软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(4):1711−1731 [doi: 10.13328/j.cnki.jos.006700]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

深度学习时代下的 RGB-D 显著性目标检测研究进展*

丛润民[1,2], 张 晨[1,2], 徐 迈[3], 刘鸿羽[1,2], 赵 耀[1,2]

[1](北京交通大学 信息科学研究所, 北京 100044)

[2](现代信息科学与网络技术北京市重点实验室, 北京 100044)

[3](北京航空航天大学 电子信息工程学院, 北京 100191)

通信作者: 徐迈, E-mail: maixu@buaa.edu.cn

摘 要: 受人类的视觉注意力机制启发, 显著性目标检测任务旨在定位给定场景中最吸引人注意的目标或区域. 近年来, 随着深度相机的发展和普及, 深度图像已经被成功应用于各类计算机视觉任务, 这也为显著性目标检测技术提供了新思路. 通过引入深度图像, 不仅能使计算机更加全面地模拟人类视觉系统, 而且深度图像所提供的结构、位置等补充信息也可以为低对比度、复杂背景等困难场景的检测提供新的解决方案. 鉴于深度学习时代下 RGB-D 显著目标检测任务发展迅速, 旨在从该任务关键问题的解决方案出发, 对现有相关研究成果进行归纳、总结和梳理, 并在常用 RGB-D SOD 数据集上进行不同方法的定量分析和定性比较. 最后, 对该领域面临的挑战及未来的发展趋势进行总结与展望.

关键词: 显著性目标检测; RGB-D 图像; 跨模态信息交互; 深度质量感知

中图法分类号: TP311

中文引用格式: 丛润民, 张晨, 徐迈, 刘鸿羽, 赵耀. 深度学习时代下的 RGB-D 显著性目标检测研究进展. 软件学报, 2023, 34(4): 1711−1731. http://www.jos.org.cn/1000-9825/6700.htm

英文引用格式: Cong RM, Zhang C, Xu M, Liu HY, Zhao Y. Research Progress of RGB-D Salient Object Detection in Deep Learning Era. Ruan Jian Xue Bao/Journal of Software, 2023, 34(4): 1711−1731 (in Chinese). http://www.jos.org.cn/1000-9825/6700.htm

Research Progress of RGB-D Salient Object Detection in Deep Learning Era

CONG Run-Min[1,2], ZHANG Chen[1,2], XU Mai[3], LIU Hong-Yu[1,2], ZHAO Yao[1,2]

[1](Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China)

[2](Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China)

[3](School of Electronic and Information Engineering, Beihang University, Beijing 100191, China)

Abstract: Inspired by the human visual attention mechanism, salient object detection (SOD) aims to detect the most attractive and interesting object or region in a given scene. In recent years, with the development and popularization of depth cameras, depth map has been successfully applied to various computer vision tasks, which also provides new ideas for the salient object detection task at the same time. The introduction of depth map not only enables the computer to simulate the human visual system more comprehensively, but also provides new solutions for the detection of some difficult scenes, such as low contrast and complex backgrounds by utilizing the structure information and location information of the depth map. In view of the rapid development of RGB-D SOD task in the era of deep learning, this studyaims to sort out and summarize the existing related research outputs from the perspective of key scientific problem solutions, and conduct the quantitative analysis and qualitative comparison of different methods on the commonly used RGB-D SOD datasets.

* 基金项目: 中央高校基本科研业务费专项资金(2022JBMC002); 北京市自然科学基金(4222013, JQ20020); 科技创新 2030——"新一代人工智能"重大项目(2021ZD0112100); 北京市科技新星计划(Z201100006820016); 国家自然科学基金(62002014, U1936212, 62120106009, 61922009, 61876013, 62050175); 中国科协青年人才托举工程(2020QNRC001); 北京市科协青年人才托举工程

收稿时间: 2021-09-23; 修改时间: 2022-02-05; 采用时间: 2022-04-14; jos 在线出版时间: 2022-07-22

Finally, the challenges and prospects are summarized for the future development trends.

Key words: salient object detection; RGB-D images; cross-modality information interaction; depth quality perception

人类视觉系统可以通过独特的感知单元和神经结构，从复杂的场景中快速关注到感兴趣的目标或区域，这被称为视觉注意力机制[1]. 一般来说，人类的视觉注意力包含了自底向上和自顶向下两条通路[2]：自底向上的注意力机制是由数据驱动的，是人脑对外界事物的非自主性感知；而自顶向下的注意力机制是由意识驱动的，是人脑主动下发指令去捕获感兴趣目标的过程. 受此启发，人们希望计算机也具有这种能力. 于是，显著性目标检测(salient object detection, SOD)任务应运而生，旨在让计算机自动地从给定场景中检测出最吸引人注意的区域或目标[3,4]，已经被广泛应用在了包括图像分割[5]、视频压缩[6]、目标检测[7]在内的大量计算机视觉任务中. 经过 10 余年的发展，显著性目标检测任务已经衍生出包括面向 RGB 图像的显著性目标检测[8,9]、面向高分辨率 RGB 图像的显著性目标检测[10]、面向 RGB-D 图像的显著性目标检测[11]、面向图像组的协同显著性目标[12]、面向 RGB-T 图像的显著性目标检测[13]、面向光场图像的显著性目标检测[14,15]、面向全景图像的显著性目标检测[16]、面向遥感图像的显著性目标检测[17]、面向视频序列的显著性目标检测[18]在内的众多分支.

虽然基于 RGB 图像的显著性目标检测已经取得了长足的发展，但是在某些低对比度、复杂背景场景下，单纯凭借 RGB 图像仍然不能取得令人满意的检测效果. 实际上，人眼不仅能够感知场景中的颜色、形状、纹理等外观信息，还可以通过双目视觉系统捕获场景的深度信息，形成立体感. 而深度图像则是深度信息的直观表现形式，其数值大小反映了场景中物体与深度传感器之间的距离. 通过引入深度图像到显著性目标检测任务，一方面可以模拟人类的立体视觉感知能力，加强计算机对不同目标之间的距离辨识度；另一方面，深度图像一些优良的特性(如目标内部一致性、形状先验等)也为解决某些困难场景的检测提供了新的思路. 近年来，深度相机的兴起(如 Kinect, RealSense)，使得深度图像的获取变得越来越便捷，这也为 RGB-D 显著性目标检测任务的兴起奠定了数据基础. 从 2012 年 Lang 等人[11]初探深度信息对视觉显著性的影响之后，基于 RGB-D 图像的显著性目标检测受到了越来越多的关注. 尤其是近几年来深度学习技术在该领域的成功应用，大量基于卷积神经网络的方法相继被提出. 图 1 统计了近 3 年来发表在部分顶级会议和期刊上的基于深度学习的 RGB-D 显著性目标检测论文情况，其反映了 RGB-D 显著性目标检测研究与日俱增的受关注程度. 因此，为了更全面地了解当前发展趋势、探明未来发展方向与研究重点，有必要对现有的基于深度学习的 RGB-D 显著性目标检测算法进行全面综述.

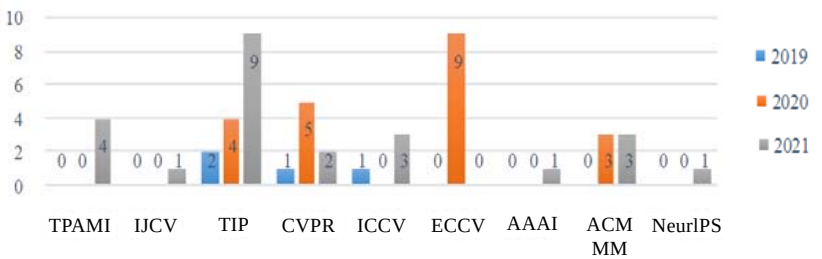

图 1　近 3 年 RGB-D 显著性目标检测领域的顶级学术论文发表情况

显著性目标检测技术发展的 20 年间，已经有一些学者撰写了中文或英文的综述论文对特定子领域的研究进展进行总结，其中，中文综述论文主要集中在 RGB 显著性目标检测领域[19]、协同显著性目标检测领域[20]、视频显著性目标检测领域[21]、全景图像显著性检测领域[22]和光场显著性目标检测领域[23]. 比较发现：针对 RGB-D 图像的显著性目标检测的中文技术综述仍是一个空白，这也是我们撰写本文的主要原因之一. 此外，与目前英文撰写的 RGB-D 显著性目标检测综述[24]相比，本文重点关注深度学习背景下的 RGB-D 显著性目标检测发展，并以一种新的视角重新审视当前发展状况. 具体不同点及创新点可概括为以下几个方面.

(1) 鉴于目前研究人员对深度图质量问题的关注越来越多，本文首次从 RGB-D 显著目标检测任务面临的跨模态特征融合和深度图质量感知两个关键问题入手进行文献综述，重新对现有的算法进行了

归纳总结, 并给出了一些关键性的思考;

(2) 对于跨模态融合问题, 与现有的根据融合位置划分的方式不同, 本文首先从网络结构出发, 按单流/双流/三流结构对现有方法进行了归纳, 尤其对于目前采用最为广泛的双流结构, 本文创新性地进一步划分为了等重要双向交互模式和深度辅助交互模式, 这将引导研究人员重新思考 RGB 和深度信息在跨模态交互中角色问题;

(3) 对于深度图质量感知问题, 本文将现有解决方案总结为深度质量评价和深度质量优化两类, 在归纳总结的同时, 也对该问题未来的发展进行了展望;

(4) 本文结合所阐述的两个重点问题, 重点对比了近期发表的 RGB-D 显著性目标检测模型, 以确定当前技术水平.

此外, 本文也总结了该领域常用的数据集(包含新提出的两个较大型数据集)及评价指标.

本文第 1 节针对 RGB-D 显著性目标检测的两个关键问题, 结合具体的解决方案, 对现有的基于深度学习的模型进行介绍. 第 2 节归纳总结 RGB-D 显著性目标检测常用的数据集及评价指标. 第 3 节对近期发表的模型进行了定量分析与定性比较. 最后, 在第 4 节中, 对该领域的挑战进行了总结,并对未来的发展进行了展望.

1 RGB-D 显著性目标检测算法

通过对现有方法的调研, RGB-D 显著性目标检测任务中的两个关键性问题可以总结为:

跨模态特征融合问题: RGB 图像和深度图像分属于两个不同的模态, 其中, RGB 图像反映了一幅场景的颜色、形状、边界、纹理等外观信息, 而深度图像描述的则是不同物体之间的景深差异及结构信息. 因此, 如何充分挖掘两种模态间的互补属性并充分利用它们各自的优势, 是需要着重解决的关键问题之一.

深度图质量感知问题: 目前, 深度传感器的种类繁多, 成像质量也参差不齐, 这会导致捕获的深度图像可能存在稀疏、孔洞、模糊等问题. 如果将这些低质量的深度图引入到显著性目标检测模型中,不仅难以起到应有的促进作用, 甚至会带来负面影响. 因此近几年来, 部分研究人员开始关注深度图的质量问题.

基于上述分析, 本文以这两个关键问题作为基本分类依据, 根据不同的解决方案, 对现有基于深度学习的 RGB-D 显著目标检测相关研究成果进行了归纳总结, 具体如图 2 所示, 下文将进行详细介绍.

