CCF优博丛书

认知规律启发的显著性物体检测方法与评测

Cognitive-inspirited Salient Object Detection Models and Benchmarks

范登平——著

机械工业出版社
CHINA MACHINE PRESS

显著性物体检测（Salient Object Detection，SOD）技术以人类视觉认知机制为基础，模拟人类视觉系统的注意力机制。该技术涉及计算机视觉、机器学习、认知心理学、脑科学等多个学科，是典型的交叉学科技术，在现实生活中有着广泛的应用基础。

本书从数据采集、模型构建和评价标准设计三个方面对 SOD 技术展开了系统的研究，包括开放环境下的图像 SOD 技术、动态场景下的 SOD 视觉转移建模技术以及符合人类认知规律的评价指标设计。

本书可以作为高等院校计算机视觉及模式识别相关专业的本科生、研究生，以及计算机相关领域科研工作者的参考书。

图书在版编目（CIP）数据

认知规律启发的显著性物体检测方法与评测/范登平著. —北京：机械工业出版社，2022.12
（CCF 优博丛书）
ISBN 978-7-111-71502-3

Ⅰ. ①认… Ⅱ. ①范… Ⅲ. ①计算机视觉②机器学习 Ⅳ. ①TP302.7②TP181

中国版本图书馆 CIP 数据核字（2022）第 157783 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：梁 伟 责任编辑：游 静
责任校对：贾海霞 王 延 封面设计：鞠 杨
责任印制：单爱军
北京虎彩文化传播有限公司印刷
2023 年 1 月第 1 版第 1 次印刷
148mm×210mm · 6.125 印张 · 4 插页 · 115 千字
标准书号：ISBN 978-7-111-71502-3
定价：39.00 元

电话服务 网络服务
客服电话：010-88361066 机 工 官 网：www.cmpbook.com
010-88379833 机 工 官 博：weibo.com/cmp1952
010-68326294 金 书 网：www.golden-book.com
封底无防伪标均为盗版 机工教育服务网：www.cmpedu.com

CCF 优博丛书编委会

丛书序

博士研究生教育是教育的最高层级，是一个国家高层次人才培养的主渠道。博士学位论文是青年学子在其人生求学阶段，经历“昨夜西风凋碧树，独上高楼，望尽天涯路”和“衣带渐宽终不悔，为伊消得人憔悴”之后的学术巅峰之作。因此，一般来说，博士学位论文都在其所研究的学术前沿点上有所创新、有所突破，为拓展人类的认知和知识边界做出了贡献。博士学位论文应该是同行学术研究者的必读文献。

为推动我国计算机领域的科技进步，激励计算机学科博士研究生潜心钻研，务实创新，解决计算机科学技术中的难点问题，表彰做出优秀成果的青年学者，培育计算机领域的顶级创新人才，中国计算机学会（CCF）于 2006 年决定设立“中国计算机学会优秀博士学位论文奖”，每年评选不超过 10 篇计算机学科优秀博士学位论文。截至 2021 年已有 145 位青年学者获得该奖。他们走上工作岗位以后均做出了显著的科技或产业贡献，有的获国家科技大奖，有的获评国际高被引学者，有的研发出高端产品，大都成为计算机领域国内国际知名学者、一方学术带头人或有影响力的企业家。

博士学位论文的整体质量体现了一个国家相关领域的科技发展程度和高等教育水平。为了更好地展示我国计算机学科博士生教育取得的成效，推广博士生科研成果，加强高端学术交流，中国计算机学会于 2020 年委托机械工业出版社以“CCF 优博丛书”的形式，陆续选择 2006 年至今及以后的部分优秀博士学位论文全文出版，并以此庆祝中国计算机学会建会 60 周年。这是中国计算机学会又一引人瞩目的创举，也是一项令人称道的善举。

希望我国计算机领域的广大研究生向该丛书的学长作者们学习，树立献身科学的理想和信念，塑造“六经责我开生面”的精神气度，砥砺探索，锐意创新，不断摘取科学技术明珠，为国家做出重大科技贡献。

谨此为序。

中国工程院院士

2022 年 4 月 30 日

推荐序 I

视觉认知中的注意力机制是人眼高效感知周围环境的重要机制之一，《认知规律启发的显著性物体检测方法与评测》的研究内容以人类认知规律为理论基础，以感兴趣的目标为研究对象，深入研究了基于视觉注意力机制的建模中存在的三个重要问题：图像数据存在选择性偏差，视频标注与视觉注意转移具有不一致性，非结构化度量指标具有非完备性。针对这三个问题，作者从数据采集、模型构建、损失函数构造、评价标准设计和开源评测平台搭建方面入手，做出了系统性的创新贡献。作者积极践行开放、共享的科研记录行动倡议，通过在线演示系统、开源代码和公开讨论平台积极服务学术共同体。

该书研究内容紧密结合人类视觉认知机制与显著性计算技术，所提出的核心技术为计算机视觉的诸多任务提供了重要的技术基础。其主要贡献包括：揭示了视频显著性物体检测领域中长期使用的数据标注方式存在的本质缺陷，通过引入眼动仪来记录人眼视觉注意的变化，从而巧妙地解决了数据标注不一致的问题；设计了一个面向显著性转移的长短时记忆卷积网络，可通过学习人类注意力转移行为来有效地捕

获视频动态显著性，该研究工作具有引领性的科学价值，并被计算机视觉顶级国际会议 IEEE CVPR 评为“Best Paper Finalist Award”；阐明了传统像素级评测指标的非完备性，设计的两项结构敏感的评测指标 S-measure 和 E-measure 均被显著性检测领域广泛采用，为计算机视觉前景目标分割系统的设计、研发、测试、部署与运维提供了一套更完备的评价标准。该书可为计算机视觉及模式识别领域的本科生、研究生以及科研工作者提供宝贵经验，值得推荐。

