

# 基于复杂背景语义建模的协同显著目标检测

王子泰<sup>1)</sup> 许倩倩<sup>1)</sup> 曹宇辰<sup>2),3)</sup> 柳洋<sup>4)</sup> 黄庆明<sup>1),4),5)</sup>

<sup>1)</sup>(中国科学院计算技术研究所智能算法安全全国重点实验室 北京 100190)

<sup>2)</sup>(中国科学院信息工程研究所 北京 100085)

<sup>3)</sup>(中国科学院大学网络空间安全学院 北京 101408)

<sup>4)</sup>(中国科学院大学计算机科学与技术学院 北京 101408)

<sup>5)</sup>(中国科学院大学大数据挖掘与知识管理重点实验室 北京 101408)

**摘要** 协同显著目标检测旨在从一组图像中识别出共同的显著目标。现有研究大多提取图像间的共享表示，从而挖掘协同显著区域。近年来，部分方法开始关注背景区域的建模，但通常直接沿用前景模块，忽视了背景语义的复杂性。具体而言，一方面，一些背景区域可能在外观上与协同显著目标高度相似，即存在语义模糊性，导致特征难以区分；另一方面，不同图像的背景区域往往差异显著，即存在语义异质性，导致捕捉图像间共性模式的前景模块效果不佳。为解决上述问题，本文提出了一种面向背景复杂语义的协同显著目标检测新方法，其中两个子模块分别关注模糊性与异质性问题：模糊背景检索模块(Ambiguous Background Retrieval, ABR)和异质背景检索模块(Heterogeneous Background Retrieval, HBR)。通过显式建模背景表示与协同前景表示之间的关系，所提方法能够有效剔除复杂背景区域，从而提升协同显著图的精度。最后，在CoCA、CoSOD3k和CoSal2015三个主流基准数据集上开展了实验。结果表明，在背景复杂数据集上，所提方法较基线模型在结构指标S-measure上提升了2.7%，在平均E值检测指标mean E-measure上提升了3%，优于现有最优方法。

**关键词** 协同显著性检测；背景建模；全局池化

中图法分类号 TP18 DOI号 10.11897/SP.J.1016.2026.00610

## Co-Salient Object Detection Via Complex Background Semantic Modeling

WANG Zi-Tai<sup>1)</sup> XU Qian-Qian<sup>1)</sup> CAO Yu-Chen<sup>2),3)</sup> LIU Yang<sup>4)</sup> HUANG Qing-Ming<sup>1),4),5)</sup>

<sup>1)</sup>(State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

<sup>2)</sup>(Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085)

<sup>3)</sup>(School of Cyber Security, University of Chinese Academy of Sciences, Beijing 101408)

<sup>4)</sup>(School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408)

<sup>5)</sup>(Key Laboratory of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing 101408)

**Abstract** Co-salient object detection aims to identify common salient objects from a group of related images. Most prior arts focus on extracting shared representations across images to localize co-salient

收稿日期：2025-06-25；在线发布日期：2025-11-11。本课题得到国家自然科学基金(62525212,62236008,62441232,U21B2038,U23B2051,62502500)、中国科学院青年促进会优秀会员项目、中国科学院战略性先导科技专项(XDB0680201)、博士后创新人才支持计划(国家资助博士后研究人员计划A档)(BX20240384)、北京市自然科学基金(L252144)、中国博士后科学基金面上资助(2025M771558)资助。王子泰，博士，助理研究员，中国计算机学会(CCF)会员，主要研究方向为机器学习与数据挖掘。E-mail: wangzitai@ict.ac.cn。许倩倩(通信作者)，博士，研究员，中国计算机学会(CCF)高级会员，主要研究方向为统计机器学习及其在多媒体和计算机视觉领域的应用。E-mail: xuqianqian@ict.ac.cn。曹宇辰，硕士，主要研究方向为计算机视觉。柳洋，博士研究生，中国计算机学会(CCF)学生会会员，主要研究方向为计算机视觉和自监督学习。黄庆明(通信作者)，博士，讲席教授，中国计算机学会(CCF)会士，主要研究领域为多媒体计算、图像处理、计算机视觉、模式识别。E-mail: qmhuang@ucas.ac.cn。

regions. In recent years, some methods attempt to explore backgrounds. However, they often directly reuse foreground modules, overlooking the semantic complexity of background regions. Specifically, some background areas may appear visually similar to co-salient objects, i.e., semantic ambiguity, making it difficult to discriminate features. Meanwhile, background areas across different images often vary significantly, i.e., semantic heterogeneity, which compromises the effectiveness of foreground modules designed to capture inter-image commonalities. To address these issues, we propose a novel co-salient object detection method tailored to complex background semantics. The method introduces two dedicated modules: the Ambiguous Background Retrieval (ABR) module and the Heterogeneous Background Retrieval (HBR) module, respectively. By explicitly modeling the relationship between background representations and co-salient representations, the proposed method effectively filters out complex background areas, thereby improving the accuracy of co-saliency maps. Extensive experiments are conducted on three widely-used benchmark datasets: CoCA, CoSOD3k, and CoSal2015. Results show that, on datasets with complex backgrounds, the proposed method improves the structural similarity metric by 2.7% and the mean enhanced alignment metric by 3.0% over the baseline, outperforming existing state-of-the-art approaches.

**Key words** co-salient object detection; background modeling; global pooling

## 1 引言

协同显著目标检测(Co-salient Object Detection, CoSOD)旨在从一组图像中分割出共同存在的显著目标。该任务在视频目标共定位以及图像匹配等下游任务中具有重要应用价值。与传统的显著性目标检测(SOD)仅处理单张图像不同,协同显著目标需满足图像间一致性(inter-image commonality)和图像内显著性(intra-image saliency)两方面要求。由于图像背景区域往往包含一些具有显著性但不具备共性的物体(例如酒杯),这类物品与协同显著目标(例如酒瓶)频繁共现,难以被准确剔除。

现有研究主要聚焦于挖掘图像间的共性表示(co-representation)以提升协同目标的检测性能。例如,作为该类方法代表性工作,全局池化(global pooling)策略通过聚合图像中潜在显著区域的特征来提取共性表示。典型方法如 CORP<sup>[1]</sup>提出在多次迭代中逐步剔除与共性偏离的噪声区域,从而优化协同显著特征。尽管该方法可有效建模协同目标之间的相关性,但在处理协同目标与背景之间的非相关性时仍存在不足。为此,DMT<sup>[2]</sup>提出了一种显式背景区域建模方法。然而,该方法采用单一模块同时处理前景与背景,未能充分考虑背景区域在语义上的异质性与模糊性差异。如图1所示,第一幅图像中的黄色沙发与第二幅图像中的黄色玩具鸭在颜色上极为相似(语义模糊性),尽管它们出现在不同图像中。该语义模糊性使得前景与背景在表征空

间中难以区分。此外,多幅图像之间背景差异巨大(语义异质性),导致图像间共性提取模块难以有效聚焦背景区域。

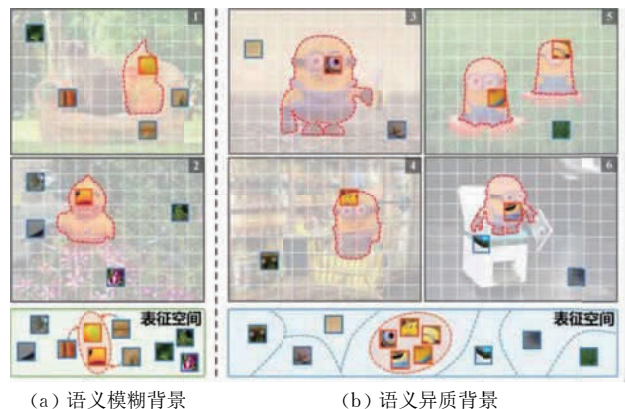


图1 复杂背景中存在的两种有害属性((a)语义模糊背景:某些背景中的物体(如黄色沙发)在外观模式上与显著目标(如黄色玩具鸭)相似,尽管它们出现在不同图像中;(b)语义异质背景:不同图像之间的背景区域存在显著差异。)

为解决上述问题,本文提出一种面向背景复杂语义建模的协同显著目标检测方法,命名为模糊与异质背景挖掘(Ambiguous and Heterogeneous Background Exploration, AHBE)。该方法基于全局池化机制,额外引入两个专用子模块,分别用于处理背景语义模糊性(Ambiguous Background Retrieval, ABR)与语义异质性(Heterogeneous Background Retrieval, HBR)。具体而言,所提方法首先计算协同目标与背景区域的代理特征表示,并以此为依据分别检索出与前景特征最相似或最差异的背景区域。随后,通