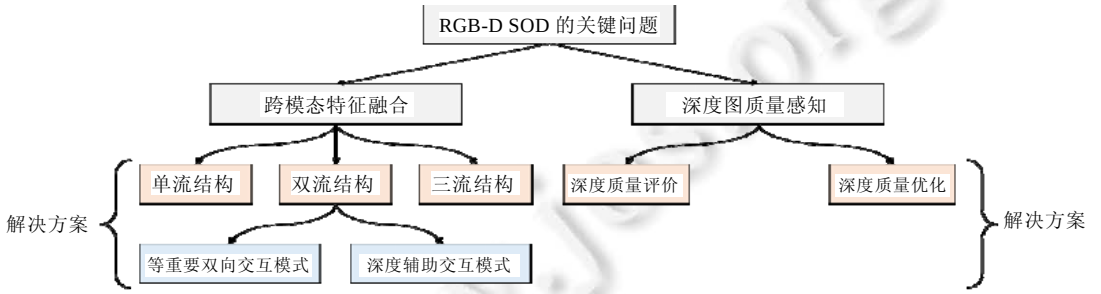

图 2 RGB-D 显著性目标检测任务的关键问题及解决方案

1.1 跨模态特征融合问题的解决方案

作为一个像素级的预测任务, 几乎所有基于深度学习的 RGB-D 显著性目标检测模型都采用了编码器-解码器的网络架构, 如图 3 所示. 其中, 编码器一般采用预训练的卷积神经网络(如 VGG[25], ResNet[26]等), 将输入图像映射到隐空间, 形成多层级视觉特征表示; 解码器则更多地关注到显著性导向的特征, 同时恢复图像分辨率, 得到最终的预测显著图. 目前, 大多数研究人员根据 RGB 图像和深度图像在编解码网络中融合的位置不同, 将现有的算法分为早期融合、中期/多尺度融合和晚期融合这 3 类[24,27,28]. 具体来说, 早期融合是在输入端融合, 一般指将 RGB 图像和深度图像直接组合为四通道作为输入; 中期/多尺度融合融是在编解码过程中融合不同模态、不同尺度的特征; 晚期融合是在输出端融合, 一般指多个输出显著图的融合. 本文创新性地

从网络结构形式和交互模式出发，首先将现有模型根据特征编码器的数量划分为单流结构、双流结构和三流结构；然后，将目前采用最为广泛的双流结构进一步根据 RGB 特征和深度特征(无特殊说明，本文中的深度分支、深度网络、深度特征中的“深度”均指代深度图像，而非传统意义上的深度神经网络)在网络中角色的不同，划分为深度辅助交互模式和等重要双向交互模式两种. 表 1 对现有模型的跨模态特征融合问题解决方案进行了梳理和归纳.

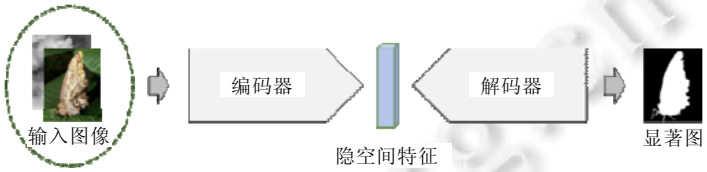

图 3 编码器-解码器网络示意图

表 1 跨模态特征融合问题解决方案总结

模型分类 \ 成果年份			2016	2017	2018	2019	2020	2021
单流结构			–	DF[29]	SPC[31]	SSRCNN[32], MLF[33]	DANet[30], DASNet[35], CoNet[37]	RD3D[34], UTANet[36]
双流结构	深度辅助交互模式	编码阶段辅助	–	–	–	PDNet[38], MMCI[39], CPFP[41]	CMWNet[40], ICNet[42]	–
		解码阶段辅助	–	–	–	–	S2MA[28], A2dele[43], PGAR[44], HDFNet[45]	MobileSal[46]
	等重要双向交互模式	对称式结构	DLFI[51]	MV-CNN[52], M^{3}Net[62]	PCF[59], ACCF[61]	TSRN[50], TANet[54], DCFF[55], AFNet[64], CAFM[65]	JL-DCF[47,48], CoCNN[58], FRDT[60], GFNet[63], GAS-GNN[68], CmSalGAN[69], DCMF[70]	MCINet[49], SwinNet[53], BiANet[56], ASIFNet[57], BTSNet[66], DSNet[67], SPNet[71], EBFNet[72], CMINet[73]
		非对称式结构	–	–	–	DSD[76] DMRA[77]	ATSA[74], cmMS[78]	CDINet[75], CCAFNet[79], DSA2F[80]
三流结构			–	–	–	–	D3Net[81], TCSDN[82], DRLF[83]	–

1.1.1 单流结构

单流结构(如图 4 所示)指利用单个编码器提取图像特征的网络架构，常见的单流结构可以分为深度图输入形式和深度图监督形式两种.

(a) 深度图作为输入参与网络训练　　(b) 深度图作为监督参与网络训练

图 4 单流结构示意图

典型的深度图输入形式类似于早期融合方案，其直接将 RGB 图像和深度图像同时为作为网络输入,以单个编码网络同时处理两个模态的输入数据. 文献[29]是基于深度学习的 RGB-D 显著性目标检测的开篇之作，其核心贡献在于首次引入了深度学习技术来提取 RGB 图像和深度图像的高级特征表示,并学习它们的交互机制. 具体地，作者首先将输入图像划分为数个超像素区域并生成初级显著性向量；然后，通过单流深度卷积

神经网络自动学习模态间的交互机制以生成超特征表示; 最后, 结合 Laplacian 传播框架获得具备空间一致性的显著图. 通过引入卷积神经网络, 在很大程度上缓解了特征表达能力不足、跨模态交互不充分等问题, 为 RGB-D 显著性目标检测任务提供了新的思路. 但是由于它采用的是非端到端的优化方式, 在模型设计上也较为简单, 所以在性能上仍不令人满意. 最近, Zhao 等人[30]基于单流结构设计了一个实时的 RGB-D 显著性检测模型, 在处理 384×384 大小的输入图像时, 实现了 32FPS 的推理速度. 在该方法中, 单通道的深度图像和三通道 RGB 图像在通道维度上拼接后, 输入到预训练的 VGG 网络提取特征. 此外, 正如图 4(a)所展示的: 该工作创造性地将深度图像引入到了网络内部, 通过提出的轻量级深度增强双重注意力模块引导显著性目标的定位, 使得深度信息得到了更充分利用, 在网络性能和推理速度上都获得了较大提升. 此外, 文献[31–33]也采用了四通道 RGB-D 输入的单流结构同时提取颜色和深度特征, 然后通过空间一致性约束、多级特征融合、显著性对象感知的数据增强等方案, 进一步提高模型性能. 但是直接在输入层面上结合很容易带来信息利用不充分的问题, 为此, 文献[34]以配备 3D 卷积的骨干网络代替普通 2D 卷积的骨干网络作为特征编码器进行 RGB 和深度图的预融合, 然后通过含有多条特征回传路径的 3D 解码器进行深层次的特征融合, 在一定程度上改善了模态不一致性.

深度图监督形式是指利用深度图像作为监督信息, 使得网络在学习过程中逐步具备深度信息的感知能力, 并在网络测试阶段可省去深度图像输入, 这在一定程度上提升了模型的推理速度. 在文献[35]中, 作者首次提出了将深度图像作为监督信号而非直接输入到网络中的思想. 在具体实现上, 作者设计了一个深度感知的显著性目标检测框架. 一方面, 该框架通过多级深度感知正则化进行显著性特征优化; 另一方面, 深度信息也作为误差加权图来纠正分割过程. 通过以上设计, 所提出的网络仅用 36.68M 的参数量, 便达到了先进的性能水平. 与之类似, 文献[37]提出了一种协作学习框架, 以显著图、深度图和边缘图共同作为监督信息, 深度特性和显著性特性以互利的方式集成到高层特征的学习过程中, 同时, 可以利用边缘图强调显著性目标的边界区域.

总的来说, 单流结构的优越性在于较低的参数量及计算代价, 但是仅仅在输入端组合 RGB 和深度图像, 或者将深度图作为监督信息, 很难解决跨模态数据不一致问题和特征融合不充分问题. 除了在网络的输入输出阶段使用深度信息, 作为备选方案, 还可考虑将深度信息融入到卷积、池化等操作中[84].

1.1.2 双流结构

双流结构指通过两个编码器分别处理输入 RGB 图像和深度图像的网络架构. 基于双流结构的编码器-解码器网络是目前最为流行的 RGB-D 显著性目标检测解决方案, 这是因为 RGB 图像和深度图像作为两种不同模态的数据, 往往需要不同的网络来分别进行处理, 以便更加充分地提取各自的优势特征, 进而为跨模态特征融合作准备. 典型的双流结构模型通常利用独立的骨干网络分别提取 RGB 和深度图像的多级特征, 然后, 通过设计跨模态特征融合方案实现不同模态、不同尺度的特征融合, 进而得到预测输出. 因此, 跨模态特征融合过程中, 两个模态所起的作用成为了需要关注的重点问题. 本文根据两个模态在网络设计中扮演角色的不同, 将双流结构进一步划分为深度辅助交互模式和等重要双向交互模式两种, 如图 5 所示.

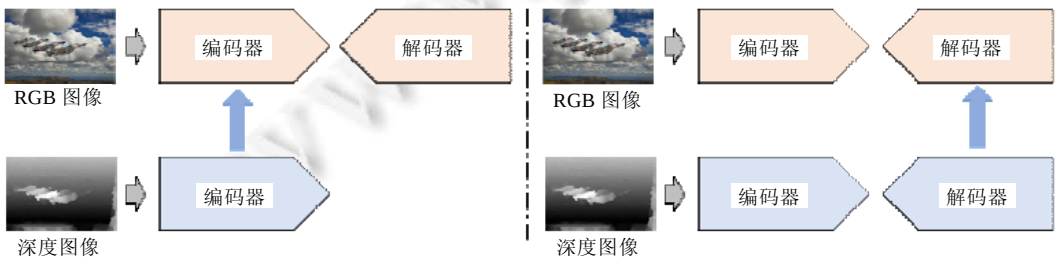

图 5 双流结构-深度辅助交互模式示意图

1.1.2.1 深度辅助交互模式

相比于深度图像, RGB 图像包含了更加丰富的信息. 因此, 深度图像通常被看作是 RGB 图像的补充和辅助, 进而在网络中信息的传递方向就定义为深度向 RGB 的传递. 根据深度信息在编码器-解码器网络中辅助增强的位置不同, 可以将此类辅助交互模式进一步划分为编码阶段辅助和解码阶段辅助两种方案.

(a) 编码阶段辅助方案

图 5(a)为编码阶段辅助方案, 深度分支的编码特征被用于辅助 RGB 分支编码特征的学习, 然后将增强的 RGB 特征用于解码过程. 在辅助策略上, 较简单的做法可以通过乘法、加法或通道级串联的方式实现. 在采用该方案的代表性文献[38]中, Zhu 等人通过使用一个辅助子网络处理深度图像, 证明了其性能优于四通道输入的方法. 具体地, 该工作提出了由一个主网络和一个子网络组成先验模型引导的深度增强网络: 先验模型引导的主网络遵循编码器-解码器的网络结构, 用于处理 RGB 图像输入; 辅助子网络仅包含编码器, 用于处理深度图像输入. 然后, 将得到的深度图像编码特征以通道级串联的方式附着到 RGB 特征上作为补充, 最后通过卷积上采样得到显著图. 明显地, 该方法引入了一个额外的深度子网络, 更容易获得良好的深度特征. 但是仅仅在编码器最高层上补充不能完成充分的模态交互, 例如还可考虑多尺度信息交互等问题. 在文献[39]中, Chen 等人提出了一个多路径、多模态融合框架, 将深度编码特征以元素级相加的方式传递到 RGB 分支. 在文献[40]中, Li 等人分别在编码器的低、中、高层设计了跨模态交互模块, 深度特征和 RGB 特征的融合以相乘后相加的方式进行.

