 (SOD) 的一个新兴且增长迅速的研究分支，其目的是检测多张图像中同时出现的显著物体。然而，现有的 CoSOD 数据集往往存在着严重的数据偏见，即假设每一组图像中都包含视觉上表观相似的显著物体。这种偏见会导致过于理想化的设置并影响算法模型的有效性，因为相似度通常是指语义上或概念上的相似度，所以在现有数据集上训练的模型可能会在实际情况中表现不佳。为了解决这一问题，本工作首先收集一个新的高质量数据集（名为 CoSOD3k），其包含的大量语义内容比现有的其他 CoSOD 数据集更具挑战性。我们的 CoSOD3k 数据集包含了 160 组，共 3,316 张图像，并提供了多个级别的标注信息。这些图像囊括了多种类别、形状、以及不同的物体尺寸以及背景。此外，我们基于现有 SOD 方法建立一个统一的、可训练的 CoSOD 框架，这类框架是本领域所缺少的。具体来说，本文提出了一种新颖的 CoEG-Net 框架，它采用协同注意投影策略来扩展本文先前的 EGNet 模型，从而能够实现快速地学习协同信息。CoEG-Net 充分利用了以前的大规模 SOD 数据集并且大大提高了模型的可扩展性和稳定性。第三，本文广泛地总结了 40 种最先进的算法，在三个极具挑战性的数据集 (MSRC，iCoSeg，和本文的 CoSOD3k) 上对其中的 18 个模型进行了评测，并呈现了详细的性能分析。最后，本文讨论了当前 CoSOD 研究的挑战以及未来潜在的发展方向。希望本文的研究可以大大促进 CoSOD 研究社区的发展。基准测评工具箱和显著性结果图可在本文的项目主页上找到，网址为：http://dpfan.net/CoSOD3K/.

Index Terms—协同显著性检测，协同注意力投影，CoSOD 数据集，基准评测.

1 引言

在过去的几十年里，RGB 显著物体检测 (SOD) [3]–[7]，RGB-D 显著物体检测 [8]–[12] 和视频显著物体检测 [13]–[15] 已经成为计算机视觉领域一个热门的研究方向 [16]–[23]。显著物体检测旨在模拟人类视觉系统，从单张图像中检测出最吸引人的物体，如图 1 (a) 所示。作为 SOD 任务的扩展，最近出现的协同显著物体检测（CoSOD）则从一组图像中挖掘信息。CoSOD 旨在从单张图像 (例如, 图 1 (c) 中的红衣足球运动员) 或多张图像 (例如, 图 1 (b) 中的蓝衣体操运动员) 中提取出共同的显著物体。协同显著物体的两个重要特性是局部显著性和全局相似性。CoSOD 具有多方向的应用潜力，例如面向集合的图像裁剪 [24]、协同分割 [25], [26]、弱监督学习 [27]、图像检索 [28], [29] 和视频前景检测 [30] 等领域都得到了认可。

因此，CoSOD 任务在近几年里得到了迅速的发展 [19], [35]，例如从 2010[1] 年起，就有几百篇相关刊物相继被发表。大多数 CoSOD 数据集倾向于关注物体之间外观的相似性，来确定多张图片间的协同显著物体。然而，这可能会导致模型对数据选择出现偏差 [3], [36]，在实际应用中这并不合适。因为，即使出现在一组图像中的显著物体，虽然它们属于同一类，但是其通常在纹理、颜色、场景和背景等方面均有所不同 (请参阅图 1 (d) 中的 CoSOD3k 数据集)。除了数据选择的偏差，CoSOD 方法面临着另外两个主要的局限：

(A) 完整性： ϵ (平均绝对误差) [37] 和 F-measure [38] 是在 CoSOD/SOD 模型评价中常用的两个指标。如 [39] 中所讨论的，这些指标有其固有的局限性。为了提供详尽可靠的结论，因此需要引入更准确的指标例如，基于结构的评估指标或基于感知的评估指标。

(B) 公平性： 为了使用 F-measure 进行评估，第一步是对显著性图进行二值化使用一组不同的阈值分成多张前景图。二值化策略 [40] 有很多，例如自适应阈值，固定阈值等。但是，不同的策略将得到不同的 F-measure 性能。此外，以前的工作很少提供有关其二值化策略的详细信息，这会导致来自不同研究机构使用的 F-measure 评价不一致的现象。

为了解决上述问题，本文认为亟需整理各种开源的 CoSOD 算法、数据集和指标，然后呈现一个完整的、统一的评测基准。为此，本文在这项工作的贡献体现在以下四个方面：

- 范登平，林铮和程明明来自中国天津，南开大学计算机学院。(邮箱：dengpfan@gmail.com, frazer.linzheng@gmail.com, cmm@nankai.edu.cn)
- 李腾鹏来自中国南京，南京信息工程大学江苏省大数据实验室 (邮箱: ltpfor1225@gmail.com)
- 季葛鹏来自中国湖北，武汉大学计算机学院。(邮箱：gepengai.ji@gmail.com)
- 张鼎文在中国西安，西北工业大学自动化学院脑与人工智能实验室工作 (邮箱: zhangdingwen2006yyy@gmail.com)
- 付华柱和沈建冰在阿拉伯联合酋长国阿布扎比，起源人工智能研究院工作。(邮箱: {huazhu.fu, jianbing.shen}@inceptioniai.org)

- 这项工作的初步版本已经在 CVPR 2020 [1] 上发表。
- 本文为 TPMAI2021 [2] 论文的中文翻译版本。

- 通讯作者：程明明。

1. 一些代表作可以在 https://hzfu.github.io/proj_cosal_review.html 中找到。

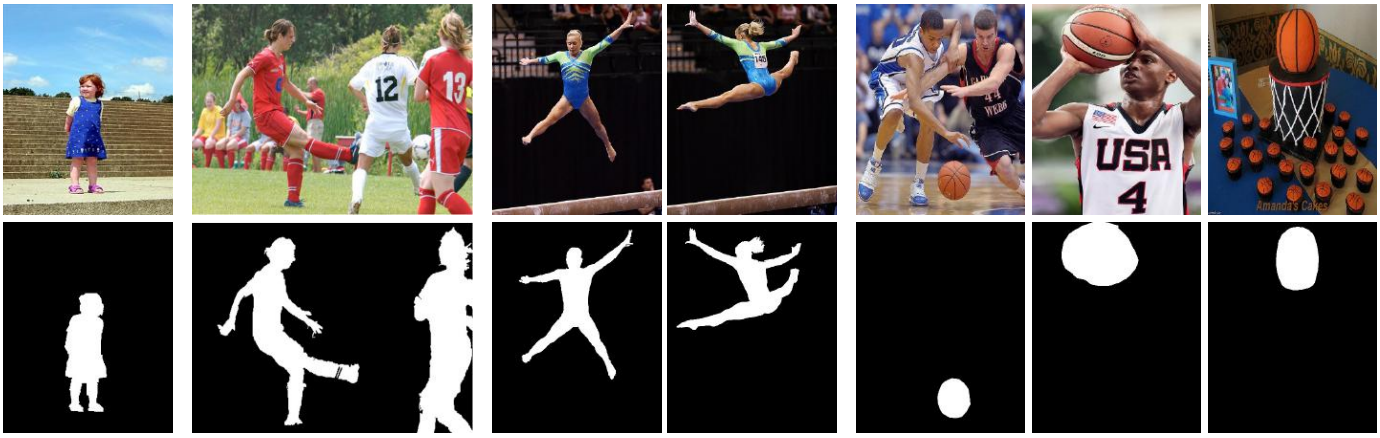

图 1. 不同的显著性物体检测 (SOD) 任务。(a) 传统的 SOD [31]。(b) 图内 CoSOD [32] 旨在从一幅图像内寻找相似的显著物体。(c) 现有的 CoSOD 旨在从一对图像 [33] 或一组图像 [34] 中寻找外观近似的显著物体。(d) 本文提出的开放环境下的 CoSOD 任务，需要模型理解丰富的上下文语义，这将比现有的 CoSOD 任务更具挑战性。

- **首先，本文构建了一个更贴近真实场景的、极具挑战性的 CoSOD3k 数据集。** 本文的 CoSOD3k[2]数据集在规模上是最大的，主要体现在以下两方面: 1）它包含了 13 个超类，160 个组别，共计 3,316 张图像，每个超类都经过了严格的筛选以涵盖各种场景，例如，车辆、食物、工具等；2) 如图 2所示，每幅图像都提供了层次化的标注，包括类别、边界框、物体和实例，这将会最大限度地促进各种视觉任务（例如，物体检测、协同定位、协同分割、协同实例检测等）的发展。
- **其次，本文是第一个大规模地调研协同显著物体检测的研究，** 调研了 40 个最新的模型，并在其中三个大规模的 CoSOD 数据集（iCoSeg，CoSal2015 和本文提出的 CoSOD3k）上评测了其中的 18 个模型。本文还提供了一个便捷的基准测试工具箱，并利用多个指标以便更好地在现有的开源 CoSOD 数据集上进行评估。基准工具箱和结果已开放访问：https://dpfan.net/CoSOD3K/.
- **第三，本文针对 CoSOD 任务提出了一个简单有效的 *CoEG-Net* 基线模型**，它利用协同注意投影和一个基本的 SOD 网络统一在一起从而嵌入外观和语义特征。广泛的基准评测结果表明了 *CoEG-Net* 优于 18 个前沿模型。并且它在视觉效果上更胜一筹，使其成为 CoSOD 任务一个有效的解决方案。
- **最后，本文发现了一些有趣的现象，讨论了一些由基准测试引出的重要问题，并提出了未来的方向。** 本文的研究将为未来 CoSOD 大规模模型的研究提供了催化剂。

2. 与 SOD 数据集相比，收集 CoSOD 数据集更加困难，这就是为什么在过去的 15 年中，最大规模的 CoSOD 数据集，比如 [41] 只含有 2K 图像。即便对于本文的 3K 数据集，就已经花费了 1 年的时间收集，才得到了如此高质量的数据集。同时，为了促进相关视觉任务的发展，本文侧重于提供高质量的层次化标注（例如，图像级别和物体/实例级别）而不仅仅是数据集的规模。

表 1
现有 CoSOD 数据集和本文提出的 CoSOD3k 的统计信息表明 CoSOD3k 提供了更高质量和更丰富的标注。#Gp: 图像组的数量。#Img: 图像的数量。#Avg: 平均每组的图像数量。IL: 是否提供实例级标注。Ceg: 是否为每个组提供类别标签。BBx: 是否为每张图像提供边界框标注。HQ: 高质量标注。

数据集	年份	#Gp	#Img	#Avg	IL	Ceg	BBx	HQ	输入
MSRC [34]	2005	8	240	30					图像组
iCoSeg [42]	2010	38	643	17				✓	图像组
Image Pair [33]	2011	105	210	2		✓*			图像对
CoSal2015 [41]	2015	50	2,015	40		✓*		✓	图像组
WICOS [32]	2018	364	364	1				✓	单张图像
CoSOD3k	2020	160	3,316	21	✓	✓	✓	✓	图像组

