基于深度学习的RGB-D图像显著性目标检测前沿进展

黄年昌[1),2)] 杨 阳[1),2)] 张 强[1),2)] 韩军功[3)]

[1)](西安电子科技大学高性能电子装备机电集成制造全国重点实验室 西安 710071)

[2)](西安电子科技大学机电工程学院 西安 710071)

[3)](谢菲尔德大学计算机科学系 谢菲尔德 英国 S10 2TN)

摘 要 显著性目标检测是计算机视觉领域的基础问题之一,旨在对图像中最吸引人注意的目标进行检测和分割。随着深度学习技术的发展,基于RGB(Red-Green-Blue)图像的显著性目标检测算法取得了巨大进步,在简单场景下已经取得较为满意的结果。然而,局限于可见光相机的成像能力,RGB图像易受到光照条件的影响,且无法捕捉场景的三维空间信息。相应地,基于RGB图像的显著性目标检测算法通常难以在一些复杂场景下取得较好的检测效果。近年来,随着深度成像技术不断发展和硬件成本不断降低,深度相机得到了广泛应用。其捕获的场景空间信息,与可见光图像获取的场景细节信息相互补充,有助于提升复杂场景下显著性目标检测性能。因此,RGB-深度(RGB-Depth,RGB-D)图像显著性目标检测引起了学者广泛研究。本文对近期基于深度学习的RGB-D图像显著性目标检测算法进行了整理和分析。首先,分析了多模态RGB-D图像显著性目标检测所面临的关键问题,并以此对现有算法解决这些关键问题的主要思路和方法进行了总结和梳理。然后,介绍了用于RGB-D图像显著性目标检测算法研究的主流数据集和常用性能评价指标,并对各类主流模型进行了定量比较和定性分析。最后,本文进一步分析了RGB-D图像显著性目标检测领域有待解决的问题,同时对今后可能的研究趋势进行了展望。

关键词 显著性目标检测;RGB图像;深度图像;深度学习;多模态图像处理

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.00284

Progress of RGB-D Salient Object Detection Based on Deep Learning

HUANG Nian-Chang[1),2)] YANG Yang[1),2)] ZHANG Qiang[1),2)] HAN Jun-Gong[3)]

[1)](State Key Laboratory of Electromechanical Integrated Manufacturing of High-performance Electronic Equipments, Xidian University, Xi'an 710071)

[2)](School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071)

[3)](Department of Computer Science, University of Sheffield, Sheffield S10 2TN UK)

Abstract Salient object detection (SOD) is one of the fundamental research tasks for computer vision, aiming at detecting and segmenting the most visual attractive objects in an image. The RGB-based SOD algorithms have achieved promising performance in those simple scenarios. However, limited by the susceptibility to lighting conditions and the inability to capture 3D spatial information from the scenes, those RGB-based SOD algorithms cannot work well in some complex scenarios. Recently, with the rapid development of depth imaging technologies, the

收稿日期:2023-12-20;在线发布日期:2024-09-29。本课题得到中国博士后科学基金(2023M742745)、国家资助博士后研究人员计划(GZB20230559)、广东省基础与应用基础研究基金委员会、粤穗联合基金青年基金项目(2023A1515110165)、陕西省创新团队项目(2018TD-012)、国家自然科学基金(61773301))、河北工业大学电气设备可靠性与智能化国家重点实验室基金项目(EERI_KF2022005)资助。**黄年昌**,博士,助理研究员,主要研究领域为计算机视觉中的多模态图像处理和深度学习。E-mail:huangnianchang@163.com。**杨 阳**,博士研究生,主要研究领域为多模态图像处理和深度学习。**张 强**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为计算机视觉与智能图像处理。E-mail:qzhang@xidian.edu.cn。**韩军功**,博士,教授,博士生导师,主要研究领域为计算机视觉、人工智能和机器学习。

hardware costs of depth cameras are constantly decreasing and depth cameras have been widely applied. Different from RGB images that mainly provide some color and texture information, depth images can provide additional geometric structures, such as spatial cues and 3D layouts, which are robust to light and color changing. This helps an SOD model to deal with those complex scenarios. Accordingly, RGB-D SOD has attracted extensive interests. This paper provides a comprehensive review for the progress of deep learning based RGB-D models. Specifically, this paper first analyzes and summarizes some major issues and then categorizes those existing algorithms according to their primarily addressed issues. After that, we introduce some widely-used evaluation metrics as well as the datasets for RGB-D SOD, and provide quantitative comparisons and qualitative analyses of various mainstream algorithms. Based on the results, we point out some limitations of current methods. At last, we outline some challenges as well as and future research trends in this field.

Keywords salient object detection; RGB image; depth image; deep learning; multi-modal image processing

1 引 言

显著性目标检测旨在检测并分割出图像中最吸引人注意的目标，作为重要的图像预处理算法之一，已被广泛应用于计算机视觉的多个子任务中，如图像跟踪[1]、图像压缩[2]、语义分割[3]、目标检测[4]和目标分割[5]。近年来，深度学习方法强大的特征提取能力有效地推动了RGB(Red-Green-Blue)图像显著性目标检测的发展。然而，RGB图像具有一定局限性，即RGB图像易受到光照条件的影响，且无法捕捉场景的三维空间信息，使得现有RGB图像显著性目标检测算法[6]在一些复杂场景下，如图像包含多个显著目标、图像背景较为复杂、目标边界复杂、不同类别目标外观相似、同类别目标外观差异大等，难以取得较好的检测效果。近年来，随着成像技术快速发展，多模态相机应用日益广泛，如RGB-深度相机(RGB-D相机)和RGB-红外相机(RGB-T相机)等。其中，RGB-D相机利用深度图像对RGB图像进行补充，其能够捕捉场景的三维空间信息，同时对场景光照变化具有鲁棒性，能够在上述复杂场景中取得更好的效果。且相比于RGB-T相机，RGB-D相机成本更加低廉，被广泛应用于机器人、无人机、自动驾驶等领域中。因此，研究如何利用RGB图像和深度图像之间的互补信息提升显著性目标检测性能引起了学者广泛兴趣。

传统的RGB-D图像显著性目标检测方法通常利用传统的手工特征和不同的先验信息检测显著目标[7-9]。例如，Guo等人[9]首先对可见光图像和深度图像进行超像素分割。在此基础上，分别利用颜色先验和深度先验得到可见光图像显著图和深度图像显著图，之后，通过融合并细化单模态显著图得到初始显著性结果。最后，利用元胞自动机(Cellular Automata)在初始显著性图上迭代传播显著性，生成更完整的显著性目标检测结果。然而，手工特征主要为低层级细节特征，难以表达高层级语义信息，进而导致传统方法在复杂场景下检测效果较差。同时，手工特征的设计依赖于有限的先验知识，往往针对特定的场景有效，难以拓展到通用场景的检测中[10]。

相比于传统方法，深度学习算法能够自动地从大量数据中学习辨别的低层级细节特征和高层级语义特征[11-12]，逐渐成为计算机视觉领域的主流算法。基于此，Qu等人[13]于2017年提出了首个RGB-D图像显著性目标检测(DFNet)的深度学习方法。相比于传统算法，该算法取得了显著性能提升。此后，基于深度学习的RGB-D图像显著性目标检测算法飞速发展，并逐步成为RGB-D图像显著性目标检测领域的主流。因此本文主要综述基于深度学习的RGB-D图像显著性目标检测的前沿进展。

目前已有学者对基于深度学习的RGB-D[14-16]显著性目标检测方法进行总结。Tao等人[14]分别从传统方法/深度方法、融合策略、网络结构、注意力模型和轻量化网络5个方面分析和介绍RGB-D图像显著性目标检测现状。例如，如图1所示，根据融合策略不同，其将RGB-D图像显著性目标检测模型分为前期融合[17-19]、后期融合[20-21]和中间级融合[22-25]三类。Ren等人[15]重点介绍了RGB-D图像显著性检测的各个子领域，包括基于深度图的显著性目标

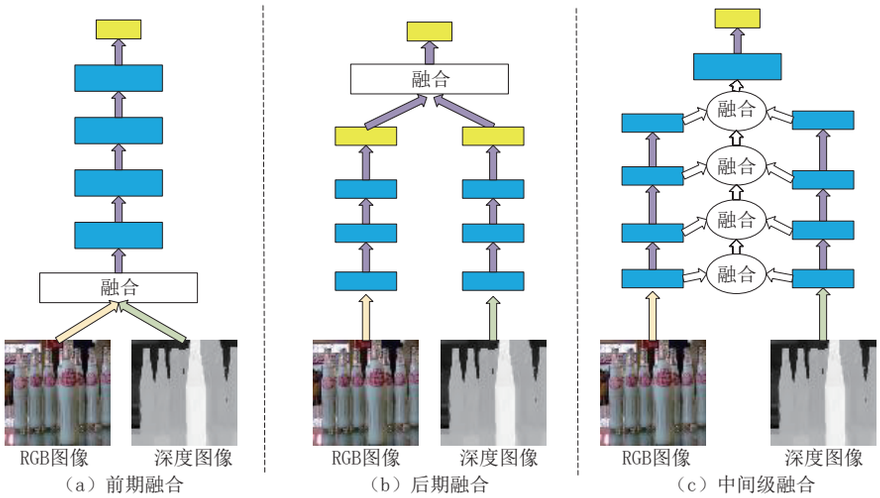


图1 RGB-D图像显著性目标检测模型分类示意图

检测、RGB-D图像显著性目标检测和RGB-D图像协同显著性目标检测。

上述文献有助于读者明晰现有RGB-D图像显著性目标检测方法的主流设计结构和任务目标，然而其不便于读者了解和掌握现有基于深度学习的RGB-D图像显著性目标检测方法模型首要关注的关键问题，以及不同方法模型解决这些关键问题所采用的设计思路。为此，本文以RGB-D图像显著性目标检测所面临的关键问题为出发点，对现有基于深度学习的RGB-D显著目标检测所面临的关键问题及相应的解决方案进行归纳和总结，为初学者及相关领域的学者提供新的参考。具体来说，本文归纳了现有的RGB-D图像显著性目标检测算法并将其所面临的关键问题，并总结为以下五点：干扰信息问题、特征提取问题、多模态信息融合问题、上下文信息挖掘问题和模型复杂问题。基于所总结的问题，本文整理现有基于深度学习的RGB-D图像显著性目标检测解决这些问题的主要研究思路。随后，本文给出RGB-D图像显著性目标检测领域主流的数据集和常用的性能评价指标，并在此基础上，对现在方法的性能进行评估。最后，进一步分析RGB-D图像显著性目标检测领域有待解决的问题，并分析本领域未来可能发展方向。

2 RGB-D图像显著性目标检测的关键问题和代表性方法

本节首先对现有RGB-D图像显著性目标检测方法主要关注的五个问题，即干扰信息问题、特征提取问题、多模态信息融合问题、上下文信息挖掘问题和模型复杂问题进行介绍，并进一步归纳分析现有方法解决上述问题的主要思路，并对相关典型模型进行了介绍。需要特别说明的是，相比于现有的分类方法，其不同类别中的模型具有明显的区别。然而，如图2所示，若从关键问题的角度对现有方法进行分类，不同类别中的模型也不可避免地存在交集，即同一个显著目标检测模型可能同时涉及对多个关键问题的研究，某模型中某个模块亦有可能同时解决多个关键问题。例如，可以通过提高其上下文信息提取能力的角度，优化特征提取模块，进而使得特征提取过程能够同时解决特征提取问题和上下文信息挖掘问题。我们在表1中对现有方法所面临的关键问题、解决关键问题的主要思路、优缺点和对应的模型方法进行了简要汇总，详细内容如下。

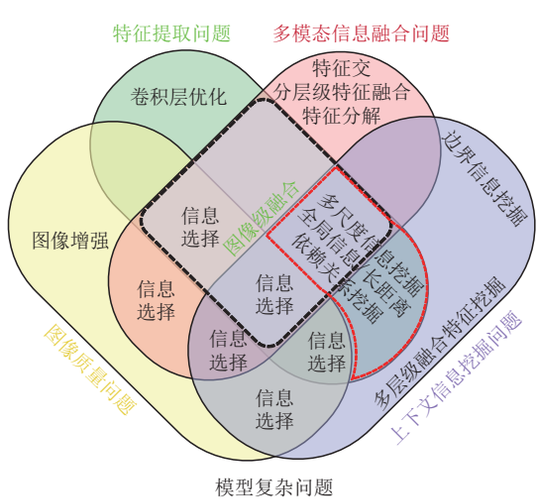


图2 从关键问题的角度对RGB-D图像显著性目标检测模型进行分类的可视化示意图

表1　RGB-D图像显著性目标检测模型总体研究范畴：关键问题、解决思路、优缺点和模型

关键问题	解决思路			优点		缺点		典型模型
干扰信息问题	信息选择	图像级选择		结构简单		局部质量差会导致整个图像被舍弃		D^3Net[26]
		特征级选择	基于空间注意力的方法	较好解决局部质量问题	设计灵活，即插即用	学习过程不可控，难以对干扰信息建模并监督		DQM[21]、DPANet[27]、AFINet[28]、GFNet[29]、DUF[30]、DQSD[31]
			基于通道注意力的方法	筛选辨别特征类型				ACCF[32]、TANet[33]、JCU[34]、MCMFNet[35]、MGU[36]、MADNet[37]
			基于混合注意力的方法	同时关注局部质量和特征质量				BPGNet[38]、DSNet[39]、SSF[40]、MFUR-Net[41]、STA[42]
	图像增强	跨模态信息生成		理论完善		结果取决于跨模态图像生成效果，而模态差异导致高质量跨模态生成十分困难		FCFNet[43]、CDNet[44]、ISDNet[45]、ISP[46]
特征提取问题	基于CNN的方法			-		难以有效提取全局信息		DFNet[13]
	基于图网络的方法			能够获取大尺度信息和关系信息	发展成熟，能够利用预训练数据的先验信息	网络结构设计复杂，计算量大	缺少针对深度数据设计和预训练的模型	Cas-GNN[47]
	基于Transformer的方法			能够获取长距离信息和全局信息		计算量大		SwinNet[48]、VST[50]、HRTRansNet[51]
多模态信息融合问题	图像级融合	通道级联		简单方便，简化网络结构		效果较差		SCDLF[52]、UC-Net[17]、SSRCNN[53]
		时间序列级联		考虑深度图像特性，融合效果较好		结构相对复杂，参数量更大		DGD[54]、RGBD3D[18]、RGBD3D++[55]
		3D空间转换		在3D空间中融合深度信息和RGB信息		相关操作相对复杂		MVSalNet[56]
	特征级融合	特征交互		设计灵活，即插即用	效果优于图像级融合和决策级融合	网络结构设计复杂，计算量大		PCFNet[22]、BTS-Net[23]、IRFR-Net[57]、ASIF-Net[58]、LI-ANet[59]、CCNet[60]、ACMF[61]、EBFS[24]、CAN[62]、PDD[63]、SP-SN[64]、CIRNet[65]、C2DFNet[66]、CAVER[67]
		分层级特征融合		根据不同层级特征特性细粒度融合				CCAFNet[68]、DCPG-Net[69]、FRDT[70]、CMWNet[71]、ACF-Net[72]
		特征分解		根据不同模态特征类型细粒度融合				DCMF[73]、CMIM[74]、SP-Net[75]
	决策级融合方法			结构简单		效果较差		AFNet[20]、KDQA[76]
	混合融合方法（其他）			能够结合不同融合方式的优点		网络结构更加复杂，计算量更大		TMFNet[77]、Triple-Net[78]

续表

关键问题	解决思路		优点	缺点		典型模型
上下文信息挖掘问题	多尺度特征挖掘		挖掘不同尺度特征，能够较好地应对显著目标尺度变化			PMFNet[85]、MMNet[36]、PGFNet[86]、AILNet[87]、HDFNet[88]、DMRA[89]、GCENet[90]
	多层级融合特征挖掘		挖掘不同层级特征间互补信息，准确定位显著目标和恢复显著目标边界	网络模型设计复杂，计算量相对较大		BBSNet[91]、CFID-Net[92]、CPFP[93]、MCI-Net[94]、MC-Net[95]、MRINet[96]、FANet[97]、PGHF-Net[98]、HINet[99]、BBSNet[100]
	全局信息/长距离依赖关系挖掘		挖掘全局信息和长距离依赖关系，能够较好地应对显著目标尺度和位置变化			MMCI[101]、Tri-TransNet[102]、M2RNet[103]、CM-LCG[104]
	边界信息挖掘		物理意义明确，提高显著目标轮廓预测效果	网络模型设计复杂，计算量相对较大，且需要引入额外信息		CENet[105]、Swin-Net[48]、EGA-Net[106]
模型复杂问题	单分支轻量化模型	网络结构轻量化	大幅降低网络参数量	发展相对成熟，优化空间小，难以进一步优化		DANet[113]
		跨模态信息迁移	通过生成模态信息迁移的方式减少输入模态，简化网络结构	跨模态迁移过程中会损失大量辨别信息		CoNet[114]、A2dele[115]、DKDNet[116]
	双分支轻量化模型	模块轻量化	大幅降低网络参数量	发展相对成熟，优化空间小，难以进一步优化	基于双分支框架，网络结构更加复杂	JL-DCF[117]
		框架轻量化 异构框架	针对RGB图像和深度图像特性进行轻量化优化设计	网络模型相对复杂		ATS[119]、GRNet[120]、PGAR[121]、MoAD-Net[122]、MobileSal[118]
		框架轻量化 中间层级特征融合框架	能够结合图像级融合和特征级融合的优势，显著降低参数量			MLF[123]

2.1 干扰信息问题

如图3所示，干扰信息问题指的是：由于图像特性不同、图像质量差异或者背景干扰信息等因素使得某个模态图像中存在一定的干扰信息，导致显著目标检测不准确。例如，在不同模态图像成像的过程中，因成像环境（如低光照、雾霾等特殊场景），或成像技术（如低分辨率相机和外界设备干扰等）的限制，某个模态的成像传感器不可避免地受到噪声影响，导致生成低质量图像（图3的第一行），进而在显著目标检测过程中引入干扰信息。又或者在RGB图像中前景目标和背景目标存在相似的外貌或者颜色信息（图3的第二行），在显著目标检测过程，这些区域同样会引入干扰信息。