除了上述较简单的做法, 更进一步, 部分研究人员考虑了更为高效的设计. Zhao 等人[41]首先利用传统算法中常见的对比度先验增强深度信息, 然后将增强后的深度信息作为注意力图, 对每一层的 RGB 编码特征加权. Li 等人[42]提出了一个跨模态深度加权组合模块, 它在编码器每个层级, 用深度特征产生的空间注意力图辅助增强 RGB 特征.

(b) 解码阶段辅助方案

图 5(b)为解码阶段辅助方案, 此方案中, 编码器仅用来提取单一模态特征, 模态间的相互作用在解码器中进行. 文献[43]最先将知识蒸馏[85]技术引入到了此任务, 所提出深度蒸馏网络的编码器采用两个独立的 VGG 网络分别提取 RGB 图像和深度图像的多级特征, 但不进行任何模态间的信息交互. 在解码阶段, 深度分支解码输出的显著图及中间层生成的注意力图作为深度流蕴含的知识, 以蒸馏的方式转移到 RGB 流, 在测试阶段不再需要深度图输入, 加快了模型推理速度, 但在性能上要劣于同时期的很多算法.

受非局部模型[86]启发, Liu 等人[28]首先通过自模态和跨模态的注意力机制来捕获长程依赖, 并进行跨模态信息交互; 然后在解码过程中, 通过设计的残差融合模块将深度分支的特征传递到 RGB 分支. Chen 等人[44]致力于开发一个高效的检测网络, 因此设计了一个轻量级的深度分支来从头学习深度图像特征, 并在解码阶段提出了一种交替式修正策略, 深度特征被渐进地加入到解码过程进行由粗到细的预测. Pang 等人[45]首次将动态滤波的思想引入到 RGB-D 显著性目标检测任务中, 他们将骨干网络生成的 RGB 特征和深度特征结合在一起, 生成区域感知的动态滤波器来指导 RGB 为主干的解码过程. Wu 等人[46]提出了一个预测速度达 450FPS 的 RGB-D 显著性检测模型, 该网络仅把最高层的深度特征用于辅助 RGB 为主干的解码过程.

1.1.2.2 等重要双向交互模式

RGB 图像和深度图像作为不同模态的输入, 所反映的信息侧重点是不同的, 提取到的特征也往往具有一定的差异性和互补性. 面对不同的数据, 哪种模态起主要作用不能一概而论, 有的数据仅通过深度图即可获得较好的检测结果, 而有的数据要更依赖于颜色特征. 因此, 一些研究工作不再将信息传递方向固定在深度分支到 RGB 分支, 而是将两种模态信息同等对待, 在网络学习过程自动地挖掘它们之间的互补性, 我们称这一类方法为等重要双向交互模式. 以此为出发点, 在具体实施上, 可以细分为对称式(图 6 左)和非对称式两种结构(图 6 右).

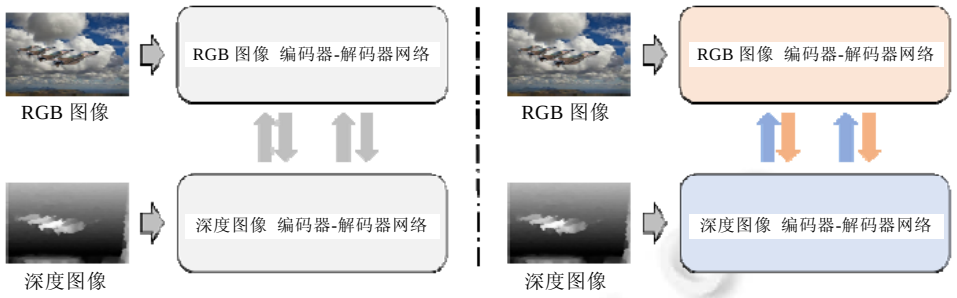

图 6 双流结构-等重要双向交互模式示意图

(a) 对称式结构

对称式结构是指通过两个完全相同的网络分别提取 RGB 图像和深度图像各自模态的特征, 同时, 特征交互也采用相同的处理方式. 这种结构关注的一个重点问题是, 如何融合两种模态特征以生成更全面的跨模态特征. 一类较简单的做法是元素级的特征相乘或相加[47−53], 用于强调两个模态的共同显著性区域; 另一类方法首先在通道维度上并联两个模态的特征[54−58], 然后通过后续的学习过程完成更深层次的融合. 这两类方法虽然被广泛采用, 但从特征融合的充分性和可解释性来看仍有不足. 与之不同, 文献[59]最先意识到了更加明确地建模跨模态互补信息的重要性. 为此, 作者设计了一种新的互补感知融合模块, 该模块作为解码过程的重要组件被嵌入到解码器的每一层, 它通过引入跨模态残差函数和互补感知监督将互补信息学习问题建模为残差函数的渐进逼近问题, 减少了融合歧义性并增加了融合充分性. 该文献所探索的明确的建模方式启发了之后的许多工作, 目前已在 Google Scholar 获得了近 200 次的引用.

还有些方法考虑如何自适应地、有选择地融合两个模态的特征. 例如: Chen 等人[61]提出了一个注意力感知的跨模态、跨层级融合模块, 它能够从每个层级和每个模态中自动地选择出最具判别力的特征用于后续解码过程; Chen 等人[62]提出了一个多路径、多模态融合网络, 自适应地融合了全局和局部信息; Zhou 等人[63]提出了一种基于 Res2Net 的门控融合网络, 以门控机制自动调节跨模态融合过程; Wang 等人[64]首先利用双流卷积网络分别预测 RGB 图像和深度图像对应的显著图, 然后利用一个自适应融合模块生成最终预测; Zhou 等人[65]提出了基于内容感知的融合模块, 该模块可以在保持特征图空间结构的同时, 自适应地利用融合特征强调在通道维度上的关键信息; Zhang 等人[66]认为, 大部分网络缺乏编码阶段的特征双向交互, 因此提出了一个基于空间注意力和通道注意力的双向传输和选择模块, 可以较早获得更为鲁棒的编码特征; Wen 等人[67]认为, 大多数方法忽略了两种模态之间的内在差异, 继而提出了一个动态选择网络, 它不仅可以自动选择和融合跨模态特征, 还可以自主优化和加强多层次和多尺度特征, 为跨模态的自适应融合贡献了一种新方案.

除了上述的工作, 还有些较为特殊的做法. 例如: Luo 等人[68]设计了一个级联的图神经网络, 通过图推理模块实现跨模态信息交互, 从而从不同模态中获取有用的知识; Jiang 等人[69]通过一个跨模态生成对抗网络, 从 RGB 特征和深度特征中学习最佳视角不变性及像素级的一致性表示, 从而同时融合模态内信息和跨模态图像相关信息; Chen 等人[70]设计了一个分离的跨模态融合网络, 通过跨模态重构来揭示两种模态的结构和内容表示, 在简化了融合过程的基础上, 提高了融合的充分性; Zhou 等人[71]提出了一个特异性保留网络, 该网络通过探索共享信息和特定于模态的特性来提高显著性检测性能; Huang 等人[72]认为, 一般的线性融合策略不能完全捕获跨模态互补信息, 因此提出了一个新的多模态融合方案, 通过联合使用一些简单的线性融合策略和双线性融合策略, 更好地捕获了跨模态互补信息, 这也是双线性融合策略在 RGB-D 显著性检测任务的首次探索; Zhang 等人[73]提出了一个基于互信息量最小化的学习框架, 显式地建模了两个模态之间地冗余信息, 使得网络能够挖掘出各自模态的有用特征.

(b) 非对称式结构

非对称式结构是指在同等对待两个模态的前提下, 针对两个模态各自的特性进行有差别的网络设计, 即

针对 RGB 特征和深度特征的处理是不一致的. Zhang 等人[74]提出的 ATSA 方法可以看作此类结构的一个典型代表, 他们认为, RGB 图像和深度图像的内在差异决定了它们不适用于同一个网络. 因此, 本文的主要贡献就在于提出了一个非对称的双流结构, 其中, 为 RGB 图像设计的自底向上的流梯型结构可以提取全局和局部信息, 为深度图像设计的深度注意力模块保证了深度图像特征与 RGB 特征密切结合和自适应引导. 其非对称式的设计思想更加适合处理不同模态的输入, 但是本文在交互方案的设计上仅通过像素级乘法进行特征融合并不一定是最优的选择. 近期, Zhang 等人[75]提出了一种跨模态差异性交互模式, 根据各层级不同的特征表示, 设计了 RGB 引导的细节增强模块和深度引导的语义增强模块, 明确地建立了模态间的依赖关系.

另有一些工作, 如 Ding 等人[76]提出了一种深度感知显著性模型, 其中, 深度显著网络采用颜色显著网络的权值初始化, 并设计了多层特征金字塔结构, 提高了深度特征提取的能力. Piao 等人[77]提出了一个深度诱导的多尺度循环注意力网络, 通过所提出的深度诱导的多尺度上下文加权模块, 首次探索了深度信息和不同尺度物体之间的关联关系, 为彩色信息和深度信息的融合提供了一个新的方案. Li 等人[78]提出了一个跨模态特征调制模块, 该模块从深度特征中学习像素级的仿射变换参数, 在每个层级调制相应的 RGB 特征, 完成互补关系的建模. Zhou 等人[79]提出了一个跨模态和跨尺度的自适应融合网络, 其中的通道融合模块遵循 RGB 特征为主、深度特征为辅的融合方案, 重点关注高层特征的利用; 而空间融合模块遵循深度特征为主、RGB 特征为辅的融合方案, 重点关注底层特征的利用. 最终, 在显著性目标定位和细节信息优化上都取得了良好的效果. Sun 等人[80]除了深度图用来提取多级特征外, 还将原始深度图按深度区间分层后作用到 RGB 特征上来减少背景的干扰, 最后以网络架构搜索的方式确定 RGB 和深度信息的融合方案.

1.1.3 三流结构

三流结构(如图 7 所示)是指通过 3 个编解器进行特征提取和显著性检测的网络架构.

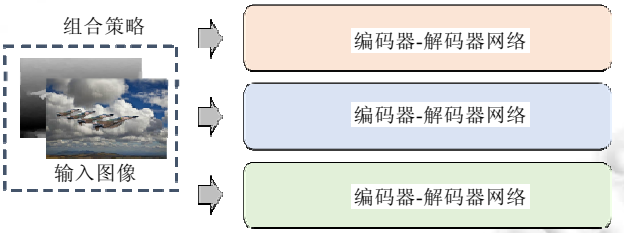

图 7 三流结构示意图

例如, Fan 等人[81]提出了一个简单的通用框架, 称为深度净化器网络. 该网络由一个深度净化单元和一个三流的特征学习模块组成, 分别进行低质深度图过滤和多模态特征学习. 其中, 三流特征学习网络分别以 RGB、深度、RGB-D 图像作为输入, 经过相同的类空间金字塔结构分别预测显著图, 在测试阶段, 通过所提出的深度净化器, 根据深度图质量选择较优输出. 该方法所提出的三流结构通过独立的单模态流进行和混合模态流分别进行预测, 使得可以根据深度图的质量选择合适的输出, 如果能在此架构下探讨更为充分的跨模态融合方案, 可以获得进一步的性能提升.

文献[82]提出了一个三流互补网络, 前两个子网络用于从 RGB 图像和 RGB-D 图像中提取显著性图, 第 3 个子网络通过综合考虑 RGB 图像、深度图像和前两个子网络的显著图, 以获得更精细化的结果. 文献[83]提出一种新的数据重组策略, 在图像输入之前, 将四维 RGB-D 特征转换为 DGB, RDB 和 RGD 这 3 种输入形式, 然后通过一个轻量级三流融合网络, 充分利用了新生成的输入数据, 保证了 RGB 和 D 的最佳互补融合状态.