刘青山

南京信息工程大学教授

2022 年 5 月 21 日

推荐序II

2017年，国际计算机视觉大会（International Conference on Computer Vision，ICCV）在意大利水城威尼斯举办，我带着学生参加会议，碰到了范登平博士，他发表的论文引起了我的关注，当时他是南开大学的在读博士生。后来我受邀去程明明教授的媒体计算课题组访问交流，我和范博士再一次有了近距离的交谈。我们交流了很多关于显著性目标检测（Salient Object Detection，SOD）未来发展的问题。

SOD涉及计算机视觉、机器学习、认知心理学以及脑科学等多个领域，属于热门的交叉学科研究领域，也是人工智能领域中的基础共性问题。我研究SOD课题已经有十来年了，一转眼认识范博士也已经有五个年头了，他顺利地从学生蜕变成一名优秀的学者，将我们当年探讨的课题变成了一篇篇顶级学术会议论文，如今又获得了CCF优秀博士学位论文奖。翻开《认知规律启发的显著性物体检测方法与评测》，可以看到他对SOD领域亟需解决的问题的深刻洞察，这些问题可以被概括为：①图像数据中的采集偏差性；②视频数据中的注意焦点转移性；③评价标准的局限性。针对这些科学难题，他从数据采集、模型设计和评价指标三个方面

系统化地提出了解决方案。

例如，最令人兴奋的是“富上下文环境下的显著性物体检测数据集与评测”工作率先系统地梳理了 SOD 领域的模型，发现了当前数据采集的偏差性问题，并通过引入复杂背景下的显著目标图像，更真实地还原了 SOD 任务本身。评测模型的方式也从传统的整体性能评测过渡到了属性级（遮挡、光照变化、运动模糊等）性能评测，这为更加精细化的模型设计、参数调试和部署提供了重要参考依据。

第 4 章“基于注意力转移机制的视频显著性物体检测”，首次揭示了当前视频数据逐帧标注时目标不变性与人类在动态场景中关注目标时注意焦点转移的矛盾，作者利用眼动追踪仪器来实时标定受试者感兴趣的对象，从而构建了符合人类认知机制的目标检测新任务。该工作因其研究方向的前瞻性被国际计算机视觉与模式识别会议（Computer Vision and Pattern Recognition Conference， CVPR）评选为“Best Paper Finalist”（最佳入围论文）。

更值得一提的是，范博士针对当前像素级评价方式的局限性，设计了更符合人类认知规律的结构性指标 S-measure 以及基于整体-局部评价的增强型指标 E-measure。它们已经成为 SOD 领域评测模型的黄金标准，为该领域的学术共同体提供了更加全面、客观的结果。 E-measure 由于其广泛的学术影响力，入选了 Paper Digest 学术平台筛选的 2018 年度国际人工智能联合会议（International Joint Conferences on Artifi-

cial Intelligence，IJCAI）最有影响力的 10 篇论文。

基础性、原创性和前沿性是该书的三个亮点，书中各章节都有对应的开源代码以及项目主页，我相信该书非常值得阅读，可为有志从事计算机视觉方向研究的学生提供良好的借鉴。

卢湖川

大连理工大学教授

2022 年 5 月 29 日

导师序

本人是南开大学计算机学院的教师，研究方向为计算机视觉和计算机图形学，特向读者推荐《认知规律启发的显著性物体检测方法与评测》，该书的内容获得了2021年度中国计算机学会评选的“CCF优秀博士学位论文奖”。作者范登平博士于2015年考入媒体计算实验室（https://mmcheng.net/），在本人的指导下进行研究，于2019年6月以优秀毕业生的荣誉身份获得博士学位，先后加入阿联酋起源人工智能研究院（IIAI）和瑞士苏黎世联邦理工学院（ETH Zurich）继续从事科研工作。

人类获取的80%以上信息由视觉系统处理，而其能始终轻松应对的原因之一是具备视觉注意力机制。视觉注意力机制的研究涉及生物学、脑科学、计算机视觉以及深度学习等多个领域。研究如何让机器人系统具备类人的强大视觉感知能力，甚至超越人类视觉系统，在更加开放的场景下表现出优异的场景认知能力，是“新一代人工智能”技术体系中的技术难点，也是计算机视觉领域的研究重点，这有利于推动军事、医疗、农业和商业等领域的科技发展。

作者针对上述难题，从生物视觉认知机理入手，结合类

脑计算与深度学习的最新研究成果，从数据采集、模型构建、学习函数设计、评测标准制定方面展开研究，在视觉注意机制领域形成了系统性的创新成果。例如，作者提出了开放环境下的显著性目标检测任务并搜集大规模的数据集进行属性级别的评测，从而将该领域的研究从实验室环境推进到更加真实的场景；首次揭示了视频显著性目标检测中注意焦点转移的问题，并突破性地提出了基于注意力转移的视频显著目标检测关键技术。值得一提的是该书第 5 章提出的基于结构相似度的评测标准 S-measure 被证明更加符合人类认知规律，特别是将与人的主观评价一致性的性能从低于 50% 提升到了 77%，该成果被国际计算机视觉顶级会议 ICCV 2017 录用，并受邀做大会焦点论文报告。第 6 章提出的基于整体-局部相似度的评价标准 E-measure 相比国际最先进的评测算法，性能提高了 19%，成为显著性和伪装目标检测两大领域的黄金指标。 E-measure 被进一步推广应用到新冠肺炎诊断系统性能评估中，极大地提高了系统诊断效率，该系统获得了第二十二届中国国际工业博览会高校展区优秀展品特等奖。

该书内容属于典型的交叉学科研究课题，系统地展示了提出问题、构建新任务和解决问题的三步曲。作者为各个章节的内容提供了开源代码（https://dengpingfan.github.io/）、数据集以及在线演示系统等资源来更好地服务学术共同体，这种行为是非常值得称赞的。希望本书能够为有志从事人工