* 表示类别信息是广义类而不是细分类。

本文基于并扩展了早前发表的 CVPR2020 版本 [1]。1) 本文实现了一个简单而有效的 CoSOD 框架，该框架通过稀疏卷积和基本的 SOD 网络提供统一框架得以同时嵌入外观和语义特征。更重要的是，本文还设计了一个通用的、即插即用的特征检测器。2）本文在原有工作的基础上竭尽全力进行重新措辞和排版（例如，数据集、框架和重要的结果）。本文还添加了新的章节并从模型定义、对应的技术内容以及更深入的实验（如，与基线和运行时间的比较）这几个方面来描述本文的新框架。此外，本文重写了几个章节以提高可读性，并在引言、CoSOD 模型、定量/定性比较和讨论的等部分给出了更详细的表述。3）本文为 CoSOD 任务建立了第一个标准的基准和模型库，它将各种开源的 CoSOD 数据集整理到一起并且提供了统一的输入/输出格式 (比如, 图像采用 JPEG 格式编码; 真值图采用 PNG 格式编码)。所收集的基于传统特征或基于学习的代码也将尽快开源。

表 2

40 种经典和前沿 CoSOD 方法的总结。**训练集:** PV = PASCAL VOC07 [43]. CR = Coseg-Rep [44]. DO = DUT-OMRON [45]. COS = COCO-subset. **主成分:** IMC = 图像内对比度。IGS: 组内可分离性。IGC: 组内一致性。SPL: 自主学习。CH: 颜色直方图。GMR: 基于图的流形排序。CAE: 卷积自编码器。HSR: 高空间分辨率。FSM: CBCS [30], RC [46], DCL [17], RFCN [47] 和 DWSI [32] 五个显著性模型. **SL.** = 监督级别。W = 弱监督。S = 全监督。U = 无监督。**Sp.:** 是否使用超像素技术。**Po.:** 是否使用 Proposal 算法。**Ed.:** 是否显式地使用边缘特征。**Post.:** 是否引入后处理方法，比如 CRF [48], 图割 (GCut), 或自适应/固定阈值 (THR)。‡ 表示深度模型。有关这些模型的更多详细信息，请参见最新的综述论文。 [1], [19], [35].

序号	模型	出版社年份	数据量	训练集	主模块	SL	Sp	Po	Ed.	Post.
1	WPL [24]	UIST2010			Morphological, Translational Alignment	U				
2	PCSD [49]	ICIP2010	120,000	8*8 image patch	Sparse Feature [50], Filter Bank	W				
3	IPCS [33]	TIP2011			Ncut, Co-multilayer Graph	U	✓			
4	CBCS [30]	TIP2013			Contrast/Spatial/Corresponding Cue	U				
5	MI [51]	TMM2013			Feature/Images Pyramid, Multi-scale Voting	U	✓			GCut
6	CSHS [52]	SPL2013			Hierarchical Segmentation, Contour Map [53]	U			✓	
7	ESMG [54]	SPL2014			Efficient Manifold Ranking [55], OTSU [56]	U				
8	BR [57]	MM2014			Common/Center Cue, Global Correspondence	U	✓			
9	SACS [58]	TIP2014			Self-adaptive Weight, Low Rank Matrix	U	✓			
10	DIM‡ [59]	TNNLS2015	1,000 + 9,963	ASD [38] + PV	SDAE Model [59], Contrast/Object Prior	S	✓			
11	CODW‡ [60]	IJCV2016		ImageNet [61] pre-train	SermaNet [62], RBM [63], IMC, IGS, IGC	W	✓	✓		
12	SP-MIL‡ [64]	TPAMI2017	(240+643)*0.1	MSRC-V1 [34] + iCoSeg [42]	SPL [65], SVM, GIST [66], CNNs [67]	W	✓			
13	GD‡ [68]	IJCAI2017	9,213	MSCOCO [69]	VGGNet16 [70], Group-wise Feature	S				
14	MVSRC‡ [71]	TIP2017			LBP, SIFT [72], CH, Bipartite Graph		✓	✓		
15	UMLF [73]	TCSVT2017	(240 + 2015)*0.5	MSRC-V1 [34] + CoSal2015 [60]	SVM, GMR [45], Metric Learning	S	✓			
16	DML‡ [74]	BMVC2018	10,000 + 6,232 + 5,168	M10K [46] + THUR15K [29] + DO	CAE, HSR, Multistage	S				
17	DWSI [32]	AAAI2018			EdgeBox [75], Low-rank Matrix, CH	S		✓		
18	GONet‡ [76]	ECCV2018		ImageNet [61] pre-train	ResNet-50 [77], Graphical Optimization	W	✓			CRF
19	COC‡ [78]	IJCAI2018		ImageNet [61] pre-train	ResNet-50 [77], Co-attention Loss	W		✓		CRF
20	FASS‡ [79]	MM2018		ImageNet [61] pre-train	DHS [80]/VGGNet, Graph Optimization	W	✓			
21	PJO [81]	TIP2018			Energy Minimization, BoWs	U	✓			
22	SPIG‡ [82]	TIP2018	10,000+210 +2015+240	M10K [46]+IPCS [33] + CoSal2015 [60] + MSRC-V1 [34]	DeepLab, Graph Representation	S	✓			
23	QGF [83]	TMM2018		ImageNet [61] pre-train	Dense Correspondence, Quality Measure	S	✓			THR
24	EHL‡ [84]	NC2019	643	iCoSeg [42]	GoogLeNet [85], FSM	S	✓			
25	IML‡ [86]	NC2019	3624	CoSal2015 [60] + PV + CR	VGGNet16 [70]	S	✓			
26	DGFC‡ [87]	TIP2019	>200,000	MSCOCO [69]	VGGNet16 [70], Group-wise Feature	S	✓			
27	RCANet‡ [88]	IJCAI2019	>200,000	MSCOCO [69] + iCoSeg [42] + CoSal2015 [60] + COS + MSRC [34]	VGGNet16 [70], Recurrent Units	S				THR
28	GS‡ [89]	AAAI2019	200,000	COCO-SEG [89]	VGGNet19 [70], Co-category Classification	S				
29	MGCNet‡ [90]	ICME2019			Graph Convolutional Networks [91]	S	✓			
30	MGLCN‡ [92]	MM2019	N/A	N/A	VGGNet16, PiCANet [93], Inter-/Intra-graph	S	✓			
31	HC‡ [94]	MM2019	N/A	N/A	VAE-Net [95], Hierarchical Consistency	S	✓	✓		CRF
32	CSMG‡ [96]	CVPR2019	25,00	MB [97]	VGGNet16 [70], Shared Superpixel Feature	S	✓			
33	DeepCO^{3}‡ [98]	CVPR2019	10,000	M10K [46]	SVFSal [99] / VGGNet [70], Co-peak Search	W		✓		
34	GWD‡ [100]	ICCV2019	>200,000	MSCOCO [69]	VGGNet19 [70], RNN, Group-wise Loss	S				THR
35	CAFCN‡ [101]	TCSVT2020	200,000	MSCOCO [69]	VGGNet16 [70], Co-Attention, FCN	S				
36	GSPA‡ [102]	TNNLS2020	200,000	COCO-SEG [102]	VGGNet19 [70], Group Semantic, Pyramid Attention	S				
37	GOMAG [103]	TMM2020	N/A	N/A	General Optimization, Adaptive Graph Learning	U	✓			
38	AGC‡ [104]	CVPR2020	200,000	MSCOCO [69]	VGGNet16 [70], Graph Convolution & Clustering	S				
39	GICD‡ [105]	ECCV2020	8,250	DUTS [31]	VGGNet19 [70], Gradient Inducing, Attention Retaining	S				
40	CoEG-Net‡	2020	10,553	DUTS [31]	VGGNet16, Co-attention Projection	S			✓	CRF

2 相关工作

2.1 CoSOD 数据集

目前，如表 1中所示，该领域仅有少量的 CoSOD 数据集 [29], [32]–[34], [41], [42], 其中 *MSRC* [34] 和 *Image Pair* [33] 是最早的两个数据集。*MSRC* 用于识别图像中的物体类别，并在过去几年激发了许多有趣的想法。该数据集含有 8 组图像，图像总共 240 张，并提供了手工像素级标注的真值图。Li 等人 [33] 提出 *Image Pair* 数据集，是专门为一对图像设计的且总共包含 210 张图像（105 个图像组）的数据集。*iCoSeg* [42] 数据集于 2010 年发布, 是一个总共包含 38 个类别，643 张图像的相对较大的数据集。该数据集中的每个图像组包含 4 到 42 张图像，而不像 *Image Pair* 数据集那样仅包含 2 张图像。*THUR15K* [29] 和 *CoSal2015* [41] 是两个大规模的开源数据集，与 CoSal2015 一起广泛用于协同显著物体检测算法的评估。与上述数据集不同，*WICOS* [32] 数据集旨在从单张图像中检测协同显著物体，因此可将每张图像视为一组。

尽管上述数据集都不同程度地推进了 CoSOD 任务的发展，但是由于仅仅包含几十组图像，因此，在多样性方面受到严重限制。在如此小规模的数据集上，模型的扩展性便无法得到充分评估。此外，这些数据集仅提供物体级别的标注，忽略了诸如边界框，实例等更丰富的标注，而这对于促进其他视觉任务以及多任务建模至关重要。尤其在当前的深度学习时代，模型经常需要大量的训练数据。鉴于这些原因，本文将专注于两个规模相对较大的数据集 (比如，*iCoSeg* [42] 和 *CoSal2015* [41]) 连同本文提出的具有挑战性的数据集一起进行更深入的分析。

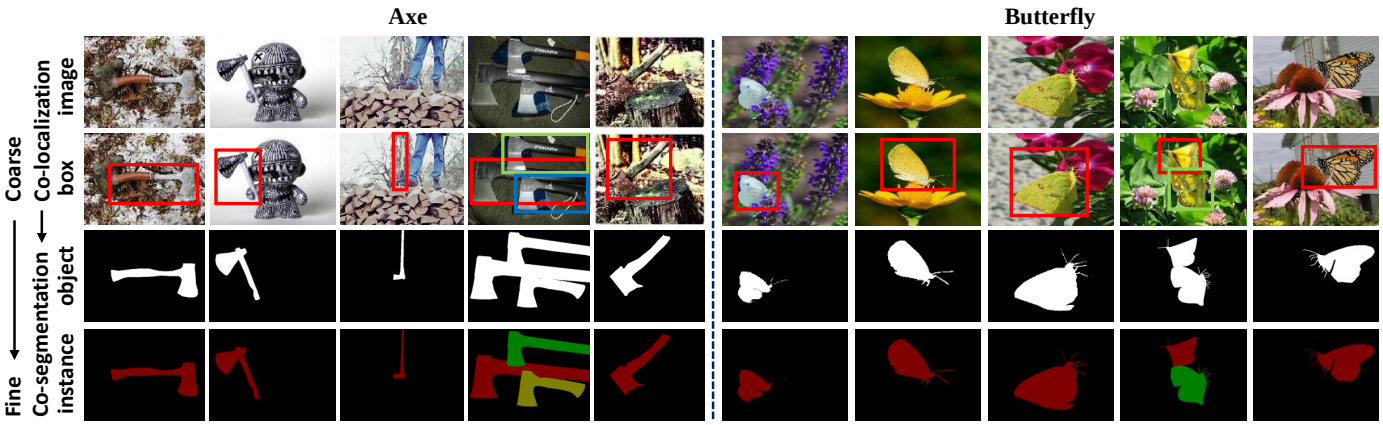

图 2. 本文的 CoSOD3k 数据集中的样本图像。它具有丰富的标注，例如图像级类别（第一行）、边界框、物体级标注和实例级标注。本文的 CoSOD3k 将为 CoSOD 任务提供坚实的基础并可以使更多的相关领域受益，例如，协同分割和弱监督定位任务。