1.2 深度图质量感知问题的解决方案

受深度成像设备硬件技术限制及采集过程中环境因素影响, 采集得到的深度图像质量参差不齐, 有时会存在像素值缺失、伪影、模糊等低质量问题, 甚至完全无法从深度图中判读目标. 引入此类低质量深度图到训练过程, 将会带来大量的噪声干扰, 使得网络难以优化, 进而影响显著性检测算法性能. 这一问题首次在

2016 年的 DCMC 算法[87]中被关注, 2020 年起, 陆续有基于深度学习的解决方案被提出. 既然深度图质量不好, 那么其解决方案无外乎两种: 一种是对现有深度图进行质量评估, 以此来控制低质量深度图在模型中的作用, 我们称为深度质量评价类方法; 另一种则是对现有深度图进行优化, 可以设计一些专门的图像增强类模块来提高深度图质量, 也可以直接通过深度估计生成新的深度图参与模型训练, 这一类方法我们称为深度质量优化类方法. 相关算法总结见表 2, 需要注意的是: 大多数方法并不仅仅解决深度图质量问题, 其可能也在表 1 的分类框架下. 为了避免重复性, 我们仅在此列举一次.

表 2 深度图质量感知方法分类

类型 \ 年份	2020	2021
深度质量评价	D3Net[81], DQSD[89]	DPANet[88], DFMNet[90]
深度质量优化	BBS-Net[27,91], MMNet[93], UC-Net[96,97], SSF[98], TPSD[100]	EF-Net[92], HAINet[95], TriTransNet[99], DCFNet[102], CDNet[101]

1.2.1 深度质量评价

此类方法通过设计某种度量方案得到深度图或深度特征的质量评级, 并以此控制低质深度图的引入量, 进而降低低质深度图带来的负面影响. 根据评价的对象不同, 大体可以分为图像级别的质量评价和特征级别的质量评价两种类型.

对于图像级别上的评价方案, 这里重点对 Chen 等人[88]提出的 DPANet 方法进行介绍, 该论文也入选了 ESI 高被引论文, 其主要贡献在于: 通过统一的模型, 同时解决了深度图质量感知问题和跨模态特征融合问题. 该方法首先计算阈值分割后的深度图和显著性真值图之间的相似性, 得到深度图质量的伪标签, 在训练过程中, 利用此伪标签作为监督信息, 使网络能够以学习的方式感知深度图的质量, 获得深度图质量的表征因子, 并以此控制深度信息引入的多少, 最终获得了先进的实验结果.与之不同, Fan 等人[84]在输出端进行了质量评价, 他们首先通过三流网络分别得到以 RGB、深度、RGBD 为输入的显著图 S_{RGB}, S_D, S_{RGBD}, 然后通过度量 S_D 和 S_{RGBD} 之间的相似性, 选择出最终的输出:如果差异较小, 则以 S_{RGBD} 为最终输出; 反之, 则抛弃深度信息以 S_{RGB} 为最终输出. 以上两种方法都是基于整体图像来进行质量评价, 考虑到每个区域对于最终检测结果的影响是不一致的, Chen 等人[89]设计了一个弱监督的深度贡献评价子网络, 通过比较 RGB 和深度分支的显著图与真值图之间的关系, 选择出深度图中的有价值区域, 并作为伪标签进行网络训练.

Zhang 等人[90]首先提出了特征级别的评价方案, 作者提出了一种新颖的深度图质量启发的特征控制过程, 其独特的门控机制可以用于过滤深度图, 大大提高了识别的准确性. 在具体实现上, 包括深度图启发的加权组件和深度整体注意力组件, 其中: 前者将两个模态的低级特征表示组合起来, 并通过多层感知机学习得到一个门控向量来指导融合过程中深度特征“量”的引入; 后者利用最高层的语义信息做指导, 通过生成空间注意力图来突出深度图中需要关注的区域. 在特征级别上做质量评价的优点在于对深度信息的引入可以进行更细致的控制, 但是如何定义深度特征的质量好坏, 是一个难以解决的问题.

1.2.2 深度质量优化

正如前文所提到的, 深度质量优化一方面可以采取图像增强类方法, 其主要通过一些设计提高深度图的质量, 以此降低低质深度图的影响. 具体可分为两类: 自身增强和利用 RGB 辅助增强. 自身增强的方式一般是指利用注意力机制等方式对深度特征进行增强, 例如 Fan 等人[27,91]引入了深度增强模块, 通过级联的通道注意力机制和空间注意力机制捕获深度图像中的有用信息, 提高 RGB 和深度特征的兼容性. 利用 RGB 辅助增强深度信息是目前采用更为广泛的方案, 简单而常规的做法是以乘法的方式利用 RGB 特征对深度信息中的噪声进行过滤. 例如, Chen 等人[92]首先利用 RGB 图像提取粗糙显著图, 然后利用此显著图和原始深度图相乘, 以抑制背景噪声并锐化边界. 文献[93]提出了一个跨模态引导的注意力模块, 该模块将 RGB 特征作为引导信息, 通过元素级的相乘, 从深度特征中选择出重要和可靠的部分. Liu 等人[94]认为: 深度流的侧输出特征无法提供准确的边界信息, 甚至会干扰最终的显著图, 但其所提供的空间结构信息也是不可缺少的. 为此, 他们设计了一种简单直接的方法, 将 RGB 流的侧输出特征添加到深度流的每个侧输出特征中, 在一定程度上改

善了深度特征质量. Li 等人[95]提出了一个交替式交互单元: 首先, 利用 RGB 特征生成一个空间注意力图与深度特征相乘, 以完成干扰信息的过滤; 然后, 利用增强后的深度特征进一步辅助 RGB 特征. 除了乘法的方式, Zhang 等人[96]设计了一个基于 VGG16 的深度校正网络. 该网络假设深度图边缘应与 RGB 图边缘对齐,通过使用边界 IoU 作为该网络的正则化器, 引导深度图像的增强. Zhang 等人[98]认为, 深度图像中目标的边界容易出现随机噪声及区域缺失. 因此, 将深度图划分为多个二值掩码, 通过一个初始的显著图为每个掩码分配重要性权重, 以信息过滤的方式降低低质量深度图的影响. 最后, 我们对 Liu 等人[99]近期提出的方法进行展开介绍, 以便更加直观地确定当前技术水平. 为了缓解深度图质量差的问题, 作者在编码器的每个阶段都引入了一个深度净化模块, 该模块首先利用当前层 RGB 特征和深度特征生成混合模态特征表示; 然后, 利用此混合特征生成空间和通道注意力权重对深度特征进行重新加权, 进而强调深度图中的重要信息; 最后, 将增强后的深度特征作为 RGB 分支的补充. 总的来说, 由于深度图像和 RGB 图像是互补而不是包含的关系, 简单的乘法等方式在去除深度图中的背景噪声的同时, 也很容易削弱模态间的互补信息. 因此, 如何在保证深度互补信息不损失的前提下去除干扰噪声, 是一个有待研究的问题.

另一方面, 也可以直接通过深度估计生成新的深度图参与模型训练. 这是由于深度图质量不佳是源自数据集本身的问题, 因此, 通过深度估计获得更高质量的深度图成为一种自然而然的解决方案. Chen 等人[100]首先检索一部分与给定输入相似的图像, 利用图像间的关系及深度转移策略粗略地估计整体深度; 然后, 通过细粒度的对象级对应进一步提高所估计深度图的质量; 最后, 将估计得到的深度图与原始深度图输入到显著性检测网络中. 文献[101]选取深度显著性图和真值图之间 IoU 大于 0.9 的样例作为训练样本, 设计了一种动态融合方案以融合原始深度图和估计深度图. 最后, 我们对 Ji 等人[102]近期提出的方法进行展开介绍, 该工作的核心贡献在于: 提出了一个深度校准和融合的框架, 从本质上提升了深度图的质量. 作者首先通过两个独立的网络分别预测 RGB 和深度显著图; 然后计算两个显著图和真值图的交并比 IoU_{RGB} 和 IoU_{Depth}, 并选取深度图预测优于 RGB 预测($IoU_{Depth}>IoU_{RGB}$)和深度图本身预测优异(IoU_{Depth} 前 20%)的样本训练一个深度图质量判别网络和深度估计网络; 最终, 便可以通过动态加权原始深度图和估计深度图得到校正深度图以用于后续检测过程. 该方法在不增加新的训练数据的前提下, 依靠现有的优质样本训练出了一个深度估计网络. 但是, 仅凭借 RGB-D 显著性目标检测任务较小的数据规模是否还能取得进一步的提升有待验证.

2 数据集及评价标准

2.1 数据集

数据对于深度学习算法来说至关重要, 自 2012 年以来, 已经陆续提出了多个 RGB-D 显著性目标检测数据集(见表 3).

表 3 RGB-D 显著性目标检测数据集

名称	年份	出处	数量	深度图获取方式	分辨率	数据集划分
STEREO[103]	2012	CVPR	797/1000	立体图像+深度估计	[251~1200]×[222~900]	仅测试
DES[104]	2014	ICIMCS	135	Microsoft Kinect	640×480	仅测试
NLPR[105]	2014	ECCV	1 000	Microsoft Kinect	480×640, 640×480	训练-700, 测试-300
LFSD[106]	2014	CVPR	100	Lytro Illum	360×360	仅测试
NJUD[107]	2014	ICIP	1 985	立体图像+深度估计	[231~1213]×[274~828]	训练-1485, 测试-500
SSD[108]	2017	ICCVW	80	立体图像+深度估计	960×1080	仅测试
DUT[80]	2019	ICCV	1 200	Lytro2 camera+	400×600	训练-800, 测试-400
SIP[84]	2020	TNNLS	929	Huawei Mate10	992×774	仅测试
ReDWeb-S[109]	2021	TPAMI	3 179	立体图像+深度估计	[133~937]×[132~996]	训练-2179, 测试-1000
COM15K[76]	2021	ICCV	15 625	立体图像+深度估计	–	训练-8025, 测试-7600

一些数据集示例如图 8 所示, 具体介绍如下.