智能研究的专家、学者提供更系统的研究思路，并将相关技术进一步推广到更广阔的领域中。

程明明

南开大学教授

2022 年 7 月 20 日

摘　要

显著性物体检测技术起源于认知学中人类的视觉注意行为，即人类视觉系统能够快速地将注意力转移到视觉场景中最具信息量的区域而有选择性地忽略其他区域。该技术在现实生活中有着广泛的应用基础，如自动驾驶、人机交互、视频分割、视频字幕和视频压缩等。由于图像和视频数据（遮挡、模糊和运动模式等）自身存在的挑战以及人类在动态场景中注意行为（选择性注意分配和注意转移）固有的复杂性，显著性物体检测技术面临着巨大挑战。受制于采集设备，早期构建的显著性物体检测数据集表达真实场景的能力非常有限。同时，这一领域的评价指标是基于像素级误差的，完全忽略了人类认知规律的特性。上述问题严重制约了显著性物体检测技术的发展。

本书围绕图像、视频显著性物体检测，研究了基于人类认知规律的数据集构建、模型构建、评价指标三个方面的问题，主要创新点包括：

1）针对现有图像显著性物体检测公开测试存在的各种偏差问题，构建了一个富上下文环境下的图像显著性物体检测数据集 SOC，并首次从属性层面对现有方法进行了大量评

测和深入分析。

2）针对视频显著性物体检测中注意力转移的问题，构建了第一个高质量、稠密标注的视频显著性物体检测（DAV-SOD）数据集；提出了基于注意力转移的 SSAV 模型，取得了国际领先的检测性能；提供了当前最大规模、最完整的视频显著性物体评测结果。

3）针对非二进制显著性物体检测质量评价的问题，提出了符合人类认知规律的度量指标 S-measure，使得评价方法从像素级过渡到结构级，特别是将与人的主观评价相一致的性能从 23%提升到了 77%。

4）针对二进制显著性物体检测质量评价的问题，提出了符合人类认知规律的度量指标 E-measure，使得评价方法在一个紧凑项中同时考虑了全局和局部信息，上述方法的性能比国际最先进算法提高了 19%。

关键词：显著性物体检测；评价指标；数据集；视频显著性；图像显著性

ABSTRACT

Salient Object Detection (SOD) originates from the cognitive studies of human visual attention behavior, *i. e.*, the astonishing ability of the human visual system to quickly orient attention to the most informative parts of visual scenes and ignore the other parts. SOD is thus significantly instrumental to a wide range of real-world applications, *e. g.*, autonomous driving, robotic interaction, video segmentation, video captioning, and video compression. Besides its academic value and practical significance, SOD presents great difficulties due to the challenges carried by video data (*e. g.*, occlusions, blur, large object-deformations, diverse motion patterns) and the inherent complexity of human visual attention behavior (*i. e.*, selective attention allocation, attention shift) during dynamic scenes. Subject to the limitation of the acquisition device, the early build salient object detection datasets do not represent the real scene well. Moreover, the evaluation metrics in this field ignore the properties of the human visual system and are all based on pixel-level error. The above problems have seriously restricted the development of salient object detection technology.

This dissertation is based on the cognitive theory and focuses on image and video salient object detection, the research directions including the collection of the dataset, the creation of the models, and the design of evaluation metrics. The major contributions of the dissertation are:

1) My analysis points out various serious data biases in existing SOD datasets. I built a new SOD dataset, called SOC which contains diverse contexts in a realistic environment. Then, a set of attributes (*e. g.*, Appearance Change) is proposed in an attempt to obtain a deeper insight into the SOD problem. I also present the currently largest scale performance evaluation of CNNs based SOD models.

2) To further advance the research of the saliency-shift issue, I elaborately collected a high-quality Densely Annotated Video Salient Object Detection (DAVSOD) dataset. The proposed SSAV model performs better against other top competitors over the five large-scale datasets. To further contribute to the community with a complete and the largest-scale benchmark, I systematically assess several representative video salient object detection algorithms.

3) To address the evaluation problem of the non-binary map, I propose a structure similarity-based SOD measure, called S-measure. Rather than based on pixelwise error, the new measure is

based on structural similarity. Especially, the performance of human consistency has improved from 23% to 77%.

4) I propose a novel and effective Enhanced-alignment measure (E-measure) for binary salient object detection map. The motivation from the cognitive vision studies which have shown that human vision is highly sensitive to both global information and local details in scenes. Thus, the new measure achieve the largest improvement of 19% compared with other popular measures in terms of specific meta-measures.

Key Words: Salient Object Detection (SOD); evaluation metric; dataset; video saliency; image saliency

目 录

第 3 章

富上下文环境下的显著性物体检测数据集与评测

本章主要研究富上下文环境下的显著性物体检测数据集问题。3.1 节介绍背景知识、研究动机及解决方案概要；3.2 节介绍构建的开放环境下的显著性物体检测数据集；3.3 节给出评测结果和结果分析；3.4 节对本章进行小结。

3.1 引言

3.1.1 背景知识

在引入**显著性物体检测**这个概念之前，先要简单介绍一下人眼与生俱来的一个重要机制——视觉注意力机制。人类经过漫长的进化，其视觉系统形成了一种能够对外界信息进行选择性处理的功能。简而言之，人类面对场景时会将视觉注意力分配给那些更加重要的区域，而有选择地忽略其他区

域[1-2]。早期研究阶段，不同学者为重要区域赋予了各种名词，如感兴趣区域（region of interest）、重要性区域（important region）和显著性区域（saliency region）。对于静态图像来说，没有给观察者特定暗示但能引起观察者注意的刺激方式称为被动注意，它具有“自底向上、数据驱动、任务无关”的特性。而对于带任务的刺激信号，比如暗示受试者搜索场景中某些特定对象，这种情况称为主动注意，它具有“自顶向下、目标驱动、任务相关、慢速”的特点。