2.2 CoSOD 方法.

先前的 CoSOD 研究 [33], [58], [73], [81] 表明，通过将输入图像分割成若干计算单元 (例如，超像素区域 [106] 或者像素簇 [30])，可以建模图像间的对应关系，相似的观察结果同样可以在 [19], [35] 中找到。在这些方法中，启发式特征 (例如：轮廓 [52]、颜色和亮度) 可以从图像中提取，并通过不同方式捕获高级特征以表达语义属性，例如，采用度量学习 [73] 或自适应加权 [58] 方式。一些方法还研究了如何通过各种诸如翻译对齐 [24]、高效的流形排序 [54] 和全局相关性 [57] 的计算机制，以获得图像间约束。还有一些方法（例如 PCSD [49]，仅使用滤波技术）甚至不需要执行输入图像对之间的对应匹配，就能够在协同注意呈现前获得 CoSOD 结果。

最近，基于深度学习的 CoSOD 模型通过联合学习协同显著物体的表示特征取得了良好的性能。例如, Zhang 等人 [59] 引入了领域适应模型来转换 CoSOD 的先验知识。Wei 等人 [68] 在协作学习框架中使用了一组输入和输出来挖掘图组和单图特征表示之间的协作和关联关系。沿着另一套思路，MVSRCC 模型 [71] 采用了如 SIFT、LBP 和颜色直方图这类经典特征来作为多视角特征。此外，其他几种方法 [78], [82], [84], [87], [89], [96], [98] 都是基于更强大的卷积神经网络模型 (例如：ResNet [77]、Res2Net [107]、GoogLeNet [85] 和 VGGNet [70])，以此达到了前沿的效果。这些深度模型通常通过以下两种方法获得更好的性能：一种是基于弱监督学习 (例如：CODW [60]、SP-MIL [64]、GONet [76] 和 FASS [79])，另一种是基于全监督学习 (例如：DIM [59]、GD [68] 和 DML [74]。在本文初稿提交后，还出现了一些同期的新工作 [101]–[105]。表 2中呈现了现有的 CoSOD 模型。

3 CoSOD3k 数据集

3.1 图像采集

本文图像来自大规模物体识别数据集 ILSVRC [108]，从而建立了高质量的数据集 CoSOD3k。本文使用 ILSVRC 数据集来构建 CoSOD3k 数据集有诸多好处：首先，ILSVRC 是从 *Flickr* 网站中使用场景级关键词查询收集的。因此，它包括各种物体类别，多样的真实场景以及不同外观的图像，从而涵盖了 CoSOD 任务中大部分挑战，为 CoSOD 建立具有代表性的基准数据集提供了合理依据。不过，更重要的是，每个目标物体类别都提供了轴对齐的边界框标注信息，这便于我们在标注实例级物体时能够做到准确无误。

3.2 层次化标注

类似于文献 [109], [110]，数据标注以分层（从粗糙到精细）的方式执行（请参见图 2）。

• **类别标签：** 本文为 CoSOD3k 数据集建立了一个分级 (即：三级) 分类系统。首先选择了 160 个常用类别（请参见图 3) 以生成子类（例如：*Ant*、*Fig*、*Violin* 和 *Train* 等等)，这些与 ILSVRC 数据集中的原始类别一致。然后，为每个子类分配一个上级类。最后，本文将上级类集成到 13 个超类中。CoSOD3k 数据集的分类结构在图 4中给出。

• **边界框标签：** 标注的第二级是边界框标记，它广泛用于物体检测和定位。尽管 ILSVRC 数据集提供了边界框标注，但标记的物体不一定显著。遵循许多著名 SOD 数据集 [31], [38], [46], [97], [111]–[117] 的标准，本文要求三位受试者在每幅图像中那些吸引了他们的注意力的物体周围重新绘制新的包围盒。然后，将三个观看者标记的边界框合并，再邀请两名 CoSOD 领域的资深研究人员检查标注两次。接着，借鉴文献 [118] 中的做法，丢弃包含六个以上物体的图像。最后，本文在 160 个类别中收集了 3,316 张图像。例子可以在图 2中找到。

• **物体/实例级标注。** 对于 CoSOD 数据集，高质量的像素级标注是必需的。本文聘请了 20 名专业标注人员，并预先对他们进行了 100 张图像标注的培训。然后指示它们根据先前

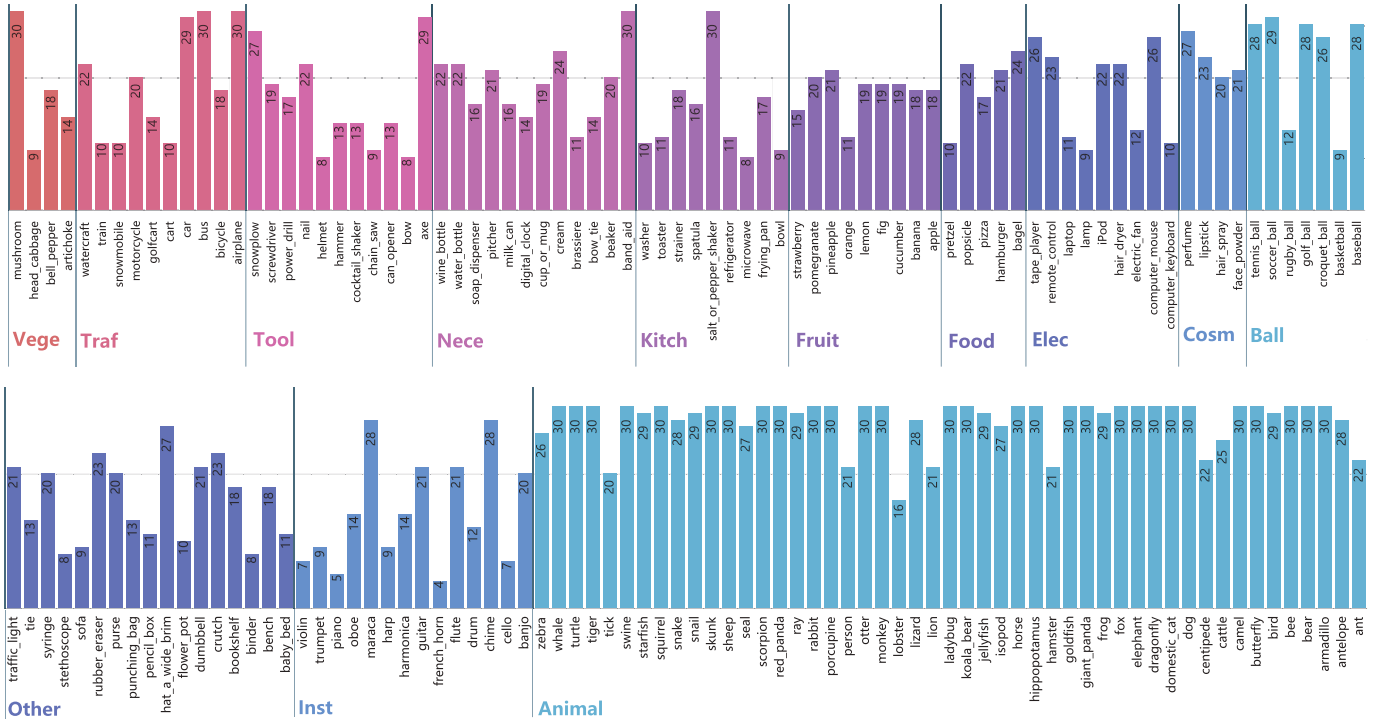

图 3. 本文数据集中 160 个子类中的图像数量。请在屏幕上放大查看，以获得最佳的视觉效果和详细信息。

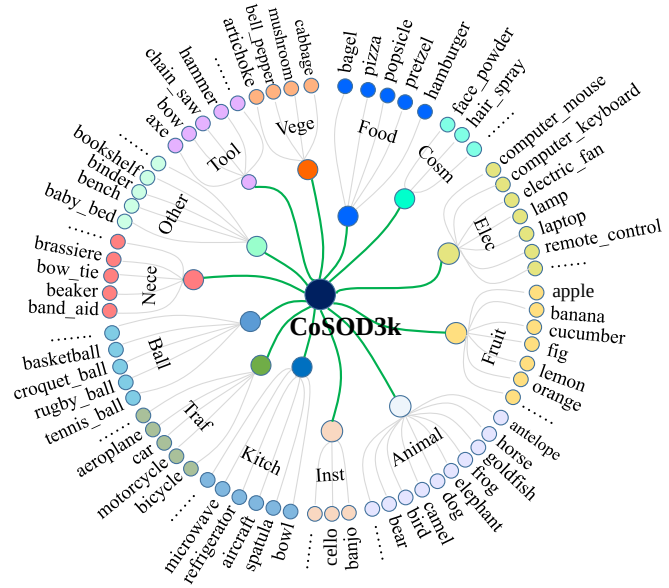

图 4. 本文数据集的分类结构，它由 13 个超类和 160 个子类组成。

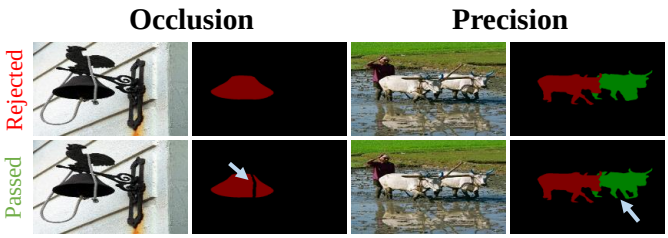

图 5. 本文的 CoSOD3k 数据集中有一些通过和拒绝的案例（例如，遮挡和精度）。

的包围盒用物体级别和实例级别的掩膜标注图像。对于物体级别和实例级别的标签，每幅图像的平均标注时间分别约为 8 分钟和 15 分钟。此外，本文还有 3 名志愿者对整个过程进行了交叉检查（3 次以上），以确保高质量的标签（请参见图 5）。通过这种方式，本文获得了准确而具有挑战性的数据集，其中包含总共 3,316 个物体级标注和 4,915 个实例级标注。请注意，本文的最终边界框标签会根据实例级别的标注进行进一步细化以收紧物体范围。

3.3 数据集的特色以统计信息

为了对 CoSOD3k 数据集提供更深入的了解，本文呈现以下几个重要特征。

• **特定类别的掩膜信息叠加：** 图 7展示了单个类别和整个数据集的平均真值标签。可以看出，某些具有独特形状的类别（例如：飞机、斑马和自行车）呈现形状偏置的模样，而具有非刚性或凸形形状的类别（例如：金鱼、鸟和公共汽车）则没有清晰的形状偏置。整体数据集标签（图 7 的右侧）趋向于显示为没有形状偏置的中心偏置图。众所周知，在拍摄照片时，人们通常倾向于更加关注视觉场景的中心。因此，在显著性物体检测算法中采用高斯函数进行平滑处理，更易于获得较高的分数。由于篇幅所限，本文将在补充材料中展示全部 160 种特定类别叠加后的掩膜图像。