现有方法解决干扰信息问题的主要思路可以概括为信息选择[26-42]和图像增强[43-46]两种。其中，信息选择是通过选择输入图像中的辨别信息，抑制非辨别信息的方式，缓解干扰信息问题。与之不同，图像增强则是通过跨模态信息生成、生成对抗等理论或方法，提高RGB图像或深度图像质量，或对低质量RGB图像/深度图像进行补充，以降低低质量图像对检测过程的影响。

2.1.1 信息选择

根据其选择对象不同，信息选择的方法可以进一步分为图像级选择方法和特征级选择方法。其中，图像级选择方法在将图像输入到RGB-D显著目标检测模型之前，首先辨别输入图像的质量，然后通过抑制甚至移除低质量单模态图像的方式，减少低质量图像

图3　包含干扰信息的RGB-D图像示例

中干扰信息对显著性目标检测的影响。与图像级选择方法不同，特征级选择方法旨在通过对RGB特征和深度特征（或融合特征）中辨别力强的特征赋予较高的权重（特征增强），对其中辨别力差的特征赋予较低的权重（特征抑制）的方式，缓解干扰信息问题。

（1）图像级选择：在基于单分支网络结构的RGB-D图像显著性目标检测模型中，RGB图像与深度图像往往通过级联的方式进行融合。因此，这些模型主要通过设计图像级选择策略决定是否利用某一模态图像（深度图像或者RGB图像）进行显著性目标检测。如Fan等人[26]观察到高质量的深度图像往往具有较好的边界区域，进而在其分布中呈现明显的双峰现象。因此，如图4所示，他们提出了一个深度图像净化单元（Depth Depurator Unit，DDU），该单元通过计算深度图像的分布，决定是否保留深度图像进行显著性目标检测。

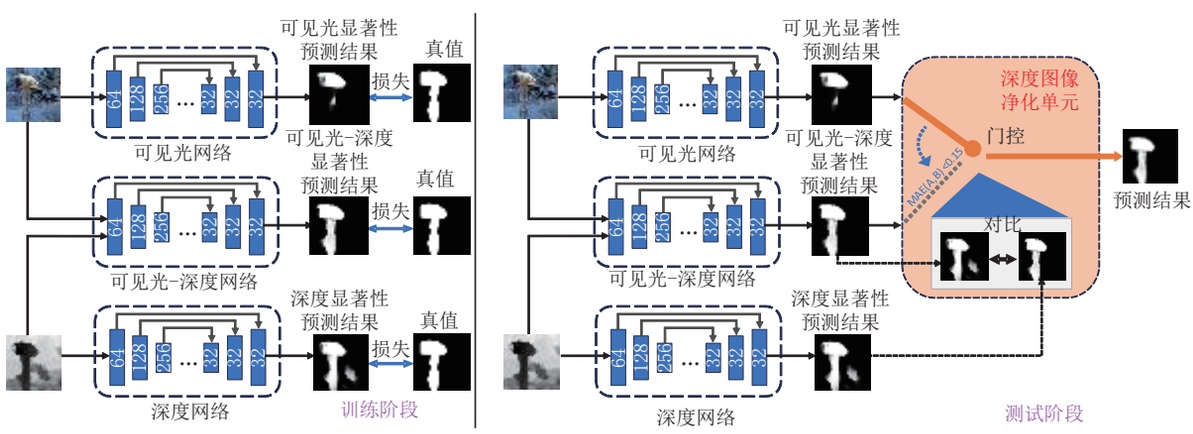


图4　D³Net网络结构示意图[26]

（2）特征级选择：相较于图像级选择方法，特征级选择方法由于其优异的信息挖掘能力往往能取得更好的效果，因此受到了广泛的关注和研究。进一步地，根据选择的方式不同，特征级选择方法可以被进一步划分为基于空间注意力的方法、基于通道注意力的方法以及基于混合注意力的方法。

①基于空间注意力的方法：本类方法主要是受到空间注意力机制的启发，通过生成空间权重图的方式，在空间维度对RGB特征和深度特征进行选择[27-31]。例如，Chen等人[27]考虑到深度图像质量的不可靠性，在其提出的深度图像质量感知网络（Depth Potentiality-Aware Gated Attention Network，DPANet）中，通过阈值分割的办法探究深度图像与基准显著图之间的关系，进而确定深度图像的可靠性。具体地说，其通过探究深度图像与基准显著图之间的关系生成深度图像的可靠性标签，然后利用深度图像的可靠性标签对RGB特征和深度特征进行选择，从而减少低质量深度图像对显著性目标检测的影响。Li等人[28]提出了一个新型的基于注意力机制的特征增强模块（Attention Module，AM）对提取到的RGB特征和深度特征进行增强。首先，利用RGB特征和深度特征分别生成对应的权重图，然后利用生成的权重图对RGB特征和深度特征进行加权以得到增强特征。

②基于通道注意力的方法：本类方法主要是受到通道注意力机制的启发，通过生成通道权重向量的方

式，从通道维度对RGB特征和深度特征进行选择[32-35]。例如，Chen等人[32]提出了一个基于通道注意力的选择模块（Attention-aware Cross-modal Cross-level Fusion，ACCF）。ACCF模块首先通过级联的方式融合当前层级的RGB特征、深度特征和融合特征，然后通过全局池化操作将级联后的特征压缩为向量特征，并利用全连接层和Softmax函数生成通道权重，并对输入特征进行选择，从而保留其辨别特征，抑制其非辨别性特征，进而更好地进行显著性目标检测。Chen等人[33]提出了三分支注意力关注网络（Three-stream Attention-aware Network，TANet），其首先在特征提取阶段使用一个独立分支融合RGB特征和深度特征，然后在显著图推理阶段通过设计一种基于通道注意力机制的特征选择模块融合多层级特征。与上述方法仅利用融合多模态特征不同，如图5所示，Huang等人[34]同时保留跨模态融合特征和单模态特征，在某一模态图像质量相对较差时，通过另一质量相对较好的单模态图像特征中的辨别信息，补充多模态融合特征，以便更好地抑制低质量图像中的干扰信息。同时，提出一种新型的基于通道注意力的特征选择模块（Feature Selection Module，FSM），旨在自适应性地选择高辨识度的跨模态特征和单模态特征实现最终的显著性目标检测，以降低低质量图像对显著性目标检测结果的影响。

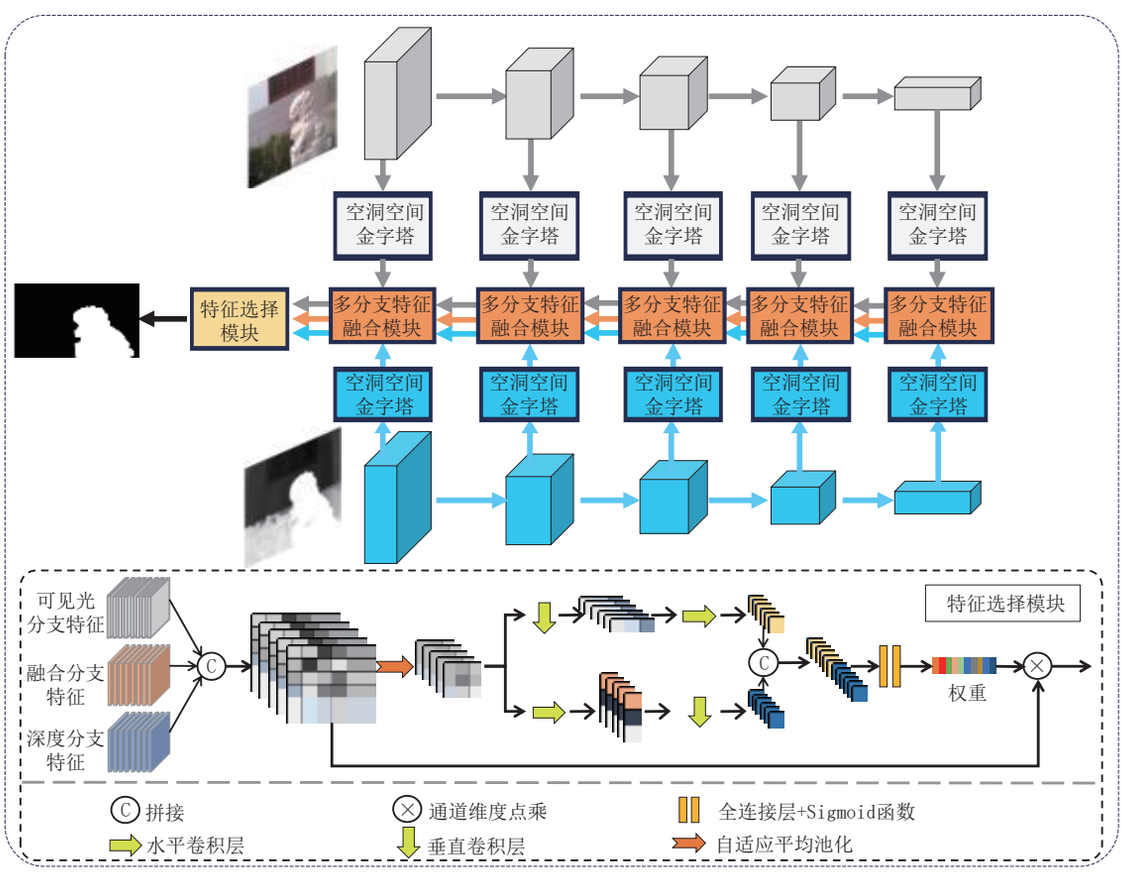


图5 JCU网络结构及特征选择模块示意图[34]

③基于混合注意力的方法：本类方法则是联合使用前两种方法进行特征选择[36-42]。例如，Yang等人[38]提出的双向渐进指导网络（Bi-directional Progressive Guidance Network，BPGNet），首先设计了一种全局上下文信息感知模块（Global Context Awareness Module，GCA），其利用RGB图像信息生成粗糙显著图，指导深度特征提取。然后，设计一种新型辅助特征提取模块（Auxiliary Feature Extraction Module，AFE），在深度特征提取过程中通过空间注意力和通道注意力选择辨别特征。最后，利用深度图像信息指导RGB图像的解码，并通过设计跨模态特征增强模块（Cross-modality Feature Enhancement Module，CFE），再次利用空间注意力和通道注意力选择辨别特征机型显著性目标检测。Wen等人[39]提出的动态选择网络（Dynamic Selective Network，DSNet）首先通过多模态特征全局信息挖掘模块

（Cross-modal Global Context Module，CGCM）提取输入图像的全局信息，然后设计动态选择模块（Dynamic Selective Module，DSM）以挖掘输入特征的多尺度信息，并建立不同模态特征之间的相关性，进而生成通道权重和空间权重对输入特征中的辨别信息进行选择。

2.1.2　图像增强

深度信息在RGB-D显著性目标检测中起着至关重要的作用，它直接决定了后续显著性目标检测的性能。然而，受成像设备和的限制，相较于RGB图像，深度图像往往包含更多的干扰信息。为此，一些学者尝试通过图像增强的方式缓解低质量深度图像对显著性目标检测的影响[43-46]。如图6所示，图像增强的方法旨在通过跨模态信息生成、生成对抗等技术获得高质量的深度图像，并基于生成的深度图像获得更具鉴别性的深度信息用于最终的显著性目标检测。

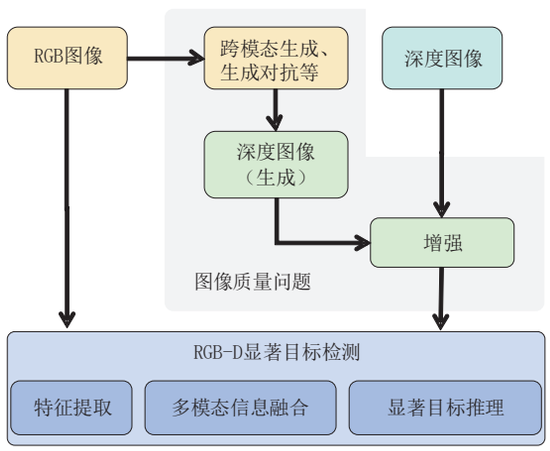


图6　图像增强方法示意图

Jin等人[44]提出的深度信息互补网络（Complementary Depth Network，CDNet）首先通过深度估计技术，利用RGB图像生成估计深度图像，然后通过孪生网络分别从原始深度图像和估计深度图像中提取多层级深度特征并进行融合，在对深度图像进行增强的同时，抑制深度图像中的干扰信息。最后，分别通过高层级多模态图像特征融合模块和低层级多模态图像特征融合模块将融合后的高层级深度特征和低层级深度特征与RGB特征进行融合，实现显著性目标检测。Chen等人[45]提出的两阶段网络同样通过RGB图像生成深度图像对原始深度图像进行增强，以缓解深度图像的质量对显著性目标检测结果带来的影响。在第一阶段提出了一个深度图像估计网络实现伪深度图像生成，并利用生成的深度图像改善原始深度图像，然后在第二阶段将RGB图像、原始深度图像和生成深度图像输入到RGB-D图像显著性目标检测网络，并通过特征选择的方式挖掘RGB-D图像中互补信息进行显著性目标检测。Zuo等人[46]首先利用深度估计方法从RGB图像中生成估计深度图像，然后提出深度恢复模块（Depth Recovery，DR）融合真实深度图像信息和估计深度图像信息，以提高深度图像质量，减少显著目标检测过程中，深度图像带来的干扰信息。

2.2　特征提取问题

在RGB-D图像显著性目标检测方法中，特征提取问题是指如何设计合理的图像特征提取网络，以挖掘更具鉴别性的单模态信息，从而提升显著性目标检测结果。考虑到深度学习算法的数据驱动特性，早期的算法大都采用成熟且经过预训练的卷积神经网络（基于CNN的方法），如VGGNet[11]、ResNet[12]，实现辨别特征提取。然而，CNN更加关注输入图像局部区域的细节/语义特征提取，难以有效捕捉场景的全局信息。事实上，全局信息包含着场景中相关目标信息以及目标之间的相互关系，对显著性目标检测模型的性能起着至关重要的作用。针对CNN的局限性，部分方法[47-50]通过增强其局部/全局信息提取能力的方式解决特征提取问题。这些方法大体可分为：基于图网络的方法和基于Transformer的方法。

（1）基于图网络的方法：Luo等人[47]提出的级联图神经网络（Cascade Graph Neural Network，Cas-GNN）算法，通过在骨干网络中引入图卷积神经网络，充分地挖掘单模态图像特征中包含的关系特征以改善CNN的局限性，进而得到更具辨别性的单模态图像特征。

（2）基于Transformer的方法：Liu等人[48]提出的SwinNet，则是直接利用SwinTransformer[49]替代CNN作为其骨干网络。相比于CNN，SwinTransformer能有效地挖掘RGB-D图像中的上下文信息，进而更好地挖掘每个模态图像中包含的场景信息。Liu等人[50]提出了一个完全基于视觉Transformer的RGB-D图像显著性目标检测模型，即视觉显著性Transformer（Visual Saliency Transformer），其不仅仅利用Transformer实现了多模态图像特征提取，而且实现了多模态图像特征融合以及显著图推理，充分利用了Transformer对上下文信息挖掘的优势。

2.3 多模态信息融合问题

多模态信息融合问题是RGB-D图像显著性目标检测的核心问题[17,20-24,29,51-83]，即如何融合RGB图像和深度图像信息，以充分挖掘和利用RGB图像和深度图像中包含的互补信息，进而弥补单模态RGB图像的缺点。从融合方式上，现有方法可以分为图像级融合方法、特征级融合方法和决策级融合方法。

2.3.1 图像级融合

图像级融合旨在通过一定的融合策略，直接融合RGB图像和深度图像。最常用的图像级融合策略就是将RGB图像和深度图像进行简单级联，得到4通道的输入RGB-D图像[17,52-53]。然而由于RGB图像和深度图像之间的模态差异，通道级联的方式虽然简单高效，但是无法充分地挖掘RGB图像和深度图像之间的互补信息。

部分方法对简单的级联策略进行改进，提出将RGB图像和深度图像按照时间序列级联的方式进行图像级融合[18,54]。例如，Li等人[54]首先将RGB图像和深度图像按照时间序列级联的方式进行图像级融合，然后利用3D卷积提取级联图像的特征。同时，为更充分地提取和挖掘RGB图像和深度图像之间的互补信息，他们还对现有3D卷积进行了一定的优化，其利用深度信息动态地调整3D卷积的感受野，以更好地挖掘RGB-D图像的互补信息。Chen等人[18]则是首先利用时间序列级联的方式级联RGB图像和深度图像，得到伪3DRGB-D图像，并进一步设计一个新型的基于3D卷积神经网络的编码-解码结构，利用3D卷积神经网络提取融合特征，并利用融合特征进行最终的显著性目标检测。他们的后续工作[55]通过提出一种新型的通道模态注意力模块（Channel-Modality Attention module，CMA），进一步优化了上述时间序列级联框架的多模态信息挖掘能力。

此外，Zhou等人[56]提出了一种基于3D空间转换的融合方式，这种方式利用深度图像提供的3D坐标，将RGB图像转换到真实世界的3D空间中，实现了深度模态和RGB模态的融合。在此基础上，Zhou等人进一步提出了一种多视角显著目标检测方法（MVSalNet），MVSalNet将从多个视角从转换后的图像中采样多张RGB图像，然后利用多视角RGB图像预测显著目标。

2.3.2 特征级融合

不同于图像级融合，特征级融合旨在通过一定的融合策略实现RGB特征和深度特征融合。相比于图像级融合方法，特征级融合方法通常能够取得更好的融合效果，因此得到了更为广泛的关注。早期的特征融合方法主要以相加或级联策略为主，具有操作简单的优点，但是难以充分挖掘RGB图像和深度图像中的冗余和互补信息。因此，设计更加高效的特征级融合策略是许多方法的研究重点，总体而言，现有方法主要基于特征交互、分层级特征融合或特征分解的思路，优化特征级融合策略，实现高效多模态信息融合。