过解码器对相关图进行重建,实现对协同显著图的迭代优化,从而逐步剔除具有复杂语义的背景区域。

本文在 CoSal2015、CoSOD3k 和 CoCA 三个主流数据集上开展了广泛实验,实验结果验证了所提出框架在处理背景复杂场景下的有效性与适用性。例如,在背景复杂数据集上,所提方法较基线模型在结构指标 S-measure 上提升了 2.7%,在平均 E 值检测指标 mean E-measure 上提升了 3%,优于现有最优方法。

综上所述,本文贡献主要包括:

(1)首次系统性探讨了协同显著目标检测中背景区域所具有的两类复杂语义,即语义模糊性与语义异质性。

(2)设计了两个用于建模背景复杂语义的子模块:模糊背景检索模块(ABR)与异质背景检索模块(HBR),从而更有效地建模协同目标与背景之间的非相关性。

(3)在三个主流基准数据集上开展的一系列实验,证实了所提出方法对于建模背景区域中复杂语义的有效性。

## 2 相关工作

### 2.1 显著目标检测

显著性目标检测旨在从图像中识别最能吸引人类注意力的区域,其起源与发展可参见综述[3-6]。在该领域,早期方法大多基于手工设计的特征<sup>[7-11]</sup>。随着深度学习的发展,研究者开始采用卷积神经网络提取显著区域的高级语义特征<sup>[12-15]</sup>。考虑到上下文信息在显著性检测中的重要性,近年来也涌现出基于 Transformer 的多尺度  $N$  建模方法<sup>[16-19]</sup>,有关最新进展可参考综述[20]。此外,显著性目标检测衍生出了多个子任务,例如多目标显著性检测<sup>[21]</sup>以及 RGB-D 显著性检测<sup>[22]</sup>,以及本文关注的协同显著目标检测<sup>[23]</sup>。

### 2.2 协同显著目标检测

协同显著目标检测旨在从一组图像中检测出共同的显著目标。早期方法多采用聚类策略建模图像之间的关系<sup>[24]</sup>,或通过轮廓等启发式特征进行协同区域的提取<sup>[25]</sup>。近年来研究大多基于神经网络挖掘图像组内的协同信息与特征间关系<sup>[26-28]</sup>。部分方法通过深入挖掘协同目标的显著性线索,在性能上取得了显著进展<sup>[1-2,23,29-33]</sup>。例如,文献[32]针对复

杂图像中多显著目标共存设计一系列专用模块,能够有效区分属于不同组别的目标,从而提升协同区域的特征表示能力。文献[33]在此基础上进一步提出精细化模块设计,增强模型在跨图像组中识别共同目标的能力,从而在协同显著目标检测的准确性、一致性和计算效率方面均带来了明显提升。此外,文献[34]采用两阶段自监督框架,先通过局部 ViT 特征对应与自适应阈值获得初步分割,再利用区域级特征一致性精炼得到高质量共显著图。尽管已有研究如 DMT<sup>[2]</sup>与 CORP<sup>[1]</sup>考虑了背景区域在显著图中的干扰问题,但它们仍主要集中于协同特征的挖掘,缺乏对复杂背景进行精细化建模的能力。

近年来,视觉语言大模型被尝试应用于协同显著目标检测中,例如 Zero-Shot Co-SOD<sup>[35]</sup>利用大模型的零样本泛化能力在无训练条件下完成检测;VCP<sup>[36]</sup>通过视觉提示调优机制,引导冻结模型聚焦于图像组的共显著区域;ConceptCoSOD<sup>[37]</sup>则引入文本概念作为语义提示,提升共显著对象识别的鲁棒性。然而,在协同显著目标检测这种底层视觉任务中,以 VGG、ResNet 为主干网络的传统方法相较于大模型依然具备显著效率优势。具体而言,该任务的应用场景,例如监控等高通量场景,通常要求覆盖范围广、数据处理量大,传统方法因此在硬件资源消耗与部署成本方面更具优势。

## 3 面向复杂背景语义的协同显著性检测

### 3.1 预备知识

给定张输入图像  $\mathcal{I} = \{\mathbf{I}_n \in \mathbb{R}^{3 \times H \times W}\}_{n=1}^N$ , 协同显著目标检测旨在输出  $N$  个共显著性图  $\mathcal{M} = \{\mathbf{M}_n \in [0,1]^{H \times W}\}_{n=1}^N$ , 其中 1 和 0 分别表示协同显著区域和背景。为此,该领域通常首先使用预训练编码器,例如 VGG-16<sup>[38]</sup>,抽取输入图像的高阶语义  $\mathcal{F} = \{\mathcal{F}_n \in \mathbb{R}^{D \times H \times W}\}_{n=1}^N$ <sup>[1-2,32]</sup>。接着,使用显著性检测头生成初始的粗粒度显著性图  $\mathcal{M}^0 = \{\mathbf{M}_n^0\}_{n=1}^N$ <sup>[1,23,39]</sup>。

给定高阶语义  $\mathcal{F}$  和当前的共显著性图  $\mathcal{M}^{t-1}$ , 以下介绍现有方法如何使用全局池化技术提取图片之间共性,从而优化共显著性图。如图 2 中共显著性分支所示,给定潜在显著区域的特征,即  $\mathbf{M}_n^{t-1} \odot \mathbf{F}_n$ , 首先在每张图像中执行全局平均池化操作,之后进行跨图像平均,从而得到共显著性区域表示的代理:

$$cp^t = \text{AVG}(\text{GAP}(\mathbf{M}_n^{t-1} \odot \mathbf{F}_n)) \in \mathbb{R}^D \quad (1)$$

其中,  $GAP(\cdot) = \frac{1}{HW} \sum_{p=1}^W \sum_{q=1}^H (\cdot)$  表示全局平均池化操作,  $AVG(\cdot)$  为图像之间的均值操作。随后,  $topk$  操作保留与该代理最相似的  $K$  个特征:

$$Co^t = \text{argtopk}(\text{Sim}(cp^t, e); K) \in \mathbb{R}^{K \times D} \quad (2)$$

其中,  $\mathcal{E} = \{e \in \mathcal{E}_{F_n}^e | F_n \in \mathcal{F}\}$  表示所有图像中图像块级特征的集合,  $\text{Sim}(\cdot, \cdot)$  为预定义的相似度函数, 如内积。之后, 将  $Co^t$  作为共显著性模式, 与特征进行相关性计算以搜索共显著区域:

$$R_n^t = \text{Corr}(Co^t, F_n) \in \mathbb{R}^{K \times H \times W'} \quad (3)$$

为了提取背景区域的相关性图(correlation map), 即  $B_n^t$ , 已有方法通常采用与上文相同的流程, 只不过将输入换为  $\tilde{M}_n^{t-1} = 1 - M_n^{t-1}$ 。然后将  $R_n^t$  与  $B_n^t$  一同送入解码器以生成更新后的共显著性图  $M_n^t$ 。然而, 该策略在处理复杂背景时存在如下问题:

一方面, 式(3)基于的前提是假设  $Co^t$  能完美地估计共显著性模式。但由于存在歧义性, 某些背景区域难免会与  $Co^t$  表现出相似模式, 尤其在  $K$  设置过大时更为显著。因此, 相关性图  $R_n^t$  可能在歧义背景区域上也产生强响应, 进而导致假阳性预测。

另一方面, 式(1)旨在提取图像间共性, 无法捕捉图像的个体特征。因此, 当图像间的背景表现出高度

异质性时, 相关性图  $B_n^t$  将难以有效地定位背景区域。随着现实数据集中背景复杂度的提升, 这一问题将愈发严重。

鉴于上述问题, 下一节将详细说明所提模块如何更好地建模复杂背景区域。

## 3.2 面向复杂背景语义的协同显著性检测框架

### 3.2.1 框架总览

如图2所示, 为了精细化共显著性图, 将高层特征  $\mathcal{F}$  和当前共显著性图  $\mathcal{M}^t$  分别送入两个不同的分支中进行迭代处理。在每一次迭代中, 共显著性分支采用全局池化技术以挖掘特征间的共性, 并获得共表示代理  $cp^t$ 。随后, 纯净共表示搜索模块<sup>[1]</sup>被用于从特征集  $\mathcal{F}$  中搜索与代理最相似的  $K$  个图像块级特征, 即最符合代理的特征, 从而得到纯净共表示特征  $Co^t \in \mathbb{R}^{K \times D}$ 。在背景分支中, 本文提出了两个模块, 分别处理模糊背景与异质背景, 核心思想是基于共表示代理  $cp^t$  与背景代理  $g^t \in \mathbb{R}^{N \times D}$ , 检索出最相似/最不同的图像块级特征。最终, 这两个分支的输出将用于生成相关性图, 即  $\{Z_n^t\}_{n=1}^N$ , 并通过解码器解码, 获得精细化的共显著性图  $\mathcal{M}^t$ 。经过  $T$  次迭代后, 最终将  $\mathcal{M}^T$  作为共显著性预测结果。