• **丰富的物体多样性：** 如表 5（第 2 行）和图 3所示，本文的 CoSOD3k 涵盖了各种各样的超类，包括蔬菜、食物、水

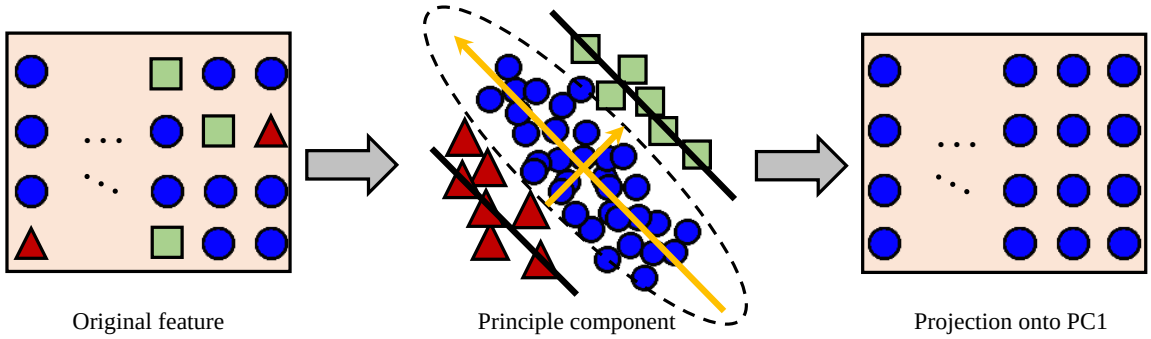

图 6. 本文协同注意投影操作示意图。在给定的涵盖了一般物体（圆形）、带噪声的前景（三角形）和杂乱的背景（正方形）的原始特征空间中，协同注意投影可以识别一般物体的特征主成分，从而有助于在保留一般物体特征的同时消除干扰特征。通过采用协同注意投影操作，本文投影了主成分并获得了新的特征表示。请参考章节4.2以获取更多细节。

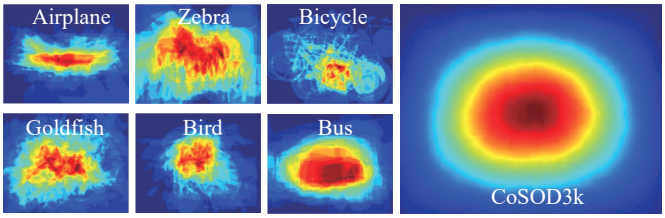

图 7. CoSOD3k 数据集中特定类别和整体数据集标注掩膜叠加的可视化结果图。

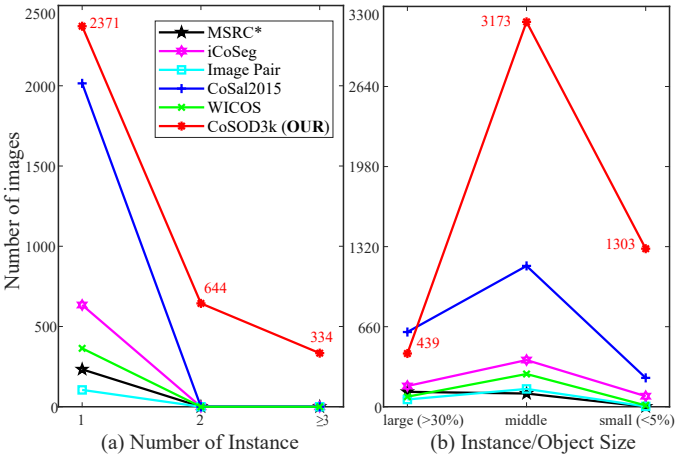

图 8. MSRC、iCoSeg、Image Pair、CoSal2015、WICOS 和本文的 CoSOD3k 数据集的图像数量，以实例数（a）和实例/物体大小（b）分别表示。

果、工具、必需品、交通、化妆品、球、乐器、厨具、动物和其它，这有利于更全面地在真实场景中探索。

• **实例的数量：** 将物体进一步解析为实例，对于人类理解、分类以及与世界交互的过程是至关重要的。为了使学习方法能够理解实例，那么提供实例标签就势在必行。考虑到这一点，与现有的 CoSOD 数据集不同，本文的 CoSOD3k 包含具有实例级注释的多实例场景。如图 8 (a) 中所示，实例数为 1 个、2 个以及大于 3 个的图像数比例为 7 : 2 : 1。

• **实例的大小：** 实例大小定义为前景实例像素与总图像像素之比。图 8 (b) 以小型、中型和大型实例/物体的形式展示了本文 CoSOD3k 数据集中不同大小的实例。其分布为 0.02%～86.5%（平均：13.8%），覆盖了很广的范围。

4 本文方法

在这项工作中，通过扩展先进的显著性物体检测模型 EGNet [119]，并以无监督的方式引入协同注意信息，本文还提出了一个简单且有效的 *CoEG-Net* 基准方法。

4.1 方法简介

对于一组 N 张关联的图像 $\{\mathbf{I}^n\}_{n=1}^N$，协同显著性检测任务旨在分割存在注意力的共同前景物体并生成优化后的代表了输入图像之间的协同显著物体的协同显著图。为了预测协同显著性物体，本文提出了一个二分支检测框架，以乘法独立的方式分别捕获共同依赖和显著前景。图 9 展示了本文提出方法的框架，它独立地在顶部分支输出协同注意图 $\{\mathbf{A}^n\}_{n=1}^N$，在底部分支中输出显著图 $\{\mathbf{S}^n\}_{n=1}^N$。然后，协同注意图 $\mathbf{A}^n$ 和显著性先验图 $\mathbf{S}^n$ 通过逐元素相乘得到最终协同显著性预测 $\mathbf{A}^n \otimes \mathbf{S}^n$。

为了获取输入图像 $\mathbf{I}^n$ 的显著性先验图 $\mathbf{S}^n$，本文仅使用边缘导向的显著物体检测方法 EGNet [119] 来获取多尺度的显著性先验。EGNet 模型在大规模单图像显著性物体检测数据集 DUTS [31] 上进行了训练，这有助于识别没有交叉图像信息的图像显著物体区域。然后，真正的挑战变成了如何以无监督的方式生成协同注意图 $\mathbf{A}^n$，本文将在下节中介绍它。

4.2 协同显著性学习中的协同注意力投影

协同注意力学习的设计（请参见图 6）受 Zhou 等人 [120] 提出的类激活映射（CAM）技术的启发。给定输入图像 $\mathbf{I}^n$ 以及利用标准分类网络 (例如 VGGNet [70]) 中最后的卷积层获得的对应特征**激活图** $\mathbf{X}^n$。更多细节参见表 3。

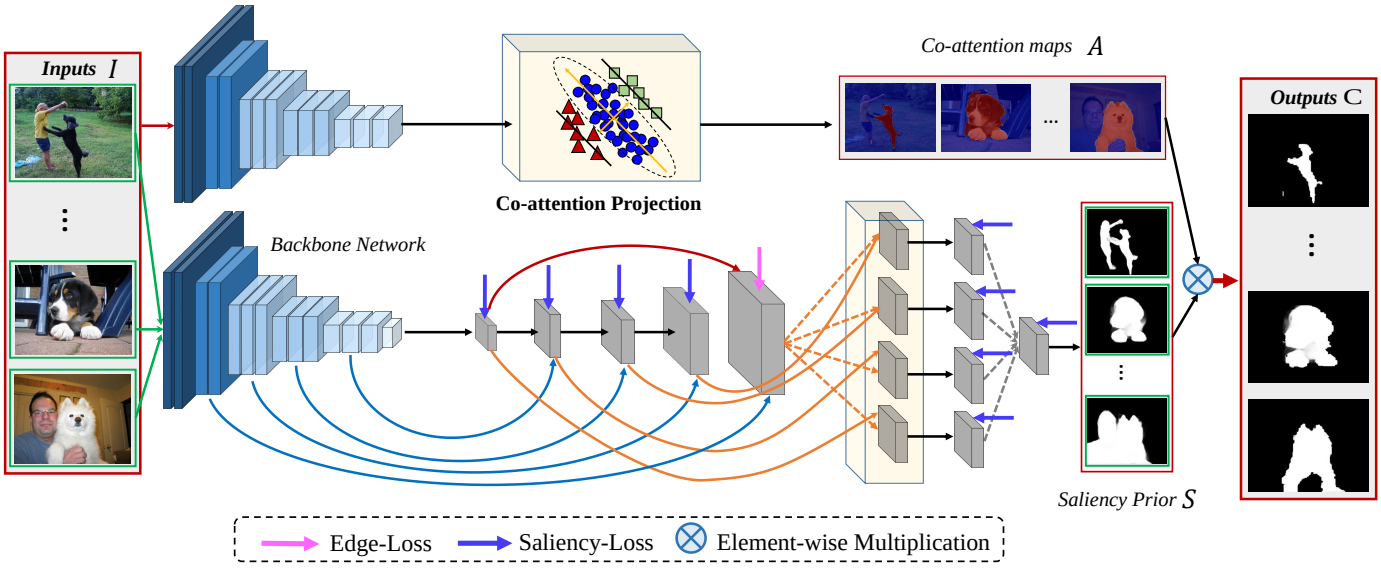

图 9. 本模型的框架流程，它包含两个独立的分支。对于一组输入图像 $\{\mathbf{I}^n\}_{n=1}^N$，顶部分支首先提取高级图像特征，并将其输入到协同注意投影模块，从而为每张输入图像 $\mathbf{I}^n$ 生成协同注意图 $\mathbf{A}^n$。在底部分支中，每张图像 $\mathbf{I}^n$ 被输入到边缘引导的显著性检测网络（EGNet [119]）中以生成显著性先验映射 $\mathbf{S}^n$。最后，$\mathbf{A}^n$ 和 $\mathbf{S}^n$ 可以简单地使用逐元素相乘以得到优化的输出 $\mathbf{A}^n \otimes \mathbf{S}^n$。详情见章节4。

表 3
符号、维度、索引和含义表。

符号	维度	索引	含义
$\mathbf{A}^n$	$H \times W$	(i,j)	$\mathbf{I}^n$ 的联合显著图
$\mathbf{S}^n$	$H \times W$	(i,j)	$\mathbf{I}^n$ 的显著性先验
$\mathbf{X}^n$	$H \times W \times K$	(i,j,k)	最后一个 conv 层的激活映射图集合
$\mathbf{X}_k^n$	$H \times W$	(i,j)	通道 $\mathbf{X}^n$ 的特征映射图
H	1×1	标量	空间高度
W	1×1	标量	空间宽度
K	1×1	标量	特征通道数量
$\mathbf{x}^n(i,j)$	$K \times 1$	k	位置 (i,j) 的 $\mathbf{X}^n$ 描述子
$\mathbf{M}_c^n$	$H \times W$	(i,j)	类别 c 的注意力图
ω^c	$K \times 1$	k	类别 c 的通道级权重
$\bar{\mathbf{x}}$	$K \times 1$	k	向量 $\mathbf{x}^n(i,j)$ 的平均值
$\hat{\mathbf{x}}^n(i,j)$	$K \times 1$	k	零均值操作 $\hat{\mathbf{x}}^n(i,j) = \mathbf{x}^n(i,j) - \bar{\mathbf{x}}$
$Cov(\hat{\mathbf{x}})$	$K \times K$	-	$\{\hat{\mathbf{x}}^n(i,j)\}$ 的协方差矩阵
ξ^*	$K \times 1$	k	$Cov(\hat{\mathbf{x}})$ 的本征值