（1）特征交互：简单的级联或者相加策略忽略了RGB特征和深度特征之间的信息的交互，进而不能充分地挖掘RGB图像和深度图像之间的互补信息。因此，一些方法提出通过建立RGB特征和深度特征之间的交互关系以充分地挖掘和利用RGB图像和深度图像之间的互补信息[22-24,29,57-63]。

部分方法在级联或相加策略的基础上，通过构建基于RGB信息和深度信息的双向结构，实现RGB信息和深度信息的特征交互[22-23,29,57-60]。例如，Chen等人[22]设计了一种渐进互补感知融合网络（Progressively Complementarity-aware Fusion Network，PCFNet），通过多个跳连接分别实现RGB图像信息和深度图像信息的互相引用，进而实现RGB特征和深度特征的互相感知与信息融合，有效地挖掘了模态间的互补信息。Zhang等人[23]提出了一种新型的双向转移-选择模块（Bidirectional Transfer-and-Selection modules，BTS），该模块通过设计一种双向结构建立RGB特征和深度特征之间的交互关系，利用RGB特征和深度特征彼此之间的信息交互，调整RGB特征深度特征本身信息，进而更好地融合了RGB特征和深度特征。Zhou等人[57]提出了一种交互式递归特征重塑网络（Interactive Recursive Feature-Reshaping Network，IRFR-Net），其通过设计一个上下文信息提取模块、交互融合模块和带权尺度金字塔模块以实现相邻层级RGB特征和深度特征的递归交互和相互指导增强，进而实现了对RGB-D图像互补特征的充分挖掘。

部分方法则是通过改进现有的级联和相加策略，来更好地建立RGB信息和深度信息之间的交互[24,61-63]。如Liu等人[61]提出跨模态注意力融合网络（Attentive Cross-modal Fusion Network，ACMF），通过利用跨模态注意力机制建立了RGB特征和深度特征之间的交互，进而实现了RGB-D图像互补信息更加充分的挖掘。Huang[24]提出了一种MFI（Multi-modal Feature Interaction module，MFI）模

块，其在现有线性融合策略的基础上，引入一种非线性融合策略，并通过联合线性和非线性融合策略，捕捉RGB-D图像中的线性和非线性关系，更好地挖掘多模态互补信息实现跨模态信息融合。Liang等人[62]考虑了CNN无法有效地提取输入图像的远距离上下文信息的缺点，提出了一种基于上下文信息感知的多模态图像特征融合网络（Context-aware Network，CAN），该网络利用了LSTM[63]能够有效地捕捉远距离上下文信息的特性，建立RGB图像和深度图像之间的联系，实现了更好的RGB-D图像特征融合。

（2）分层级特征融合：现有RGB-D图像显著性目标检测模型通常采用统一的融合策略实现不同层级特征融合，然而低层级特征主要提供细节信息（如空间、纹理和边缘信息）而高层级特征主要提供具有高鉴别力的语义信息，使用统一的融合策略可能难以充分挖掘不同层级特征间的互补信息。基于此，部分方法提出针对不同层级的特征设计不同的多模态图像特征融合策略，旨在更好地挖掘RGB-D图像中包含的互补信息[68-72]。

部分方法针对低层级特征（第1、2和3层特征）的空间细节信息和高层级特征（第4、5层特征）的语义信息，设计不同的融合策略[68-69,71]。例如，Zhou等人[68]首先针对低层级特征设计了一种基于空间信息的特征融合模块（Spatial Fusion Module，SFM），以充分挖掘低层级RGB特征和深度特征的互补空间细节信息，然后针对高层级特征，设计了一种基于通道信息的特征融合模块（Channel Fusion Module，CFM），以充分地挖掘高层级RGB特征和深度特征之间的互补语义信息，进而取得了较好的融合效果。Yao等人[69]同样针对低层级特征设计了一种多模态图像特征增强融合模块（Cross-Modality Enhance Module，CMEM），对单模态图像特征通道和空间细节信息进行选择并融合；对于高层级特征，则是设计了一种双扩张合并模块（Double Dilated Merge Module，DDMM）以充分地挖掘和利用高层级特征中包含的互补多尺度上下文信息。Zhang等人[70]提出的特征差异化整合策略（Feature Reintegration over Differential Treatment，FRDT），同样首先设计一种多模态图像特征选择融合策略，通过门控机制，选择并融合低层级RGB特征和深度特征间的辨别细节信息，然后设计了一种交互融合策略，建立高层级特征间的交互，以充分挖掘它们的互补语义信息。

在上述方法的基础上，一些方法[71-72]进一步引入了中间层级特征，即低层级特征、中间层级特征和高层级特征，其中低层级特征主要包含模态细节信息，高层级特征主要包含模态语义信息，中间层级特征同时包含部分细节信息和部分语义信息。如图7所示，Li等人[71]提出的跨模态加权网络（Cross-modal Weighting Network，CMWNet），针对低层级特征、中间层级特征以及高层级特征特点，分别设计了浅层特征融合（CMW-L）、中间层特征融合（CMW-M）和深层特征融合（CMW-H）3种融合策略，以充分挖掘RGB特征和深度特征中所包含的互补细节信息和互补语义信息。Zhu等人[72]提出的自适应协同融合网络（Adaptively-Cooperative Fusion

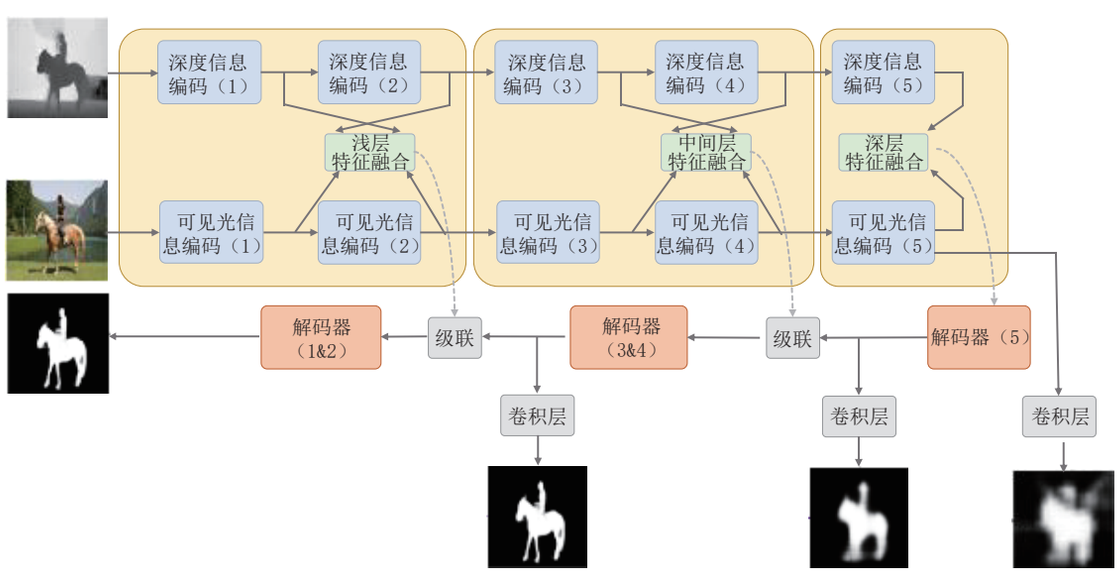


图7　CMWNet网络结构示意图[71]

Network，ACFNet)则同样是将特征分为低层级特征、中间层级特征和高层级特征，分别对应早期融合策略、中间融合策略和后期融合策略，进一步其结合了早期融合和后期融合的优点，设计了一种两阶段融合策略，在第一阶段使用两个独立的网络分别对RGB特征和深度特征进行预测，然后在第二阶段采用另一网络融合第一阶段的输出和高层级特征、低层级特征以更好地挖掘RGB-D图像的互补信息，进行显著性目标检测。

(3)特征分解：基于特征分解的方法则是希望将RGB-D图像中所包含的信息进行分解(如分解为模态特有信息和模态共有信息等)，根据分解后特征的特点，按照需求进行融合，以更加充分地挖掘RGB图像和深度图像中包含的场景信息进而实现更好的显著性目标检测[73-75]。Chen等人[73]设计了跨模态信息分解再融合(Disentangled Cross-Modal Fusion，DCMF)策略，其通过结合网络结构设计和对抗学习过程，将RGB-D图像信息分解为模态不变结构信息和模态特有信息，然后通过设计融合网络将分解后的特征进行融合，以挖掘RGB-D图像中包含的场景信息。Zhang等人[74]认为现有的融合策略间接地挖掘了RGB图像和深度图像包含的互补信息，不能明确地确定RGB图像和深度图像的贡献。为此，他们提出了一种交互信息最小化方法(Cascaded Mutual Information Minimization，CMIM)，该方法明确地挖掘RGB图像中的外观信息和深度图像中的几何信息，并通过交互信息最小化实现RGB特征和深度特征的融合，进而充分地挖掘RGB-D图像信息。Zhou等人[75]则认为RGB图像和深度图像中包含的模态特有信息和模态共有信息具有不同的特点，应该设计不同的处理方式。因此，一种新型的模态特有信息保留网络(Specificity-Preserving Network，SP-Net)被提出，该网络包含3个不同的分支，分别用于处理RGB图像特有信息，RGB图像和深度图像共有信息和深度图像特有信息。

2.3.3 决策级融合

与图像级融合和特征级融合不同，决策级融合通常分别获得RGB图像显著性结果和深度图像显著性结果，然后通过一定的融合策略实现结果融合。如图8所示，Wang等人[20]提出了一种动态融合框架，该框架首先利用两个独立的显著性目标检测网络分别从RGB图像和深度图像中预测2个初级显著图，然后通过一个子网络动态地生成权重图融合生成的2个初级显著图。类似地，Wang等人[21]提出一种新的多级深度质量评估方法，自适应地引导RGB图像显著性结果和深度显著性结果融合。与上述方法不同，Wang等人[76]则是通过强化学习的方法生成2个初级显著图的融合权重。

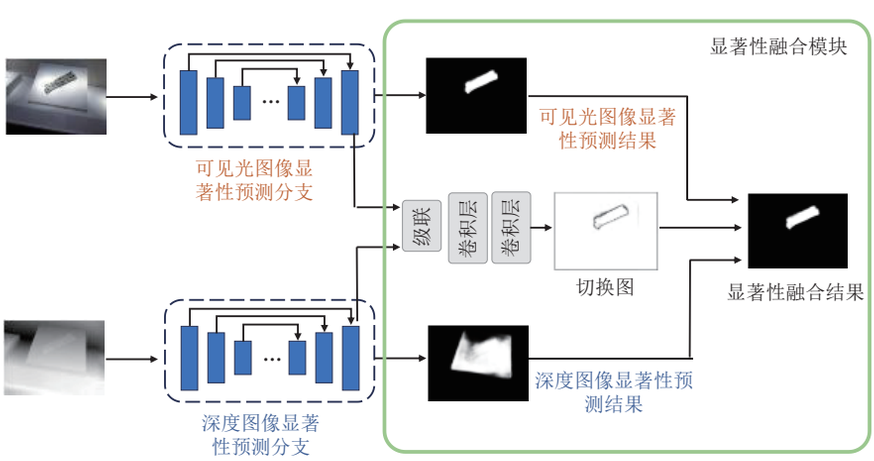


图8 AFNet网络结构示意图[20]

2.3.4 其 他

部分方法在图像级融合和特征级融合的基础上，提出图像级融合和特征级融合的联合策略，用于进一步挖掘RGB-D图像中的互补信息。如Zhou等人[77]提出的三输入多级融合网络(Three-input Multilevel Fusion Network，TMFNet)采用了一种三分支网络结构，其对RGB图像、深度图像和级联为4通道的RGB-D图像分别采用一个单独的分支提取特征，然后通过设计融合模块融合提取到的特征，以充分挖掘RGB-D图像中的互补信息。Huang等人[78]同样提出了一种三分支网络，即三重互补网络(Triple-complementary Network，Triple-Net)，该网络采用了两阶段策略，在第一阶段，通过双分支网络结构预测粗糙的显著目标，在第二阶段则是将粗

糙显著图、RGB图像、深度图像进行级联后，输入到单分支网络中，进行显著性目标检测。

2.4　上下文信息挖掘问题

上下文信息挖掘问题指的是：在显著目标检测过程中，如何有效地挖掘和利用RGB-D图像中的上下文信息，以准确地定位显著目标，并恢复显著目标的边界。现有方法[36,48,84-106]解决上下文信息挖掘问题主要思路有4种，即多尺度特征挖掘、多层级融合特征挖掘、全局信息/长距离依赖关系挖掘和边界信息挖掘。

2.4.1　多尺度特征挖掘

在不同的场景中，显著目标出现的位置、尺度和大小不一，是显著性目标检测的难点之一。通过挖掘场景中的多尺度特征，能够有效地缓解这一难点对显著性目标检测结果的影响，进而更准确地检测显著目标。因此一些模型关注于场景多尺度信息的挖掘[36,84-90]。

部分方法通过池化、洞卷积等操作获取固定尺度的多尺度特征[36,85-86]。如图9所示，Ren等人[85]提出的渐进多尺度融合网络（Progressive Multi-scale Fusion Network，PMFNet）则首先通过空间空洞金字塔池化（Atrous Spatial Pyramid Pooling，ASPP）提取多尺度特征，然后通过设计基于掩码引导的特征优化模块对多尺度特征中的干扰信息进行抑制，同时，对多尺度特征中的语义信息进行保留，进而得到更加具有辨别性的多尺度特征。与Ren等人[85]不同，Liao等人[36]提出的多阶段多尺度融合网络（Multi-stage and Multi-scale Fusion Network，MMNet）专门设计了一种新型的双向多尺度特征解码器，其采用的多模态多尺度特征融合模块通过使用注意力机制、平均池化以及最大池化操作，对多尺度特征进行了充分的提取和挖掘，进而得到了较好的检测效果。Wu等人[86]提出的渐进引导融合网络（Progressive Guided Fusion Network，PGFNet）首先设计多模态和多尺度特征注意力融合模块，通过该模块实现对单模态图像多尺度特征的提取和挖掘，并通过注意力机制建立不同模态特征间的交互，进一步通过设计多模态特征调整机制，使得模型能够利用高层级特征增强低层级特征，进而更好地进行显著性目标检测。

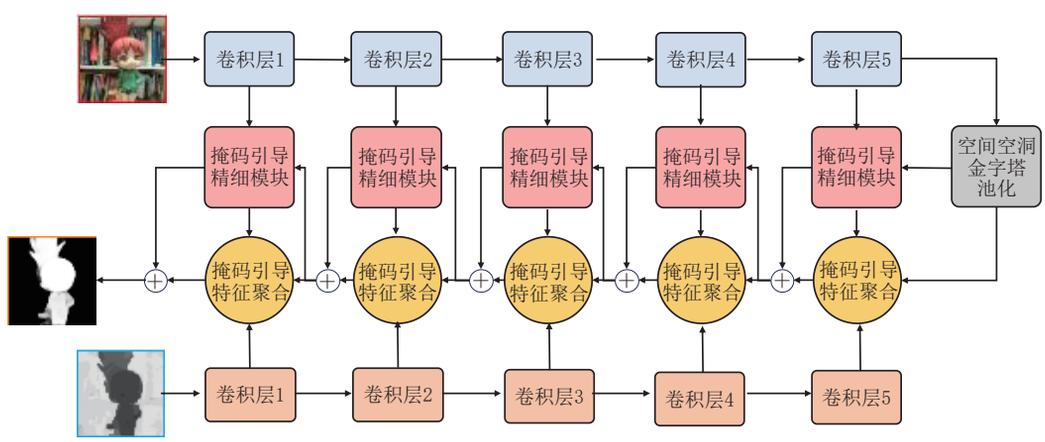


图9　PMFNet网络结构示意图[85]

一些方法则是通过动态卷积等操作获取可变尺度的多尺度特征[87-89]。Wu等人[87]在其提出的聚合交互融合网络（Aggregate Interactive Learning Network，AILNet）中，首先提出一种互动学习策略来充分地挖掘边缘特征、深度特征和显著特征，然后进一步通过可变形卷积的提取多尺度特征，以应对显著目标尺度和位置的多样性。Pang等人[88]在提出的分层动态滤波网络（Hierarchical Dynamic Filtering Network，HDFNet）中设计了一种新型的动态扩展多尺度特征提取模块，其利用输入RGBD图像融合特征的互补信息，动态地提取不同尺度的多尺度特征，进而实现了更好地挖掘了多尺度特征。Piao等人[89]提出了一种新型的基于深度信息诱导的多尺度循环注意力网络（Depth-induced Multi-scale Recurrent Attention Network，DMRA）用于RGB-D图像显著性目标检测。他们首先设计基于深度信息调整的多模态图像特征融合模块以充分地挖掘RGB-D图像的互补信息，然后通过设计基于深度信息诱导的多尺度特征加权模块以充分地挖掘融合特征中包含的多尺度信息，进而更好地应对显著目标的尺度多样性。

2.4.2 多层级融合特征挖掘

在显著图预测过程中，本类方法主要通过挖掘不同层级间融合特征的互补信息，即将低层级融合特征中包含的场景细节信息和高层级融合特征中包含的场景语义信息进行融合，来进一步挖掘场景的上下文信息，以实现更好的显著性目标检测[91-98]。

部分方法通过改进现有逐级或逐阶段的融合策略，挖掘不同层级特征间的互补信息[91-92]。例如，Zhai等人[91]在其提出的双骨干网络(Bifurcated Backbone Strategy Network，BBSNet)中采用了一种新型的两阶段逐级优化策略以充分地挖掘多层级特征之间的互补信息。该策略首先采用教学网络融合多层级特征并初步预测显著目标，然后以此为指导，利用子网络挖掘多层级特征之间的互补信息，达到最终显著目标。Chen等人[92]提出的堆叠特征交互网络(Cascaded Feature Interaction Decoder Network，CFIDNet)设计了一种级联特征交互解码器，其使用了多个级联的解码器逐级预测显著目标，同时在每个解码器中，提出一种多层级特征优化模块，通过自注意力机制融合相邻3个层级的融合特征，以充分地挖掘RGB-D图像中包含的上下文信息。