从1998年美国加州理工大学的Laurent Itti等人[3]提出第一个基于生物启发的模型到2019年南开大学Zhao等人[4]提出最新的基于深度学习技术的检测模型，显著性物体检测有了长足的发展。早期阶段（1998—2012年）的工作经常被称为显著性检测，这一阶段工作的特点是集中于视点预测（fixation prediction），旨在利用计算机模型预测出与人眼注视点相一致的区域。2012年，Cheng等人[5]提出的全局显著性物体检测模型改变了传统视点预测的方向，该工作旨在定位并分割出显著的对象而不是区域。理由是，在现实世界中，人们经常需要对图像进行编辑，此时，以对象的方式来处理图像中的元素将更加自然且高效。此后，显著性物体检测（salient object detection）与显著性检测（fixation detection）齐头并进。随着深度学习技术的兴起，这一领域也从传统的启发式模型逐渐转变到以深度学习为主流的模型。1998—2015年出现了众多传统模型，模型性能的评估可以参考

Borji 等人的评测工作[6]。在静态图像上利用深度学习进行显著性物体检测始于 2015 年大连理工大学的 Wang 等人的工作[8]。从 2015 年至今，基于深度学习的显著性物体检测模型的详细评测可以参考南开大学 Fan 等人的最大规模评测工作[9]。

3.1.2 研究动机

本章的工作主要受到两个观察的启发。首先，现有的显著性物体检测（SOD）数据集[5,33,35-40,111-112] 在数据收集过程或数据质量方面存在缺陷。具体而言，大多数数据集假设图像至少包含一个显著物体，因此它们丢弃了不包含显著物体的图像，我们称为**数据选择偏见**。此外，现有数据集主要包含具有单个物体的图像或简单环境中的多个物体（而且通常以人为主）。这些数据集不能充分反映现实场景的复杂性，因为现实世界的场景经常是杂乱的，包含多个物体。这就导致在现有数据集上训练的、表现最佳的模型几乎达到了饱和的性能（例如在大多数数据集上，模型性能 F-measure > 0.9），但它们在现实场景中的表现却无法令人满意（例如表 3.1 中 F-measure<0.45）。这是因为在之前数据集上训练出来的模型更加偏向较为理想的场景，所以当它们应用于现实世界中的场景时，其有效性可能会受到极大削弱。为了解决该问题，有必要构建更接近实际条件的数据集。

其次，在当前的数据集上只能分析模型的整体性能，这

些数据集都缺乏反映现实场景中所面临挑战的各种属性。因此，引入这些属性有助于①更深入地了解 SOD 问题，②研究 SOD 模型的优缺点，③从不同的角度客观地评价模型的性能，对不同的应用场景，其评价结果可能是不同的。

3.1.3 解决方案概要

针对上述两个问题，作者做了两个贡献。第一个贡献是构建了一个新的高质量的 SOD 数据集，将其命名为 SOC（Salient Objects in Clutter）。迄今为止，SOC 是最大的实例级 SOD 数据集，它包含来自 80 多个常见类别的 6 000 张图像。它与现有数据集的不同之处在于 3 个方面：①显著物体具有类别注释，可用于诸如弱监督 SOD 任务之类的新型研究；②包含非显著图像，使该数据集更接近真实世界场景，并且比现有数据集更具挑战性；③显著物体具有反映真实世界中面临的特定情况的属性，例如运动造成的模糊、遮挡和杂乱的背景。因此，SOC 数据集缩小了现有数据集与现实世界场景之间的差异，并提供了更合理的基准测试（如图 3.1 所示）。

此外，本章针对几种最先进的卷积神经网络（CNN）模型进行了综合评估[8,40,48-58]。为了评估模型性能，作者引入了 3 个评估指标来度量检测结果的区域相似性、分割的像素精度以及结果的结构相似性。此外，作者还提供了基于属性的性能评估。这些属性使得更深入地理解模型成为可能并且进

图 3.1　SOC 数据集中的样本图像包括非显著物体图像（第 1 行）**和显著物体图像**（第 2~4 行）（见彩插）

注：对于显著物体图像，作者提供了实例级真值图（不同颜色表示不同实例）、物体属性和类别标签。

一步指出了具有潜力的研究方向。作者相信，该数据集和基准测试会对未来的 SOD 研究，特别是对于面向应用的模型开发产生非常大的影响。完整的数据集和分析工具详见作者主页（https://dengpingfan.github.io/）。

3.2　SOC 数据集

本节将介绍本书构建的旨在反映真实世界场景的、具有

挑战性的 SOC 数据集。来自 SOC 的样例图像如图 3.1 所示。此外，关于 SOC 的类别和属性的统计信息分别如图 3.5 和图 3.7 所示。基于对现有数据集优缺点的分析，作者明确了全面和平衡的数据集应该满足七个重要方面。

3.2.1 存在非显著物体

几乎所有现有的 SOD 数据集都假设图像至少包含一个显著物体并丢弃了不包含显著物体的图像。但是，这种假设是导致**数据选择偏见**的过于理想化的设定。在真实场景的设定中，图像并不总是包含显著物体。一些背景图像中无特定形状的物体，如天空、草地和纹理等根本不包含显著的物体[113]。非显著物体或背景“元素”可能占据整个场景，因此严重限制了显著物体的可能位置。Xia 等人[41] 通过判断什么是显著物体和什么不是显著物体，提出了先进的 SOD 模型，说明非显著物体对推理显著物体至关重要。这表明非显著物体应该和显著物体受到同等重视。包含一定数量的非显著物体图像会使得数据集更接近真实场景，同时也使得 SOD 任务变得更具挑战性。因此，作者将“非显著物体”定义为没有显著物体的图像或具有“元素”性质的图像。如文献［41，113］中所述，“元素”类别包括密集分布的相似物体、形状模糊的区域和没有语义的区域，分别如图 3.2a～图 3.2c 所示。

a）密集分布的相似物体　b）形状模糊的区域　c）没有语义的区域

图 3.2　一些非显著图像的示例（更多示例请见图 3.3）

基于非显著物体的定义，作者从 DTD[114] 数据集中收集了 783 张纹理图像。为了增强数据集的多样性，作者又从互联网和其他数据集中收集了 2 217 张图像，包括极光、天空、人群、商店和许多其他类型的真实场景[7,39,45,111]。相信纳入足够多的非显著物体会为未来的研究工作开辟一个有希望的方向。