利用仅带有关键字标签的图像，CAM 技术旨在使用**特征图** $\{\mathbf{X}_k^n\}$ 为每个类 c 生成特定类别的关注图 $\mathbf{M}_c^n$：

$$\mathbf{M}_c^n = \sum_{k=1}^{K} \omega_k^c \mathbf{X}_k^n, \tag{1}$$

其中权重 ω^c 可以使用关键字级别的弱监督方式 [120] 进行训练。请注意，可以使用权重 ω^c 和 $\mathbf{X}^n$ 中的逐通道**描述符**在每个空间元素的空间位置 (i,j) 独立估计类激活图 $\mathbf{M}_c^n$

$$\mathbf{M}_c^n(i,j) = (\omega^c)^\top \cdot \mathbf{x}^n(i,j) \tag{2}$$

因此，CAM [120] 技术本质上起到了线性变换的作用，该变换使用学习过的类别特定权重 ω^c 将图像特征 $\mathbf{x}^n(i,j)$ 转换为特定类别的激活分数 $\mathbf{M}_c^n(i,j)$。

不幸的是，在协同显著性检测问题设定中，无法实现关键字级别的监督。因此，本文不得不以无监督的方式寻找协同物体的权重 ω，以揭示图像特征的内部结构。理想情况下，一组关联图像 $\{\mathbf{I}^n\}_{n=1}^N$ 中未知的协同物体类别应对应于一个线性投影，该线性投影会导致协同区域中的激活得分高，而在图像其他区域中的类别激活分数较低。从另一个角度来看，相同的物体类别应对应于在最终的类激活图中生成最高方差（信息量最多）的线性变换。借鉴粗略定位任务 [124] 中的思想，本文通过探索经典主成分分析（PCA）[125] 来实现这一目标，这是以最能解释数据差异的方式揭示数据内部结构的最简单方法。

具体来说，给定关联的图像 $\{\mathbf{I}^n\}_{n=1}^N$，对于每个具有相应的特征激活 $\mathbf{X}^n$ 的图像 $\mathbf{I}^n$，本文的目标是找到 $\mathbf{X}^n$ 的线性变换，从而产生方差最高的协同注意力图 $\{\mathbf{A}^n\}$。这可以通过分析特征描述符 $\{\mathbf{x}^n(i,j)\}$ 的协方差矩阵来实现。令 $\bar{\mathbf{x}} = \frac{1}{Z}\sum_n \sum_{i,j} \mathbf{x}^n(i,j)$，其中 $Z = N \times H \times W$。本文的描述符的零均值形式为 $\hat{\mathbf{x}}^n(i,j) = \mathbf{x}^n(i,j) - \bar{\mathbf{x}}$。协方差矩阵可以记作

$$Cov(\hat{\mathbf{x}}) = \frac{1}{Z}\sum_n \sum_{i,j} (\hat{\mathbf{x}}^n(i,j) - \bar{\mathbf{x}})(\hat{\mathbf{x}}^n(i,j) - \bar{\mathbf{x}})^T. \tag{3}$$

然后，可以使用特征向量 ξ^* 建立期望的线性投影，该特征向量对应于 $Cov(\hat{\mathbf{x}})$ 的最大特征值。因此，可以将协同注意投影设计为以最丰富的视角呈现其特征的投影

$$\mathbf{A}^n(i,j) = \xi^{*\top} \cdot \hat{\mathbf{x}}^n(i,j). \tag{4}$$

表 4
在两个经典 [41], [42] 和本文的 CoSOD3k 上对 18 种最先进的 CoSOD 方法进行基准测试的结果。符号 "○" 表示代码或结果无法获得。请注意，UMLF 采用了 MSRC 和 CoSal2015 的一半图像来训练其模型。下划线表示由已在相应数据集中训练的模型（例如 SP-MIL 和 UMLF）生成的分数。更多训练细节请参考 表 2。

	指标	CBCS [30]	ESMG [54]	RFP [121]	RCS [52]	HSSA [58]	CSC [122]	ODR [73]	UMLF	DIM [59]‡	CODW [60]‡	MIL [65]‡	IML [86]‡	GONet [76]‡	SP-MIL [64]‡	CSMG [96]‡	CPD [123]‡	GSPA [102]‡	AGC [104]‡	EGNet [119]‡	***CoEG-Net*** 本文方法‡
iCoSeg	$E_\phi \uparrow$	.797	.784	.841	.841	.817	.889	.827	.864	.832	.799	.895	.864	.843	.889	.900	.818	.897	.911	**.912**
	$S_\alpha \uparrow$	.658	.728	.744	.750	.752	.815	.703	.758	.750	.727	.832	.820	.771	.821	.861	.784	.821	.875	**.875**
	$F_\beta \uparrow$	.705	.685	.771	.765	.770	.823	.761	.797	.782	.741	.846	.832	.794	.850	.855	.718	.837	.875	**.876**
	$\epsilon \downarrow$	.172	.157	.170	.179	.154	.114	.226	.179	.184	.186	.104	.122	.174	.106	**.057**	.098	.079	.060	.060
CoSal2015	$E_\phi \uparrow$	.656	.640	○	.685	.749	.749	.769	.695	.752	.720	-	.805	○	.842	.841	.855	**.890**	.843	.882
	$S_\alpha \uparrow$	.544	.552	○	.592	.694	.689	.662	.592	.648	.673	-	.751	○	.774	.814	.797	.823	.818	**.836**
	$F_\beta \uparrow$	.532	.476	○	.564	.650	.634	.690	.580	.667	.620	-	.740	○	.784	.782	.779	.831	.786	**.832**
	$\epsilon \downarrow$	.233	.247	○	.313	.194	.204	.271	.312	.274	.210	-	.160	○	.130	.098	.099	.090	.099	**.077**
CoSOD3k	$E_\phi \uparrow$	.637	.635	○	.656	○	.700	.758	.662	○	○	.773	○	○	.804	.791	.800	.823	.793	**.825**
	$S_\alpha \uparrow$	.528	.532	○	.563	○	.630	.632	.559	○	○	.720	○	○	.711	.757	.736	.759	.762	**.762**
	$F_\beta \uparrow$	.466	.418	○	.484	○	.530	.639	.495	○	○	.652	○	○	.709	.699	.682	.729	.702	**.736**
	$\epsilon \downarrow$	.228	.239	○	.309	○	.229	.285	.327	○	○	.164	○	○	.157	.120	.124	.094	.119	**.092**

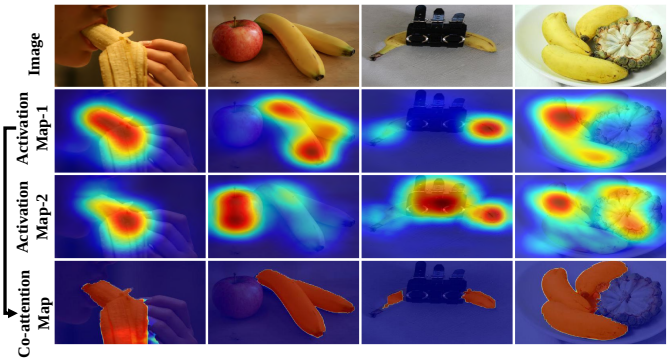

图 10. 可视化协同激活图（第二行和第三行），使用最大和第二大的特征值，以及从 CoSal2015 的 "香蕉" 组中选择的它们对应的后处理（流形排序和 DenseCRF）协同注意图 $\mathbf{A}^n$（第四行） [41]。

图 10展示了一些由式 (4) 生成的协同激活图（第二行和第三行）的可视化样本。图组中包含了诸如括香蕉、苹果、瓶子和菠萝多个类别多个物体，这增加了分割出正确区域的难度，而使用最大特征值 $Cov(\hat{\mathbf{x}})$（第二行）就可以高效地定位协同物体，并在一开始就将它们分离出来（最后一行）。

4.3 实现细节

为进行公平比较，移除顶端分类层后，将标准的 VGGNet16 [70] 作为本文的骨干网络。训练过程在 30 个轮回内完成，学习率在 15 个轮回后除以 10。对于边缘导向的上下文显著性网络，设置与 [119] 相同。请注意，在训练阶段，损失函数与 EGNet 模型保持相同。和文献 [76] 中的后处理过程类似，在融合协同注意力图 $\mathbf{A}^n$ 和显著性先验图 $\mathbf{S}^n$ 之前，本文使用 DenseCRF [48] 和流形排序算法 [126] 进一步优化协同注意力图。这些示例展现在图 10的第三行中。

5 基准实验

5.1 实验设置

• **评估指标：** 为了提供全面的评估，本文采用了四个广泛使用的度量标准来评估 CoSOD 模型的性能，包括最大 F 指标 F_β [38]、平均绝对误差（MAE, ϵ）[37] 、S 指标 S_α [127] 和最大 E 指标 E_ϕ [128]。完整的评估工具箱见：https://github.com/DengPingFan/CoSODToolbox。

F 指标（F_β）[38] 用以评估精确率和召回率的加权调平均。那么，显著性图就要使用不同的阈值进行二值化，其中每个阈值对应于二值化显著性预测结果。本文比较预测图和真值图，以获取精准率和查全率的值。F 度量 F_β 分数经常被使用，该分数对应于整个数据集的最佳固定阈值。

MAE（ϵ）[37] 是一种非常简单的评估指标，无需任何二值化要求，即可直接测量真实值与预测值之间的绝对差。F-measure 和 MAE 均以像素为单位评估预测结果。

S 指标（S_α）[127] 用于评估显著性图和相应的真值图之间的结构相似性。它可以直接评估无需将连续的显著性预测结果进行二值化，并且同时考虑大规模结构相似性。

E 指标（E_ϕ）[128] 是一种可同时评估预测图和真值图之间的局部和全局相似性的感知指标。

• **对比方法：** 在 CoSOD 实验中，本文评估/比较了 16 种前沿的 CoSOD 模型，包括 7 种传统方法 [30], [52], [54], [58], [73], [121], [122] 和 9 种深度学习模型 [59], [60], [64], [65], [76], [86], [96], [119], [123]。这些方法是根据两个标准选择的：（1）具有代表性，以及（2）代码或结果开源。

• **评测协议：** 本文评估了两个现有的 CoSOD 数据集，比如, *iCoSeg* [42] 和 *CoSal2015* [41]，以及本文的 CoSOD3k。据

图 11. CoSOD3k 数据集的示例。本文展示了来自 13 个超类的代表性物体类别的细分示例图。

表 5
每个超类在本文 CoSOD3k 数据集上的平均 E 指标性能 E_ϕ。Vege. = 蔬菜, Nece. = 必需品, Traf. = 交通, Cosm.= 化妆品, Inst. = 乐器, Kitch. = 厨具, Elec. = 电子产品, Anim. = 动物, Oth. = 其它。 "全部" 是指整个数据集的分数。本文仅使用公开的代码评估 10 个最新模型。请注意, CPD 和 EGNet 是 SOC 基准评测 (http://dpfan.net/socbenchmark) 上排名前 2 位的 SOD 模型。