(1) STEREO[103]是第 1 个公开用于 RGB-D 显著性目标检测的标准数据集, 包含了 1 000 对 RGB 和深度图像. 对于数据集的构建, 作者首先从 3 个网络平台收集了 1 250 幅立体图像, 然后由 3 名用户分别

标注每幅图像中最显著的目标, 并根据投票的一致性排序结果, 选择前 1 000 幅样本作为数据集基本构成, 最后通过光流估计算法, 从左右视图中得到深度信息;

(2) DES[104]数据集包含了 135 对室内场景样本, 又称 RGBD135, 其中, 深度图像直接使用 Kinect 深度相机获得. 3 名用户被要求分别标记显著性目标, 选取重叠区域作为最终的真值图;

(3) NLPR[105]数据集包含了 1 000 对样本. 作者首先使用 Kinect 深度相机从室内和室外的多种场景(例如办公室、超市、校园、街道等)收集了 5 000 幅 RGB 图像及对应的深度图像, 随后对深度图像进行校准, 以对齐 RGB 图像并提高深度图质量. 在标注阶段, 首先手动选择出 2 000 对更适合显著性目标检测的备选样本, 随后, 5 名用户被要求进行显著性目标标注, 最后选择前 1 000 幅意见更为一致的样本作为最终数据集构成;

(4) LFSD[106]数据集包含了 100 对样本, 其中, 室内场景 60 幅、室外场景 40 幅, 深度信息由光场相机获得. 3 名用户被要求手动分割出全焦点图像的显著对象, 且只有重叠率大于 90%的区域才作为最终的显著性区域;

(5) NJUD[107]数据集包含了 1 985 对样本, 作者首先从互联网、3D 电影和立体相机收集到原始立体图形, 然后利用光流估计方法得到深度图像. 此外, 为了最大限度地还原人眼识别的真实场景, 使用了 Nvidia 3D Vision 来在模拟的 3D 环境中进行显著性目标标注;

(6) SSD[108]是一个小型数据集, 包含了 80 对样本. 首先, 从 3 部立体电影收集左右视图自然图像, 包含室内和室外场景; 然后, 利用光流估计方法得到深度图像, 标注显著性目标时仍遵循少数服从多数的原则;

(7) DUT[80]数据集包含了 1 200 对样本, 其中, 室内场景 800 个、室外场景 400 个, 深度信息由光场相机获得, 具有透明对象、低对比度、复杂背景等多种困难场景;

(8) SIP[84]是一个以人为中心的室外场景数据集, 包含了 929 幅图像, 深度图像是由配备双摄像头的 Huawei Mate10 真实采集得到. 值得注意的是, 该数据集首次提供了实例级的标注信息;

(9) ReDWeb-S[109]是基于 ReDWeb 构建的数据集, ReDweb 中的深度图像是通过最新的光流估计算法 FlowNet2.0 并经后处理得到. 作者通过去除原始数据集中不含显著前景对象的图像, 并进行像素级标注后得到了包含 3 179 幅高质量深度图的新数据集;

(10) COME15K[76]是最新提出的目前规模最大的数据集, 包含了 15 625 对样本, 原始图像收集自 Holopix 平台, 深度图像通过光流估计算法得到. 同时, 该数据集提供了涂鸦标注、实例级标注等多种标注信息, 如图 9 所示.

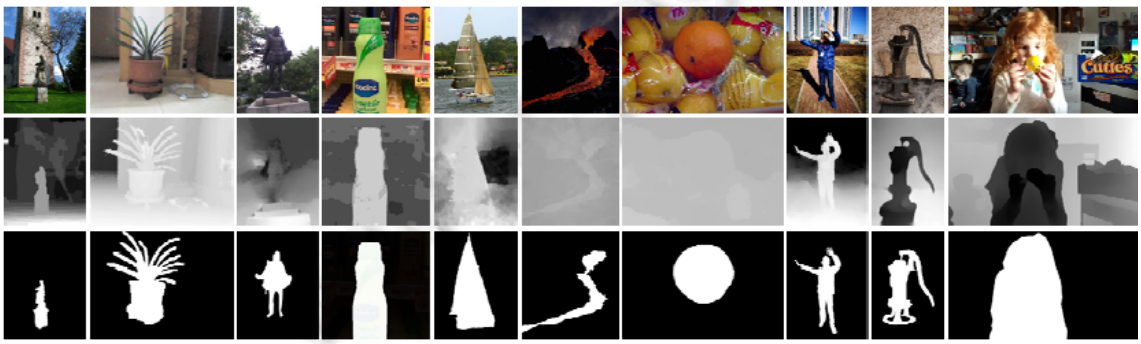

图 8 数据集示例, 从上至下依次为 RGB 图像、深度图像和真值图

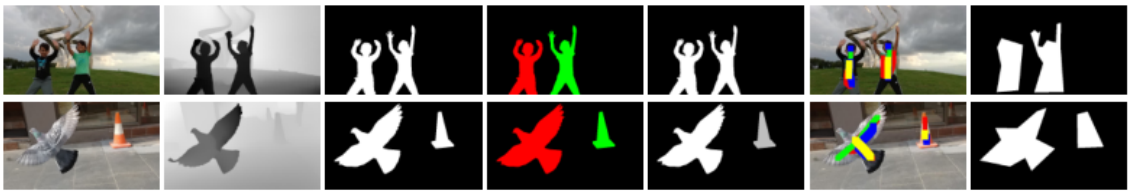

图 9 COME15K 的标注示例，图引用自文献[76]

总的来说，2020 年之前的数据集规模都不大，且标注方式比较单一. 从 2020 年开始，新提出的数据集在数据规模上有了较大提升，且标注方式也更加多样(如图 9 所示)，这为 RGB-D 显著性目标检测接下来的发展提供了良好的数据支撑. 此外，对于深度图像的获取，有深度估计和直接采集两类方式，从深度图像质量来看，深度估计方法相较于深度相机实际采集得到的深度图往往含有较少的噪声且边界比较清晰，但是受限于估计算法的好坏，可靠性较弱.

2.2 评价指标

预测显著图通常为一幅像素值在 0–255 之间的灰度图，其值越大，代表该位置显著的程度越高. 真值图为一幅二值图像，显著性区域被标记为 255，背景被标记为 0. 为了评价算法模型的优劣，除了主观地对预测显著图进行比较外，客观的评价指标是更为可靠的度量标准. 本小节对 RGB-D 显著性目标检测任务中广泛使用的 5 种定量评价指标进行了介绍.

(1) MAE (mean absolute error)

MAE (平均绝对误差)是一种线性的评价指标，以逐像素的方式计算了显著性预测图与真值图之间绝对误差的平均值，此值越低，代表预测结果越好，其具体定义式如下:

$$MAE=\frac{1}{H\times W}\sum_{y=1}^{H}\sum_{x=1}^{W}|S(x,y)-G(x,y)| \tag{1}$$

其中，H 和 W 分别代表了图像的高度和宽度，S 和 G 分别代表了预测显著图和真值图.

(2) F-measure

F-measure 是机器学习中一个常见的统计量，其数值越大，表明模型效果越好. 在计算该指标之前，需要将预测显著图按照不同的阈值进行二值化，然后利用得到的二值显著图和真值图分别计算准确率(precision)和召回率(recall)的加权调和平均值，得到相应的 F-measure. 根据阈值选取的不同，可以将 F-measure 细分为 3 个具体指标: ① 自适应 F-measure (adaptive F-measure)选择图像平均像素值的两倍作为分割阈值; ② 平均 F-measure(mean F-measure)分别以 0 到 255 作为分割阈值计算 F-measure，然后求平均值得到; ③ 最大 F-measure(max F-measure)分别以 0 到 255 作为分割阈值计算 F-measure，然后取最大值得到. F-measure 的一般定义式如下:

$$F_\beta=(\beta^2+1)\cdot\frac{precision\cdot recall}{\beta^2\cdot precision+recall} \tag{2}$$

其中，$precision=\frac{TP}{TP+FP}$，$recall=\frac{TP}{TP+FN}$，$TP$, FP, FN 分别代表了真阳性、假阳性和假阴性，是机器学习中常用的几个统计量; β^2 用来权衡准确率和召回率的重要性，一般被设置为 0.3 用于强调准确率.

(3) S-measure

S-measure (结构相似性度量)是 Fan 等人[110]在 2017 年提出的评价指标，是为非二值图像的评估而设计的. 他们认为: 之前的显著性评价指标都是基于像素级误差的，没有考虑到显著性目标的结构相似性. 因此，通过综合物体结构相似性 S_o 和区域结构相似性 S_r 定义了一种新的评价指标. 此值越高，代表预测结果越好，其定义式如下:

$$S=\alpha*S_o+(1-\alpha)*S_r \tag{3}$$

其中, α被用来权衡 S_o 和 S_r 的重要性, 一般被设置为 0.5.

(4) *E*-measure

E-measure (增强匹配指标)是 Fan 等人[111]在 2018 年提出的评价指标, 与之前的指标独立地考虑像素级匹配和和图像级信息不同, 该指标将局部像素与图像级平均值结合, 综合考虑了局部信息和全局信息.

与 *F*-measure 类似, *E*-measure 是一种二值前景图评估指标, 需要首先将显著图根据不同的阈值分割为多个二值显著图分别计算, 因此存在自适应 *E*-measure (adaptive *E*-measure)、平均 *E*-measure (mean *E*-measure)和与最大 *E*-measure (max *E*-measure). *E*-measure 的一般定义式如下:

$$E_\phi=\frac{1}{H\times W}\sum_{y=1}^{H}\sum_{x=1}^{W}\phi_{FM}(x,y) \tag{4}$$

其中, ϕ代表了增强的对齐矩阵, 用来捕获像素级匹配和图像级统计这两个重要属性.

(5) *P*-*R* 曲线

P-*R* 曲线描述了准确率(precision)和召回率(recall)的关系, 是以召回率为横坐标、准确率为纵坐标, 根据不同阈值绘制的一条曲线, 位于坐标系第一象限的(0,0)到(1,1)的区域内, 曲线越靠近(1,1)点, 代表预测结果越好.

3 方法比较

本小节对近年来部分方法进行了定性和定量比较, 旨在更直观地确定该领域发展水平. 图 10 通过可视化的显著性图, 分别列举了单流结构、双流结构和三流结构的部分方法的检测结果, 包括了低对比度(第 1 列)、多目标(第 2 列和第 3 列)、复杂背景(第 4 列、第 6 列、第 8 列)和小目标(第 7 列)场景类型. 总的来说, 近期提出的检测模型在大多数场景下都能成功地定位到显著性目标, 一方面得益于卷积神经网络强大的特征提取能力, 另一方面是因为深度图像的引入提供了一定程度的互补性. 特殊地, 在第 1 列所展示的低对比度场景中, 深度图像可以提供很好的补充信息, 因此几乎所有的方法都可以准确地定位出该目标. 但是 A2dele 算法的检测结果存在较大的目标残缺, 正如前文所提及的, 该方法通过知识蒸馏的思想将深度分支的信息在训练过程中传递到 RGB 分支, 在测试时不再使用深度图输入, 但是受限于网络的学习能力, 深度图的所表达的信息并不能被完全隐式地学到, 因此导致了图示的预测结果. 对于第 2 列所展示的多目标问题, 所有参与测试的算法都成功将所有目标检出, 但是单个目标的完整性却不尽相同. 例如, DASNet 所检出的右侧两物体下半部缺失. 当涉及多目标检测问题时, 如何建立目标之间的联系, 保证所有目标都能达到相近的检测精度, 是一个值得研究的问题. 对于第 3 列所展示的情况, 作为显著性目标的人与背景的墙体具有相近的深度值, 原本在 RGB 图像能够很容易检测出的显著性目标却受到深度图影响, 导致了检测结果的假阳性, 如 S2MA, SPNet 等算法. 在第 4 列中, RGB 图像中的显著性目标内部具有很明显的颜色变化, 深度图却能够维持良好的目标内部一致性, 但是从检测结果来看, 大部分方法都未能取得令人满意的效果. 结合第 3 列和第 4 列这两种情况, 如何根据场景的难易程度, 自适应地从 RGB 和深度图像中获取有助于显著性目标检测的信息, 是一个值得研究的问题.

与此同时, 也存在一些尚未解决的共性问题, 例如显著性目标的边界模糊问题, 如倒数第 5 列、第 7 列、第 8 列所示. 其实, 不止对于 RGB-D 显著性目标检测任务, 任何的像素级预测任务都面临着相似的问题. 造成该问题的原因主要归咎于以下两种操作: (1) 为了减少网络计算量, 通常在将输入图像送入编码器之前需要进行统一的图片大小调整(如 256×256, 352×352 等), 在网络完成所有计算之后, 再通过插值上采样恢复到初始分辨率; (2) 目前所采用的基于卷积神经网络的编码器通常需要交替地堆叠卷积层和池化层, 在减小计算量的同时, 使得网络能够获得更大的感受野, 然后在解码阶段, 通过反卷积或上采样恢复图像细化特征并恢复图像分辨率. 从信息论的角度来看, 上述两种操作都会损失一定的信息量, 属于不可逆的有损操作. 目前常见的操作是将编码器的各级特征引入到解码阶段, 辅助分辨率的恢复, 但仍需要关注的一个重点问题是: 如何

在引入有效信息的同时，减少噪声引入量.