3.2.2　图像的数量和类别

相当数量的图像对于捕捉现实世界场景的多样性和丰富性至关重要。此外，大量的数据可以让 SOD 模型避免过拟合并增强其泛化能力。为此，作者收集了来自 80 多个类别（典型的类别见图 3.4）的 6 000 张图像，其中包含 3 000 张带有显著物体的图像和 3 000 张没有显著物体的图像。作者将数据集分为训练集、验证集和测试集，比例为 6∶2∶2。为确保公平性，测试集通过网站提供在线测试㊀。图 3.5a 展示了每个类别的显著物体的数量。它表明“person（人物）”

㊀ https://github.com/DengPingFan/SODBenchmark。

类别占很大比例，这是合理的，因为人通常与其他对象一起出现在日常场景中。

图 3.3 SOC 数据集中不包含显著物体的图片示例

注：完整的数据集请查阅项目主页：https://github.com/DengPingFan/SODBenchmark。

图 3.4　SOC 数据集中包含实例级显著对象的示例

注：完整的数据集请查阅项目主页：https://github.com/DengPingFan/SODBenchmark。

3.2.3　显著物体的全局/局部颜色对比

如文献［39］所述，术语“显著”与前景和背景的全局/局部对比度有关。因此，检查显著物体是否易于检测是非常重要的。首先，分别计算每个物体前景和背景的 RGB 颜色直方图。然后，利用 χ^2 距离来测量两个直方图之间的距

离。全局和局部颜色对比度分布分别如图 3.5b 和图 3.5c 所示。与 ILSO 数据集相比，本书构建的 SOC 数据集包含了更多低全局颜色对比度和局部颜色对比度的物体。

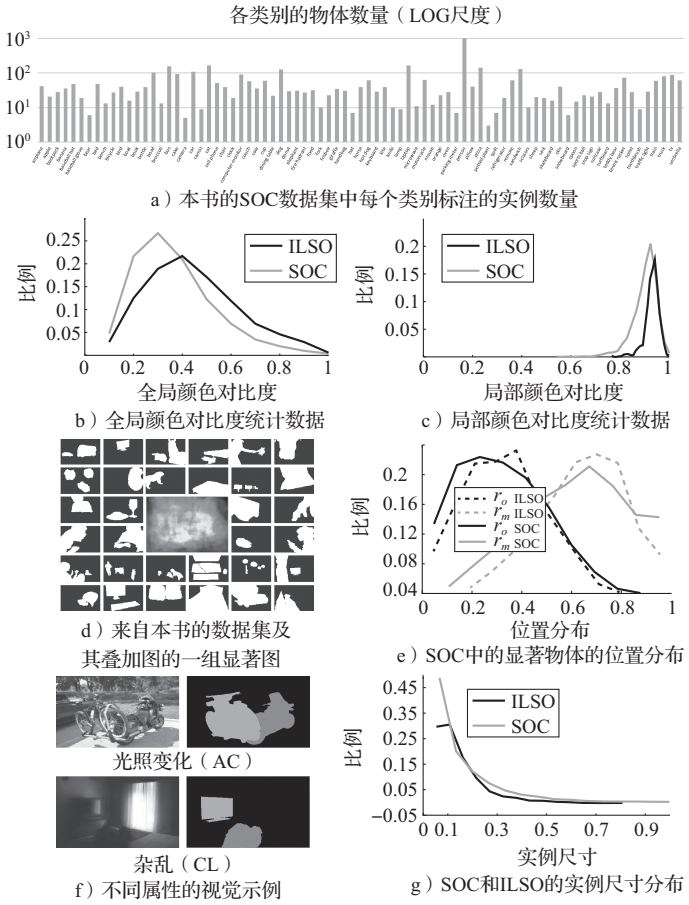


图 3.5　SOC 数据集统计信息、属性示例以及真值图示例

3.2.4 显著物体的位置

中心偏见被认为是显著性检测数据集中影响最大的偏见之一[6,11,39]。图 3.5d 展示出了一组图像及其叠加图。可以看出，虽然显著的物体位于不同的位置，但是叠加图仍然表明这组图像是存在中心偏见的。以前的基准测试通常采用这种不准确的方式来分析显著物体的位置分布。为了避免这种误导现象，作者绘制了图 3.5e 中两个量 r_o 和 r_m 的统计情况，其中 r_o 和 r_m 分别表示物体中心和物体中最远（边缘）点离图像中心有多远。将 r_o 和 r_m 除以图像对角线长度的一半以进行归一化，使得 r_o，$r_m \in [0, 1]$。从这些统计数据中可以观察到 SOC 数据集中的显著物体受中心偏见影响的情况。

3.2.5 显著物体的大小

每个显著物体实例的大小被定义为物体面积占图像总面积的比例[39]。如图 3.5g 所示，与仅有的实例级 ILSO 数据集[44] 相比，SOC 中显著物体的大小的变化范围更广泛。此外，SOC 包含了更多中等尺寸的物体。

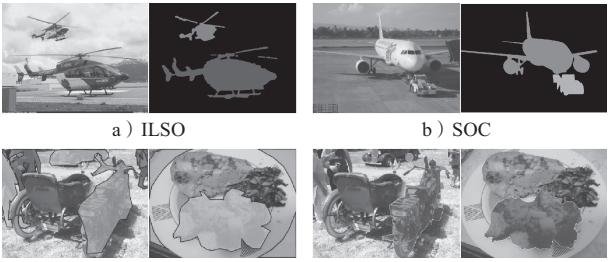
3.2.6 高质量的显著对象标签

文献［55］的实验显示，模型在 ECSSD 数据集（具有 1 000 张图像）上训练会比在其他数据集（例如 MSRA10K，具有 10 000 张图像）上训练获得更好的泛化性能。这表明除