超类	Vege.	Food	Fruit	Tool	Nece.	Traf.	Cosm.	Ball	Inst.	Kitch.	Elec.	Anim.	Oth.	All
子类数量	4	5	9	11	12	10	4	7	14	9	9	49	17	160
ESMG [54]	.577	.635	.735	.625	.546	.673	.633	.559	.655	.631	.629	.687	.592	.635
CBCS [30]	.680	.621	.739	.617	.603	.666	.664	.619	.627	.625	.640	.672	.594	.637
CSHS [52]	.613	.591	.733	.677	.585	.691	.677	.563	.637	.651	.665	.715	.624	.656
CODR [122]	.682	.682	.774	.679	.634	.756	.678	.580	.671	.686	.695	.771	.638	.700
DIM‡ [59]	.622	.687	.773	.650	.604	.708	.633	.577	.665	.612	.641	.709	.623	.662
UMLF [73]	.781	.777	.781	.694	.779	.836	.714	.668	.711	.763	.748	.810	.690	.758
IML‡ [86]	.802	.725	.808	.740	.714	.867	.753	.653	.734	.795	.729	.855	.663	.773
CPD‡ [123]	.805	.763	.818	.734	.758	.894	.763	.629	.638	.848	.784	.892	.693	.791
EGNet‡ [119]	**.833**	.761	.815	.746	.767	.890	.769	.632	.654	.841	.771	**.893**	.697	.793
CSMG‡ [96]	.755	**.872**	**.854**	.722	.744	**.908**	.766	.778	.690	**.849**	.840	.885	.690	.804
CoEG-Net(Ours)‡	.802	.842	.840	**.811**	**.790**	.897	**.795**	**.780**	**.746**	.844	**.842**	.881	**.739**	**.825**

本文所知, 本文提供了规模最大且最全面的基准。为了进行比较, 本文直接在默认设置下运行公布的代码 (例如: CBCS [30]、ESMG [54]、RFPR [121]、CSHS [52]、SACS [58]、CODR [122]、UMLF [73]、DIM [59]、CPD [123] 和 EGNet [119]) 或使用作者提供的 CoSOD 预测结果 (例如, IML [86]、CODW [60]、GONet [76]、SP-MIL [64] 和 CSMG [96])。

5.2 定量比较

5.2.1 *iCoSeg* 数据集上的表现。

iCoSeg 数据集 [42] 最初用于图像协同分割, 但现在已广泛用于 CoSOD 任务。有趣的是, 如表 4所示, 两个 SOD 模型 (即 EGNet [119] 和 CPD [123]) 获得了最好的性能。CoSOD 方法 (例如 CODR [122]、IML [86] 和 CSMG [96]) 也获得了与顶级 SOD 模型 (即 EGNet [119] 和 CPD [123]) 非常接近的性能。本文的 *CoEG-Net* 在 E_ϕ, S_α 和 F_β 中获得最佳性能, 但是结果非常接近 SOD 方法, 即 EGNet [119]。一个可能的原因是 iCoSeg 数据集包含许多带有单个物体的图像, 这些图像很容易被 SOD 模型检测到。协同显著特征在 iCoSeg 数据集中不是重要角色, 这也表明 iCoSeg 数据集可能不适合在深度学习时代用于评估 CoSOD 方法。一些示例可以在图 12中找到。

5.2.2 *CoSal2015* 数据集上的表现。

表 4列出了 CoSal2015 数据集 [41] 的评估结果。一个有趣的观察结果是, 现有的显著物体检测方法, 例如, EGNet [119] 和 CPD [123] 获得了比大多数 CoSOD 方法具有更高的性能。这意味着某些性能最高的显著物体检测框架可能更适合于 CoSOD 任务的扩展。CoSOD 方法 CSMG [96] 在 E_ϕ (0.842) 和 F_β (0.784) 中达到了有竞争力的性能, 但在 S_α (0.774) 和 ϵ (0.130) 中得分更差。这表明现有的 CoSOD 方法无法很好地解决任务。本文的 *CoEG-Net* 可获得最佳结果, 远胜过 SOD 和 Co-SOD 基准。

5.2.3 *CoSOD3k* 数据集上的表现。

CoSOD3k 的总体结果显示在表 4中。不出所料, 本文的模型仍然可以达到最佳性能。为了更深入地了解每组性能, 本文在表 5中报告了模型在 13 个超类上的表现。本文观察到, 模型在包含复杂结构的真实场景, 比如其它 (例如: 婴儿床和铅笔盒)、乐器 (例如, 钢琴、吉他和大提琴等), 必需品 (例如: 玻璃水瓶)、工具 (例如: 斧子、钉子、链锯、等) 和球 (例如: 足球、网球等) 类别上获得更低的平均分。注意到, 几乎所有基于深度的模型 (例如, EGNet [119]、CPD [123], IML [86] 和 CSMG [96]) 的性能均优于传统方法 (CODR [122], CSHS [52]、CBCS [30]、和 ESMG [54]), 这证明了利用深度学习技术解决 CoSOD 问题具有潜在优势。另一个有趣的发现是, 边缘特征可以帮助改善结果图以获得精细的边界。举例来说, 传统 (CSHS [52]) 和深度学习模型 (例如 EGNet [119]) 的最佳方法都引入了边缘信息来辅助检测。最后, 本文的方法 *CoEG-Net* 获得最佳平均性能, 其 E_ϕ 为 0.825, 远高于第二好方法 CSMG [96] 的 0.804。此外, 所有方法在 CoSOD3k 上的性能

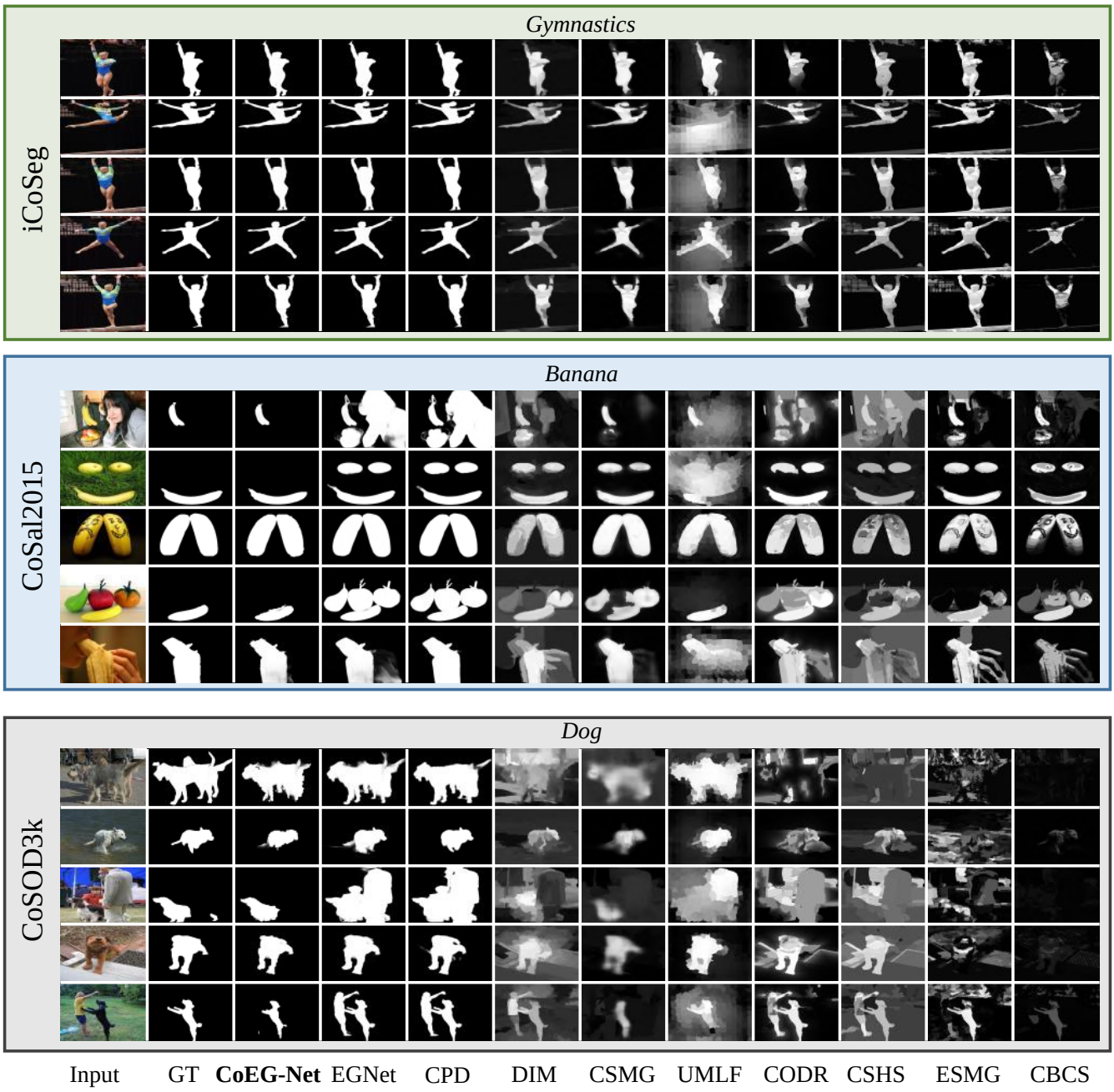

图 12. 在 iCoSeg [42]、CoSal2015 [41] 和本文的 CoSOD3k 数据集上评估的 10 个代表性模型的定性示例。

（表 4）都比其他两个数据集（iCoSeg 和 CoSal2015）差，这清楚地表明提出的 CoSOD3k 数据集具有挑战性，并为进一步的研究提供了充足的空间。

5.3 定性比较

图 12展示了在 iCoSeg、CoSal2015 和本文的 CoSOD3k 上的一些定性示例。可以看出，SOD 模型例如 EGNet [119] 和 CPD [123]，可检测所有显著物体并获得清晰的边界，其性能优于其他基准，但是这些 SOD 模型会忽略上下文信息。

例如，CoSal2015 数据集中的“香蕉”组包含其他几个不相关的物体，例如橘子、菠萝和苹果，此时 SOD 模型无法辨别出它们的区别。在本文的 CoSOD3k 的狗组中的图像也发生了另一种类似的情况，其中与狗一起检测到了人类（第三和第五张图像）。另一方面，CoSOD 方法，例如 CSMG [96] 和 DIM [59] 可以识别常见的显著物体并排除其他物体（例如人）。但是，这些 CoSOD 方法特别是在物体边界附近无法生成准确的预测结果。相比之下，本文的 *CoEG-Net* 保留了 SOD 和 CoSOD 方法的优势，并在所有数据集中获得了最佳的视觉效果。

5.4 与基准比较

本文的基线模型 *CoEG-Net* 由协同注意力预测和基本的 SOD 模型组成。为了探索协同注意力投影的效率，本文（1）采用相同的训练数据集（即 DUTS [31]）并在三个数据集（即

表 6
在三个基准数据集上对本文的模型进行了消融研究，其中 Ours-A，Ours-P，Ours-E 分别代表了 Amulet、PiCANet 和 EGNet 在本文基准上的协同显著预测结果。