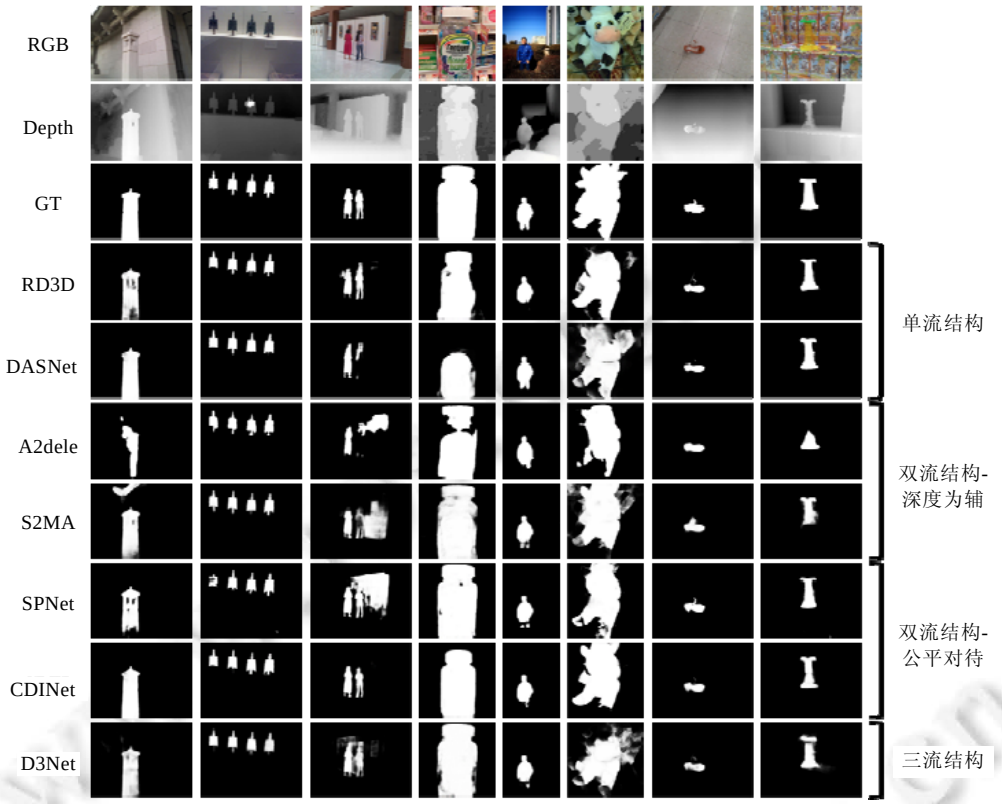

图 10　不同跨模态融合方案的可视化结果图

图 11 中给出了深度图质量问题部分解决方案的可视化结果图. 比较后，我们发现了一个有趣的现象：质量感知类模型能够较好地处理深度图质量较差的场景图 11(b)、图 11(c)，反而在深度图质量较好但 RGB 图像检测困难的场景图 11(a)中效果不好. 本文分析最有可能的原因是：模型在设计过程中过分强调了 RGB 信息的重要性，进而使得模型产生了数据偏好，即使深度图质量较好时，仍然未充分利用深度信息. 归根结底，这仍是跨模态交互过程需要解决的范畴. 但需要注意的是：跨模态交互和深度质量感知这两方面应该联动考虑，统筹解决，进而使模型能够自适应地在较好深度图的作用下，实现更加全面的跨模态交互.

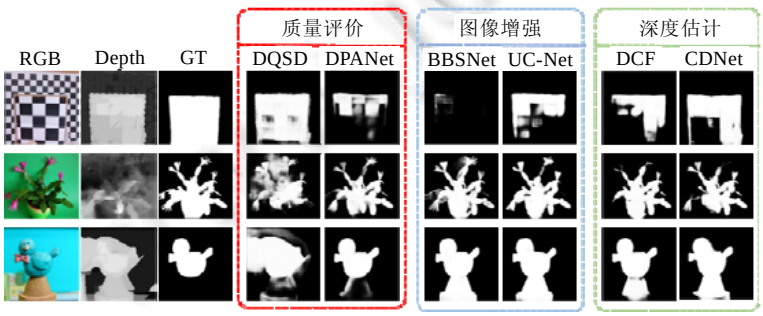

图 11　不同深度图质量问题部分解决方案的可视化结果图

最后，在表 4 中列举了近 3 年来提出的部分基于深度学习的 RGB-D 显著性目标检测方法在 6 个数据集上根据 3 个指标的定量评价结果，每个数据集上的最好性能由粗体标注出来.

表 4　在部分数据集上的定量评价结果

	模型	年份	STEREO				NLPR			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	MAE	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	MAE
单流结构	CoNet	ECCV-2020	0.882	0.890	0.932	0.051	0.896	0.915	0.949	0.027
	RD3D	AAAI-2021	0.906	0.911	0.947	0.037	0.919	0.930	0.965	0.022
双流结构-深度为辅	S2MA	CVPR-2020	0.882	0.890	0.932	0.051	0.902	0.915	0.953	0.030
	A2dele	CVPR-2020	0.874	0.878	0.915	0.044	0.878	0.896	0.945	0.028
	PGAR	ECCV-2020	0.880	0.907	0.919	0.041	0.885	0.930	0.955	0.024
	HDFNet	ECCV-2020	0.900	0.900	0.943	0.042	0.917	0.923	0.963	0.023
双流结构-同等重要	BTSNet	ICME-2021	0.911	0.915	0.949	0.038	0.923	0.934	**0.965**	0.023
	DSA2F	CVPR-2021	0.907	0.905	0.916	0.036	0.906	0.919	0.956	0.024
	CDINet	MM-2021	0.903	0.905	0.943	0.041	0.916	0.928	0.960	0.024
	CMINet	ICCV-2021	0.895	**0.921**	**0.959**	**0.034**	0.909	**0.941**	0.964	**0.019**
	SPNet	ICCV-2021	**0.915**	0.907	0.944	0.037	**0.925**	0.927	0.959	0.021
三流结构	D3Net	TNNLS-2020	0.891	0.899	0.938	0.046	0.897	0.912	0.953	0.030
	DRLF	TIP-2020	0.878	0.888	0.929	0.050	0.880	0.903	0.939	0.032
深度图质量感知	CDNet	TIP-2021	0.898	0.906	0.942	0.040	0.920	0.931	0.964	0.025
	HAINet	TIP-2021	0.906	0.907	0.944	0.040	0.915	0.924	0.960	0.024
	DCF	CVPR-2021	0.890	0.905	0.931	0.037	0.907	0.922	0.956	0.023
	DFM	MM-2021	0.893	0.898	0.941	0.045	0.908	0.923	0.957	0.026

表 4　在部分数据集上的定量评价结果(续 1)

	模型	出处	NJUD				DUT			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	MAE	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	MAE
单流结构	CoNet	ECCV-2020	0.892	0.895	0.937	0.047	0.908	0.918	0.941	0.034
	RD3D	AAAI-2021	0.914	0.916	0.947	0.036	**0.939**	**0.932**	**0.960**	0.031
双流结构-深度为辅	S2MA	CVPR-2020	0.889	0.894	0.930	0.053	0.901	0.903	0.937	0.043
	A2dele	CVPR-2020	0.874	0.869	0.897	0.051	0.890	0.886	0.924	0.043
	PGAR	ECCV-2020	0.893	0.909	0.916	0.042	0.914	0.920	0.944	0.035
	HDFNet	ECCV-2020	0.922	0.908	0.944	0.039	0.915	0.908	0.945	0.041
双流结构-同等重要	BTSNet	ICME-2021	0.924	0.921	0.954	0.036	–	–	–	–
	DSA2F	CVPR-2021	0.907	0.904	0.918	.0396	0.930	0.922	0.945	**0.030**
	CDINet	MM-2021	0.922	0.919	0.951	0.035	0.937	0.927	0.959	**0.030**
	CMINet	ICCV-2021	0.925	**0.939**	**0.956**	0.032	–	–	–	–
	SPNet	ICCV-2021	**0.935**	0.925	0.954	**0.028**	–	–	–	–
三流结构	D3Net	TNNLS-2020	0.900	0.900	0.950	0.041	0.786	0.815	0.868	0.085
	DRLF	TIP-2020	0.883	0.886	0.926	0.055	–	–	–	–
深度图质量感知	CDNet	TIP-2021	0.918	0.919	0.950	0.036	0.935	0.927	0.958	**0.030**
	HAINet	TIP-2021	0.915	0.912	0.944	0.038	0.917	0.910	0.940	0.037
	DCF	CVPR-2021	0.905	0.903	0.922	0.038	0.926	0.924	0.952	**0.030**
	DFMNet	MM-2021	0.910	0.906	0.947	0.042	–	–	–	–

表 4　在部分数据集上的定量评价结果(续 2)

	模型	出处	LFSD				SIP			
			$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	MAE	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	MAE
单流结构	CoNet	ECCV-2020	0.848	0.862	0.896	0.071	0.842	0.858	0.909	0.063
	RD3D	AAAI-2021	0.854	0.858	0.890	0.074	0.889	0.885	0.924	0.048
双流结构-深度为辅	S2MA	CVPR-2020	0.835	0.837	0.873	0.094	0.884	0.878	0.920	0.054
	A2dele	CVPR-2020	0.828	0.825	0.866	0.084	0.825	0.826	0.892	0.070
	PGAR	ECCV-2020	0.852	0.853	0.889	0.074	0.854	0.876	0.908	0.055
	HDFNet	ECCV-2020	0.858	0.846	0.889	0.085	0.894	0.886	0.930	0.048
双流结构-同等重要	BTSNet	ICME-2021	0.874	0.867	0.906	0.070	0.901	0.896	0.933	0.044
	DSA2F	CVPR-2021	**0.889**	**0.883**	**0.920**	**0.055**	–	–	–	–
	CDINet	MM-2021	0.875	0.870	0.914	0.063	0.884	0.876	0.915	0.054
	CMINet	ICCV-2021	0.862	0.877	0.911	0.064	0.887	**0.894**	**0.933**	0.044
	SPNet	ICCV-2021	–	–	–	–	**0.916**	0.894	0.930	**0.043**
三流结构	D3Net	TNNLS-2020	0.810	0.825	0.862	0.095	0.861	0.860	0.909	0.063
	DRLF	TIP-2020	–	–	–	–	–	–	–	–
深度图质量感知	CDNet	TIP-2021	0.861	0.858	0.896	0.073	0.892	0.879	0.924	0.053
	HAINet	TIP-2021	–	–		–	0.892	0.880	0.922	0.053
	DCF	CVPR-2021	0.860	0.856	0.892	0.071	0.877	0.874	0.920	0.051
	DFMNet	MM-2021	0.864	0.865	0.903	0.072	0.887	0.883	0.926	0.051

整体来看，双流结构的等重要双向交互模式获得最佳性能的方法最多，这也是目前方法中最常用、最主流的设计思路. 通常情况下，三流网络相较于双流结构通常有更大的模型容量，理论上应该表现出更好的效果. 但这并不是绝对的，因为算法的性能高低与模型的设计有着密切的联系. 例如，表 4 中很多双流模型的效果要优于三流模型，其主要原因有以下两点: (1) 文献[84,86]均为 2020 年正式刊出在 TNNLS 和 TIP 上的论文，方法提出时间较早，模型优化水平、数据增强方式与现在相比有一定差距，因此性能相对较低; (2) 性能的好坏与网络的设计有着很大关系. 例如表 4 中所对比的“三流结构-D3Net”方法，在该论文中，作者提出了 SIP 数据集，同时对已有方法进行了全面比较，因此在模型设计上相对简单，只通过提出的深度净化单元进行输出显著图的选择，缺少跨模态融合过程，故效果较差. 文献[86]致力于提出一个轻量级三流融合网络，因此作者设计了一个简易的特征融合方案，在当时获得了先进的水平，同时保持了较高的推理速度. 综上所述，研究人员在进行模型设计的过程中，应该采用辩证的眼光选定设计的主体架构，并进行精细化的模块设计，达到最佳的检测效果.