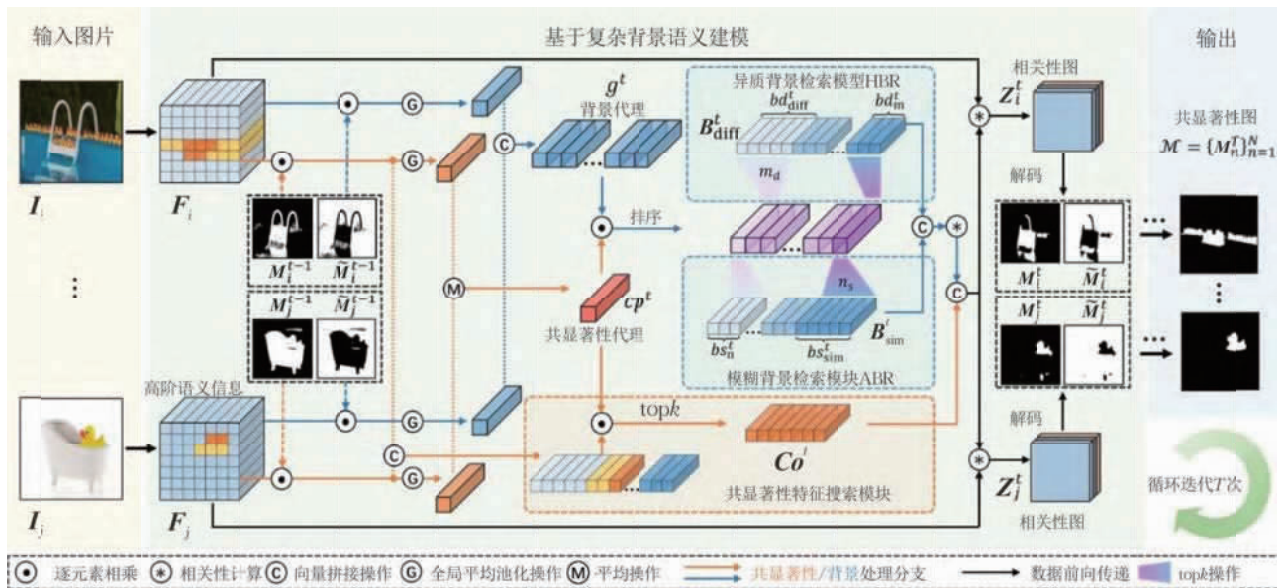


图2 本文提出的“模糊与异质背景挖掘”(AHBE)共显著性目标检测模型的框架图(在共显著性分支中, 使用基于共表示代理提取纯净的共表示特征。在背景分支中, 设计了两个模块, 即模糊背景检索和异质背景检索, 用于捕捉背景区域中的复杂上下文信息。最终, 两个分支的输出被用于计算相关性图, 并进一步通过解码器获得精细的共显著性图)

接下来, 将详细介绍所提出的 ABR 与 HBR 模块如何利用这些复杂上下文以更好地定位具有复杂语义的背景区域。

### 3.2.2 模糊背景检索模块

尽管不同图像的背景可能差异较大, 但是同一张图像背景的不同区域往往存在相似性。基于此,

本文首先利用全局池化技术提取背景特征。不同于共显著性代理  $cp^t$ ，背景代理不进行跨图像平均操作，从而保留复杂的上下文信息：

$$g_n^t = \text{GAP}(\tilde{\mathbf{M}}_n^{t-1} \odot \mathbf{F}_n) \in \mathbb{R}^D, n = 1, 2, \dots, N \quad (4)$$

如前所述，模糊性指的是背景模式与共显著性模式高度相似。为发现此类模式，检索与共表示代理  $cp^t$  最相似的背景代理：

$$n_s = \arg \max_{n \in \{1, 2, \dots, N\}} \text{Sim}(cp^t, g_n^t) \quad (5)$$

其中，为简洁起见，省略上标  $t$ 。接着，使用该代理从图像中检索出共计  $Q = Q_1 + (N-1)Q_2$  个潜在的模糊背景图像块特征，其中  $Q_1$  和  $Q_2$  分别表示从索引为  $n_s$  的图像及其余图像中检索的图像块数目，且  $Q_1 > Q_2$ 。具体而言：

$$q^* = \text{argtopk}_q(\text{Sim}(g_{n_s}^t, e_q); Q_1),$$

$$bs_{\text{sim}}^t = \mathbf{F}_{n_s}(q^*) \in \mathbb{R}^{Q_1 \times D} \quad (6)$$

其中， $e_q \in \mathbb{R}^D$ 、 $\mathbf{F}_{n_s}(q)$  表示图像  $n_s$  中第  $q$  个图像块特征，分别提取自  $\mathbf{F}_{n_s} \odot \tilde{\mathbf{M}}_{n_s}^{t-1}$  和  $\mathbf{F}_{n_s}$ 。同时，为避免噪声干扰，从其余  $N-1$  张图像中检索  $Q_2$  个图像块特征：

$$q^* = \text{argtopk}_q(\text{Sim}(g_n^t, e_q); Q_2),$$

$$bs_n^t = \mathbf{F}_n(q^*) \in \mathbb{R}^{Q_2 \times D} \quad (7)$$

最终，将  $bs_{\text{sim}}^t$  与其他图像中提取的  $bs_n^t$  进行拼接，构成模糊背景的整体特征表示：

$$\mathbf{B}_{\text{sim}}^t = \text{Concat}(bs_{\text{sim}}^t, bs_n^t) \in \mathbb{R}^{Q \times D}$$

$$n = \{1, 2, \dots, N\} \setminus \{n_s\} \quad (8)$$

### 3.2.3 异质背景检索模块

异质性要求所检索的背景区域之间尽可能不同。为此，在每张图像中，提取与共显著性代理  $cp^t$  差异最大的图像块特征：

表 1 本文采用五种广泛认可的指标来评估模型性能

方法	训练集	CoCA					CoSOD3k					CoSal2015				
		$S_u$	$E_\zeta$	$F_\beta$	$F_\zeta$	$\varepsilon$	$S_u$	$E_\zeta$	$F_\beta$	$F_\zeta$	$\varepsilon$	$S_u$	$E_\zeta$	$F_\beta$	$F_\zeta$	$\varepsilon$
ICNet	1	0.656	0.686	0.514	0.489	0.148	0.795	0.843	0.765	0.751	0.095	0.855	0.896	0.860	0.844	0.058
GCoNet	1	0.673	0.739	0.544	0.531	0.105	0.802	0.857	0.777	0.770	0.071	0.845	0.884	0.847	0.838	0.068
DCFEM	1	0.710	0.763	0.598	0.589	0.085	0.810	0.871	0.805	0.800	0.067	0.838	0.889	0.856	0.850	0.067
UFO	1	0.697	0.762	0.571	0.555	0.095	0.819	0.855	0.797	0.783	0.073	0.860	0.889	0.865	0.848	0.064
CORP**	1	0.706	0.745	0.583	0.567	0.118	0.835	0.877	0.812	0.800	0.067	0.875	0.911	0.881	0.867	0.052
GCoNet+	1	0.717	0.765	0.605	0.580	0.098	0.819	0.856	0.796	0.784	0.075	0.853	0.883	0.857	0.839	0.073
MCCL**	2,3	0.706	0.760	0.582	0.552	0.101	0.850	0.882	0.831	0.808	0.061	0.888	0.911	0.887	0.867	0.051
GCoNet+**	2,3	0.733	0.776	0.630	0.607	0.084	0.840	0.867	0.830	0.810	0.064	0.875	0.896	0.882	0.861	0.058
CADC	1,5	0.681	0.690	0.548	0.503	0.132	0.801	0.823	0.759	0.742	0.096	0.866	0.874	0.862	0.825	0.064
DMT**	1,5	0.722	0.743	0.614	0.582	0.112	0.851	0.880	0.836	0.812	0.065	0.898	0.924	0.904	0.883	0.044
SCoSPARC	1,2	0.711	-	0.614	-	0.092	0.823	-	0.827	-	0.064	0.851	-	0.863	-	0.062
Ours	1	0.733	0.775	0.623	0.611	0.101	0.840	0.883	0.822	0.811	0.062	0.871	0.905	0.880	0.867	0.055
Ours	1	0.740	0.788	0.633	0.622	0.093	0.838	0.880	0.820	0.810	0.064	0.873	0.908	0.884	0.872	0.054
Ours	1,4	0.738	0.778	0.617	0.607	0.091	0.852	0.888	0.832	0.823	0.058	0.890	0.916	0.897	0.885	0.049
Ours	1,5	0.739	0.776	0.622	0.61	0.097	0.847	0.886	0.825	0.814	0.060	0.892	0.921	0.898	0.884	0.046