部分方法则是通过多层级特征密集聚合的方式，融合不同层级特征[93-95]。如图10所示，Zhao等人[93]在基于对比先验和流体金字塔融合网络(Contrast Prior and Fluid Pyramid，CPFP)中，采用了一种金字塔结构，以低层级特征为塔底，高层级特征为塔顶，在不同层级特征之间通过密集连接的方法实现多层级信息之间的聚合，进而能够充分地挖掘和利用不同层级特征之间的冗余和互补信息，实现更好的显著性目标检测。与CPFP[93]类似，Huang等人[94]提出的多层次跨模态交互网络(Multi-level Cross-modal Interaction Network，MCINet)同样是通过设计金字塔网络结构和密集跳连接来实现不同层级特征的聚集。与此同时，在每个聚集节点，他们还考虑了不同层级特征之间的交互，以及利用上一阶段信息预测的粗糙显著图指导下阶段多层级特征的融合，进而充分地挖掘了不同层级特征之间的互补信息。Chen等人[95]在其提出的模态分类网络(Modality Classification Network，MCNet)中设计了一种跨层级特征密集反馈拓扑策略，该策略首先对每个层级的多模态图像特征进行融合，然后通过密集连接将多层级信息反馈到底层级特征中，实现多层级特征交互，进而实现更好的显著性目标检测。

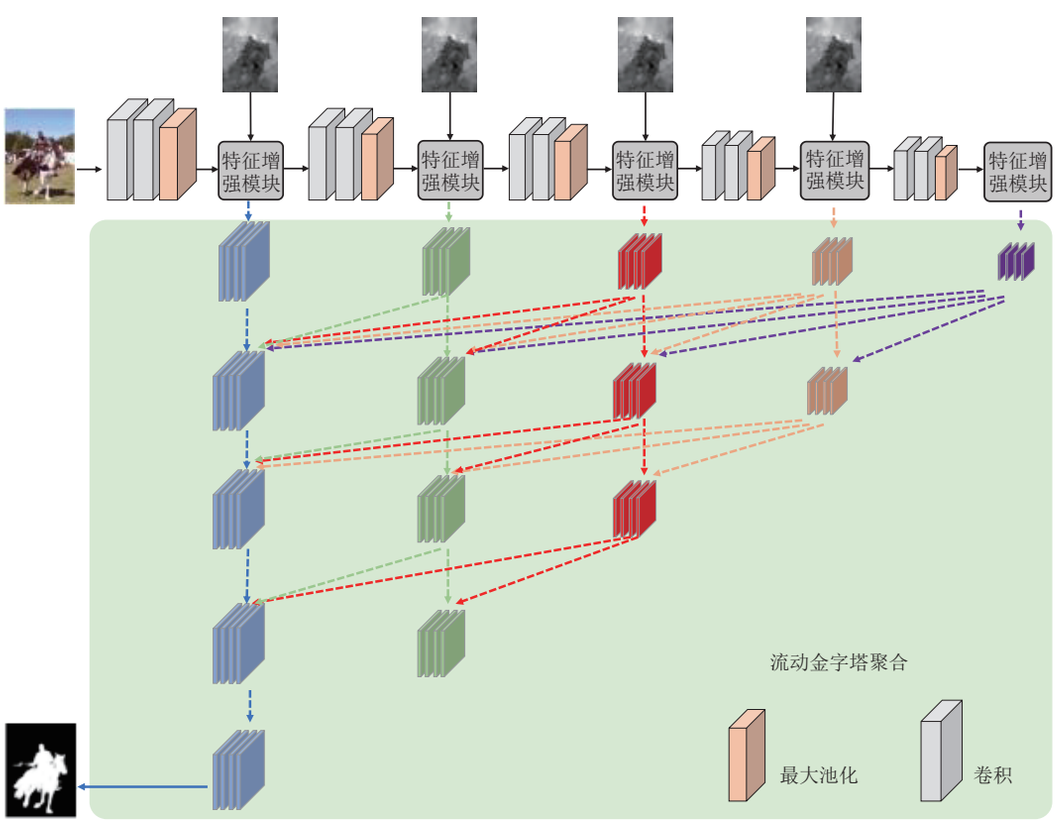


图10 CPFP网络结构示意图[93]

部分方法根据不同层级融合特征的性质，设计多层级特征融合策略[96-98]。例如，Zhou等人[96]提出的多级逆上下文交互融合网络（Multilevel Reverse-context Interactive-fusion Network，MRINet）通过考虑高层级特征和低层级特征所具有的特点（即语义性和细节性），进而设计了一种多层级特征上下文信息反向交互模块。该模块能够以较高的分辨率重构高层级特征，进而获得较好的语义信息，同时能够利用低层级特征增强显著目标的边缘信息，实现更好的多层级互补信息挖掘。Zhou等人[97]提出的特征聚合网络（Feature Aggregation Network，FANet）首先通过一个区域增强模块对融合特征的局部区域进行增强，然后通过设计一种分层融合模块整合不同层级特征间的互补信息（高层级特征的语义信息和低层级特征细节信息之间的互补关系），进而实现更好的显著性目标检测。Xiao等人[98]提出的先验引导多层级融合网络（Pre-trained Guide Hierarchical Fusion Network，PGHFNet）则是设计了一种新型的多层级特征融合结构，该结构能够充分利用不同层级特征之间的语义关系和多尺度信息，实现对多层级特征之间互补信息的挖掘，达到边界良好且空间连续性好的显著目标。

2.4.3　全局信息/长距离依赖关系挖掘

目前主流的RGB-D图像显著性目标检测方法以CNN为基础网络进行设计，然而CNN本身具有一定的局限性，难以挖掘输入特征的全局关系，进而导致无法充分地挖掘融合特征中包含的上下文信息，一些工作对这一问题进行了改进[101-104]。

早期的方法主要通过全局池化、多尺度特征以及多分支网络等方式提取全局信息。例如，Chen等人[101]设计的基于多尺度、多路径和跨模态交互的多模态融合网络（Multi-modal Fusion Network with Multi-scale Multi-path and Cross-modal Interactions，MMCI）在单模态特征提取网络中采用了全局特征提取分支和局部特征提取分支。其首先利用全局特征提取分支和局部特征提取分支分别提取RGB图像和深度图像的全局信息和局部信息。随后，将RGB图像和深度图像的全局信息和局部信息分别进行融合，并进一步分别利用融合后的全局信息和融合后的局部信息进行单模态显著性目标检测。最后，将得到的显著图进行进一步融合，得到最终的显著目标预测结果。

近年来的一些方法则是利用Transformer、图网络等工具挖掘全局上下文信息[102-104]。例如，Liu等人[102]通过使用3个权重共享的Transformer（Triplet Transformer Embedding Network，TriTransNet）提取第3到第5个层级的融合特征中包含的长距离依赖关系，并通过融合多层级特征，以更好地捕捉RGB-D图像中场景的上下文信息，实现更好的显著性目标检测。Fang等人[103]受到自注意力机制的启发，提出了一个新型的多模态多尺度特征精细网络（Multi-modal and Multi-scale Refined Network，M2RNet），其设计了一种双注意力模块（Dual-Attention Module，NDAM），分别建立融合特征在通道方向和空间方向的远距离依赖关系，进而更好地挖掘了RGB-D图像的互补信息。Wang等人[104]提出了一种多模态图像长距离上下文信息挖掘网络（Cross-modality Long-range Context Information Gathering，CM-LCG），该网络通过建立RGB特征和深度特征在任意位置之间成对的相关关系，以捕捉长距离上下文信息，并基于此设计了一种相关融合策略，以充分地挖掘多模态图像的互补信息。

2.4.4　边界信息挖掘

场景中包含的显著性目标具有多样性，即其位置、大小和尺度是不断变化的，这就要求RGB-D图像显著性目标检测方法应对显著目标内部的变化同时，能够辨别显著目标与其他目标的边界差异，从而能够准确地对显著目标的边界进行分割。然而，这两者在一定程度上是冲突的，这就导致了RGB-D图像显著性目标检测算法难以准确地检测显著目标的边界。近年来，一些方法通过引入显著目标的边界信息解决这一问题[105-106]。如图11所示，Liu等人[105]在所提出的跨模态边缘引导网络（Cross-modal Edge-guided Network，CENet）中首先通过预测显著目标边界图的方式得到边界信息，然后将这些边界信息与多层级特征相融合，以更好地预测显著图。

2.5　模型复杂问题

模型复杂问题指的是：相比于单模态图像显著性目标检测模型，多模态图像显著性目标检测模型包含更多的输入数据，相应地，其网络结构也更为复杂，并且随着实际场景需求日益复杂。同时，现有网络结构也变得越来越庞大，进而难以满足实际应用的低参数量、低复杂度和高性能的要求。解决模型复杂问题的关键就是通过模型轻量化设计，优化模型网络框架，在一定程度上保持甚至提高模型检测性能的同时降低模型参数量和计算量。因此，RGB-D显著目标检测模型轻量化设计是一个综合

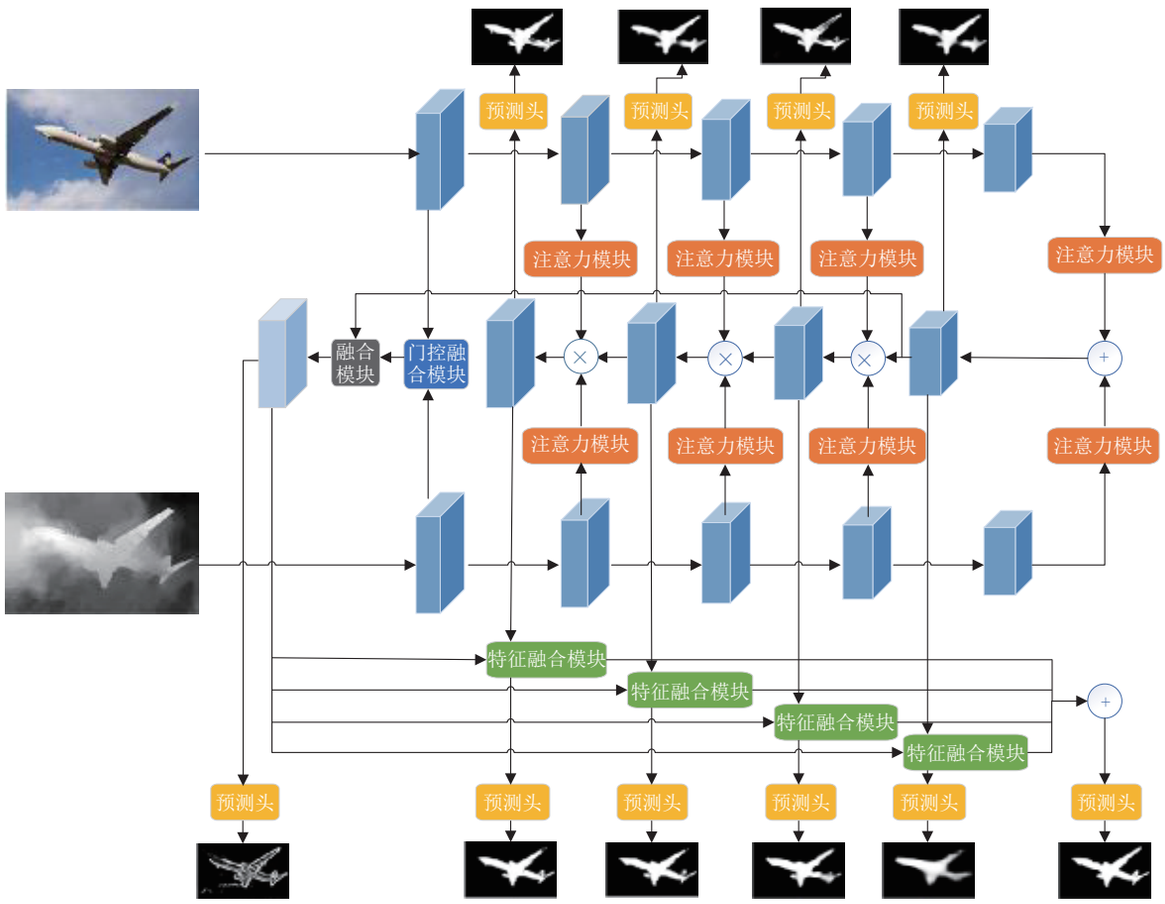
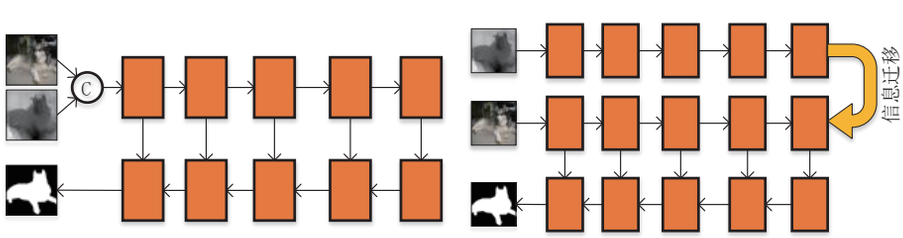


图11 CENet网络结构示意图[105]

的过程，需要兼顾RGB-D显著目标检测的其他关键问题。

目前RGB-D图像显著性目标检测方法主要以双分支网络结构为主，这通常需要两个独立的子网络实现不同模态输入图像的特征提取和多个特征融合模块实现多模态图像特征之间的融合，这大大增加了模型的参数量和计算复杂度。相比于双分支网络结构，如图12(a)所示，采用单分支网络结构的RGB-D图像显著性目标检测模型可以大幅度降低模型参数量和计算复杂度。然而采用单分支网络结构难以有效地挖掘多模态输入图像所包含的丰富场景信息，势必降低模型性能。如何通过有效的轻量化设计平衡模型复杂度和模型参数量是缓解模型复杂问题的核心。根据不同网络结构的特点，现有的RGB-D图像显著性目标检测方法的轻量化模型主要分为：针对单分支网络框架模型的轻量化方法(单分支轻量化模型)和针对双分支网络框架模型的轻量化方法(双分支轻量化模型)。



(a) 网络结构轻量化设计 (b) 跨模态信息转移

图12 单分支模型轻量化思路

2.5.1 单分支轻量化模型

相比于双分支网络模型，单分支网络模型仅用一个特征提取网络进行特征提取，并且不需要多个特征融合模块进行特征融合，因此单分支网络框架的轻量化设计有着天然的优势。如图12所示，针对单分支网络的轻量化方法主要包括网络结构轻量化

和跨模态信息转移。

（1）网络结构轻量化：此类方法的主要思路是利用轻量化的特征提取网络（MobileNet[107-109]、ShuffleNet[110-111]等）、设计轻量化的网络模块、使用轻量化的卷积核（如组卷积、深度可分离卷积[112]等）在一定范围内保证模型性能的同时，实现网络轻量化设计。例如，Zhao等人[113]在提出的实时的深度信息增强注意力网络（Depth-enhanced Attention Network，DANet）中，采用单分支网络框架、将RGB图像和深度图像级联作为网络的输入，实现对RGB-D图像中信息的初步挖掘，并通过特征降维的方式减少网络模型的参数量。同时，为了弥补单分支网络模型无法充分挖掘RGB-D图像互补信息的缺点，提出了一种基于深度信息增强的双注意力模块（Depth-Enhanced Dual Attention module，DEDA），通过建立深度图像信息和RGB图像外观信息之间的交互，实现RGB-D图像信息的挖掘，进而保证模型的性能。

（2）跨模态信息迁移：如图12（b）所示，跨模态信息转移的思路是采用迁移学习策略，使得网络在训练过程中能够将深度图像中包含的信息迁移到RGB特征提取过程中，间接地实现RGB特征和深度特征的融合，从而有效地结合RGB-D图像中包含的互补信息[114-116]。例如，Ji等人[114]在提出的协同学习网络（Collaborative Learning Network，CoNet）中，采用一种新型的协同学习框架。此框架将RGB图像作为网络的输入，在训练过程中实现RGB图像的边缘检测、显著目标初步预测和深度图像估计3个任务的协同学习。通过在学习过程中实现跨模态信息迁移，该框架能够更好地挖掘多模态互补信息，并实现了网络轻量化的目标。Piao等人[115]提出的自适应深度蒸馏模型（Adaptive and Attentive Depth Distiller，A2dele）采用了类似的思路，在训练过程中，A2dele通过两个独立的网络来分别从RGB图像和深度图像中预测显著目标，并通过蒸馏学习的方法，将基于深度图像的显著目标预测分支中的深度信息转移到RGB图像分支中。在测试过程中，仅利用RGB图像分支进行预测，达到了网络轻量化的目标。

2.5.2　双分支轻量化模型

与单分支模型轻量化不同，针对双分支网络结构的RGB-D图像显著性目标检测模型的轻量化设计则相对较难。现有方法主要通过模块轻量化计[117-118]或者框架轻量化的方法[119-123]，实现网络模型的轻量化。

（1）模块轻量化：此类方法的主要思路是通过优化网络的特征提取模块、融合模块和推理模块等模块实现网络的轻量化设计[117-118]。例如，Fu等人[117]在提出的联合学习和密集型合作融合结构（Joint Learning and Densely Cooperative Fusion，JL-DCF）中，通过孪生网络的形式，同时提取RGB图像特征和深度图像特征，实现了减少网络模型参数的目标。同时，他们进一步提出了一种多层级特征密集协作融合策略，有效地挖掘并利用了多层级特征之间的互补信息，保证了模型的性能。

（2）框架轻量化：此类方法的主要思路是通过设计异构的双分支网络结构实现网络轻量化设计[119-122]。例如，Zhang等人[119]提出了一种异构的双分支网络框架（Asymmetric Two-Stream Architecture，ATS）。对于RGB图像，该网络使用VGG网络实现特征提取，并利用一个多尺度模块从RGB图像提取辨别性的局部和全局特征。对于深度图像，其设计了一个轻量级的深度网络实现特征提取，同时结合特征降维和减少卷积层等方式，实现网络轻量化。Wang等人[120]采用了一种新型的残差引导网络（Guided Residual Network，GRNet）实现了RGB-D图像显著性目标检测算法的轻量化设计，其采用标准的卷积网络提取RGB特征，然后采用小型的卷积网络提取深度特征，同时设计了深度特征选择模块和多尺度特征提取模块以保证模型的性能。