4 总结与展望

基于深度学习的 RGB-D 显著性目标检测仍处于方兴未艾之时. 作为必须考量的关键问题，跨模态特征融合已经进行了比较充分的研究，诸如注意力机制[28]、图神经网络[71]、对抗生成网络[72]、动态卷积[48]、知识蒸馏[44]等技术都已被成功应用在该领域. 但是对于深度图质量感知问题的探索仍处于起步阶段，虽有尝试，但仍值得深挖. 除此之外，对于 RGB-D 显著性目标检测任务未来的发展，可从以下几方面着手.

(1) 任务扩展: 除了直接应用深度数据外，光场数据也包含了场景的深度信息，探索 RGB-D 显著性目标检测任务与光场图像显著性检测任务之间的关系，是一个值得探索的研究点. 此外，与 RGB-D 数据类似，RGB-T 数据同时包含了 RGB 和热红外图像两种模态，同样可以借鉴 RGB-D 显著性目标检测的相关技术.

(2) 监督方式: 目前，RGB-D 显著性目标检测模型大多基于像素级的全监督信息，但真图的标注成本是非常巨大的. 因此，探究如何在更少的标注信息下完成 RGB-D 显著性目标检测，是一项非常有意义的工作，如弱监督/半监督/自监督的 RGB-D 显著性目标检测研究. 近期，Li 等人[112]提出了第一个基于文本注释的弱监督的 RGB-D 显著性目标检测算法，以及 Zhu 等人[113]提出的第一个半监督 RGB-D 显著性目标检测算法，都为该方向接下来的发展提供了借鉴.

(3) 实例级: 现有的方法输出均是二值化的显著性图，并不能区分不同显著性目标个体. 因此，实例级的 RGB-D 显著性目标检测作为更进一步的任务有待探索. 目前，已有部分数据集进行了实例级别的标注，例如 SIP, COME20K，为相关研究奠定了数据基础.

(4) 实时性: 显著性目标检测任务作为一种预处理技术，在实际应用中对实时性的要求较高. 现有方法的研究重点主要集中在如何提高模型性能上，而对模型的实时性关注度不够. 如何在保证模型准确性的同时兼顾效率，是一个值得研究的方向.

(5) 数据集: 在追求大规模数据集构建的同时，数据集之间的标注鸿沟问题变得越来越突出. 图 12 给出了一类标注鸿沟示例. 这主要是由数据集的采集方式不同、显著性目标检测任务本身的主观性、标注标准的差异性造成的. 这无疑对模型泛化能力提出了更大的挑战. 如何平衡不同数据集之间的差异性有待解决.

(a) STEREO (b) NJUD

图 12 数据集标注鸿沟问题示例

(6) 应用扩展: 显著性目标检测任务作为一种前端技术可以赋能多种视觉任务，如常规的分割、检测、识别等. 但是，寻找更多的真实应用点仍然值得进一步探索. 比如，在手机智能拍照场景中，应用显著性检测

技术可以帮助快速定位以人为中心的区域, 也可以进一步延伸至大光圈虚化等应用. 总之, 借用一句古诗“纸上得来终觉浅, 绝知此事要躬行”, 在面向实际问题的应用过程中, 还需要不断确定新的需求点, 进而对显著性检测技术进行改进和升级.

References:

[1] Wang WG, Shen JB, Jia YD. Review of visual attention detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 416−439 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

[2] Katsuki F, Constantinidis C. Bottom-up and top-down attention: Different processes and overlapping neural systems. The Neuroscientist, 2014, 20(5): 509−521.

[3] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254−1259.

[4] Cong R, Lei J, Fu H, *et al*. Review of visual saliency detection with comprehensive information. IEEE Trans. on Circuits and Systems for Video Technology, 2018, 29(10): 2941−2959.

[5] Zeng Y, Zhuge Y, Lu H, *et al*. Joint learning of saliency detection and weakly supervised semantic segmentation. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7223−7233.

[6] Guo C, Zhang L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. on Image Processing, 2009, 19(1): 185−198.

[7] Xiao DG, Xin C, Zhang T, *et al*. Saliency texture structure descriptor and its application in pedestrian detection. Ruan Jian Xue Bao/Journal of Software, 2014, 25(3): 675−689 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4438. html [doi: 10.13328/j.cnki.jos.004438]

[8] Fan DP, Zhang J, Xu G, *et al*. Salient objects in clutter. arXiv:2105.03053, 2021.

[9] Chen Z, Xu Q, Cong R, *et al*. Global context-aware progressive aggregation network for salient object detection. In: Proc. of the AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 34(7): 10599−10606.

[10] Zeng Y, Zhang P, Zhang J, *et al*. Towards high-resolution salient object detection. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7234−7243.

[11] Lang C, Nguyen TV, Katti H, *et al*. Depth matters: Influence of depth cues on visual saliency. In: Proc. of the European Conf. on Computer Vision. Florence: Springer, 2012. 101−115.

[12] Zhang Q, Cong R, Hou J, *et al*. CoADNet: Collaborative aggregation-and-distribution networks for co-salient object detection. In: Proc. of the Advances in Neural Information Processing Systems. Virtual Conf.: MIT Press, 2020. 33: 6959−6970.

[13] Zhou W, Guo Q, Lei J, *et al*. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection. IEEE Trans. on Circuits and Systems for Video Technology, 2022, 32(3): 1224−1235.

[14] Zhang M, Ji W, Piao Y, *et al*. LFNet: Light field fusion network for salient object detection. IEEE Trans. on Image Processing, 2020, 29: 6276−6287.

[15] Fu K, Jiang Y, Ji GP, *et al*. Light field salient object detection: A review and benchmark. arXiv:2010. 04968, 2020.

[16] Zhang Y. ASOD60K: Audio-induced salient object detection in panoramic videos. arXiv:2107.11629, 2021.

[17] Zhang Q, Cong R, Li C, *et al*. Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans. on Image Processing, 2020, 30: 1305−1317.

[18] Fan DP, Wang W, Cheng MM, *et al*. Shifting more attention to video salient object detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8554−8564.

[19] Shi CJ, Zhang WM, Chen HR, *et al*. Survey of salient object detection based on deep learning. Journal of Frontiers of Computer Science and Technology, 2021, 15(2): 219−232 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.2007074]

[20] Qian XL, Bai Z, Chen Y, *et al*. A review of co-saliency detection. Acta Electronica Sinica, 2019, 47(6): 1352−1365 (in Chinese with English abstract). [doi: 10.3969/j.issn.0372-2112.201906024]

[21] Cong RM, Lei JJ, Fu HZ, *et al*. Research progress of video saliency detection. Ruan Jian Xue Bao/ Journal of Software, 2018, 29(8): 2527−2544 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5560.htm [doi: 10.13328/j.cnki.jos.005560]

[22] Ding Y, Liu YW, Liu JX, *et al*. An overview of research progress on saliency detection of panoramic VR images. Acta Electronica Sinica, 2019, 47(7): 1575−1583 (in Chinese with English abstract). [doi: 10.3969 /j.issn.0372-2112.2019. 07.024]

[23] Liu YM, Zhang J, Zhang XD, *et al*. Review of saliency detection on light fields. Journal of Image and Graphics, 2020, 25(12): 2465−2483 (in Chinese with English abstract). [doi: 10.11834/jig.190679]

[24] Zhou T, Fan D, Cheng MM, *et al*. RGB-D salient object detection: A survey. Computational Visual Media, 2021, 7(1): 37−69.

[25] Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[26] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770−778.

[27] Fan DP, Zhai Y, Borji A, *et al*. BBS-net: RGB-D salient object detection with a bifurcated backbone strategy network. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 275−292.

[28] Liu N, Zhang N, Han J. Learning selective self-mutual attention for RGB-D saliency detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13756−13765.

[29] Qu L, He S, Zhang J, *et al*. RGBD salient object detection via deep fusion. IEEE Trans. on Image Processing, 2017, 26(5): 2274−2285.

[30] Zhao X, Zhang L, Pang Y, *et al*. A single stream network for robust and real-time RGB-D salient object detection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 646−662.

[31] Huang P, Shen CH, Hsiao HF. RGB-D salient object detection using spatially coherent deep learning framework. In: Proc. of the IEEE 23rd Int'l Conf. on Digital Signal Processing. Beijing: IEEE, 2018. 1−5.

[32] Liu Z, Shi S, Duan Q, *et al*. Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing, 2019, 363: 46−57.

[33] Huang R, Xing Y, Wang ZZ. RGB-D salient object detection by a CNN with multiple layers fusion. IEEE Signal Processing Letters, 2019, 26(4): 552−556.

[34] Chen Q, Liu Z, Zhang Y, *et al*. RGB-D salient object detection via 3D convolutional neural networks. In: Proc. of the AAAI Conf. on Artificial Intelligence. Virtual Conf.: AAAI, 2021. 35(2): 1063−1071.

[35] Zhao J, Zhao Y, Li J, *et al*. Is depth really necessary for salient object detection? In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1745−1754.

[36] Zhao Y, Zhao J, Li J, *et al*. RGB-D salient object detection with ubiquitous target awareness. IEEE Trans. on Image Processing, 2021, 30: 7717−7731.

[37] Ji W, Li J, Zhang M, *et al*. Accurate RGB-D salient object detection via collaborative learning. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 52−69.

[38] Zhu C, Cai X, Huang K, *et al*. PDNet: Prior-model guided depth-enhanced network for salient object detection. In: Proc. of the IEEE Int'l Conf. on Multimedia and Expo. Shanghai: IEEE, 2019. 199−204.

[39] Chen H, Li Y, Su D. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognition, 2019, 86: 376−385.

[40] Li G, Liu Z, Ye L, *et al*. Cross-modal weighting network for RGB-D salient object detection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 665−681.

[41] Zhao JX, Cao Y, Fan DP, *et al*. Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3927−3936.

[42] Li G, Liu Z, Ling H. ICNet: Information conversion network for RGB-D based salient object detection. IEEE Trans. on Image Processing, 2020, 29: 4873−4884.

[43] Piao Y, Rong Z, Zhang M, *et al*. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9060−9069.

[44] Chen S, Fu Y. Progressively guided alternate refinement network for RGB-D salient object detection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 520−538.

[45] Pang Y, Zhang L, Zhao X, *et al*. Hierarchical dynamic filtering network for RGB-D salient object detection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 235−252.

[46] Wu YH, Liu Y, Xu J, *et al*. MobileSal: Extremely efficient RGB-D salient object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021.

[47] Fu K, Fan DP, Ji GP, *et al*. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3052−3062.

[48] Fu K, Fan DP, Ji GP, *et al*. Siamese network for RGB-D salient object detection and beyond. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021.

[49] Huang Z, Chen HX, Zhou T, *et al*. Multi-level cross-modal interaction network for RGB-D salient object detection. Neurocomputing, 2021, 452: 200−211.

[50] Liu D, Hu Y, Zhang K, *et al*. Two-stream refinement network for RGB-D saliency detection. In: Proc. of the IEEE Int'l Conf. on Image Processing. IEEE, 2019. 3925−3929.