注：为简化表示，将 COCO9k、DUTS Class、COCO-SEG、Jigsaw-DUTS 和基于合成策略的 DUT-Class 分别记作训练集-1、2、3、4 和 5。对于对比方法，使用其官方实现并遵循推荐的超参数设置，带有\*\*的结果表示通过多次重复实验获得。

$$bd_m^t = \text{argtopk}_{e \in \mathcal{E}_m}(\text{Diff}(cp^t, e); Q') \in \mathbb{R}^{Q' \times D} \quad (9)$$

其中， $\text{Diff}(\cdot, \cdot)$  是预定义的差异函数， $\mathcal{E}_m = \{e \in \mathbf{F}_m \mid m \in \{1, L, N\}\}$  表示图像  $m$  中所有图像块特征的集合， $m_d = \arg \max_m \text{Diff}(cp^t, g_m^t)$ 。对于每张图像，根据是否为最异质背景图像，设定提取 patch 数：

$$Q' = \begin{cases} Q_1, & m = m_d; \\ Q_2, & \text{otherwise} \end{cases} \quad (10)$$

最终，将所有图像中的异质背景特征拼接，得到异质背景的整体表示：

$$\mathbf{B}_{\text{diff}}^t = \text{Concat}(bs_1^t, bs_2^t, \dots, bs_N^t) \in \mathbb{R}^{Q \times D} \quad (11)$$

### 3.2.4 相关性图生成

与共显著性挖掘模块类似，下一步将生成背景区域的相关性图。具体地，先使用卷积操作  $\text{Conv}(\cdot)$  融合  $\mathbf{B}_{\text{diff}}^t$ 、 $\mathbf{B}_{\text{sim}}^t$ 、 $\mathbf{B}_{\text{diff}}^t$ 、 $\mathbf{B}_{\text{sim}}^t$ ，然后将融合结果视为复杂背景的模式表征，并据此搜索各图像中的背景区域：

$$\mathbf{BG}_n^t = \text{Corr}(\text{Conv}(\mathbf{B}_{\text{diff}}^t, \mathbf{B}_{\text{sim}}^t), \mathbf{F}_n) \in \mathbb{R}^{Q \times H' \times W'} \quad (12)$$

### 3.2.5 共显著性图精化

目前为止，已经为每张图像获得了三种相关性图：(1) 共显著性相关图  $\mathbf{R}_n^t \in \mathbb{R}^{K \times H' \times W'}$ ，由式(3)定

义；(2)粗粒度背景相关图  $\mathbf{B}_n^t \in \mathbb{R}^{K \times H' \times W'}$ ，由共显著性流程生成；(3)  $\mathbf{BG}_n^t \in \mathbb{R}^{Q \times H' \times W'}$ ，由式(12)定义。为了生成精化后的共显著性图，首先在图像内对上述每一类相关性图进行聚合。具体而言：

$$\begin{aligned} \bar{\mathbf{R}}_n^t &= \text{Conv}(\mathbf{R}_n^t) \in \mathbb{R}^{D \times H' \times W'}, \\ \bar{\mathbf{B}}_n^t &= \text{Conv}(\mathbf{B}_n^t) \in \mathbb{R}^{D \times H' \times W'}, \\ \bar{\mathbf{BG}}_n^t &= \text{Conv}(\mathbf{BG}_n^t) \in \mathbb{R}^{D \times H' \times W'} \end{aligned} \quad (13)$$

接着，将上述三种相关性图拼接为每张图像的最终特征表示，记为  $\mathbf{Z}_n^t$ 。最后，将  $\mathbf{Z}_n^t$  输入解码器，以生成精化后的共显著性图：

$$\mathbf{M}_n^t = \text{Decode}(\mathbf{Z}_n^t) \in \mathbb{R}^{H \times W} \quad (14)$$

## 4 实验

### 4.1 数据集与评价指标

本文在四个训练集上训练模型，包括 COCO9k<sup>[40]</sup>、DUTS class<sup>[41]</sup>、COCO-SEG<sup>[42]</sup>和 Jigsaw-DUTS<sup>[43]</sup>，分别记为 Train-1、Train-2、Train-3 和 Train-4。具体而言，COCO9k 包含来自 65 个类别的 9213 张图片，DUTS class 包含来自 291 个组的 8250 张图片，COCO-SEG 包含来自 78 个组的 20 万张图片，Jigsaw-DUTS 是 DUTS class 的增强版。此外，还引入图片合成策略<sup>[31]</sup>构建了 Train-5，以对齐更多现有方法。

本文在三个广泛使用的共显著性检测基准数据集上评估模型性能，分别为 CoCA<sup>[43]</sup>(包含 80 个类别，共 1295 张图片)、CoSOD3k<sup>[44]</sup>(覆盖 13 个超类，共 160 个图像组、共计 3316 张图片)以及 CoSal2015<sup>[45]</sup>(包含 50 个类别，共 2015 张图片)。这三个数据集可按照背景复杂度排序为：CoCA > CoSOD3k > CoSal2015。需特别指出的是，CoCA 是最具代表性的现实场景数据集，因为每张图像至少包含一个无关前景目标以及一个共显著目标。此外，该数据集中图像组的类别在常用训练集之外，使得该任务更具挑战性。

本文采用五种常用指标对模型性能进行评估，包括结构相似度 S-measure( $S_\alpha$ )<sup>[12]</sup>、平均增强度 E-measure( $E_\xi$ )<sup>[46]</sup>、最大 F-measure( $F_\beta$ )<sup>[7]</sup>、平均 F-measure( $F_\xi$ )<sup>[7]</sup>以及平均绝对误差 MAE( $\varepsilon$ )<sup>[47-48]</sup>。

### 4.2 实现细节

本文采用三种训练策略，分别为 Train-1、Train-1,4 和 Train-1,5。所有方案均包含 Train-1，这是因为该数据集不仅具有背景复杂性，还包含较少图像数量。然而，Train-1 仅涵盖 65 个类别，可

能限制模型在包含大量类别的测试集上的性能。因此，引入两个合成训练集，即 Train-4 和 Train-5，以平衡背景复杂度、训练效率以及类别覆盖范围。相比之下，Train-2 的背景相对简单，纳入该数据集可能会降低整体背景复杂度，从而削弱模型性能。此外，Train-3 图像数量众多(200k+)，训练成本较高。为简化训练流程，设置  $Q_2 = Q/N$ 。

为突出所提模块的优势，主干网络与训练参数设定与 CORP<sup>[1]</sup>保持一致，总参数量为 106 MB，包括 VGG-16 编码器、所提组件、解码器等；解码器也同 CORP<sup>[1]</sup>和 ICNet<sup>[23]</sup>保持一致，解码四个共显著性特征和剩余的浅层特征，以预测共显著性图。以 Train-1 为例，在 COCO9k<sup>[40]</sup>和 DUTS<sup>[41]</sup>上分别训练 CoSOD 主干网络与 SOD 头，其中图像大小均调整为  $224 \times 224$ 。在每轮训练中，从 CoSOD 数据集中采样 10 张图片，从 SOD 数据集中采样 8 张图片。优化目标采用 IoU 损失<sup>[49]</sup>，其中 CoSOD 损失权重为 90%，SOD 损失权重为 10%。优化器选用 Adam<sup>[50]</sup>，训练轮数设为 70，初始学习率为  $1 \times 10^{-5}$ ，权重衰减为  $1 \times 10^{-4}$ 。所有实验均基于 PyTorch<sup>[51]</sup>与 Jitter<sup>[52]</sup>实现，在 Nvidia RTX 3080 GPU 上运行，每轮训练约耗时 6 min，推理速度约为 40 FPS，在批大小为 10 的时候显存占用为 2144 MB。

### 4.3 与现有先进方法的比较

为了验证所提方法的有效性，本文与十种当前先进的共显著性目标检测方法进行了全面比较，包括 CORP<sup>[1]</sup>、DMT<sup>[2]</sup>、ICNet<sup>[23]</sup>、GCoNet<sup>[29]</sup>、CADC<sup>[31]</sup>、GCoNet+<sup>[32]</sup>、MCCL<sup>[33]</sup>、SCoSPARC<sup>[34]</sup>、UFO<sup>[38]</sup>、DCF<sup>[53]</sup>。