Chen等人[121]同样通过精简深度特征提取网络实现模型的轻量化设计，并通过增强特征融合模块实现对多模态互补信息更好地挖掘，以保证模型性能。与之类似，Wu等人[118]提出的MobileSal首先将轻量化特征提取网络MobileNet[104]作为骨干网络，然后设计了一种新型的隐式深度信息重构技术（Implicit Depth Restoration Technique，IDR）以增强骨干网络的特征提取能力。同时，其仅对最高层级RGB特征和深度特征进行融合，通过设计一种新型的紧凑金字塔特征细化模块（Compact Pyramid Refinement，CPR）挖掘不同层级特征之间的互补信息，在实现模型轻量化的同时，保证了模型的性能。

此外，与上述方法不同，如图13所示，Huang等人[123]在双分支网络框架的基础上，提出了一种新的、适应于RGB-D图像显著目标检测轻量化设计的网络框架，即中间层级特征融合框架。该框架仅采用一个融合模块对某一特定中间层级单模态图像

表2 10个RGB-D基准数据集(按年份、发布会议/期刊、数据集大小、场景类型和图像分辨率分类)

数据集	年份	发布会议/期刊	数据集大小	场景类型	分辨率
STERE[124]	2012	CVPR	1000	互联网	[251-1200] × [222-900]
NLPR[10]	2014	ECCV	1000	室内/室外	640 × 480, 480 × 640
NJU2K[125]	2014	ICIP	1985	影视/互联网/摄像	[231-1213] × [274-828]
DUT-RGBD[89]	2019	ICCV	1200	室内/室外	400 × 600
SIP[26]	2020	TNNLS	929	室外行人	992 × 744
ReDWeb-S[126]	2021	TPAMI	3179	多元场景	[133-937] × [132-996]
GIT[127]	2013	BMVC	80	居家环境	640 × 480
LFSD[128]	2014	CVPR	100	室内/室外	360 × 360
DES[129]	2014	ICIMS	135	室内	640 × 480
SSD[130]	2017	ICCVW	80	影视	960 × 1080

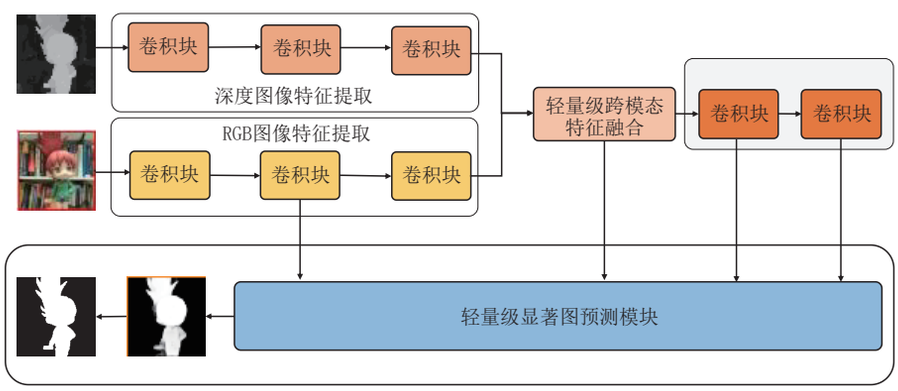


图13 中间级特征融合网络结构示意图[123]

特征进行融合,在大大减少模型参数量的同时,通过设计更加高效的多模态信息融合模块和多层级特征融合模块,挖掘多模态图像中所包含的更加丰富的场景信息,弥补因模型轻量化带来的性能损失。

3 数据集和评价指标

3.1 数据集介绍

表2总结了目前主流的RGB-D图像显著性目标检测数据集,其中包括6个数据规模相对较大的数据集:STERE[124]、NLPR[10]、NJU2K[125]、DUT-RGBD[89]、SIP[26]、ReDWeb-S[126]和4个数据规模相对较小的数据集:GIT[127]、LFSD[128]、DES[129]、SSD[130]。接下来我们将提供每个数据集的详细信息。

(1) STERE[124]:数据集从三个公开网站(Filckr、NVIDIA3DVisionLive 和 Stereoscopic Image Gallery)收集1250幅样本图像,通过人工标注的方式标注显著目标,并从中选择了1000幅样本构成数据集,最后通过光流估计算法,获取了配对的RGB-D图像。

(2) NLPR[10]:本数据集首先通过 Microsoft Kinect深度相机从各种室内和室外场景,如办公室、校园、街道等,采集了5000对RGB-D图像,然后通过人工标注的方式进行显著性目标标注,最终选择了1000对RGB-D图像构成最终的数据集。

(3) NJU2K[125]:本数据集包含了1985对样本,其中部分RGB-D图像是从网络中的3D电影中进行采集,其他RGB-D图像则是通过FujiW3深度相机进行成像。本数据集包含很多具有挑战性的复杂场景,如存在多个显著目标,复杂背景等。

(4) DUT-RGBD[89]:本数据集同样为近几年新产生的数据集,其从室内场景中采集了800对RGBD图像,从室外场景中采集了400对RGB-D图像,共1200张RGB-D图像,其中包含很多具有挑战性的场景,如多显著目标、前景背景目标相似等。

(5) SIP[26]:本数据集是近年来新产生的数据集,其通过华为魅族10智能手机自带深度相机进行数据采集。其以图像中的人物为显著目标,并通过人工标注了925张高清RGB-D图像。

(6) ReDWeb-S[126]:本数据集包含从许多网络立体图像中选择的3179张图像,具有各种现实生活中的场景。它的官方数据集划分包括一个包含

2179个RGB-D图像对的训练集和一个包含剩余1000个图像对的测试集。

相比于上述数据集，小数据集只包括少量RGB-D图像，往往用于测试模型性能。

GIT[127]：本数据集包含80张RGB-D图像。

LFSD[128]：本数据集通过Lytro Lightfield相机采集了100张RGB-D相机。其通过多个标注者进行显著目标标注，并将多个标注者的标注结果取交集，得到最后的显著目标。

DES[129]：本数据集通过Kinect深度相机从室内场景中采集135张RGB-D图像。同样通过多个标注者进行标注，选取标注结果的交集作为最终的真值图。

SSD[130]：本数据集的图像从3个不同3D电影的室内室外场景采集。该数据集共包括80张RGB-D图像。

总体而言，尽管RGB-D显著目标检测数据集数量较多，但RGB-D图像总量较少，大多数数据集为小规模数据集。在上述数据集中，GIT、LFSD、DES和SSD数据量过少，仅用于RGB-D显著目标检测模型的测试。数据集STERE包含大量的含噪声的合成数据。这些数据有助于模拟真实世界中的复杂场景，使得STERE成为测试模型鲁棒性和泛化能力的理想选择。数据集SIP中的显著目标仅限于人类，其目标类型相对单一。考虑到STERE和SIP数据集各自的特点，当前这两个数据集也主要用于模型性能测试。由于可供使用训练数据相对较少，目前研究人员通常将NJU2K、NLPR、DUT-RGBD和ReDWeb-S等数据库的训练数据混合，以增加训练数据多样性，支撑RGB-D显著目标检测模型的训练需求。相比于NJU2K和NLPR数据集以简单场景为主，DUT-RGBD和ReDWeb-S数据集包含更多的复杂场景，更加接近真实场景，因此也更加具有挑战性。

3.2　数据标注

不同于通用目标检测或语义分割任务的研究对象更为客观，显著性目标依赖于人类对场景中目标的认知。因此，显著性目标检测任务具有一定的主观性，能够在一定程度上反映/模拟人类视觉注意力机制，有助于人们进一步理解人类视觉注意力机制。显著性目标检测任务在虚拟现实、人机交互、图像压缩等一些需要参考人类注意力机制或者需要对人类注意力机制建模的应用中有着重要作用。相对应的，相比于通用目标检测或语义分割任务，RGB-D图像显著目标检测数据的标定更为困难，其标注过程应满足最长视觉停留准则、重复性准则、从众性准则和RGB图像优先原则。

最长视觉停留准则指的是在一定时间内，标注者在给定图像中视觉停留时间最长的区域，为该标注者感兴趣区域。重复性准则指的是：对于同一标注者而言，在一定时间间隔后，其对同一图像的感兴趣区域应该是相同的。从众性准则指的是：显著目标检测数据的标注者应该有多人，并且对于给定图像，其最终显著目标应为标注者们感兴趣次数最多的目标。RGB图像优先原则指的是：标注者们应分别标注RGB图像和深度图像，若RGB图像和深度图像标注一致，则为最终标注，若RGB图像和深度图像标注不一致，且RGB图像成像质量较好，考虑到人类视觉系统对颜色信息更为敏感，则优先采用RGB图像标注，若RGB图像质量较差，则采用深度图像标注。

3.3　评价指标

本节主要介绍RGB-D图像显著性目标检测领域最常用的模型性能评价指标，包括：平均绝对误差(Mean Absolute Error，MAE)[131]、综合评价指标(F-Measure，F_β)[132]、结构评价指标(S-Measure，$S\alpha$)[133]和增强配准指标(E-Measure，E_ξ)[134]。

平均绝对误差：其用于评价基准显著图S与预测图Y之间的误差，其计算公式为

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

其中，W和H表示基准显著图和预测显著图的宽度和高度。

综合评价指标(F-Measure，F_β)[132]：其主要用于评价模型的综合预测性能，计算公式为

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

其中β为平衡参数，一般设为0.3。

结构评价指标(S-Measure，S_α)[133]：主要用于评价基准显著图与预测显著图之间的结构相似性，其计算公式为

$$S_\alpha=\alpha\times S_o+(1-\alpha)\times S_r \tag{3}$$

其中，$\alpha\in[0,1]$表示平衡参数，其值设为0.5。S_r表示面向区域(region-aware)的结构相似性度量，S_o表示面向目标(object-aware)的结构相似性度量。更详细的度量计算细节请查阅文献[133]。

增强配准指标(E-Measure，E_ξ)[134]：该指标将像素级评估和图像级评估进行统合整一，主要计算基准显著图和预测显著图在图像级别的统计特性和局

表3 不同RGB-D图像显著性目标检测模型算法结果比较

模型细节		DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
方法名称	年份	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓
ACCF[32]	2018	0.767	0.859	0.791	0.113	0.852	0.915	0.858	0.079	0.815	0.913	0.856	0.059	0.822	0.928	0.848	0.065	0.771	0.839	0.787	0.132	0.863	0.927	0.873	0.068
PCF[22]	2018	0.771	0.858	0.801	0.1	0.871	0.896	0.877	0.059	0.841	0.916	0.874	0.044	0.804	0.912	0.842	0.049	0.779	0.842	0.794	0.112	0.859	0.897	0.875	0.064
TANet[33]	2019	0.79	0.861	0.808	0.093	0.874	0.925	0.878	0.06	0.863	0.941	0.886	0.041	0.827	0.91	0.858	0.046	0.796	0.847	0.801	0.111	0.861	0.923	0.871	0.06
MMCI[101]	2019	0.767	0.859	0.791	0.113	0.852	0.915	0.858	0.079	0.815	0.913	0.856	0.059	0.822	0.928	0.848	0.065	0.771	0.839	0.787	0.132	0.863	0.927	0.873	0.068
DMRA[89]	2019	0.898	0.933	0.889	0.048	0.872	0.908	0.886	0.051	0.855	0.942	0.899	0.031	0.857	0.945	0.901	0.029	0.849	0.899	0.847	0.075	0.868	0.92	0.886	0.048
SSF[40]	2020	0.924	0.951	0.915	0.033	0.896	0.935	0.899	0.043	0.896	0.953	0.914	0.026	0.883	0.941	0.905	0.025	0.867	0.9	0.859	0.044	0.88	0.936	0.893	0.044
DQSD[31]	2020	0.827	0.878	0.845	0.072	0.9	0.936	0.899	0.05	0.898	0.952	0.916	0.029	0.927	0.973	0.935	0.021	0.847	0.878	0.851	0.085	0.886	0.935	0.892	0.051
ASIFNet[58]	2020	-	-	-	-	0.877	0.907	0.891	0.047	0.869	0.944	0.909	0.029	0.915	0.974	0.934	0.019	0.85	-	0.81	0.069	0.852	0.908	0.874	0.051
ACMF[61]	2020	0.874	-	0.881	0.059	0.089	-	0.904	0.044	0.876	-	0.914	0.028	-	-	-	-	0.865	-	0.864	0.084	-	-	-	-
FRDT[70]	2020	0.903	0.941	0.91	0.039	0.879	0.917	0.898	0.048	0.867	0.945	0.914	0.029	0.868	0.942	0.902	0.028	0.855	0.899	0.857	0.073	0.88	0.927	0.901	0.043
CMWNet[71]	2020	-	-	-	-	0.902	0.936	0.903	0.046	0.903	0.951	0.917	0.029	0.93	0.969	0.934	0.022	0.883	0.912	0.876	0.066	0.9011	0.944	0.905	0.043
DCMF[73]	2020	0.931	0.958	0.928	0.035	0.915	0.948	0.913	0.043	0.906	0.954	0.922	0.029	-	-	-	-	0.875	0.909	0.878	0.068	0.906	0.946	0.91	0.043
BBSNet[100]	2020	-	-	-	-	0.92	0.949	0.921	0.035	0.918	0.961	0.93	0.023	0.927	0.966	0.933	0.021	0.858	0.901	0.964	0.072	0.903	0.942	0.908	0.041
ATSA[119]	2020	0.92	0.948	0.918	0.032	0.893	0.921	0.901	0.04	0.876	0.945	0.907	0.028	0.885	0.952	0.907	0.024	0.862	0.905	0.865	0.064	0.884	0.921	0.897	0.039
Cas-GNN[47]	2020	0.912	0.932	0.891	0.042	0.903	0.933	0.911	0.035	0.904	0.952	0.919	0.025	0.906	0.947	0.905	0.028	0.864	0.877	0.849	0.073	0.901	0.930	0.899	0.039
CDNet[44]	2021	0.934	0.955	0.93	0.029	0.918	0.95	0.918	0.036	0.919	0.96	0.929	0.023	0.929	0.973	0.936	0.019	0.879	0.911	0.877	0.061	0.907	0.947	0.912	0.037
CCNet[60]	2021	-	-	-	-	0.929	0.948	0.917	0.037	0.922	0.966	0.926	0.023	0.936	0.966	0.922	0.022	0.894	0.907	0.875	0.061	0.913	0.944	0.908	0.037
BTSNet[23]	2021	-	-	-	-	0.924	0.954	0.921	0.036	0.923	0.965	0.934	0.023	0.94	0.979	0.943	0.018	0.874	0.906	0.867	0.07	0.911	0.949	0.915	0.038
CCAFNet[68]	2021	0.915	0.941	0.905	0.036	0.91	0.943	0.909	0.037	0.908	0.956	0.921	0.026	0.937	0.977	0.938	0.018	0.832	0.876	0.826	0.087	0.887	0.934	0.892	0.044
SP-Net[75]	2021	-	-	-	-	0.935	0.954	0.925	0.028	0.925	0.959	0.927	0.021	0.95	0.98	0.945	0.014	-	-	-	-	0.915	0.944	0.907	0.037
TMFNet[77]	2021	-	-	-	-	0.882	0.91	0.91	0.041	0.867	0.944	0.921	0.027	0.892	0.968	0.936	0.021	0.846	0.865	0.849	0.084	-	-	-	-
M2RNet[103]	2021	0.925	0.935	0.903	0.042	0.922	0.904	0.91	0.049	0.921	0.941	0.918	0.033	0.937	0.971	0.934	0.019	0.861	0.874	0.842	0.088	0.913	0.929	0.899	0.042
TriTansNet[102]	2021	0.938	0.957	0.933	0.025	0.919	0.925	0.92	0.03	0.909	0.96	0.928	0.02	0.936	0.981	0.943	0.014	-	-	-	-	0.893	0.927	0.908	0.033

续表

模型细节		DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
方法名称	年份	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$S_\alpha\uparrow$	$MAE\downarrow$
CENet[105]	2021	0.922	-	0.924	0.031	0.892	-	0.908	0.039	0.875	-	0.929	0.025	0.896	-	0.923	0.022	0.862	-	0.866	0.063	0.886	-	0.895	0.043
MobileSal[118]	2021	0.912	0.94	0.896	0.041	0.914	0.939	0.905	0.041	0.916	0.95	0.92	0.025	-	-	-	-	0.867	0.88	0.847	0.08	0.906	0.916	0.903	0.041
GRNet[120]	2022	0.932	0.956	0.929	0.029	0.905	0.926	0.914	0.036	0.906	0.961	0.932	0.02	-	-	-	-	0.878	0.912	0.879	0.064	0.897	0.928	0.911	0.034
BPGNet[38]	2022	0.938	0.958	0.93	0.031	0.926	0.953	0.923	0.034	0.914	0.959	0.927	0.024	0.932	0.973	0.937	0.02	0.875	0.908	0.874	0.066	0.904	0.944	0.907	0.04
FANet[97]	2022	-	-	-	-	0.892	0.914	0.899	0.044	0.885	0.951	0.913	0.023	0.874	0.925	0.894	0.026	0.855	0.882	0.85	0.076	0.863	0.908	0.881	0.047
CFIDNet[92]	2022	-	-	-	-	0.923	0.913	0.914	0.038	0.915	0.95	0.922	0.026	-	-	-	-	0.884	0.893	0.87	0.07	0.908	0.924	0.901	0.043
DCMF[104]	2022	0.932	0.958	0.928	0.035	0.915	0.948	0.913	0.043	0.906	0.954	0.922	0.029	-	-	-	-	0.875	0.909	0.878	0.068	0.906	0.945	0.91	0.043
MoADNet[122]	2022	0.92	0.945	0.907	0.033	0.907	0.929	0.901	0.042	0.908	0.95	0.918	0.024	-	-	-	-	0.873	0.902	0.859	0.064	0.901	0.931	0.896	0.043
GCENet[90]	2022	0.923	0.951	0.918	0.035	0.915	0.947	0.914	0.038	0.907	0.953	0.919	0.025	0.925	0.966	0.926	0.018	0.849	0.883	0.846	0.079	0.900	0.940	0.899	0.040
CIRNet[65]	2022	0.929	0.954	0.918	0.035	0.916	0.947	0.916	0.039	0.914	0.947	0.923	0.025	-	-	-	-	0.873	0.906	0.869	0.069	0.903	0.945	0.905	0.043
C2DFNet[66]	2022	0.944	0.964	0.933	0.025	0.916	0.947	0.916	0.039	0.916	0.961	0.927	0.021	0.915	0.961	0.924	0.018	0.867	0.903	0.863	0.065	0.904	0.947	0.905	0.038
RFNet[84]	2022	-	-	-	-	0.936	0.951	0.926	0.029	0.932	0.962	0.931	0.020	0.946	0.977	0.941	0.015	-	-	-	-	0.921	0.944	0.911	0.035
HINet[99]	2023	-	-	-	-	0.913	0.945	0.915	0.039	0.906	0.957	0.922	0.026	-	-	-	-	0.847	0.888	0.852	0.076	0.883	0.933	0.892	0.049
AFNet[79]	2023	-	-	-	-	0.928	0.958	0.926	0.032	0.925	0.968	0.936	0.020	0.924	0.956	0.926	0.021	0.888	0.923	0.89	0.056	0.918	0.918	0.957	0.034
TPCL[80]	2023	0.956	0.974	0.946	0.020	0.930	0.959	0.926	0.028	0.930	0.970	0.936	0.017	0.942	0.977	0.941	0.015	0.888	0.926	0.892	0.049	0.922	0.960	0.920	0.029
HiDANet[81]	2023	-	-	-	-	0.939	0.954	0.926	0.029	0.929	0.961	0.930	0.021	0.952	0.980	0.946	0.013	-	-	-	-	0.921	0.946	0.911	0.035
EM-Trans[82]	2024	-	-	-	-	0.935	0.961	0.931	0.027	0.934	0.970	0.939	0.017	-	-	-	-	-	-	-	-	0.926	0.958	0.925	0.028
DCT[83]	2024	0.952	0.969	0.948	0.023	0.934	0.959	0.932	0.031	0.923	0.965	0.934	0.023	0.944	0.978	0.948	0.017	-	-	-	-	0.919	0.955	0.922	0.035
EGANet[106]	2023	-	-	-	-	0.918	0.952	0.918	0.033	0.925	0.967	0.933	0.021	0.941	0.978	0.938	0.016	0.865	0.899	0.861	0.069	0.87	0.924	0.865	0.042