了规模之外，数据集质量也是一个重要因素。为了获得大量高质量的图像，作者从 MSCOCO 数据集[45] 中随机选择图像，这是一个大型的真实世界数据集，其中的物体用多边形标注（例如粗略标注）。高质量标注在提高 SOD 模型的准确性方面也起着关键作用[34]。为此，作者使用逐像素的标注来重新标记数据集。类似于著名的 SOD 任务导向基准测试数据集[5,33-35,37,40-44,111]，本书没有使用眼动仪设备，而是采取了多个步骤来提供高质量的注释。这些步骤包括两个阶段：①要求 5 名观众使用标定框标记他们认为的每张图像中较为显著的物体；②保留大多数观众（≥3）在显著性上意见相同的物体（标定框的 IOU>0. 8）。在第一阶段之后，得到 3 000 张用标定框标注的显著物体图像；在第二阶段，根据标定框的提示进一步手工标记显著物体的逐像素轮廓。请注意，有 10 名志愿者参与了整个过程以交叉检查标注的质量。最后，作者保留了 3 000 张具有高质量的实例级标记显著物体的图像。如图 3. 6b 和图 3. 6d 所示，本书的物体边界的标注是精确、清晰和平滑的。在标注过程中，作者还添加了一些未在 MSCOCO 数据集中标记的新类别[45]（例如计算机显示器、帽子和枕头）。

3. 2. 7 具有属性的显著对象

数据集中图像的属性信息有助于研究者客观评估模型在不同类型的参数上的性能，它还允许对模型失败情况进行检



a）ILSO　　b）SOC

c）MSCOCO　　d）SOC

图 3.6　数据集标注质量对比（与最近提出的数据集的比较），**本书的 SOC 数据集中的标注边界更平滑，质量更高**

查。为此，作者定义了一组属性来表示在真实场景中面临的特定情况，例如运动模糊、遮挡和杂乱的背景（见后文表 3.2 中的总结）。因为这些属性不是独占的，所以一张图像可以使用多个属性进行标注。

受文献［31］的启发，图 3.7 左展示了数据集图片属性的分布情况。SO 类型具有最大比例是因为精确的实例级（例如图 2.1 中的网球拍）的标注。因为现实世界的场景由不同视觉特色的材料组成，所以 HO 类型占很大比例。MB 类型在视频帧中比静态图像更常见，但有时也会出现在静态图像中。因此，MB 类型在本书的数据集中占比相对较小。由于真实图像通常包含多个属性，因此作者根据出现的频率展示了属性之间的主要依赖关系（如图 3.7b 所示）。例如，包含许多异构物体的场景可能具有大量彼此阻挡并形成复杂空间结构的物体。因此，HO 类型与 OC 类型、OV 类型和 SO

类型都具有强依赖性。

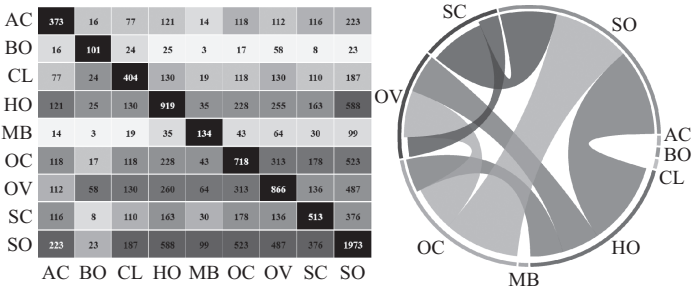


图 3.7 属性分布与依赖关系 a）SOC 数据集中显著图像的属性分布，网格中的每个数字表示图像的出现次数。b）基于出现频率绘制的属性之间的主要依赖关系，属性之间依赖度越高则宽度越大

3.3 基于深度学习的显著性检测模型评测结果

本节中，作者在 SOC 数据集上呈现了 16 个 SOD 模型的评估结果。几乎对所有基于卷积神经网络的代表性 SOD 模型都进行了评估。但是，由于某些模型的代码不公开，因此对此类模型不予考虑。此外，大多数模型都没有针对非显著物体检测进行优化。为了公平起见，作者只使用 SOC 数据集的测试集来评估 SOD 模型。本书在 3.3.1 节中描述了评估指标。SOC 数据集的整体模型性能见 3.3.2 节和表 3.1，而针对各个属性的性能评估结果（例如光照变化属性上的性能

表 3.1　三种指标下 SOD 模型的性能

指标	单任务													多任务		
	LEGS [8]	MC [48]	MDF [40]	DCL [49]	AMU [49]	RFCN [52]	DHS [51]	ELD [50]	DISC [53]	IMC [54]	UCF [58]	DSS [55]	NLDF [56]	DS [59]	WSS [42]	MSR [44]
$F_{all}\uparrow$	0.276	0.291	0.307	0.339	0.341	**0.435**	0.360	0.317	0.288	0.352	0.333	0.341	0.352	0.347	0.327	**0.380**
$S_{all}\uparrow$	0.677	0.757	0.736	0.771	0.737	0.814	0.804	0.776	0.737	0.664	0.657	0.807	**0.818**	0.779	0.785	**0.819**
$\varepsilon_{all}\downarrow$	0.230	0.138	0.150	0.157	0.185	0.113	0.118	0.135	0.173	0.269	0.282	0.111	**0.104**	0.155	0.133	0.113

注：F 代表区域相似性，ε 是平均绝对误差，S 是结构相似性。↑代表数字越高越好，反之亦然。根据式（3.3）在 SOC 数据集上通过计算得出评估结果。S_{all}、F_{all}、ε_{all} 分别表示用 S、F、ε 指标来表示的整体性能表现。加粗表示最好成绩。

表现）见3.3.3节和表3.3。作者公开了评估脚本并且在网站上提供了在线评估。

3.3.1 评估指标

在强监督评估框架中，给定由SOD模型生成的预测图 M 及人工标注 G，研究者寄希望于通过评估指标预测出究竟哪一种模型能够生成最佳结果。本书在SOC数据集上使用3种不同的评估指标来评估SOD模型。