定量比较：表 1 对比了各方法在三个基准数据集上的定量性能。从结果中，可以得出以下几点观察：

在背景复杂的数据集上性能优越：所提方法在三个数据集上都取得了优越的表现，尤其在背景复杂的数据集(如 CoCA)上效果更加显著。例如，在 CoCA 数据集上，所提方法相比次优方法在结构相似度  $S_\alpha$  上提升 0.7%，平均增强度  $E_\xi$  提升 1.2%，最大  $F$  值  $F_\beta$  提升 0.3%，平均  $F$  值  $F_\xi$  提升 1.5%。这与本文的动机一致，即现有 CoSOD 方法的瓶颈可能来自背景语义的模糊性与异质性。

训练高效且不损失性能：为获得更好的性能，已有方法通常依赖于大规模训练集(如 MCCL 和 GCoNet+)。相比之下，所提方法致力于挖掘复杂

背景中蕴含的丰富上下文信息。在仅使用 Train-1 的情况下,也可实现优越的性能。得益于模型的高效性,还可以轻松调节超参数  $T$  和  $Q$ ,进一步提升性能。例如,在仅使用 Train-1 训练的模型中,所提方法在 CoCA 数据集上相比次优方法提升 2.3% ( $S_\alpha$ )、2.3% ( $E_\xi$ )、2.8% ( $F_\beta$ ) 和 3.3% ( $F_\xi$ )。

在背景简单时同样取得较优性能:乍看之下,所提方法在背景简单的 CoSal2015 上表现并不如预期,尤其是在仅使用 Train-1 训练时。然而,深入分析发现该现象源于训练集类别覆盖不足。因此,引

入 Train-4 或 Train-5 后,可显著提升模型在该数据集上的性能,例如提升 1.7% ( $S_\alpha$ )、0.8% ( $E_\xi$ )、1.3% ( $F_\beta$ )、1.2% ( $F_\xi$ )、0.5% ( $\epsilon$ )。

定性比较:图 3 展示了所提方法与其他主流方法在 CoCA 数据集上的共显著性图预测结果。该数据集中,大多数图像都包含至少一个干扰目标(即显著但非共现的目标)。同时,一些图像包含多个共显著目标,使任务更加具有挑战性。此类图像的处理结果能够较好地体现方法在复杂背景条件下的表现能力。

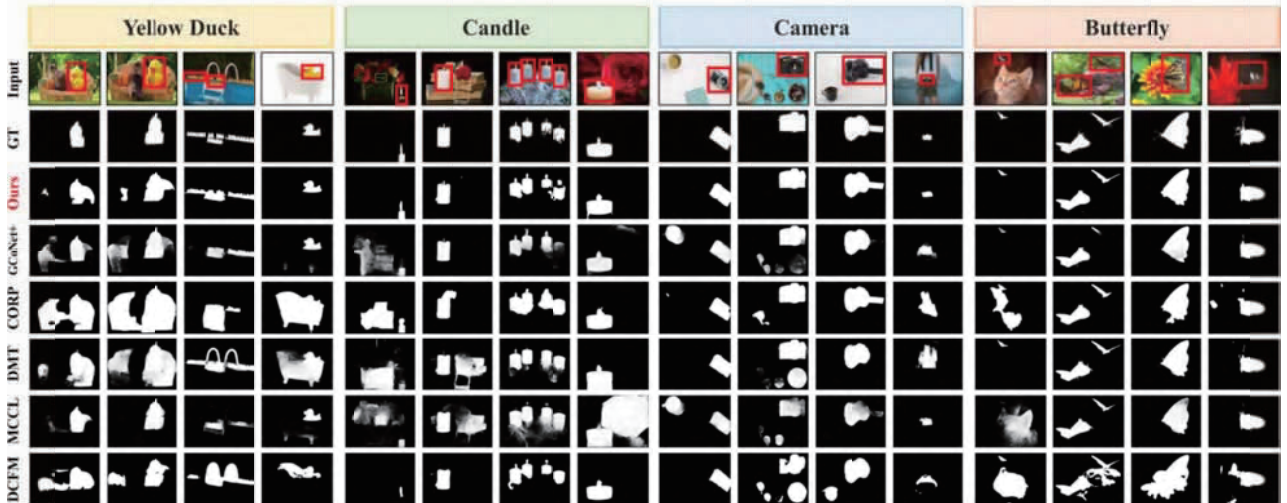


图 3 CoCA 数据集上的定性比较结果(所提方法能够更有效地去除背景干扰目标,从而生成更优质的共显著性图)

本文所提方法能够更有效地建模模糊背景模式,从而更好地去除干扰目标并同时检测多个共显著目标。例如在小黄鸭图像中,其他方法常常错误地将沙发、猫等干扰物体识别为共显著目标,或漏检图像左侧的鸭子。相比之下,所提方法在有效压制干扰目标的同时,能够完整识别出多个共显著目标。

得益于对异质背景的建模能力,所提方法在区分与共显著目标经常共现的非目标区域时表现更优。例如,玫瑰常常出现在蜡烛图像中,因此其他方法易误判玫瑰为共显著目标。所提方法能够有效避免这一误判。此外,所提方法还可以更好地去除异质背景,如相机图像中的眼镜、水杯等干扰物体。

#### 4.4 可视化分析

图 4 可视化结果显示,所提方法能够有效检索语义模糊背景和异质背景。例如,在 Pillow(枕头)图片中,蓝色虚线框表示检索到的沙发垫,其颜色与白色枕头接近,橙色实线框表示检索到的复杂背景,包括花朵、树林、沙发底座等;在 Tomato(番茄)图片中,红色的苹果被成功检索为语义模糊背

景,其他水果被检索为异质背景;在 Yellow duck(小黄鸭)中,与小黄鸭颜色接近的沙发被检索为语义模糊背景,而猫、树林、沙发上的图案被检索为异质背景。这些检索到的复杂背景信息对于优化共显著区域识别具有积极作用。

图 5 和图 6 分别基于结构相似度指标(S-measure,  $S_\alpha$ )与平均增强度指标(E-measure,  $E_\xi$ )对各类别性能进行排序分析。实验结果表明,在背景复杂的 CoCA 数据集中,本文所提方法在超过一半的类别中取得了超过 2% 的性能提升。这一现象进一步验证了本文所强调的观点:背景建模对于共显著性目标检测具有重要意义。

在图 7 所示的可视化结果中,所提方法在处理背景噪声方面相较于基线模型 CORP 展现出了显著优势。以“pineapple”(菠萝)类别为例,所提方法能够准确地区分出菠萝与其他具有相似特征的干扰物体,这得益于所设计的模糊背景检索(ABR)模块。相比之下,在“binoculars”(双筒望远镜)类别中,所提方法有效抑制了大量具有同等显著性的无关目标,此类改进正是来自异质背景检索(HBR)

模块。

综上所述，以上观察结果充分表明，本文所提出的“模糊与异质背景挖掘”机制在抑制背景噪声干扰方面是有效且具有实际意义的。



图4 AHBE模块检索结果可视化(其中蓝色虚线框表示检索到的语义模糊背景, 橙色实线框表示检索到的异质背景)

### 4.5 消融实验

进一步进行消融实验,以验证所提出的 ABR 与 HBR 模块的有效性。所有模型均在 Train-1 数据集上训练,训练超参数保持一致。从表 2 的结果中,可以得到如下观察:

单独引入 ABR 或 HBR 模块均可在背景复杂的数据集(如 CoCA)上带来显著性能提升。具体而言,使用 ABR 或 HBR 后,在  $S_\alpha$ 、 $E_\xi$ 、 $F_\beta$ 、 $F_\xi$  和  $\varepsilon$  上分别至少提升 0.9%、1.1%、0.6%、1.6%和 0.4%。然而,随着背景复杂度的降低,单一模块的优势逐渐减弱。在 CoSOD3k 数据集上,二者仍可保持与基线相当的性能,但在 CoSal2015 上略有退步。

联合使用 ABR 与 HBR 模块可在各类数据集上进一步带来显著性能增益,无论背景是否复杂。具体来说,在 CoCA 数据集上,所提方法相比基线在  $S_\alpha$ 、 $E_\xi$ 、 $F_\beta$ 、 $F_\xi$  和  $\varepsilon$  上分别提升 2.7%、3.0%、4.0%、4.4%和 1.7%。即使是在背景较为简单的 CoSOD3k 数据集上,联合使用 ABR 与 HBR 仍能维持或略优于基线的性能。