部区域的像素匹配程度，其计算公式为

$$E_{\xi}=\frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\boldsymbol{M}_{FM}(\boldsymbol{\xi}_{FM}(x,y)) \quad (4)$$

其中，$\boldsymbol{\xi}_{FM}(x,y)$表示对齐矩阵，用于衡量预测结果和标签之间的局部一致性和全局一致性。$\boldsymbol{M}_{FM}(*)$表示增强的一致性矩阵。更详细的度量计算细节请查阅文献[134]。

4 算法性能比价和分析

4.1 整体评估

根据已有文献公开结果，我们对近年来一些典型的基于深度学习的RGB-D图像显著性目标检测算法成果进行了汇总，并在表3中列举了部分方法在6个数据集上4个指标的定量评价结果。可以看出，近年来，基于深度学习的RGB-D图像显著性目标检测算法得到了飞速的发展，其检测性能在不断地提升。与此同时，图14中提供了一些典型的基于深度学习的RGB-D图像显著性目标检测算法在不同场景下的可视化结果。可以看出，这些方法在大多数简单场景中都能成功地定位到显著性目标，具有较好的检测精度和视觉效果，这进一步体现了基于深度学习的RGB-D图像显著性目标检测算法的优越性。接下来，我们将对每一类问题的典型方法展开具体分析。

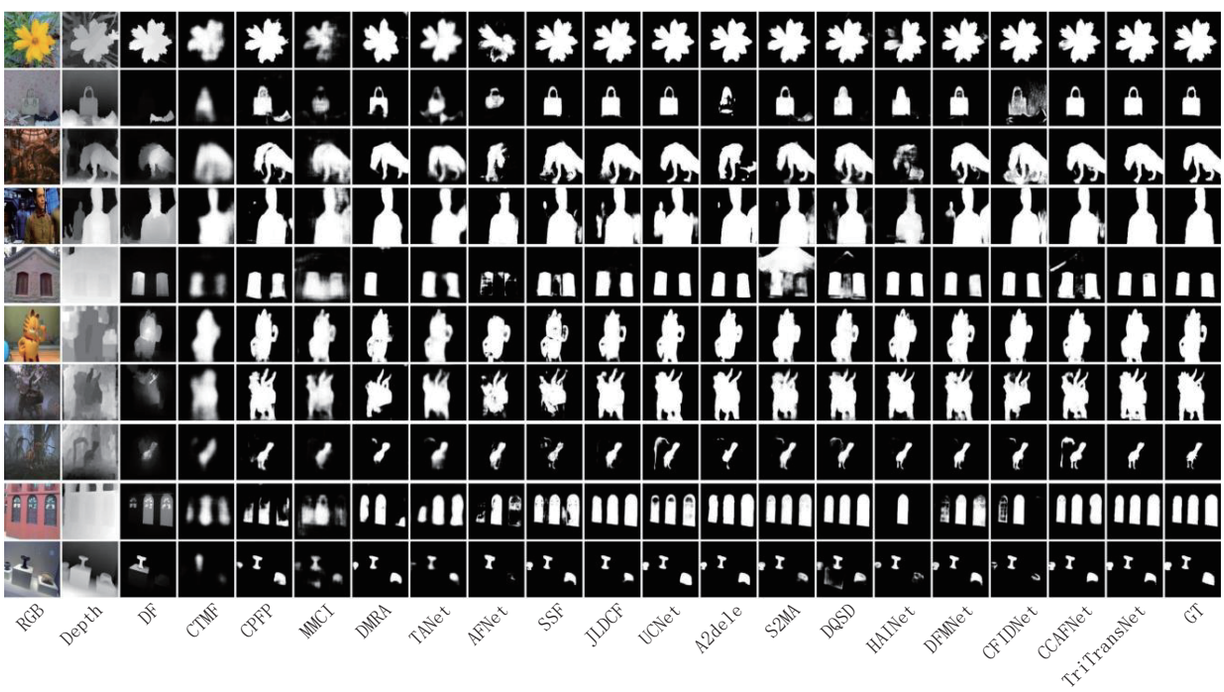


图14 基于深度学习的RGB-D图像显著性目标检测代表性算法在不同场景下的可视化结果

4.2 干扰信息问题的研究结果分析

表4给出了现有关注于干扰信息问题的典型算法的结果分析。从表4中可以看出，基于特征级选择的方法是现有解决干扰信息问题的主流思路，研究相对较多，图像级选择方法和基于图像增强的方法研究相对较少。从结果上看，基于特征级选择方法普遍取得了更好的效果。与此同时，作为一种典型的基于图像增强的方法，CDNet[44]同样能够取得具有竞争力的性能，说明图像增强的方法仍具有较大研究潜力和提升空间。此外，从网络结构看，基于信息选择的方法结构通常相对简单，使用也更加灵活，具有即插即用的特点，尤其是不同的特征选择方式具有不同的特点。其中基于空间注意力的方法能够较好地辨别输入图像的局部质量问题，基于通道注意力的方法能够较好地筛选辨别性特征，基于混合注意力的方法则是能够结合两者的优点。同时，由于难以对干扰信息建模，即提供干扰和非干扰信息的监督信息，其优化过程主要为数据驱动，可解释性差。而基于图像增强的方法网络结构通常相对复杂，且需要单独的子网络进行跨模态生成，并且其结果依赖于RGB-深度图像跨模态生成的效果。然而，相比于信息选择方法，其理论性更高，可解释性也更强。

4.3 特征提取问题研究结果分析

表5给出了现有关注于特征提取问题的典型算法结果。可以看出，现有研究特征提取问题的算法相对较少。并且相比于基于CNN的方法和基于图

表4 关注于干扰信息问题的典型算法结果分析

干扰信息	数据集	DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
	评价指标	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓
图像级选择	D3Net(2021)	0.793	0.829	0.773	0.098	0.900	0.950	0.900	0.041	0.897	0.953	0.912	0.030	0.885	0.946	0.898	0.031	0.810	0.862	0.825	0.095	0.891	0.938	0.899	0.046
特征级选择	ACCF(2018)	0.767	0.859	0.791	0.113	0.852	0.915	0.858	0.079	0.815	0.913	0.856	0.059	0.822	0.928	0.848	0.065	0.771	0.839	0.787	0.132	0.863	0.927	0.873	0.068
	TANet(2019)	0.790	0.861	0.808	0.093	0.874	0.925	0.878	0.060	0.863	0.941	0.886	0.041	0.827	0.910	0.858	0.046	0.796	0.847	0.801	0.111	0.861	0.923	0.871	0.060
	SSF(2020)	0.924	0.951	0.915	0.033	0.896	0.935	0.899	0.043	0.896	0.953	0.914	0.026	0.883	0.941	0.905	0.025	0.867	0.900	0.859	0.066	0.880	0.936	0.893	0.044
	MMNet(2020)	0.933	0.960	0.923	0.031	0.912	0.943	0.911	0.039	0.908	0.958	0.921	0.024	-	-	-	-	0.874	0.913	0.875	0.061	0.888	0.937	0.921	0.044
	DQAM(2020)	-	-	-	-	0.873	-	0.897	0.052	0.864	-	0.916	0.029	0.901	-	0.935	0.021	0.826	-	0.851	0.085	0.854	-	0.892	0.051
	DQSD(2020)	0.827	0.878	0.845	0.072	0.900	0.936	0.899	0.050	0.898	0.952	0.916	0.029	0.927	0.973	0.935	0.021	0.847	0.878	0.851	0.085	0.886	0.935	0.892	0.051
	AFINet(2021)	-	-	-	-	0.853	0.903	0.854	0.073	0.864	0.933	0.878	0.045	0.829	0.910	0.860	0.050	0.832	0.874	0.825	0.097	0.851	0.912	0.856	0.068
	MCMFNet(2021)	-	-	-	-	0.882	0.923	0.889	0.061	0.885	0.938	0.905	0.040	0.877	0.934	0.903	0.036	-	-	-	-	-	-	-	-
	GRNet(2022)	0.932	0.956	0.929	0.029	0.905	0.926	0.914	0.036	0.906	0.961	0.932	0.020	-	-	-	-	0.878	0.912	0.879	0.064	0.897	0.928	0.911	0.034
	BPGNet(2022)	0.938	0.958	0.930	0.031	0.926	0.953	0.923	0.034	0.914	0.959	0.927	0.024	0.932	0.973	0.937	0.020	0.875	0.908	0.874	0.066	0.904	0.944	0.907	0.040
图像增强	CDNet(2021)	0.934	0.955	0.930	0.029	0.918	0.950	0.918	0.036	0.919	0.960	0.929	0.023	0.929	0.973	0.936	0.019	0.879	0.911	0.877	0.061	0.907	0.947	0.912	0.037

注：最好的三个结果分别由红色，绿色和蓝色表示

表5 关注于特征提取问题的典型算法结果分析

特征提取	数据集 / 评价指标	DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
		F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓
FANet (2022)	VGG	-	-	-	-	0.888	0.914	0.893	0.050	0.881	0.945	0.904	0.030	0.874	0.941	0.895	0.029	0.831	0.862	0.808	0.097	0.854	0.903	0.857	0.060
	ResNet	-	-	-	-	0.892	0.914	0.899	0.044	0.885	0.951	0.913	0.026	0.874	0.925	0.894	0.026	0.855	0.882	0.850	0.076	0.863	0.908	0.881	0.047
Cas-GNN (2020)	GNN	0.912	0.932	0.891	0.042	0.903	0.933	0.911	0.035	0.904	0.952	0.919	0.025	0.906	0.947	0.905	0.028	0.864	0.877	0.849	0.073	0.901	0.930	0.899	0.039
SwinNet (2021)	Transformer	-	-	-	-	0.922	0.934	0.935	0.027	0.908	0.967	0.941	0.018	0.926	0.980	0.945	0.016	-	-	-	-	0.893	0.929	0.919	0.033

注：最好的三个结果分别由红色，绿色和蓝色表示

神经网络的方法，基于Transformer的方法通过挖掘RGB特征和深度特征中的长距离依赖关系，进而取得了更好的效果。例如，Liu等人[48]设计的方法“SwinNet”借助SwinTransformer从RGB图像和深度图像中分别提取了更具辨别性的单模态特征，从而在多个数据集上取得了优异的表现。总体而言，现有方法对特征提取问题研究相对较少，并倾向于使用发展较为成熟的预训练CNN、Transformer模型。在RGB-D显著目标检测领域，使用预训练模型能够利用预训练数据中的先验信息，有效地提高训练效果。然而，现有预训练模型主要是针对RGB图像数据进行优化设计并预训练，缺少针对深度图像数据优化和预训练的特征提取网络，这一领域有待进一步的探索与发展。

4.4 多模态信息融合问题的研究结果分析

表6给出了现有关注于研究多模态信息融合问题的典型方法的结果分析。从性能上看出，特征交互融合策略、分层级特征融合策略和特征分解融合策略都能取得先进的检测效果。从设计原理上来看，分层级特征融合策略和特征分解融合策略可以视为特征交互融合策略的进一步发展。相比于特征交互融合策略，分层级特征融合策略和特征分解融合策略是一种细粒度的融合策略，前者以特征层级为单位进行融合，而后者以模态特征类型为单位进行融合。相对应地，分层级特征融合策略和特征分解融合策略网络结构也更加复杂。此外，相比于特征交互融合策略和分层级特征融合策略，特征分解融合策略发展较晚，研究相对较少，然而如表5所示，Zhou等人[75]设计的方法“SP-Net”显式地联合了模态共性信息和模态特性信息，进而有效地提高了显著性目标检测性能，在NJU2K等多个数据集上取得了优异的表现。这充分体现了特征分解融合策略的有效性和发展潜力。此外，现有关注于研究多模态信息融合问题的典型方法在DUT-RGBD数据集上的检测结果相对较差，而在其他数据集上的检测结果具有相近的竞争力，这说明在一些相对复杂场景下，现有多模态信息融合策略仍未能很好地挖掘和利用多模态图像互补信息。

4.5 上下文信息挖掘问题的研究结果分析

表7给出了现有关注于研究上下文信息挖掘问题的典型方法的结果分析。从性能上看出，通过多尺度特征提取、多层级特征融合、全局信息/长距离依赖关系挖掘和边界信息挖掘等方式能较好地利用场景上下文信息，进而取得先进的检测效果。并且，近年来利用Transformer的特性，挖掘场景上下文信息，逐渐成为发展趋势。例如，TriTransNet[102]通过Transformer挖掘图像中的长距离依赖关系，弥补了CNN主要关注局部信息的缺陷，取得了较大的性能提升。从网络结构上来看，多尺度特征提取模块和全局信息/长距离依赖关系挖掘模块通常比较灵活，具有即插即用的特点；多层级特征融合模块的结构形式更加广泛，但是随着研究深入，其网络结构也逐渐复杂；引入边界信息能够明确对部分上下文信息进行建模，物理含义明确，然而其训练过程通常需要提供额外的监督信息。此外，上述上下文信息挖掘方法间并不冲突，甚至存在一定互补性，因此在实践或者未来研究中，可以进一步考虑结合使用上述策略，以更加充分地挖掘上下文信息。

4.6 模型复杂问题的研究结果分析

表8给出了现有典型的轻量级RGB-D图像显著性目标检测模型的结果。同时，在表9中，我们给出了部分非轻量化方法、单分支轻量化方法和双分支轻量化方法的参数量和推理速度。具体而言，我们首先从DUT-RGBD数据集中随机读取1000张图像并将图像尺寸统一调整为352×352。随后，我们在一张NVIDIA3090Ti显卡上以每个模型推理1000张图像的平均速度作为该模型的推理速度。可以看出，相比于标准的算法，现有轻量级算法具有更少的参数量和更快的运行速度。并且由表9可以看出，采用轻量级特征提取网络（如MobileNet）能够有效减少轻量化模型的参数量，并提高模型的推理速度。相比于现有的轻量化网络模型，MLF[123]同时从框架和模块两方面对模型进行轻量化设计，进而提出了一种基于中间层级特征融合的轻量级RGB-D显著目标检测范式，有效地平衡了性能与参数量之间的矛盾，能够为其他方法提供参考。然而，相比于标准的算法，这些轻量级算法的检测性能依然有着较大的差距。这说明RGB-D图像显著性目标检测方法的轻量化研究仍然处于起步阶段，依然有很多问题需要解决，未来应更加关注RGB-D图像显著性目标检测方法的轻量化研究。

5 总结与展望

本文对基于深度学习的RGB-D图像显著性目标检测方法进行了综述。首先，分析了RGB-D图像显著性目标检测领域中的关键问题。然后，根据这些问题对现有的RGB-D图像显著性目标检测算