数据集	指标	Amulet	**Ours-A**	PiCANet	**Ours-P**	EGNet	**Ours-E**
iCoSeg	$E_\phi\uparrow$	.877	**.878**	.906	**.907**	.911	**.912**
	$S_\alpha\uparrow$	.828	**.829**	.869	**.870**	.875	**.875**
	$F_\beta\uparrow$	.829	**.829**	.854	**.854**	.875	**.876**
	$\epsilon\downarrow$	.088	**.087**	.065	**.064**	.060	**.060**
CoSal2015	$E_\phi\uparrow$	.772	**.831**	.859	**.870**	.843	**.882**
	$S_\alpha\uparrow$	.719	**.744**	.801	**.825**	.818	**.836**
	$F_\beta\uparrow$	.684	**.758**	.799	**.818**	.786	**.832**
	$\epsilon\downarrow$	.147	**.125**	.090	**.084**	.099	**.077**
CoSOD3k	$E_\phi\uparrow$	.752	**.803**	.780	**.819**	.793	**.825**
	$S_\alpha\uparrow$	.685	**.692**	.750	**.758**	.762	**.762**
	$F_\beta\uparrow$	.629	**.700**	.682	**.724**	.702	**.736**
	$\epsilon\downarrow$	.145	**.122**	.137	**.095**	.119	**.092**

表 7
在 10 个最先进模型上的平均运行时间。

模型	CBCS [30]	ESMG [54]	CSHS [52]	CODR [122]
时间 (秒)	0.3	1.2	102	35
语言类型	Matlab	Matlab	Matlab	Matlab
Models	UMLF [73]	DIM‡ [59]	CSMG‡ [96]	CPD‡ [123]
时间 (秒)	87	25	3.2	0.016
语言类型	Matlab	Matlab	Caffe	PyTorch
Models	EGNet‡ [119]	Ours‡		
时间 (秒)	0.034	2.3		
语言类型	PyTorch	PyTorch		

iCoSeg、CoSal2015 和 CoSOD3k）测试了前沿显著性目标检测模型（即 Amulet [18]、PiCANet [93] 和 EGNet [119]）；（2）对这些模型应用相同的协同注意力投影策略，如章节4中所述，以进行此实验。表 6根据 E_ϕ、S_α、F_β 和 ϵ 指标显示三个基准的性能。根据结果，本文观察到：(i) 在相对简单的 iCoSeg 数据集上，本文的基准（即：Ours-A/-P/-E）相对于主干模型（即：Amulet、PiCANet 和 EGNet）提升微小。本文注意到，由于该数据集在每个组中包含大量具有相似外观的单个物体（图 12），因此仅使用 SOD 模型就可以实现非常高的性能。该结论与章节5.2.1中的分析一致；（ii）在经典的 CoSal2015 数据集上，本文的基准在四个指标上的表现始终优于主干网络。值得注意的是，对于这个更复杂的数据集，在 S_α 指标上仍然获得 2.5%、1.4%、和 1.8% 性能提高的结果；（iii）对于最具挑战性的数据集 CoSOD3k，本文发现改进仍然很显著（例如，相对 Amulet 方法 F_β 有 7.1% 提升）。为了进一步分析改进程度，本文还在补充材料中提供了 160 个子类的性能。本文观察到，对于常见的超类（比如："球"）的物体，例如"橄榄球"和"足球"，本文分别获得 23.5% 和 23.9% 的 F_β 的提升。本文将此归因于协同注意投影操作能够自动学习相互特征，这对于克服具有挑战性的歧义信息是至关重要的。

5.5 运行时间

本文的 *CoEG-Net* 是在 PyTorch 和 Caffe 中通过单张 RTX 2080Ti 显卡进行加速实现的。对于传统算法（CBCS [30]、ESMG [54]、CSHS [52]、CODR [122] 和 UMLF [73]），比较实验是基于配备 Intel (R) CoreTM i7-2600 CPU @ 3.4GHz 的笔记本电脑上执行的。其余的深度学习模型（DIM [59], CSMG [96], CPD [123], and EGNet [119]）在配有 Intel（R）Core（TM）i7-8700K CPU @ 3.70GHz 和 RTX 2080Ti GPU 的工作站上进行了测试。如表 7中所示，在排名前三的 CoSOD 模型中，即本文提出的 *CoEG-Net*、CSMG [96] 和 UMLF [73]，利用 E_ϕ 指标在本文提出的 CoSOD3k 数据集上进行评估，本文的模型实现了最快的推理时间。此外，与排名前 2 位最快的 CoSOD 模型（即 CBCS [30] 和 ESMG [96]）相比，尽管所提出的模型具有更长的测试时间，但它在 S_α 上的表现获得了显著提升。这部分表明本文的框架在 CoSOD 任务上效率高并且性能优越。但是，与最近发布的两个最新模型 CPD [123] 和 EGNet [119] 相比，运行时间方面仍有很大的改进空间。

6 讨论及未来方向

通过评估，本文观察到，在大多数情况下，当前的 SOD 方法（例如, EGNet [119] 和 CPD [123]）与 CoSOD 方法相比，可以获得非常有竞争力甚至更好的性能（例如, CSMG [96] 和 SP-MIL [64]）。但是，这并不一定意味着当前的数据集不够复杂或者直接使用 SOD 方法可以获得良好的性能。也就是说，SOD 方法在 CoSOD 数据集上的的性能实际要低于在 SOD 数据集上的性能。例如，EGNet 在 HKU-IS 数据集 [114] 和 ECSSD 数据集 [117] 上分别获得 0.937 和 0.943 的 F_β 分数。但是，它仅分别在 CoSal2015 和 CoSOD3k 数据集上获得 0.786 和 0.702 的 F_β 分数。因此，评估结果表明，CoSOD 中的许多问题仍未充分研究，这更造成现有 CoSOD 模型有效性的降低。在本节中，本文讨论了四个重要问题（比如：可扩展性、稳定性、兼容性和度量标准），这些问题尚未被现有的协同显著物体检测方法完全解决，并且应在未来进一步研究。最后，本文讨论了所提出的 *CoEG-Net* 框架的弱点。

• **可扩展性：** 可扩展性是设计 CoSOD 算法时需要考虑的最重要问题之一。具体来说，它表示 CoSOD 模型处理大规模图像场景的能力。据本文所知，CoSOD 任务的一个关键特性是该模型需要考虑每组中的多张图像。但实际上，一个图像组可能包含大量相关图像。在这种情况下，不考虑可扩展性的方法将具有巨大的计算成本，并且运行时间非常长，因此在实践中是不可接受的（例如：CSHS-102 和 UMLF-87）。因此，如何解决可扩展性问题，或如何减少由图像组中包含的图像数量引起的计算复杂度，成为该领域的关键问题，尤其是在将 CoSOD 方法应用于实际应用中时。

• **稳定性：** 另一个重要问题是模型的稳定性。在处理包含多张图像的图像组时，某些现有方法（例如，HCNco [129]、PCSD [49] 和 IPCS [33]）将图像组分为图像对或图像子组（例如，GD [68]）。另一类方法采用基于 RNN 的模型（例如，GWD [100]），这涉及为输入图像分配顺序。因为没有划分图像组或为相关图像分配输入顺序的原则方法，使得这些策略在整个训练过程中变得不稳定。换句话说，当按照不同的策略生成图像子组或分配输入顺序时，学习过程会产生不同的协同显著性检测器，并且测试结果也不稳定。因此，这不仅给评估所学习的协同显著性检测器的性能带来了困难，而且还影响了协同显著性物体检测的应用。

• **兼容性：** 在 CoSOD 中引入 SOD 是建立 CoSOD 框架的直接而有效的策略，因为单张图像的显著性有助于协同显著性信息的识别。但是，大多数现有的 CoSOD 工作仅将 SOD 模型的结果或特征作为有用的信息线索。本文提出的 *CoEG-Net* 基准仍遵循此两阶段框架，该框架比单个 SOD 模型花费更多的推理时间。尽管是初步尝试，但本文方案依然获得了现有 CoSOD 模型中最佳的性能。从这个角度出发，另一个利用 SOD 技术的方向是将基于卷积神经网络的 SOD 模型与 CoSOD 模型进行深度组合，以构建直接检测 CoSOD 的端到端可训练框架。为了实现这一目标，需要考虑 CoSOD 框架的兼容性，以便于集成现有的 SOD 技术。

• **新颖的指标：** 当前针对 CoSOD 的评估指标是根据 SOD 设计的，即它们直接计算每个组 SOD 得分的平均值。与 SOD 相比，CoSOD 涉及不同图像的协同显著物体之间的关系信息，这对于 CoSOD 评估更为重要。例如，当前的 CoSOD 指标假定目标物体在所有图像中具有相似的大小。由于物体在不同的图像中实际上具有不同的大小，因此这些度量（章节5中的 S_α、E_ϕ、F_β 和 ϵ）可能对检测出的大物体会更友好。此外，当前的 CoSOD 指标是基于检测单张图像中的物体，而不是识别多张图像中协同的物体。因此，如何为 CoSOD 设计合适的度量标准成为一个悬而未决的问题。

• **弱点：** 端到端 CoSOD 检测框架的输出是具有平滑精细结构的二值化预测图，相比之下，*CoEG-Net* 的预测结果受到影响使得边界粗糙，这表明 *CoEG-Net* 无法很好地保留协同显著物体的形状细节信息。图 13展示了一些检测失败的结果。

• **潜在应用：** 在这一部分，本文讨论如何从高质量 CoSOD 模型中受益的两个潜在的新应用。有关更多 CoSOD 的应用，请参考 [19], [35] 中的相关研究。

集合导向裁剪。此应用是从 Jacob 等人 [24] 的工作中派生出来的。这一工作主要探究人们在比较图像时人的注意力位置的问题，是 CoSOD 任务的源头。秉承相同的思想，本文展示了一种不仅限于图像对的更通用的潜在应用。例如，在处理图 14中的自动缩略图任务时，本文首先从 *CoEG-Net* 生成的显著性图中获得黄色边界框。之后，将使用放大的（约为 60 像素）红色框[3]用于自动识别裁剪区域。为了（从第一行的图片中）获得高质量的裁剪，还可以引入现有最先进的超分辨率技术 [130], [131] 以进一步提高视觉效果。

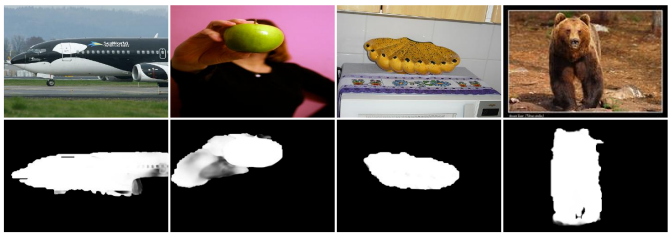

图 13. 本文的 *CoEG-Net* 的一些挑战案例。

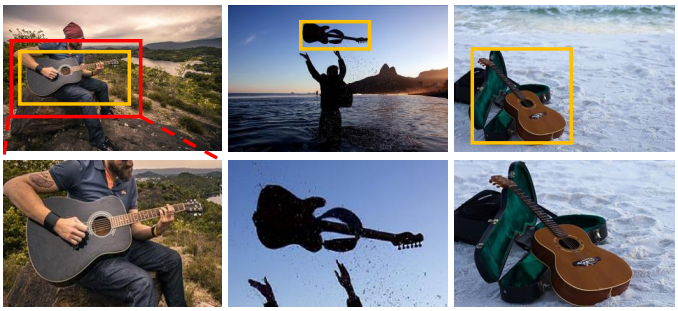

图 14. 集合导向裁剪。

物体协同定位。如 DeepCO^3 [98] 所示，协同显著性检测结果将为物体协同定位任务提供类别无关的注意力线索。将 *CoEG-Net* 引入现有的商务应用将是提高该领域性能的一种可能的解决方案。