因此,可以得出结论:同时使用 ABR 与 HBR 模块是安全且推荐的策略,在不同背景复杂度条件下均表现稳定可靠。

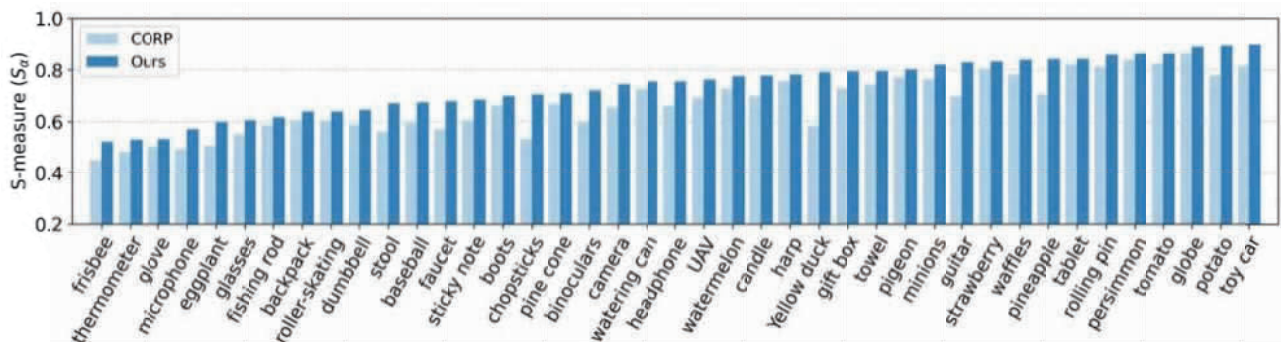


图5 所提方法相较于基线模型在 CoCA 数据集各类别结构相似度指标  $S_\alpha$  上的提升情况

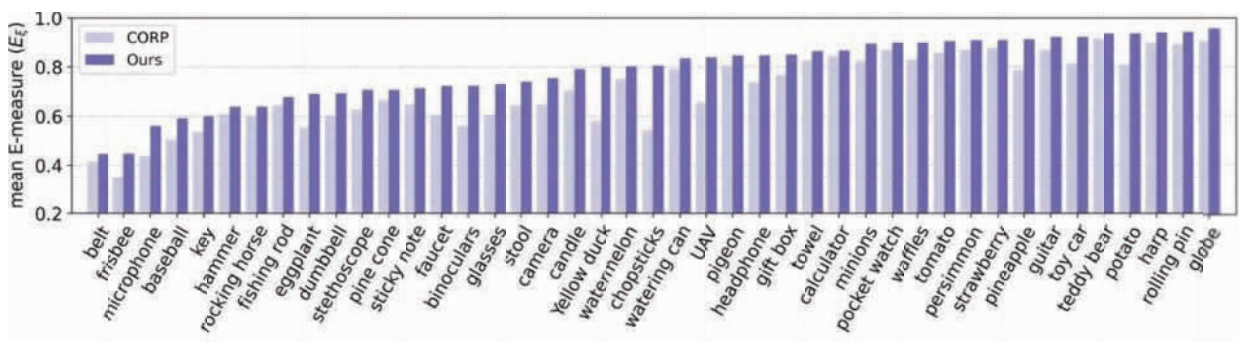


图6 所提方法相较于基线模型在 CoCA 数据集各类别结构相似度指标  $E_\xi$  上的提升情况

表 2 所提模块的消融研究

设定	CoCA					CoSOD3k					CoSal2015				
	$S_\alpha$	$E_\xi$	$F_\beta$	$F_\xi$	$\varepsilon$	$S_\alpha$	$E_\xi$	$F_\beta$	$F_\xi$	$\varepsilon$	$S_\alpha$	$E_\xi$	$F_\beta$	$F_\xi$	$\varepsilon$
基线模型	0.706	0.745	0.583	0.567	0.118	0.835	0.877	0.812	0.800	0.067	0.875	0.911	0.879	0.865	0.052
+ABR	0.721	0.757	0.589	0.587	0.104	0.834	0.877	0.817	0.808	0.065	0.865	0.900	0.875	0.865	0.058
+HBR	0.715	0.756	0.598	0.583	0.114	0.834	0.878	0.817	0.806	0.064	0.864	0.899	0.874	0.864	0.063
+ABR+HBR	0.733	0.775	0.623	0.611	0.101	0.840	0.883	0.822	0.811	0.062	0.871	0.905	0.880	0.867	0.055

注:以 CORP 为基线模型,展示所提模块在复杂背景下更加有效。

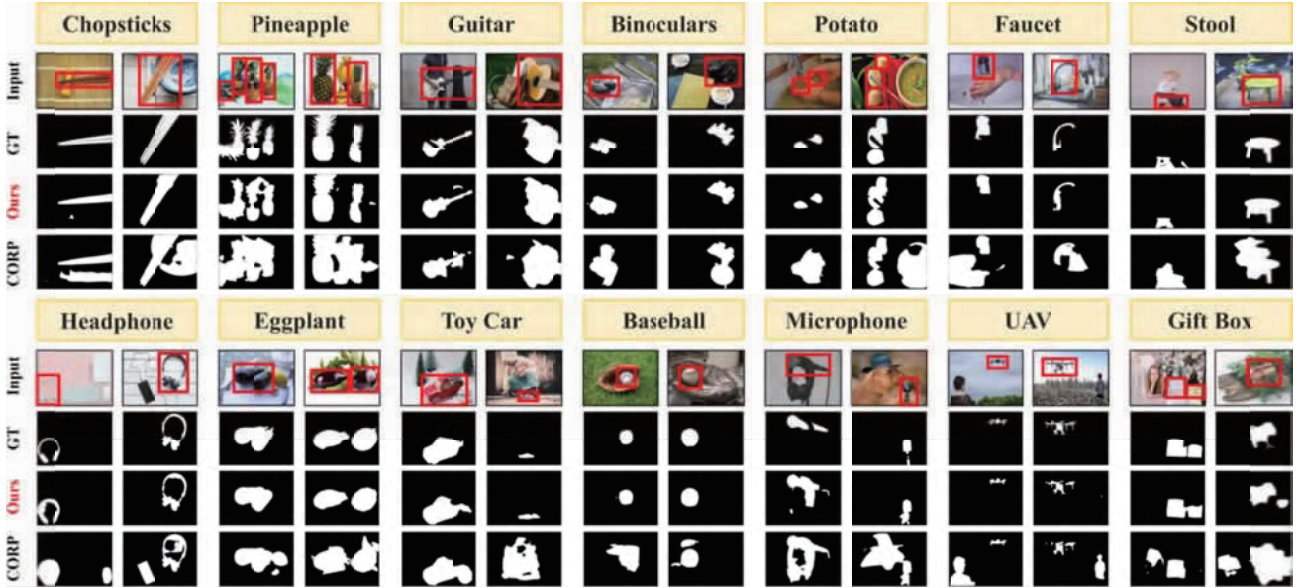


图7 所提方法在 CoCA 数据集上与基线模型的定性比较

此外,在式(2)和(12)中,文本沿用了 COPR 中相关性计算方法,即使用  $1 \times 1$  卷积实现逐元素点积计算。在表 3 中,点积被替换为余弦相似度函数,可以发现模型取得了。

表 3 相关性计算方法消融实验(其中卷积被替换为余弦相似度)

相关性计算	CoCA				
	$S_\alpha$	$E_\xi$	$F_\beta$	$F_\xi$	$\varepsilon$
点积	0.733	0.775	0.623	0.611	0.101
余弦相似度	0.715	0.749	0.591	0.577	0.114

#### 4.6 敏感性分析

本节探讨所提出模块的敏感性特性。所有模型均在 Train-1 数据集上训练,并在 CoCA 数据集上测试。参照基线模型<sup>[1]</sup>中设定,固定  $K = 32$ 。随后,分别调整迭代次数  $T$  与背景特征数量  $Q$ ,并将模型性能记录于表 4 与表 5 中。实验结果揭示了如下结论:

增加迭代次数  $T$  有助于性能提升。可以发现,随着迭代次数的增加,模型性能相应提高。然而,随着  $T$  的增大,其边际收益迅速递减并趋于稳定。因此,选择适中的  $T$  (例如 6 左右)即可在效率与性能之间取得良好平衡。

选择适中的背景特征数量  $Q$  是实现最优性能的关键。一方面,若  $Q$  设置过小,模块可能无法覆盖复杂背景中关键模式;另一方面,若  $Q$  过大,则背景模块可能错误地包含部分共显著区域,从而导