表6　关注于研究多模态信息融合问题的典型方法结果分析

特征级融合方式	数据集	DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
	评级指标	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓
特征交互	PCFNet(2018)	0.771	0.858	0.801	0.1	0.871	0.896	0.877	0.059	0.841	0.916	0.874	0.044	0.804	0.912	0.842	0.049	0.779	0.842	0.794	0.112	0.859	0.897	0.875	0.064
	ASIFNet(2020)	-	-	-	-	0.877	0.907	0.891	0.047	0.869	0.944	0.909	0.029	0.915	0.974	0.934	0.019	0.850	-	0.810	0.069	0.852	0.908	0.874	0.051
	ACMF(2020)	0.874	-	0.881	0.059	0.089	-	0.904	0.044	0.876	-	0.914	0.028	-	-	-	-	0.865	-	0.864	0.084	-	-	-	-
	IRFR(2021)	0.951	0.924	0.919	0.035	0.908	0.945	0.909	0.040	0.910	0.960	0.921	0.026	0.920	0.971	0.930	0.021	0.861	0.898	0.860	0.072	0.893	0.941	0.897	0.044
	CCNet(2021)	-	-	-	-	0.929	0.948	0.917	0.037	0.922	0.966	0.926	0.023	0.936	0.966	0.922	0.022	0.894	0.907	0.875	0.061	0.913	0.944	0.908	0.037
	BTSNet(2021)	-	-	-	-	0.924	0.954	0.921	0.036	0.923	0.965	0.934	0.023	0.940	0.979	0.943	0.018	0.874	0.906	0.867	0.070	0.911	0.949	0.915	0.038
	LIANet(2022)	-	-	-	-	0.911	0.911	0.904	0.042	-	-	-	-	0.938	0.969	0.929	0.021	0.884	0.889	0.862	0.070	0.914	0.929	0.906	0.037
分层级特征融合	FRDT(2020)	0.903	0.941	0.910	0.039	0.879	0.917	0.898	0.048	0.867	0.945	0.914	0.029	0.868	0.942	0.902	0.028	0.855	0.899	0.857	0.073	0.880	0.927	0.901	0.043
	CMWNet(2020)	-	-	-	-	0.902	0.936	0.903	0.046	0.903	0.951	0.917	0.029	0.930	0.969	0.934	0.022	0.883	0.912	0.876	0.066	0.901	0.944	0.905	0.043
	ACFNet(2021)	0.871	0.943	0.909	0.041	0.883	0.936	0.914	0.037	0.881	0.955	0.924	0.025	0.873	0.965	0.915	0.022	0.802	0.882	0.852	0.079	0.859	0.935	0.904	0.040
	CCAFNet(2021)	0.915	0.941	0.905	0.036	0.910	0.943	0.909	0.037	0.908	0.956	0.921	0.026	0.937	0.977	0.938	0.018	0.832	0.876	0.826	0.087	0.887	0.934	0.892	0.044
特征分解	DCMF(2020)	0.932	0.958	0.928	0.035	0.915	0.948	0.913	0.043	0.906	0.954	0.922	0.029	-	-	-	-	0.875	0.909	0.878	0.068	0.906	0.946	0.910	0.043
	SP-Net(2021)	-	-	-	-	0.935	0.954	0.925	0.028	0.925	0.959	0.927	0.021	0.950	0.980	0.945	0.014	-	-	-	-	0.915	0.944	0.907	0.037
其他	TMFNet(2021)	-	-	-	-	0.882	0.910	0.910	0.041	0.867	0.944	0.921	0.027	0.892	0.968	0.936	0.021	0.846	0.865	0.849	0.084	-	-	-	-

注:最好的三个结果分别由红色,绿色和蓝色表示

表7 关注于研究上下文信息挖掘问题的典型方法结果分析

上下文信息挖掘	数据集	DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
	评价指标	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓
多尺度特征提取	DMRA(2019)	0.898	0.933	0.889	0.048	0.872	0.908	0.886	0.051	0.855	0.942	0.899	0.031	0.857	0.945	0.901	0.029	0.849	0.899	0.847	0.075	0.868	0.920	0.886	0.047
	HDFNet(2020)	0.930	0.938	0.907	0.041	0.922	0.932	0.908	0.038	0.927	0.957	0.923	0.023	0.932	0.971	0.926	0.021	0.883	0.891	0.854	0.076	0.910	0.931	0.900	0.041
	PMFNet(2021)	-	-	-	-	0.919	0.926	0.910	0.045	0.924	0.942	0.923	0.030	0.933	0.964	0.926	0.027	0.894	0.907	0.874	0.072	-	-	-	-
	PGFNet(2021)	0.926	0.954	0.921	0.032	0.921	0.952	0.918	0.033	0.913	0.960	0.923	0.023	0.932	0.971	0.932	0.017	0.864	0.897	0.850	0.070	0.903	0.944	0.906	0.037
	AILNet(2022)	0.917	0.951	0.926	0.031	0.876	0.912	0.898	0.045	0.857	0.935	0.912	0.029	0.873	0.941	0.911	0.024	0.852	0.900	0.866	0.068	0.880	0.922	0.908	0.038
多层级特征融合	CPFP(2019)	0.793	0.829	0.773	0.076	0.877	0.923	0.878	0.053	0.869	0.932	0.888	0.036	0.838	0.923	0.872	0.037	0.826	0.872	0.828	0.088	0.874	0.925	0.879	0.051
	MCINet(2020)	-	-	-	-	0.873	0.920	0.900	0.050	0.890	0.947	0.917	0.027	0.897	0.957	0.927	0.024	-	-	-	-	0.872	0.929	0.901	0.042
	PGHF(2020)	-	-	-	-	0.884	0.876	0.882	0.050	0.897	0.883	0.909	0.027	0.919	0.907	0.923	0.021	0.840	0.835	0.835	0.083	-	-	-	-
	BBSNet(2020)	-	-	-	-	0.920	0.949	0.921	0.035	0.918	0.961	0.930	0.023	0.927	0.966	0.933	0.021	0.858	0.901	0.864	0.072	0.903	0.942	0.908	0.041
	MRINet(2021)	-	-	-	-	0.864	0.893	0.904	0.046	0.860	0.934	0.917	0.029	-	-	-	-	-	-	-	-	0.829	0.888	0.881	0.060
	FANet(2022)	-	-	-	-	0.892	0.914	0.899	0.044	0.885	0.951	0.913	0.026	0.874	0.925	0.894	0.026	0.855	0.882	0.850	0.076	0.863	0.908	0.881	0.047
	CFIDNet(2022)	-	-	-	-	0.923	0.913	0.914	0.038	0.915	0.950	0.922	0.026	-	-	-	-	0.884	0.893	0.870	0.070	0.908	0.924	0.901	0.043
全局信息/长距离依赖	MMCI(2019)	0.767	0.859	0.791	0.113	0.852	0.915	0.858	0.079	0.815	0.913	0.856	0.059	0.822	0.928	0.848	0.065	0.771	0.839	0.787	0.132	0.863	0.927	0.873	0.068
	M2RNet(2021)	0.925	0.935	0.903	0.042	0.922	0.904	0.910	0.049	0.921	0.941	0.918	0.033	0.937	0.971	0.934	0.019	0.861	0.874	0.842	0.088	0.913	0.929	0.899	0.042
	TriTransNet(2021)	0.938	0.957	0.933	0.025	0.919	0.925	0.920	0.030	0.909	0.960	0.928	0.020	0.936	0.981	0.943	0.014	-	-	-	-	0.893	0.927	0.908	0.033
	DCMF(2022)	0.932	0.958	0.928	0.035	0.915	0.948	0.913	0.043	0.906	0.954	0.922	0.029	-	-	-	-	0.875	0.909	0.878	0.068	0.906	0.945	0.910	0.043
边界信息	ACENet(2021)	0.922	-	0.924	0.031	0.892	-	0.908	0.039	0.875	-	0.929	0.025	0.896	-	0.923	0.022	0.862	-	0.866	0.063	0.886	-	0.895	0.043

注:最好的三个结果分别由红色,绿色和蓝色表示

表8　轻量级RGB-D图像显著性目标检测模型的结果对比

模型复杂	数据集	DUT-RGBD				NJU2K				NLPR				RGBD135				LFSD				STERE			
	评价指标	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓	F_β↑	E_ξ↑	S_α↑	MAE↓
单分支模型	A2dele(2020)	0.870	0.920	0.884	0.042	0.873	0.916	0.869	0.051	0.881	0.945	0.881	0.028	0.890	0.971	0.943	0.020	0.836	0.900	0.862	0.068	0.879	0.928	0.879	0.043
	DANet(2020)	0.883	0.934	0.899	0.043	0.871	0.922	0.899	0.045	0.870	0.949	0.915	0.028	0.889	0.967	0.937	0.021	0.822	0.874	0.849	0.079	0.868	0.921	0.895	0.043
	CoNet(2020)	0.908	0.941	0.918	0.034	0.872	0.912	0.894	0.047	0.848	0.936	0.907	0.031	0.861	0.945	0.910	0.027	0.848	0.897	0.862	0.071	0.885	0.923	0.908	0.041
	DKDNet(2021)	-	-	-	-	0.889	0.923	0.877	0.049	0.890	0.941	0.892	0.032	0.923	0.965	0.917	0.023	0.877	0.897	0.850	0.068	-	-	-	-
双分支模型	PGAR(2020)	0.852	0.889	0.853	0.074	0.893	0.916	0.909	0.042	0.885	0.955	0.930	0.024	0.880	0.939	0.914	0.026	0.852	0.889	0.853	0.074	0.880	0.919	0.907	0.041
	ATSNet(2020)	0.920	0.948	0.918	0.032	0.893	0.921	0.901	0.040	0.876	0.945	0.907	0.028	0.885	0.952	0.907	0.024	0.862	0.905	0.865	0.064	0.884	0.921	0.897	0.039
	GRNet(2022)	0.932	0.956	0.929	0.029	0.905	0.926	0.914	0.036	0.906	0.961	0.932	0.020	-	-	-	-	0.878	0.912	0.879	0.064	0.897	0.928	0.911	0.034
	MobileSal(2021)	0.912	0.940	0.896	0.041	0.914	0.939	0.905	0.041	0.916	0.950	0.920	0.025	-	-	-	-	0.867	0.880	0.847	0.080	0.906	0.916	0.903	0.041
	MoADNet(2022)	0.920	0.945	0.907	0.033	0.907	0.929	0.901	0.042	0.908	0.950	0.918	0.024	-	-	-	-	0.873	0.902	0.859	0.064	0.901	0.931	0.896	0.043
	MLF(2022)	-	-	-	-	0.900	0.941	0.913	0.035	0.898	0.952	0.923	0.025	-	-	-	-	-	-	-	-	0.881	0.936	0.903	0.039

注：最好的三个结果分别由红色、绿色和蓝色表示

法进行了分析和探究。随后，介绍了RGB-D图像显著性目标检测任务中的主流数据集和评价指标，并以此对现有方法进行了对比和分析。最后，本文结合RGB-D图像显著性目标检测的研究现状和现实需求，做出以下展望。

5.1　模型设计

近年来，RGB-D图像显著性目标检测快速发展，其模型设计相关方法和相关理论发展相对成熟，但是依然存在以下不足。

（1）特征提取：现有RGB-D图像显著性目标检测主要利用现有的、在诸如ImageNet等大规模数据集上经过预训练的基础模型，例如VGGNet、ResNet以及Transformer等，进行单模态特征提取。极少有工作围绕多模态图像的特征提取基础网络展开研究，然而目前使用的预训练模型主要是在可见光图像上进行预训练，其数据分布与深度图像的数据分布不一致，在一定程度上影响了最终显著目标检测算法性能。我们认为在多模态图像处理领域中，研究包含RGB图像、深度图像、文本、声音等多模态信息的统一模态大模型，必然会成为多模态领域研究的热点之一。事实上，今年已经有学者开展了多模态图像大模型的研究工作，如Meta-Transformer[135]，证明了多模态大模型的可行性和有效性。

（2）轻量化方法：在实际应用中，RGB-D图像显著性目标检测模型通常部署在资源受限的设备上，然而相比于单模态图像显著性目标检测模型，多模态图像显著性目标检测模型需要处理多个模态的数据，使得其具有较高的参数量和计算复杂度。同时，目前多模态图像显著性目标检测模型的轻量化方法仍然处于起步阶段，不足以支撑实际应用的需求。与此同时，随着技术的不断发展，现有RGB-D显著目标检测模型在现有数据集上的性能，逐渐进入瓶颈。对于有监督RGB-D显著目标检测方法研究而言，轻量化方法依然是未来重要的研究方向和发展趋势之一。

（3）损失函数：多模态图像显著性目标检测主要利用常见的损失函数，包括交叉熵损失函数（BCELoss）[136]和交并比损失函数（IoULoss）[137]，进行模型训练。例如，Zhai等人[91]和Xiao等人[138]利用交叉熵损失函数，对网络的某一或某些层级特征进行监督。仅有部分方法尝试在BCE损失与IoU损失的基础上，添加新的损失函数，以更好地区分前景与背景。例如，Zhang等人[43]设计了一种新的区域

表9 轻量级RGB-D图像显著性目标检测模型的参数量和推理速度结果对比

	模型	骨干网络	参数量/M↓	推理速度/FPS↑
非轻量化方法	EBFS(2022)	VGG-16	122.70	7
	DSNet(2021)	VGG-16	173.00	17
	TriTransNet(2021)	VIT	139.60	9
	SwinNet(2022)	Swin-Transformer	199.20	10
单分支轻量化方法	A2dele(2020)	VGG-16	15.10	121
	DANet(2020)	VGG-16	26.70	97
	CoNet(2020)	ResNet-50	43.70	62
双分支轻量化方法	MoADNet(2022)	MobileNetV3	5.03	62
	MLF(2022)	MobileNetV3	3.70	69
	MobileSal(2021)	MobileNetV2	6.50	141

一致性感知(Region Consistency Aware,RCA)损失,该损失通过考虑前景显著目标区域和背景区域内的局部区域一致性,进一步减少了背景干扰并获得了更加完整的检测结果。Zhao等人[139]设计了一个深度感知误差加权(Depth-aware Error-weighted)损失来挖掘模糊像素。目前,仍未有工作对损失函数的设计和使用方式进行系统研究。因此,如何优化设计损失函数以实现准确的显著性目标分割是未来的研究方向之一。

5.2 学习方法

目前,RGB-D图像显著性目标检测方法主要为有监督学习方法,其学习过程依赖于大量的、带有高质量标注训练数据的支撑。然而,如前所述,现有RGB-D图像显著目标检测数据集远少于单模态图像显著目标检测数据,不足以支撑基于有监督学习方法的RGB-D图像显著目标检测算法的进一步发展。并且,RGB-D图像显著目标检测任务为像素级标注任务,其数据标注所需的人力、物力成本相对较高。因此,研究如何突破有监督学习方法的局限性,同样是未来重要的发展方向之一,具体包括:

(1)基于弱、半、无监督学习的RGB-D图像显著目标检测方法研究:弱监督学习方法、半监督学习方法和无监督学习方法是缓解基于有监督学习的RGB-D图像显著目标检测方法对海量标注数据依赖的重要手段。其中,弱监督学习方法通过为视觉数据提供相对简单的弱标注(例如显著目标类别、框标注和涂鸦标注等),而不是完全的、相对困难的像素级标注,并通过设计弱监督学习策略,从相对较弱的标注中,自主挖掘更加详细的、视觉任务感兴趣目标的其他信息(如位置,大小,轮廓等),进而减少深度学习模型对高质量标注的依赖性。半监督学习方法则是仅对视觉数据进行少量标注,而不是对所有的视觉数据进行标注,并通过设计学习策略以充分挖掘未标注数据中的显著信息,进行算法的训练。无监督学习方法则是挖掘训练数据中的显著模式,进行算法的训练。目前,已经有少量研究[140-143]初步探索了RGB-D图像显著目标检测的弱、半、无监督学习方法,证明了相关领域的可行性。研究基于弱、半、无监督学习的RGB-D图像显著目标检测方法同样是未来的必然趋势之一。

(2)基于多模态大模型迁移学习、少样本学习、零样本学习的RGB-D图像显著目标检测方法研究:近年来,多模态图像大模型的研究逐步成为了计算机视觉领域的研究热点之一。相比于现有预训练模型,多模态图像大模型是通过海量的视觉数据进行训练,因此能够从海量训练数据中学习更加准确的知识,对不同模态数据中提取特征的辨别能力更强。利用现有多模态图像大模型[135],设计一定的迁移学习方法、少样本学习方法或者零样本学习方法,挖掘多模态图像大模型中包含的丰富知识,辅助RGB-D图像显著目标检测方法的训练,同样能够缓解基于有监督学习的RGB-D图像显著目标检测方法对数据的依赖性,这也是未来的发展方向之一。

5.3 多模态显著目标检测未来趋势

近年来,随着成像技术的快速发展,深度相机、红外相机、事件相机等多模态相机在社会生产、生活中得到广泛应用。利用多模态图像间的互补信息,推动显著目标检测的发展也成为研究热点之一。基于多模态图像的显著性目标检测任务,其未来的研究方向包括但不局限于以下几点:

(1)探索更多模态数据的联合使用:目前多模态图像显著目标检测,如RGB-D显著目标检测、RGB-T显著目标检测等,主要关注于如何挖掘和利用两个模态图像中的互补信息,从而进行显著目标

检测。然而，真实场景往往是复杂多变的，仅依靠两个模态图像间的互补信息难以充分反映场景复杂情况，进行准确的显著目标检测。因此，在复杂场景下，研究更多模态图像的显著目标检测方法，如基于RGB图像、红外图像和深度图像的三模态显著目标检测方法，突破单模态、两模态图像显著目标检测方法的局限性，同样是未来重要的发展方向之一。

(2)设计更加普适的模型：目前，多模态图像显著目标检测方法通常是根据某些固定模态图像进行设计或者训练，能够在它们的目标模态上取得较好的检测效果，然而在非其目标模态上则会取得较差的结果，如RGB-D图像显著性目标检测模型无法在RGB-T图像显著性目标检测模型上取得较好的效果，即普适性相对较差。然而，实际应用可能需要选择性地利用不同的多模态图像进行显著目标检测，如RGB-D图像或者RGB-T图像，这就要求实际应用进行设计并训练多个多模态显著目标检测模型进行切换，增加了设计成本。目前，已经有部分多模态图像显著目标检测算法可被同时应用于RGB-D图像显著性目标检测和RGB-T图像显著性目标检测[51,67]。这些算法在一定程度上提高了算法的普适性。

然而，这些算法依然需要在RGB-D图像显著性目标检测和RGB-T图像显著性目标检测数据上分别进行训练。与之相对，若能通过一个模型和一次训练便能够实现RGB-D图像显著性目标检测和RGB-T图像显著性目标检测，则能够进一步提升算法的普适性，减少实际应用的研发成本。事实上，CVPR24年的工作VSCode[144]已经实现了一个模型对RGB图像、RGB-T图像和RGB-D图像的显著目标检测和隐藏目标检测任务。Huang等人[145]更是进一步提出了任意模态显著目标检测任务，实现了通过一个模型，便能够从包括RGB图像、深度图像、红外图像、RGB-T图像、RGB-D图像、D-T图像和RGB-D-T图像等任意模态输入中检测显著目标。设计更加普适的模型已经成为了当下的研究热点。