[51] Zhang Q, Wang X, Jiang J, *et al*. Deep learning features inspired saliency detection of 3D images. In: Proc. of the Pacific RIM Conf. on Multimedia. Xi'an: Springer, 2016. 580−589.

[52] Han J, Chen H, Liu N, *et al*. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Trans. on Cybernetics, 2017, 48(11): 3171−3183.

[53] Liu Z, Tan Y, He Q, *et al*. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. IEEE Trans. on Circuits and Systems for Video Technology, 2021.

[54] Chen H, Li Y. Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2019, 28(6): 2825−2835.
[55] Chen H, Li Y, Su D. Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection. IEEE Trans. on Cybernetics, 2019, 50(11): 4808−482.
[56] Zhang Z, Lin Z, Xu J, *et al*. Bilateral attention network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 1949−1961.
[57] Li C, Cong R, Kwong S, *et al*. ASIF-net: Attention steered interweave fusion network for RGB-D salient object detection. IEEE Trans. on Cybernetics, 2021, 51(1): 88−1.
[58] Liang F, Duan L, Ma W, *et al*. CoCNN: RGB-D deep fusion for stereoscopic salient object detection. Pattern Recognition, 2020, 104: 107329.
[59] Chen H, Li Y. Progressively complementarity-aware fusion network for RGB-D salient object detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3051−306.
[60] Zhang M, Zhang Y, Piao Y, *et al*. Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 4107−4115.
[61] Chen H, Li YF, Su D. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In: Proc. of the IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Madrid: IEEE, 2018. 6821−6826.
[62] Chen H, Li YF, Su D. M3Net: Multi-scale multi-path multi-modal fusion network and example application to RGB-D salient object detection. In: Proc. of the IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Vancouver: IEEE, 2017. 4911−4916.
[63] Zhou W, Chen Y, Liu C, *et al*. GFNet: Gate fusion network with Res2Net for detecting salient objects in RGB-D images. IEEE Signal Processing Letters, 2020, 27: 800−804.
[64] Wang N, Gong X. Adaptive fusion for RGB-D salient object detection. IEEE Access, 2019, 7: 55277−55284.
[65] Zhou W, Lv Y, Lei J, *et al*. Global and local-contrast guides content-aware fusion for RGB-D saliency prediction. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2021, 51(6): 3641−3649.
[66] Zhang W, Jiang Y, Fu K, *et al*. BTS-net: Bi-directional transfer-and-selection network for RGB-D salient object detection. In: Proc. of the 2021 IEEE Int'l Conf. on Multimedia and Expo. Virtual Conf.: IEEE, 2021. 1−6.
[67] Wen H, Yan C, Zhou X, *et al*. Dynamic selective network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 9179−9192.
[68] Luo A, Li X, Yang F, *et al*. Cascade graph neural networks for RGB-D salient object detection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 346−364.
[69] Jiang B, Zhou Z, Wang X, *et al*. CmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks. IEEE Trans. on Multimedia, 2020, 23: 1343−1353.
[70] Chen H, Deng Y, Li Y, *et al*. RGBD salient object detection via disentangled cross-modal fusion. IEEE Trans. on Image Processing, 2020, 29: 8407−8416.
[71] Zhou T, Fu H, Chen G, *et al* Specificity-preserving RGB-D saliency detection. arXiv:2108.08162, 2021.
[72] Huang N, Yang Y, Zhang D, *et al*. Employing bilinear fusion and saliency prior information for RGB-D salient object detection. IEEE Trans. on Multimedia, 2021.
[73] Zhang J, Fan DP, Dai Y, *et al*. RGB-D saliency detection via cascaded mutual information minimization. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 4318−4327.
[74] Zhang M, Fei SX, Liu J, *et al*. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 374−39.
[75] Zhang C, Cong R, Lin Q, *et al*. Cross-modality discrepant interaction network for RGB-D salient object detection. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. Chengdu: ACM, 2021. 2094−2102.
[76] Ding Y, Liu Z, Huang M, *et al*. Depth-aware saliency detection using convolutional neural networks. Journal of Visual Communication and Image Representation, 2019, 61: 1−9.
[77] Piao Y, Ji W, Li J, *et al*. Depth-induced multi-scale recurrent attention network for saliency detection. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7254−7263.
[78] Li C, Cong R, Piao Y, *et al*. RGB-D salient object detection with cross-modality modulation and selection. In: Proc. of the European Conf. on Computer Vision. Glasgow: Spring, 2020. 225−241.
[79] Zhou W, Zhu Y, Lei J, *et al*. CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images. IEEE Trans. on Multimedia, 2021.
[80] Sun P, Zhang W, Wang H, *et al*. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Virtual Conf.: IEEE, 2021. 1407−1417.
[81] Fan DP, Lin Z, Zhang Z, *et al*. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. IEEE Trans. on Neural Networks and Learning Systems, 2020, 32(5): 2075−2089.

[82] Huang R, Xing Y, Zou Y. Triple-complementary network for RGB-D salient object detection. IEEE Signal Processing Letters, 2020, 27: 775−779.
[83] Wang X, Li S, Chen C, *et al*. Data-level recombination and lightweight fusion scheme for RGB-D salient object detection. IEEE Trans. on Image Processing, 2020, 30: 458−471.
[84] Barchid S, Mennesson J, jéraba C. Review on indoor RGB-D semantic segmentation with deep convolutional neural networks. In: Proc. of the Int'l Conf. on Content-Based Multimedia Indexing. Lille: IEEE, 2021. 1−4.
[85] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
[86] Wang X, Girshick R, Gupta A, *et al*. Non-local neural networks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2018. 7794−7803.
[87] Cong R, Lei J, Zhang C, *et al*. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Processing Letters, 2016, 23(6): 819−823.
[88] Chen Z, Cong R, Xu Q, *et al*. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 7012−7024.
[89] Chen C, Wei J, Peng C, *et al*. Depth-quality-qware salient object detection. IEEE Trans. on Image Processing, 2021, 30: 2350−2363.
[90] Zhang W, Ji GP, Wang Z, *et al*. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. Chengdu: ACM, 2021. 731−740.
[91] Zhai Y, Fan DP, Yang J, *et al*. Bifurcated backbone strategy for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 8727−8742.
[92] Chen Q, Fu K, Liu Z, *et al*. EF-net: A novel enhancement and fusion network for RGB-D saliency detection. Pattern Recognition, 2021, 112: 10774.
[93] Liao G, Gao W, Jiang Q, *et al*. Mmnet: Multi-stage and multi-scale fusion network for RGB-D salient object detection. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 2436−2444.
[94] Liu Z, Zhang W, Zhao P. A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. Neurocomputing, 2020, 387: 210−220.
[95] Li G, Liu Z, Chen M, *et al*. Hierarchical alternate interaction network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 3528−3542.
[96] Zhang J, Fan DP, Dai Y, *et al*. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8582−8591.
[97] Zhang J, Fan DP, Dai Y, *et al*. Uncertainty inspired RGB-D saliency detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021.
[98] Zhang M, Ren W, Piao Y, *et al*. Select, supplement and focus for RGB-D saliency detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3472−3481.
[99] Liu Z, Wang Y, Tu Z, *et al*. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. Chengdu: ACM, 2021. 4481−4490.
[100] Chen C, Wei J, Peng C, *et al*. Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion. IEEE Trans. on Image Processing, 2020, 29: 4296−4307.
[101] Jin WD, Xu J, Han Q, *et al*. CDNet: Complementary depth network for RGB-D salient object detection. IEEE Trans. on Image Processing, 2021, 30: 3376−339.
[102] Ji W, Li J, Yu S, *et al*. Calibrated RGB-D salient object detection. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Virtual Conf.: IEEE, 2021. 9471−9481.
[103] Niu Y, Geng Y, Li X, *et al*. Leveraging stereopsis for saliency analysis. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 454−461.
[104] Cheng Y, Fu H, Wei X, *et al*. Depth enhanced saliency detection method. In: Proc. of Int'l Conf. on Internet Multimedia Computing and Service. Xiamen: ACM, 2014. 23−27.
[105] Peng H, Li B, Xiong W, *et al*. RGBD salient object detection: A benchmark and algorithms. In: Proc. of the European Conf. on Computer Vision. Zurich: Springer, 2014. 92−109.
[106] Li N, Ye J, Ji Y, *et al*. Saliency detection on light field. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 2806−2813.
[107] Ju R, Ge L, Geng W, *et al*. Depth saliency based on anisotropic center-surround difference. In: Proc. of the IEEE Int'l Conf. on Image Processing. Paris: IEEE, 2014. 1115−1119.
[108] Zhu C, Li G. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: Proc. of the IEEE Int'l Conf. on Computer Vision Workshops. Venice: IEEE, 2017. 3008−3014.
[109] Liu N, Zhang N, Shao L, *et al*. Learning selective mutual attention and contrast for RGB-D saliency detection. arXiv:2010.05537, 2020.

[110] Fan DP, Cheng MM, Liu Y, *et al*. Structure-measure: A new way to evaluate foreground maps. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 4548−4557.
[111] Fan DP, Gong C, Cao Y, *et al*. Enhanced-alignment measure for binary foreground map evaluation. arXiv:1805.10421, 2018.
[112] Li J, Ji W, Bi Q, *et al*. Joint semantic mining for weakly supervised RGB-D salient object detection. In: Proc. of the Advances in Neural Information Processing Systems. Virtual Conf.: MIT Press, 2021.
[113] Zhu L, Wang X, Li P, *et al*. S3Net: Self-supervised self-ensembling network for semi- supervised RGB-D salient object detection. IEEE Trans. on Multimedia, 2021.

附中文参考文献:

[1] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述. 软件学报, 2019, 30(2): 416−439. http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]
[7] 肖德贵, 辛晨, 张婷, 等. 显著性纹理结构特征及车载环境下的行人检测. 软件学报, 2014, 25(3): 675−689. http://www.jos.org.cn/1000-9825/4438.html [doi: 10.13328/j.cnki.jos.004438]
[19] 史彩娟, 张卫明, 陈厚儒, 等. 基于深度学习的显著性目标检测综述. 计算机科学与探索, 2021, 15(2): 219−232. [doi: 10.3778/j.issn.1673-9418.2007074]
[20] 钱晓亮, 白臻, 陈渊, 等. 协同视觉显著性检测方法综述. 电子学报, 2019, 47(6): 1352−1365. [doi: 10.3969/j.issn.0372-2112.2019.06.024]
[21] 丛润民, 雷建军, 付华柱, 等. 视频显著性检测研究进展. 软件学报, 2018, 29(8): 2527−2544. http://www.jos.org.cn/1000-9825/5560.htm [doi: 10.13328/j.cnki.jos.005560]
[22] 丁颖, 刘延伟, 刘金霞, 等. 虚拟现实全景图像显著性检测研究进展综述. 电子学报, 2019, 47(7): 1575−1583. [doi: 10.3969/j.issn.0372-2112.2019.07.024]
[23] 刘亚美, 张骏, 张旭东, 等. 光场显著性检测研究综述. 中国图像图形学报, 2020, 25(12): 2465−2483. [doi: 10.11834/ jig.190679]

丛润民(1989－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为计算机视觉, 水下环境感知, 显著性检测.

张晨(1998－), 男, 硕士生, 主要研究领域为计算机视觉, 包括 RGB-D 显著性目标检测, 目标检测.

徐迈(1981－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为视频通信, 图像处理, 计算机视觉.

刘鸿羽(2000－), 男, 本科生, 主要研究领域为面向高分辨率图像的显著性目标检测.

赵耀(1967－), 男, 博士, 教授, 所长, 博士生导师, CCF 杰出会员, 主要研究领域为跨媒体智能处理, 图像视频编码.