逐像素精度 ε。区域相似性评估方法忽略了背景中的显著性分布。作为补充，作者采用 M 和 G 之间的归一化（[0, 1]）后的平均绝对误差（MAE）作为指标，其定义为：

$$\varepsilon = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H} \| M(x,y) - G(x,y) \| \qquad (3.1)$$

其中 W 和 H 分别是图像的宽度和高度。

区域相似性 F。为了测量两张图片各区域相匹配的程度，作者使用F-measure，该方法定义如下：

$$F = \frac{(1+\beta^2)\,\text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}} \qquad (3.2)$$

其中 $\beta^2 = 0.3$ 由文献［34］提出并用于平衡召回率和精度。然而，在计算召回率和精度时，F-measure并没有考虑到手工标注值全为0的特殊情况，因此，不同的前景图会得到相同的结果（0），而这显然是不合理的。从而可以得出结论，

F-measure 不适合评估非显著物体检测的结果。

综上，ε 和 F 这两个指标都是基于逐像素的计算方式，因此经常忽略结构相似性。行为视觉研究表明人类视觉系统对场景结构非常敏感[65]，在许多应用场景中都对能够保留物体结构的显著性检测模型较为青睐。

结构相似性 S。Fan 等人[65] 提出的 S-measure 同时考虑局部区域（region）和全局对象（object）两个层次上的相似性来评估检测结果的结构相似性。因此，作者也使用 S-measure 来评估 M 和 G 之间的结构相似性。需要特别注意的是，接下来整体性能表现的评估和分析都是基于 S-measure 进行的。

3.3.2 指标统计

为了获得整体结果，作者对评估指标的分数取平均值。设 $\eta \in \{F, \varepsilon, S\}$，其公式为：

$$M_\eta(D) = \frac{1}{|D|} \sum_{I \in D} \overline{\eta}(I_i) \tag{3.3}$$

其中 $\overline{\eta}(I_i)$ 是模型在图像数据集 D 中图像 I_i 上的评估得分。

单任务：对于单任务模型，在整个 SOC 数据集上性能表现（表 3.1 中的 S_{all}）最佳的模型是 NLDF[56]（$M_S = 0.818$），其次是 RFCN[52]（$M_S = 0.814$）。MDF[40] 和 AMU[57] 使用边缘线索来提升显著图提取的准确度但却未能达到理想的目标。为了使用图像的局部区域信息，MC[48]、MDF[40]、ELD[50]

和 DISC[53] 尝试使用超像素方法将图像分割成数个区域，然后从这些区域中提取特征，但这是较为复杂而耗时的。为了进一步提高性能，UCF[58]、DSS[55]、NLDF[56] 和 AMU[57] 利用全卷积网络来改善 SOD 模型的性能（表 3.3 中的 S_{sal}）。其他一些方法诸如 DCL[49] 和 IMC[54] 则尝试将超像素与全卷积网络结合起来构建一个强大的模型。此外，RFCN[52] 将包括边缘和超像素的两个相关线索组合到全卷积网络中，进而在整个数据集上获得了良好的性能（$M_F = 0.435$，$M_S = 0.814$）。

多任务：与上述模型不同，MSR[44] 使用三个密切相关的步骤去检测实例级显著物体——估计显著图、检测显著物体轮廓以及识别显著物体的实例。它创建了一个多尺度显著性检测网络，可以实现最高性能（S_{all}）。其他两个多任务模型 DS[59] 和 WSS[42] 同时利用分割和分类结果生成显著图从而获得适度的性能提升。值得一提的是，尽管 WSS 是一种弱监督的多任务模型，但它仍然可以获得与其他全监督的单任务模型相当的性能。因此，基于弱监督和多任务的模型可能是未来的研究方向。

3.3.3 基于属性的评估

如表 3.2 所示，作者为显著图像分配了属性。每个属性代表在现实世界场景的显著性检测中存在的挑战性问题。这

些属性可区分出具有主导性特征（例如杂乱属性的存在）的图像集合，它们对于解释 SOD 模型的性能以及将 SOD 与面向应用的任务相关联均是非常重要的。例如，Sketch2photo 应用[115] 青睐在大物体上具有良好性能的模型，而这可以通过基于属性的性能评估方法来辨别。

表 3.2　显著物体图像的属性列表和相应描述

属性名称	属性描述
AC	**光照变化**：物体区域中出现了明显的光照变化
BO	**大物体**：物体面积和图像面积的比值大于 0.5
CL	**杂乱**：物体周围的前景和背景区域具有相似的颜色，将全局颜色对比度值大于 0.2、局部颜色对比度值小于 0.9 的图像标记为杂乱图像
HO	**异构物体**：由视觉上独特/不相似的部分组成的物体
MB	**运动模糊**：由于相机或运动的抖动，物体具有模糊的边界
OC	**遮挡**：物体被部分或全部遮挡
OV	**超出视野**：物体的部分区域超出了图像边界
SC	**形状复杂性**：物体有纤细组件之类的复杂边界，比如动物的脚
SO	**小物体**：物体面积和图像面积的比值小于 0.1