7 结论

在这项研究中，本文对协同物体检测（CoSOD）任务进行了全面的研究。在确定当前数据集中存在严重的数据偏差（假定每个图像组包含外观相似的显著物体）后，本文建立了一个新的高质量数据集，名为 CoSOD3k，其中包含在语义或概念层面上相似的协同显著物体。值得注意的是，CoSOD3k 是迄今为止最具挑战性的 CoSOD 数据集，含有 160 组图像，合计 3,316 张图像，这些图像分别标记有类别、边框、物体级别和实例级别的注释。本文的 CoSOD3k 数据集在多样性、难度和可扩展性方面取得了重大飞跃，一些相关的视觉任务，例如协同分割、弱监督定位、实例级检测以及相关未来发展都会受益。

为了创建高效的协同显著性物体检测器，本文集成了现有的显著目标检测技术，以构建统一且可训练的 CoSOD 框架——*CoEG-Net*。具体来说，本文使用协同注意投影策略增强了先前的 EGNet 模型，从而有效地学习协同信息，并提高了协同物体检测检测框架的可扩展性和稳定性。

3. 注意，当放大的红色框触及图像边界时，本文将保留黄色框的原始宽度。

此外，本文还总结了 40 种最先进的算法，在两个经典数据集以及本文提出的 CoSOD3k 数据集上对其中的 18 个算法进行了基准测评，从而展现了全面的调研。通过评估最新的 SOD 和 CoSOD 方法，本文出乎意料地发现了 SOD 方法会表现的更好，这一有趣发现可以为进一步探索更好 CoSOD 算法提供思路。本文希望这项工作中提出的研究可以进一步促进 CoSOD 研究社区的发展。将来，本文计划增加数据集规模以激发更多新颖的想法。

参考文献

[1] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a Deeper Look at the Co-salient Object Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[2] 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 T. Pattern Anal. Mach. Intell.*, 2021.

[3] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.*, 2018, pp. 186–202.

[4] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 7234–7243.

[5] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE T. Geosci. Remote. Sens. Lett.*, vol. 57, no. 11, pp. 9156–9166, 2019.

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

[7] 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.

[8] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3927–3936.

[9] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Re-thinking RGB-D Salient Object Detection: Models, Datasets, and Large-Scale Benchmarks," *IEEE T. Neural Netw. Learn. Syst.*, 2020.

[10] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *arXiv preprint arXiv:2009.03075*, 2020.

[11] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *arXiv preprint arXiv:2008.12134*, 2020.

[12] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Comput. Vis. Media*, pp. 1–33, 2021.

[13] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8554–8564.

[14] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video Saliency Detection via Sparsity-Based Reconstruction and Propagation," *IEEE T. Image Process.*, vol. 28, no. 10, pp. 4819–4831, 2019.

[15] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE T. Pattern Anal. Mach. Intell.*, 2020.

[16] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007.

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

[18] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

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

[20] 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," *arXiv preprint arXiv:1904.09146*, 2019.

[21] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.

[22] H. Bi, K. Wang, D. Lu, C. Wu, W. Wang, and L. Yang, "C2net: a complementary co-saliency detection network," *The Vis. Comput.*, pp. 1–13, 2020.

[23] J. Ren, Z. Liu, G. Li, X. Zhou, C. Bai, and G. Sun, "Co-saliency detection using collaborative feature extraction and high-to-low feature integration," in *Int. Conf. Multimedia and Expo*, 2020, pp. 1–6.

[24] 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.

[25] W. Wang and J. Shen, "Higher-order image co-segmentation," *IEEE T. Multimedia*, vol. 18, no. 6, pp. 1011–1021, 2016.

[26] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-Based Multiple Foreground Video Co-Segmentation via Multi-State Selection Graph," *IEEE T. Image Process.*, vol. 24, no. 11, pp. 3415–3424, 2015.

[27] 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.

[28] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *IEEE Int. Conf. Inf. Sci. Cloud Comput. Companion*, 2013, pp. 728–733.

[29] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: group saliency in image collections," *Vis. Comput.*, vol. 30, no. 4, pp. 443–453, 2014.

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

[31] 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.

[32] H. Yu, K. Zheng, J. Fang, H. Guo, W. Feng, and S. Wang, "Co-saliency detection within a single image," in *AAAI Conf. Art. Intell.*, 2018, pp. 7509–7516.

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

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

[35] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans Intell Syst Technol*, vol. 9, no. 4, pp. 1–31, 2018.

[36] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1521–1528.

[37] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.

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

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

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

[41] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2994–3002.

[42] 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.*, 2010, pp. 3169–3176.

[43] 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, no. 2, pp. 303–338, 2010.

[44] J. Dai, Y. Nian Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Int. Conf. Comput. Vis.*, 2013, pp. 1305–1312.

[45] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.

[46] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global Contrast based Salient Region Detection," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.

[47] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[48] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Adv. Neural Inform. Process. Syst.*, 2011, pp. 109–117.

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

[50] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Adv. Neural Inform. Process. Syst.*, 2009, pp. 681–688.

[51] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE T. Multimedia*, vol. 15, no. 8, pp. 1896–1909, 2013.

[52] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, 2013.

[53] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2010.

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

[55] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, and J. Luo, "Efficient manifold ranking for image retrieval," in *ACM Spec. Interest Group Inf. Ret.*, 2011, pp. 525–534.

[56] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE TSMC*, vol. 9, no. 1, pp. 62–66, 1979.

[57] X. Cao, Y. Cheng, Z. Tao, and H. Fu, "Co-saliency detection via base reconstruction," in *ACM Int. Conf. Multimedia*, 2014, pp. 997–1000.

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

[59] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE T. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, 2015.

[60] 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.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[62] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Int. Conf. Learn. Represent.*, 2014.

[63] Y. Bengio *et al.*, "Learning deep architectures for ai," *FTML*, vol. 2, no. 1, pp. 1–127, 2009.

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

[65] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Int. Conf. Comput. Vis.*, 2015, pp. 594–602.

[66] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3238–3245.

[67] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Brit. Mach. Vis. Conf.*, 2014.

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

[69] 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.

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

[71] 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 T. Image Process.*, vol. 26, no. 7, pp. 3196–3209, 2017.

[72] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

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

[74] M. Li, S. Dong, K. Zhang, Z. Gao, X. Wu, H. Zhang, G. Yang, and S. Li, "Deep learning intra-image and inter-images features for co-saliency detection," in *Brit. Mach. Vis. Conf.*, 2018, p. 291.

[75] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[76] 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.

[77] 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.

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

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

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

[81] 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 T. Image Process.*, vol. 28, no. 1, pp. 56–71, 2018.

[82] D.-j. Jeong, I. Hwang, and N. I. Cho, "Co-salient object detection based on deep saliency networks and seed propagation over an integrated graph," *IEEE T. Image Process.*, vol. 27, no. 12, pp. 5866–5879, 2018.

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

[84] S. Song, H. Yu, Z. Miao, D. Guo, W. Ke, C. Ma, and S. Wang, "An easy-to-hard learning strategy for within-image co-saliency detection," *Neurocomputing*, vol. 358, pp. 166–176, 2019.

[85] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[86] J. Ren, Z. Liu, X. Zhou, C. Bai, and G. Sun, "Co-saliency Detection via Integration of Multi-layer Convolutional Features and Inter-image Propagation," *Neurocomputing*, vol. 371, pp. 137–146, 2020.

[87] 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 T. Image Process.*, vol. 28, no. 10, pp. 5052–5063, 2019.

[88] B. Li, Z. Sun, L. Tang, Y. Sun, and J. Shi, "Detecting Robust Co-Saliency with Recurrent Co-Attention Neural Network," in *Int. Jt. Conf. Artif. Intell.*, 2019, pp. 818–825.

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

[90] 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.

[91] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Int. Conf. Learn. Represent.*, 2017.

[92] 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.

[93] 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.

[94] B. Li, Z. Sun, Q. Wang, and Q. Li, "Co-saliency Detection Based on Hierarchical Consistency," in *ACM Int. Conf. Multimedia*, 2019, pp. 1392–1400.

[95] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Int. Conf. Learn. Represent.*, 2014.

[96] 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.

[97] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to Detect A Salient Object," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[98] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "DeepCO3: Deep Instance Co-Segmentation by Co-Peak Search and Co-Saliency Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8846–8855.

[99] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Int. Conf. Comput. Vis.*, 2017, pp. 4048–4056.

[100] 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.

[101] G. Gao, W. Zhao, Q. Liu, and Y. Wang, "Co-saliency detection with co-attention fully convolutional network," *IEEE T. Circuit Syst. Video Technol.*, 2020.

[102] Z.-J. Zha, C. Wang, D. Liu, H. Xie, and Y. Zhang, "Robust deep co-saliency detection with group semantic and pyramid attention," *IEEE T. Neural Netw. Learn. Syst.*, 2020.

[103] B. Jiang, X. Jiang, J. Tang, and B. Luo, "Co-saliency detection via a general optimization model and adaptive graph learning," *IEEE T. Multimedia*, 2020.

[104] 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.

[105] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, "Gradient-induced co-saliency detection," in *Eur. Conf. Comput. Vis.*, 2020.

[106] J. Zhao, R. Bo, Q. Hou, M.-M. Cheng, and P. Rosin, "Flic: Fast linear iterative clustering with active search," *Comput. Vis. Media*, vol. 4, no. 4, pp. 333–348, 2018.

[107] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[108] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[109] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 909–918.

[110] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged Object Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[111] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[112] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang, "Joint Salient Object Detection and Existence Prediction," *Front. Comput. Sci.*, pp. 778—788, 2017.

[113] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 247–256.

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

[115] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Int. Conf. Comput. Vis.*, 2001, pp. 416–423.

[116] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4142–4150.

[117] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1155–1162.

[118] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann, "The discrimination of visual number," *Am. J. Psychol.*, vol. 62, no. 4, pp. 498–525, 1949.

[119] J.-X. Zhao, J.-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. 8779–8788.

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

[121] L. Li, Z. Liu, W. Zou, X. Zhang, and O. Le Meur, "Co-saliency detection based on region-level fusion and pixel-level refinement," in *Int. Conf. Multimedia and Expo*, 2014, pp. 1–6.

[122] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2073–2077, 2015.

[123] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3907–3916.

[124] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognit.*, vol. 88, pp. 113–126, 2019.

[125] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philos Mag (Abingdon)*, vol. 2, no. 11, pp. 559–572, 1901.

[126] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Adv. Neural Inform. Process. Syst.*, 2004, pp. 321–328.

[127] 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.

[128] 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. Jt. Conf. Artif. Intell.*, 2018, pp. 698–704.

[129] J. Lou, F. Xu, Q. Xia, W. Yang, and M. Ren, "Hierarchical co-salient object detection via color names," in *IEEE Asian Conf. Pattern Recog.*, 2017, pp. 718–724.

[130] Y. Bahat and T. Michaeli, "Explorable super resolution," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2716–2725.

[131] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *Eur. Conf. Comput. Vis.*, 2020, pp. 715–732.