表 4 迭代次数的敏感性分析

参数设定	CoCA				
	$S_\alpha$	$E_\xi$	$F_\beta$	$F_\xi$	$\varepsilon$
$T=1$	0.684	0.715	0.545	0.530	0.132
$T=2$	0.712	0.749	0.588	0.574	0.113
$T=3$	0.723	0.761	0.604	0.592	0.107
$T=4$	0.728	0.768	0.613	0.601	0.104
$T=5$	0.731	0.772	0.618	0.606	0.102
$T=6$	0.732	0.774	0.621	0.608	0.101
$T=7$	0.733	0.775	0.622	0.610	0.101
$T=8^*$	0.733	0.775	0.623	0.611	0.101
$T=9$	0.733	0.775	0.623	0.611	0.101

注:随着迭代次数的增加,模型性能逐渐改善,在  $T^*=8$  时达到最佳性能。

表 5 背景表示数量  $Q$  的敏感性分析

参数设定	CoCA				
	$S_\alpha$	$E_\xi$	$F_\beta$	$F_\xi$	$\varepsilon$
$Q=22$	0.727	0.766	0.609	0.598	0.103
$Q^*=22$	0.740	0.788	0.633	0.622	0.093
$Q=28$	0.733	0.775	0.623	0.611	0.101
$Q=38$	0.727	0.767	0.610	0.597	0.100
$Q=42$	0.727	0.766	0.610	0.597	0.103
$Q=48$	0.728	0.769	0.615	0.600	0.105

注:适中的  $Q$  对于性能至关重要。若  $Q$  过小,可能无法覆盖复杂背景;若  $Q$  过大,则可能错误地引入共显著区域,导致性能下降。最优性能在  $Q^*=28$  时取得。

致性能退化。综合实验表明,当  $Q = 28$  时可获得最优性能。

#### 4.7 失效样例

虽然实验结果整体表现优异，然而所提方法在部分类别上可能失效。图8展示了几个代表性样例，并归结于以下原因：

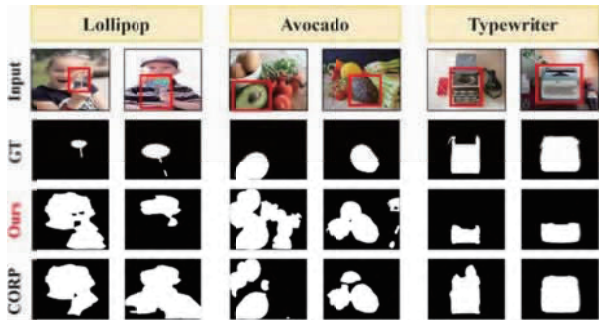


图8 三种代表性的失效样例

若背景区域与共显著物体经常共现，且显著性显著强于共显著物体，则容易导致所提方法误判。例如针对棒棒糖(Lollipop)，人体往往与其共现，且具有显著巨大的尺寸，这导致所提方法容易将此类背景误判为共显著区域。

共显著区域的模式如果存在较大差异，可能导致所提方法失效。例如针对牛油果(Avocado)，两张图片呈现出完全不同状态：一种是带核切开的，另一种是未切开。由于所提方法依赖共显著代理进行检索，当共显著区域本身存在显著差异时，所选共性表示的可靠性会降低，更容易与背景区域产生错误关联。

共显著区域如果存在更为相似的子模式，可能导致所提方法过于聚焦局部。例如针对打字机(Typewriter)，键盘区域通常高度相似，而其他区域则呈现出多样性。此时，所提方向倾向于排除其他区域，仅将键盘区域识别为共显著性区域。

## 5 总结与未来工作

本文提出了一种针对复杂背景场景中共显著性目标检测问题的模糊与异质背景挖掘框架。该方法引入了两个专门模块，分别为模糊背景检索与异质背景检索，以更有效地应对背景区域中普遍存在的模糊性与异质性问题。通过在多个共显著性检测基准数据集上的大量实验验证，所提方法在检测性能和计算效率方面均优于当前先进方法，特别是在面对具有挑战性的复杂背景时表现更加显著。

尽管本文方法已能有效应对 CoSOD 中背景的模糊性与异质性问题，但仍存在一些值得进一步研究的方向：一方面，未来可探索自适应策略，以根

据数据特性自动选择适合的超参数(如  $Q$ )；另一方面，模型性能仍可能受到检索背景中的噪声影响，因此对背景特征噪声的鲁棒建模也值得深入探讨。

## 参考文献

- [1] Zhu Zi-Yue, Zhang Zhao, Lin Zheng, et al. Co-salient object detection with co-representation purification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(7): 8193-8205
- [2] Li Long, Han Jun-Wei, Zhang Ni, et al. Discriminative co-saliency and background mining transformer for co-salient object detection// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada, 2023: 7247-7256
- [3] Harel Jonathan, Koch Christof, Perona Pietro. Graph-based visual saliency//*Proceedings of the Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, 2006: 545-552
- [4] Hou Xiao-Di, Zhang Li-Qing. Saliency detection: A spectral residual approach//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, USA, 2007: 1-8
- [5] Itti Laurent, Koch Christof, Niebur Ernst. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254-1259
- [6] Liu Tie, Yuan Ze-Jian, Sun Jian, et al. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 33(2): 353-367
- [7] Achanta Radhakrishna, Hemami Sheila, Estrada Francisco, Susstrunk Sabine. Frequency-tuned salient region detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, 2009: 1597-1604
- [8] Borji Ali, Itti Laurent. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 35(1): 185-207
- [9] Cheng Ming-Ming, Mitra Niloy J, Huang Xiao-Lei, et al. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 37(3): 569-582
- [10] Ren Zhi-Xiang, Gao Sheng-Hua, Chia Liang-Tien, Tsang Ivor Wai-Hung. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013, 24(5): 769-779
- [11] Zhang Jian-Ming, Sclaroff Stan. Saliency detection: A Boolean map approach//*Proceedings of the IEEE International Conference on Computer Vision*. Sydney, Australia, 2013: 153-160
- [12] Fan Deng-Ping, Cheng Ming-Ming, Liu Yun, et al. Structure-measure: A new way to evaluate foreground maps//*Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy, 2017: 4548-4557
- [13] Chen Zu-Yao, Xu Qian-Qian, Cong Run-Min, Huang Qing-Ming. Global context-aware progressive aggregation network for salient object detection//*Proceedings of the AAAI Conference on Artificial Intelligence*. New York, USA, 2020: 10599-10606
- [14] Hou Qi-Bin, Cheng Ming-Ming, Hu Xiao-Wei, et al. Deeply supervised salient object detection with short connections// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA, 2017: 5300-5309
- [15] Zhao Rui, Ouyang Wan-Li, Li Hong-Sheng, Wang Xiao-Gang. Saliency detection by multi-context deep learning//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA, 2015: 1265-1274
- [16] Li Gong-Yang, Bai Zhen, Liu Zhi, et al. Salient object detection in