注：通过观察现有数据集的特征，作者总结了这些属性。一些视觉示例可以在图 3.1 和图 3.8 中找到。更多示例，请参阅作者主页。

结果：在表 3.3 中，本书展示了各种 SOD 模型在特定属性表征的数据子集上的性能。接下来，作者选择一些代表性属性进行进一步分析。

大物体（BO）：当物体与相机距离很近时，经常会出现大物体场景，因此在图片中可以清楚地看到微小的文字或图

表 3.3　在 SOC 显著性物体子数据集上基于属性的性能表现

属性	单任务													多任务		
	LEGS [8]	MC [48]	MDF [40]	DCL [49]	AMU [49]	RFCN [52]	DHS [51]	ELD [50]	DISC [53]	IMC [54]	UCF [58]	DSS [55]	NLDF [56]	DS [59]	WSS [42]	MSR [44]
S_{sal}	0.607	0.619	0.610	0.705	0.705	0.709	**0.728**	0.664	0.629	0.679	0.678	0.698	0.714	0.719	0.676	**0.748**
AC	0.625	0.631	0.614	0.734	0.736	0.744	**0.745**	0.673	0.644	0.702	0.714	0.726	0.737	0.764	0.691	**0.789**
BO	0.509	0.490	0.461^{-}	0.610	0.569	0.540	0.590	0.576	0.517	$\mathbf{0.701}^{+}$	0.636	0.496^{-}	0.568	0.685	0.566	0.667
CL	0.620	0.635	0.566	0.699	0.708	0.714	**0.743**	0.658	0.635	0.696	0.704	0.677^{-}	0.713	0.729	0.678	**0.756**
HO	0.666	0.666	0.648	0.745	0.755	0.759	**0.766**	0.706	0.681	0.715	0.744	0.748	0.755	0.756	0.707	**0.777**
MB	0.543^{-}	0.603	0.615	0.693	0.706	0.715	**0.722**	0.639	0.600	0.689	0.682	0.695	0.685	0.711	0.641	**0.757**
OC	0.609	0.617	0.608	0.708^{+}	$\mathbf{0.725}^{+}$	0.711	0.716	0.658	0.630	0.672	0.701^{+}	0.689	0.709	0.725^{+}	0.672	**0.740**
OV	0.548	0.584	0.568	0.699	$\mathbf{0.708}^{+}$	0.687	0.706	0.637	0.573	0.693^{+}	0.685^{+}	0.665	0.688	0.722^{+}	0.624	**0.743**
SC	0.608	0.620	0.669^{+}	0.738	0.731	0.735	**0.763**	0.688	0.653	0.690	0.722^{+}	0.746^{+}	0.745	0.724	0.677	**0.773**
SO	0.573^{-}	0.601	0.621	0.691	0.685	0.698	**0.713**	0.644	0.614	0.648^{-}	0.650	0.696^{-}	0.703	0.696	0.659	**0.730**

注：对于每一个模型，其分数对应于在特定属性的所有测试图像上的结构相似性 M_S（见 3.3.1 节）的平均值，分数越高，性能表现越好，加粗表示最高成绩，平均显著物体检测性能 S_{sal} 在第 1 行通过结构相似性 S 呈现，+和−分别表示与平均值相比之下的性能增加和减少。

图 3.8　SOC 数据集中标注不同属性的图片示例
（完整的数据集见作者主页）

案。然而在这种情况下，倾向于关注局部信息的模型将被严重误导，导致较大的性能损失（例如，DSS[55] 损失了28.9%的性能，MC[48] 损失了20.8%的性能以及RFCN[52] 损失了23.8%的性能）。然而，IMC[54] 模型的性能表现略微上

升了3.2%。在深入了解该模型的流程后，作者得出了一个可能的解释，即IMC使用粗略预测图来表达语义，并利用过度分割的图像来补充结构信息，从而在BO类型的图像上获得了令人满意的结果。但是，过度分割的图像无法弥补缺失的细节，因此会导致此类模型在SO类型的图像上的性能下降4.6%。

小物体（SO）：对于所有SOD模型来说，识别SO类型的图像是一个较为棘手的问题。在此类图像上，所有模型都遇到了性能下降（例如，DSS[55]的性能下降了0.3%，LEGS[8]的性能下降了5.6%），因为在卷积神经网络的下采样期间很容易忽视小物体。DSS[55]是唯一一个在SO类型图像上性能仅略微下降的模型，而它在BO类型图像上的性能损失最大（28.9%）。MDF[40]使用多尺度超像素图像作为网络的输入，它能够很好地保留小物体的细节。然而，由于超像素的大小有限，MDF仍无法有效地感知全局语义，导致其在BO类型图像上出现大的识别失败概率。

遮挡（OC）：在遮挡场景中，物体被部分遮挡。因此，SOD模型需要捕获全局语义以弥补不完整的物体信息。为此，DS[59]和AMU[57]利用下采样过程中的多尺度特征生成融合显著图，UCF[58]提出了一种模糊的学习机制来学习不确定的卷积特征。所有这些方法都试图获得包含全局和局部特征的显著图。不出所料，这些方法在OC类型的图像上取得了相当不错的效果。基于上述分析，作者还发现这三个模

型在需要更多语义信息的场景上的性能表现非常好，如 AC、OV 和 CL 类型。

异构物体（HO）：该类型场景在现实生活中很常见。在 HO 类型的图像上，不同模型的性能分别比其在总数据集上的平均性能有所提升，基本在 3.9%～9.7%之间波动。作者认为这是因为 HO 类型在数据集中占据较大的比例，从而使得 SOD 模型过拟合这种类型。图 3.7 中的统计结果在一定程度上符合这样的结论。

3.4 讨论和结论

据作者所知，这项工作是目前最大规模的、针对卷积神经网络的显著性物体检测模型的性能评估方案。作者的分析指出了现有 SOD 数据集中存在严重的**数据选择偏见**，这种设计偏见使得最先进的 SOD 算法在现有数据集上几乎达到了饱和的性能。然而，在真实场景中，其效果仍远不能令人满意。基于本书的分析，作者确定了全面和平衡的数据集应该满足的 7 个重要方面。作者首先构建了高质量的 SOD 数据集 SOC，它包含来自日常生活的、自然环境中的、更接近真实环境的显著物体图像。SOC 数据集将随着时间的推移不断发展和增长，并将在多个方向上拓宽研究的可行性，例如显著物体的感数[116]、实例级显著性物体检测[44]、基于弱监督的

显著对象检测[42] 等。其次，为了更深入地了解 SOD 问题，作者研究了 SOD 算法的优缺点，并在不同的观点和要求下客观地评估模型性能，作者还提出了一组属性（例如外观变化）。最后，作者在 SOC 数据集上对现有 SOD 模型进行了基于属性的性能评估，评估的结果为未来的模型开发和模型评测开辟了充满希望的新方向。