- optical remote sensing images driven by transformer. *IEEE Transactions on Image Processing*, 2023, 32: 5257-5269
- [17] Liu Nian, Han Junwei, Yang Ming-Hsuan. PiCANet: Learning pixel-wise contextual attention for saliency detection//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA, 2018: 3089-3098
- [18] Liu Nian, Zhang Ni, Wan Kai-Yuan, et al. Visual saliency transformer//*Proceedings of the IEEE International Conference on Computer Vision*. Montreal, Canada, 2021: 4722-4732
- [19] Vaswani Ashish, Shazeer Noam, Parmar Niki, et al. Attention is all you need//*Proceedings of the Annual Conference on Neural Information Processing Systems*. Long Beach, USA, 2017: 5998-6008
- [20] Wang Wen-Guan, Lai Qiu-Xia, Fu Hua-Zhu, et al. Saliency object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 44(6): 3239-3259
- [21] Li Fei-Ran, Xu Qian-Qian, Bao Shi-Long, et al. Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection//*Proceedings of the International Conference on Machine Learning*. Vienna, Austria, 2024, 235: 28989-29021
- [22] Chen Zu-Yao, Cong Run-Min, Xu Qian-Qian, Huang Qing-Ming. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 2020, 30: 7012-7024
- [23] Jin Wen-Da, Xu Jun, Cheng Ming-Ming, et al. ICNet: Intra-saliency correlation network for co-saliency detection//*Proceedings of the Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, 2020, 33: 18749-18759
- [24] Fu Hua-Zhu, Cao Xiao-Chun, Tu Zhuo-Wen. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 2013, 22(10): 3766-3778
- [25] Liu Zhi, Zou Wen-Bin, Li Lina, et al. Co-saliency detection based on hierarchical segmentation. *IEEE Signal Processing Letters*, 2013, 21(1): 88-92
- [26] He Kai-Ming, Zhang Xiang-Yu, Ren Shao-Qing, Sun Jian. Deep residual learning for image recognition//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA, 2016: 770-778
- [27] Simonyan Karen, Zisserman Andrew. Very deep convolutional networks for large-scale image recognition//*Proceedings of the International Conference on Learning Representations*. San Diego, USA, 2015: 1-14
- [28] Szegedy Christian, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA, 2015: 1-9
- [29] Fan Qi, Fan Deng-Ping, Fu Hua-Zhu, et al. Group collaborative learning for co-salient object detection//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Nashville, USA, 2021: 12288-12298
- [30] Guo Ruo-Hao, Ying Xiang-Hua, Qi Yan-Yu, Qu Liao. UniTR: A unified transformer-based framework for co-object and multi-modal saliency detection. *IEEE Transactions on Multimedia*, 2024, 26: 7622-7635
- [31] Zhang Ni, Han Jun-Wei, Liu Nian, Shao Ling. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection//*Proceedings of the IEEE International Conference on Computer Vision*. Montreal, Canada, 2021: 4167-4176
- [32] Zheng Peng, Fu Hua-Zhu, Fan Deng-Ping, et al. GCoNet+: A stronger group collaborative co-salient object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(9): 10929-10946
- [33] Zheng Peng, Qin Jie, Wang Shuo, et al. Memory-aided contrastive consensus learning for co-salient object detection//*Proceedings of the AAAI Conference on Artificial Intelligence*. Washington, USA, 2023: 3687-3695
- [34] Chakraborty Souradeep, Samaras Dimitris. Self-supervised co-salient object detection via feature correspondences at multiple scales//Leonardis Ales, Ricci Elisa, Roth Stefan, Russakovsky Olga, Sattler Torsten, Varol Gül, eds. *Proceedings of the European Conference on Computer Vision*. Milan, Italy: Springer, 2024: 231-250
- [35] Xiao Hao-Ke, Tang Lv, Li Bo, et al. Zero-shot co-salient object detection framework//*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Republic of Korea, 2024: 4010-4014
- [36] Wang Jie, Yu Nana, Zhang Zi-Hao, Han Ya-Hong. Visual consensus prompting for co-salient object detection//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, USA, 2025: 9591-9600
- [37] Zhu Jia-Yi, Guo Qing, Juefei-Xu Felix, et al. Concept guided co-saliency objection detection. *CoRR*, 2024, abs/2412.16609
- [38] Su Yu-Kun, Deng Jing-Liang, Sun Rui-Zhou, et al. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 2023, 26: 313-325
- [39] Fan Deng-Ping, Li Teng-Peng, Lin Zheng, et al. Re-thinking co-salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 44(8): 4339-4354
- [40] Lin Tsung-Yi, Maire Michael, Belongie Serge, et al. Microsoft COCO: Common objects in context//*Proceedings of the European Conference on Computer Vision*. Zurich, Switzerland, 2014: 740-755
- [41] Wang Li-Jun, Lu Hu-Chuan, Wang Yi-Fan, et al. Learning to detect salient objects with image-level supervision//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA, 2017: 136-145
- [42] Chong Wang, Zheng-Jun Zha, Dong Liu, Hong-Tao Xie. Robust deep co-saliency detection with group semantic//*Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, Honolulu, USA, 2019: 8917-8924
- [43] Zhang Zhao, Jin Wen-Da, Xu Jun, Cheng Ming-Ming. Gradient-induced co-saliency detection//*Proceedings of the European Conference on Computer Vision*. Glasgow, UK, 2020: 455-472
- [44] Fan Deng-Ping, Lin Zheng, Ji Ge-Peng, et al. Taking a deeper look at co-salient object detection//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Seattle, USA, 2020: 2919-2929
- [45] Zhang Ding-Wen, Han Jun-Wei, Li Chao, et al. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision*, 2016, 120(2): 215-232
- [46] Fan Deng-Ping, Gong Cheng, Cao Yang, et al. Enhanced-alignment measure for binary foreground map evaluation//*Proceedings of the International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 2018: 698-704
- [47] Borji Ali, Cheng Ming-Ming, Jiang Huaizu, Li Jia. Saliency object detection: A benchmark. *IEEE Transactions on Image Processing*, 2015, 24(12): 5706-5722
- [48] Cheng Ming-Ming, Warrell Jonathan, Lin Wen-Yan, et al. Efficient salient region detection with soft image abstraction//*Proceedings of the IEEE International Conference on Computer Vision*. Sydney, Australia, 2013: 1529-1536
- [49] Lin Huai jia, Qi Xiaojuan, Jia Jiaya. AGSS-VOS: Attention guided single-shot video object segmentation//*Proceedings of the IEEE*

International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 3949-3957

- [50] Kingma Diederik P, Ba Jimmy. Adam: A method for stochastic optimization//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015
- [51] Paszke Adam, Gross Sam, Massa Francisco, et al. PyTorch: An imperative style, high-performance deep learning library//Proceedings of the Annual Conference on Neural Information Processing Systems.



**WANG Zi-Tai**, Ph.D. His research interests include machine learning and data mining.

**XU Qian-Qian**, Ph.D., professor. Her research interests include statistical machine learning, with applications in

## Background

This paper addresses a fundamental challenge in the field of computer vision, specifically in Co-Salient Object Detection (CoSOD), which aims to identify and segment common salient objects from a group of related images. CoSOD is an important subtask of salient object detection and has practical applications in areas such as image co-segmentation, visual surveillance, and object tracking.

Traditionally, CoSOD methods emphasize learning consistent co-salient representations across images by focusing on common features while suppressing irrelevant signals. State-of-the-art approaches often leverage global pooling techniques and transformer-based architectures to refine the shared representations of co-salient regions. While these techniques have advanced the field considerably, they typically treat background modeling as a secondary process, often reusing the same pipeline designed for foreground processing. However, in real-world scenarios, backgrounds frequently exhibit two critical properties: ambiguity—where background regions mimic the appearance of co-salient objects—and heterogeneity—where backgrounds vary significantly across images. These issues severely compromise the reliability of co-salient representation learning. Existing methods such as DMT attempt to explicitly incorporate background modeling but fail to distinguish the complexity of background patterns from co-salient features effectively.

To bridge this gap, the authors propose a novel Ambiguous and Heterogeneous Background Exploration (AHBE) framework. This method introduces two dedicated modules: Ambiguous Background Retrieval (ABR) and Heterogeneous Background Retrieval (HBR), each designed to isolate and suppress misleading background cues. The ABR

Vancouver, Canada, 2019: 8024-8035

- [52] Hu Shi-Min, Liang Dun, Yang Guo-Ye, et al. Jittor: A novel deep learning framework with meta-operators and unified graph execution. Science China Information Sciences, 2020, 63: 1-21
- [53] Yu Si-Yue, Xiao Ji-Min, Zhang Bing-Feng, Lim Eng Gee. Democracy does matter: Comprehensive feature mining for co-salient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 979-988

multimedia and computer vision.

**CAO Yu-Chen**, M.S. His research interests include computer vision.

**LIU Yang**, Ph.D. candidate. His current research interests include computer vision and self-supervised learning.

**HUANG Qing-Ming**, Ph.D., chair professor. His research areas include multimedia computing, image processing, computer vision and pattern recognition.

module retrieves background features that closely resemble the co-salient proxy, addressing ambiguity. Meanwhile, HBR emphasizes dissimilar background retrieval to mitigate heterogeneity. These background cues are then used to iteratively refine co-saliency maps, leading to improved segmentation performance.

Experimental results across three benchmark datasets—CoSal2015, CoSOD3k, and CoCA—demonstrate that AHBE significantly outperforms state-of-the-art methods, particularly in challenging scenarios with complex backgrounds. Moreover, the proposed framework achieves strong results with relatively limited training data, highlighting its efficiency and robustness.

In summary, this paper makes three main contributions: (1) it identifies and formalizes ambiguity and heterogeneity as critical issues in CoSOD; (2) it proposes ABR and HBR modules for fine-grained background modeling; and (3) it establishes new state-of-the-art performance across multiple datasets with a computationally efficient design.

This work was supported in part by National Natural Science Foundation of China: 62525212, 62236008, 62441232, U21B2038, U23B2051, and 62502500, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680201, in part by the China National Postdoctoral Program for Innovative Talents under Grant BX20240384, in part by Beijing Natural Science Foundation under Grant No. L252144, and in part by General Program of the Chinese Postdoctoral Science Foundation under Grant No. 2025M771558.