作者贡献声明　黄年昌、杨阳对本文贡献相同，为共同第一作者。

参 考 文 献

[1] Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 597-606

[2] Guo C, Zhang L. A novel multi resolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on Image Processing, 2009, 19(1): 185-198

[3] Mukherjee P, Lall B. Saliency and KAZE feature sassisted object segmentation. Image and Vision Computing, 2017, 61: 82-97

[4] Zhao Xing-Ke, Li Ming-Lei, Zhang Gong, et al. Object detection method based on saliency map fusion foruav-borne thermal images. Acta Automatica Sinica, 2021, 47(9): 2120-2131 (in Chinese)
(赵兴科，李明磊，张弓等. 基于显著图融合的无人机载热红外图像目标检测方法. 自动化学报，2021，47(9)：2120-2131)

[5] Liu Song-Tao, Liu Zhen-Xing, Jiang Ning. Target segmentation of infrared image using fused saliency map and efficient subwindow search. Acta Automatica Sinica, 2018, 44(12): 2210-2221 (in Chinese)
(刘松涛，刘振兴，姜宁. 基于融合显著图和高效子窗口搜索的红外目标分割. 自动化学报，2018，44(12)：2210-2221)

[6] Zhang Dong-Ming, Jin Guo-Qing, Dai Feng, et al. Salient object detection based on deep fusion of handcrafted features. Chinese Journal of Computers, 2019, 42(9): 2076-2086 (in Chinese)
(张冬明，靳国庆，代锋等. 基于深度融合的显著性目标检测算法. 计算机学报，2019，42(9)：2076-2086)

[7] 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

[8] Ren J, Gong X, Yu L, et al. Exploiting global priors for RGB-D saliency detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, USA, 2015: 25-32

[9] Guo J, Ren T, Bei J. Salient object detection for RGB-D image via saliency evolution//Proceedings of the IEEE International Conference on Multimedia and Expo. Seattle, USA, 2016: 1-6

[10] Peng H, Li B, Xiong W, et al. RGB-Dsalient object detection: A benchmark and algorithms//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 92-109

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the International Conference on Learning Representations. SanDiego, USA, 2015: 1-5

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 770-778

[13] Qu L, He S, Zhang J, et al. RGB-D salient object detection via deep fusion. IEEE Transactions on Image Processing, 2017, 26(5): 2274-2285

[14] Zhou T, Fan D, Cheng M, et al. RGB-Dsalient object detection: A survey. Computational Visual Media, 2021, 7: 37-69

[15] Ren T, Zhang A. RGB-D salient object detection: A review. RGB-D Image Analysis and Processing, 2019, 203-220

[16] Wu Lan-Hu, Li Zhi-Wei, Liu Lei-Ye, et al. A survey of salient object detection based on scene geometric information. Pattern Recognition and Artificial Intelligence, 2023, 36(2): 120-142 (in Chinese)
(吴岚虎,李智玮,刘垒烨等. 基于场景几何信息的显著性目标检测方法综述. 模式识别与人工智能, 2023, 36(2):120-142)

[17] Zhang J, Fan D, Dai Y, et al. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational auto encoders //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA, 2020: 8579-8588

[18] Chen Q, Liu Z, Zhang Y, et al. RGB-D salient object detection via 3D convolutional neural networks//Proceedings of the AAAI Conference on Artificial Intelligence, 35: 2. Online, 2021: 1063-1071

[19] Wang X, Li S, Chen C, et al. Data-level recombination and lightweight fusion scheme for RGB-D salient object detection. IEEE Transactionson Image Processing, 2020, 30: 458-471

[20] Wang N, Gong X. Adaptive fusion for RGB-Dsalient object detection. IEEE Access, 2019,7: 55277-55284

[21] Wang X, Li S, Chen C, et al. Knowing depthquality in advance: A depth quality assessmentmethod for RGB-D salient object detection.arXiv preprint arXiv:2008.04157, 2020

[22] Chen H, Li Y. Progressively complementarity-aware fusion network for RGB-D salient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. SaltLake City, USA, 2018: 3051-3060

[23] Zhang W, Jiang Y, Fu K, et al. BTS-Net: Bi-directional transfer-and-selection network for RGB-D salient object detection//Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen,China, 2021: 1-6

[24] Huang N, Yang Y, Zhang D, et al. Employing bilinear fusion and saliency prior information for RGB-D salient object detection. IEEE Transactions on Multimedia, 2021, 24: 1651-1664

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

[26] Fan D, Lin Z, Zhao J X, et al. Rethinking RGB Dsalient object detection: Models, datasets,and large-scale benchmarks. IEEE Transactionson Neural Networks and Learning Systems, 2021, 32: 2075-2089

[27] Chen Z, Cong R, Xu Q, et al. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. IEEE Transactionson Image Processing, 2020, 30: 7012-7024

[28] Li L, Zhao S, Sun R, et al. AFI-Net: Attention-guided feature integration network for RGBD saliency detection. Computational Intelligence and Neuroscience, 2021,2021(1): 8861446

[29] 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

[30] Huang N, Luo Y, Zhang Q, et al. Discriminative unimodal feature selection and fusion for RGB Dsalient object detection. Pattern Recognition,2022, 122: 108359

[31] Chen C, Wei J, Peng C, et al. Depth-quality-awaresalient object detection. IEEE Transactionson Image Processing, 2021, 30: 2350-2363

[32] Chen H, Li Y, Su D. Attention-Aware cross-modal cross-level fusion network for RGB-Dsalient object detection//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain, 2018:6821-6826

[33] Chen H, Li Y. Three-stream attention-aware network for RGB-D salient object detection. IEEE Transactions on Image Processing, 2019,28(6): 2825-2835

[34] Huang N, Liu Y, Zhang Q, et al. Joint cross-modal and unimodal features for RGB-D salient object detection. IEEE Transactions on Multimedia,2021, 23: 2428-2441

[35] Wu J, Zhou W, Luo T, et al. Multiscale multilevel context and multimodal fusion for RGB Dsalient object detection. Signal Processing,2021, 178: 107766

[36] Liao G, Gao W, Jiang Q, et al. Mmnet: Multistage and multi-scale fusion network for RGB Dsalient object detection// Proceedings of the ACM International Conference on Multimedia. Online, 2020: 2436-2444

[37] Song M, Song W, Yang G, et al. Improving RGB-D salient object detection via modality-aware decoder. IEEE Transactions on Image Processing, 2022, 31: 6124-6138

[38] Yang Y, Qin Q, Luo Y, et al. Bi-directional progressive guidance network for RGB-D salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022,32(8): 5346-5360

[39] Wen H, Yan C, Zhou X, et al. Dynamic selective network for RGB-D salient object detection. IEEE Transactions on Image Processing, 2021,30: 9179-9192

[40] Zhang M, Ren W, Piao Y, et al. Select, supplement and focus for RGB-D saliency detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3472-3481

[41] Feng Z, Wang W, Li W, et al. MFUR-Net: Multimodal feature fusion and unimodal feature refinement for RGB-D salient object detection. Knowledge-Based Systems, 2024, 299: 112022

[42] Song P, Li W, Zhong P, et al. Synergizing triple attention with depth quality for RGB Dsalient object detection. Neurocomputing, 2024, 589: 127672

[43] Zhang Q, Qin Q, Yang Y, et al. Feature calibrating and fusing network for RGB-D salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(3): 1493-1507

[44] Jin W D, Xu J, Han Q, et al. CDNet: Complementary depth network for RGB-D salient objectdetection. IEEE Transactions on Image Processing, 2021, 30: 3376-3390

[45] 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 Transactions on Image Processing, 2020,29: 4296-4307

[46] Zuo K, Xiao H, Zhang H, et al. Improving RGB-D salient

object detection by addressing inconsistent saliency problems. Knowledge-Based Systems，2024，299：111996

[47] Luo A，Li X，Yang F，et al. Cascade graph neural networks for RGB-D salient object detection//Proceedings of the European Conference on Computer Vision. Online，2020：346-364

[48] Liu Z，Tan Y，He Q，et al. SwinNet：Swin transformer drives edge-aware RGB-D and RGB-Tsalient object detection. IEEE Transactions on Circuits and Systems for Video Technology，2021，32(7)：4486-4497

[49] Liu Z，Lin Y，Cao Y，et al. Swin transformer：Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada，2021：10012-10022

[50] Liu N，Zhang N，Wan K，et al. Visual saliency transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada，2021：4722-4732

[51] Tang B，Liu Z，Tan Y，et al. HR TransNet：HR Former-driven two-modality salient object detection. IEEE Transactions on Circuits and Systems for Video Technology，2023，33(2)：728-742

[52] Huang P，Shen C，Hsiao H. RGB-Dsalient object detection using spatially coherent deep learning framework//Proceedings of the IEEE International Conference on Digital Signal Processing. Tokyo，Japan，2018：1-5

[53] 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

[54] Li F，Zheng J，Zhang Y F. Depth-guided deformable convolutions for RGB-D saliency object detection//Proceedings of the International Conference on Communication，Imageand Signal Processing. Chengdu，China，2021：234-239

[55] Chen Q，Zhang Z，Lu Y，et al. 3-D convolutional neural networks for RGB-D salient object detection and beyond. IE EE Transactions on Neural Networks and Learning Systems，2024，35(3)：4309-4323

[56] Zhou J，Wang L，Lu H，et al. MVSalNet：Multiview augmentation for RGB-D salient object detection//Proceedings of the European Conference on Computer Vision. Tel Aviv，Israel，2022：270-287

[57] Zhou W，Guo Q，Lei J，et al. IRFR-Net：Interactive recursive feature-reshaping network for detecting salient objects in RGB-D images. IEEE Transactions on Neural Networks and Learning Systems，2021，1-13

[58] Li C，Cong R，Kwong S，et al. ASIF-Net：Attention steered interweave fusion network for RGB-D salient object detection. IEEE Transactions on Cybernetics，2020，51(1)：88-100

[59] Han Y，Wang L，Du A，et al. LIANet：Layer interactive attention network for RGB-D salient object detection. IEEE Access，2022，10：25435-25447

[60] Bai Z，Liu Z，Li G，et al. Circular complement network for RGB-D salient object detection. Neurocomputing，2021，451：95-106

[61] Liu D，Zhang K，Chen Z. Attentive cross-modal fusion network for RGB-D saliency detection. IEEE Transactions on Multimedia，2020，23：967-981

[62] Liang F，Duan L，Ma W，et al. Context-aware network for RGB-D salient object detection. Pattern Recognition，2021，111：107630

[63] Song H，Wang W，Zhao S，et al. Pyramid dilated deeper convlstm for video salient object detection//Proceedings of the European Conference on Computer Vision. Munich，Germany，2018：715-731

[64] Lee M，Park C，Cho S，et al. SPSN：Super-pixel prototype sampling network for RGB-Dsalient object detection//Proceedings of the European Conference on Computer Vision. Tel Aviv，Israel，2022：630-647

[65] Cong R，Lin Q，Zhang C，et al. CIR-Net：Cross-modality interaction and refinement for RGB Dsalient object detection. IEEE Transactions on Image Processing，2022，31：6800-6815

[66] Zhang M，Yao S，Hu B，et al. C2DFNet：Criss-cross dynamic filter network for RGB Dsalient object detection. IEEE Transactions on Multimedia，2022，25：5142-5154

[67] Pang Y，Zhao X，Zhang L，et al. CAVER：Cross-modalview-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing，2023，32：892-904

[68] 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 Transactions on Multimedia，2021，24：2192-2204

[69] Yao C，Feng L，Kong Y，et al. Double cross-modality progressively guided network for RGB-D salient object detection. Image and Vision Computing，2022，117：104351

[70] 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//Proceedings of the ACM International Conference on Multimedia. Online，2020：4107-4115

[71] Li G，Liu Z，Ye L，et al. Cross-modal weighting network for RGB-D salient object detection//Proceedings of the European Conference on Computer Vision. Online，2020：665-681

[72] Zhu J. ACFNet：Adaptively-cooperative fusion network for RGB-D salient object detection. arXiv preprint arXiv：2109.04627，2021

[73] Chen H，Deng Y，Li Y，et al. RGB-D salient object detection via disentangled cross-modal fusion. IEEE Transactions on Image Processing，2020，29：8407-8416.

[74] Zhang J，Fan D，Dai Y，et al. RGB-D saliency detection via cascaded mutual information minimization//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada，2021：4338-4347

[75] Zhou T，Fu H，Chen G，et al. Specificity-preserving RGB-D saliency detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，Canada，2021：4681-4691

[76] Wang X，Sun T，Yang R，et al. Quality-aware dual-modal saliency detection via deep reinforcement learning. Signal Processing：Image Communication，2019，75：158-167

[77] Zhou W，Pan S，Lei J，et al. TMFNet：Three-input multilevel fusion network for detecting salient objects in RGB-D images.

IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 6(3): 593-601

[78] Huang R, Xing Y, Zou Y. Triple-complementary network for RGB-D salient object detection. IEEE Signal Processing Letters, 2020, 27: 775-779

[79] Chen T, Xiao J, Hu X, et al. Adaptive fusion network for RGB-D salient object detection. Neurocomputing, 2023, 522: 152-164

[80] Wu J, Hao F, Liang W, et al. Transformer fusion and pixel-level contrastive learning for RGB-Dsalient object detection. IEEE Transactions on Multimedia, 2023, 26: 1011-1026

[81] Wu Z, Allibert G, Meriaudeau F, et al. Hidanet:Rgb-d salient object detection via hierarchical depth awareness. IEEE Transactions on Image Processing, 2023, 32: 2160-2173

[82] Chen G, Wang Q, Dong B, et al. EM-Trans: Edge-aware multimodal transformer for RGB-D salient object detection. IEEE Transactions on Neural Networks and Learning Systems, 2024, 1-14

[83] Chen H, Shen F, Ding D, et al. Disentangled cross-modal transformer for RGB-D salient object detection and beyond. IEEE Transactions on Image Processing, 2024, 33:1699-1709.

[84] Wu Z, Gobichettipalayam S, Tamadazte B, et al.Robust RGB-D fusion for saliency detection//Proceedings of the International Conference on 3D Vision. Prague, Czechia, 2022: 403-413

[85] Ren G, Xie Y, Dai T, et al. Progressive multiscale fusion network for RGB-D salient object detection. Computer Vision and Image Understanding,2022, 223: 103529

[86] Wu J, Han G, Wang H, et al. Progressive guide dfusion network with multi-modal and multiscale attention for RGB-D salient object detection. IEEE Access, 2021, 9: 150608-15062.

[87] Wu J, Sun F, Xu R, et al. Aggregate interactive learning for RGB-D salient object detection. Expert Systems with Applications, 2022, 195:116614

[88] Pang Y, Zhang L, Zhao X, et al. Hierarchical dynamic filtering network for RGB-D salient object detection//Proceedings of the European Conference on Computer Vision. Online, 2020: 235-252

[89] Piao Y, Ji W, Li J, et al. Depth-induced multiscale recurrent attention network for saliency detection//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 7254-7263

[90] Xia C, Duan S, Gao X, et al. GCENet: Global contextual exploration network for RGB-Dsalient object detection. Journal of Visual Communication and Image Representation, 2022, 89: 103680

[91] ZhaiY, Fan D P, Yang J, et al. Bifurcated backbone strategy for RGB-D salient object detection. IEEE Transactions on Image Processing, 2021, 30: 8727-8742

[92] Chen T, Hu X, Xiao J, et al. CFIDNet: Cascaded feature interaction decoder for RGB-D salient object detection. Neural Computing and Applications, 2022, 34(10): 7547-7563

[93] Zhao J, CaoY, Fan D, et al. Contrast prior and fluid pyramid integration for RGB Dsalient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019:3927-3936

[94] Huang Z, Chen H X, Zhou T, et al. Multilevel cross-modal interaction network for RGB Dsalient object detection. Neurocomputing, 2021, 452: 200-211

[95] Chen H, Li Y, Su D. Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection. IEEE Transactions on Cybernetics, 2019, 50(11): 4808-4820

[96] Zhou W, Pan S, Lei J, et al. MRINet: Multilevel reverse-context interactive-fusion network for detecting salient objects in RGB-D images. IEEE Signal Processing Letters, 2021, 28: 1525-1529

[97] Zhou X, Wen H, Shi R, et al. FANet: Feature aggregation network for RGB-D saliency detection. Signal Processing: Image Communication, 2022, 102: 116591

[98] Xiao F, Li B, Peng Y, et al. Multi-modal weights sharing and hierarchical feature fusion for RGB-D salient object detection. IEEE Access, 2020, 8: 26602-26611

[99] Bi H, Wu R, Liu Z, et al. Cross-modal hierarchical interaction network for RGB-D salient object detection. Pattern Recognition, 2023, 136: 109194

[100] Fan D, Zhai Y, Borji A, et al. BBS-Net: RGB Dsalient object detection with a bifurcated backbone strategy network// Proceedings of the European Conference on Computer Vision. Online, 2020: 275-292

[101] Chen H, LiY, SuD. 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

[102] Liu Z, Wang Y, Tu Z, et al. TriTransNet: RGBD salient object detection with a triplet transformer embedding network// Proceedings of the ACM International Conference on Multimedia.Chengdu, China, 2021: 4481-4490

[103] Fang X, Jiang M, Zhu J, et al. M2RNet: Multi-modal and multi-scale refined network for RGB-D salient object detection. Pattern Recognition, 2023, 135: 109139

[104] Wang F, Pan J, Xu S, et al. Learning discriminative cross-modality features for RGB-Dsaliency detection. IEEE Transactions on Image Processing, 2022, 31: 1285-1297

[105] Liu Z, Wang K, Dong H, et al. A cross-modal edge-guided salient object detection for RGBD image. Neurocomputing, 2021, 454: 168-177

[106] Wei L, Zong G. EGA-Net: Edge feature enhancement and global information attention network for RGB-D salient object detection. Information Sciences, 2023, 626: 223-248

[107] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017

[108] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4510-4520

[109] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea,

2019：1314-1324

[110] Zhang X，Zhou X，Lin M，et al. Shufflenet：An extremely efficient convolutional neural network for mobile devices//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA，2018：6848-6856

[111] Ma N，Zhang X，Zheng H T，et al. ShufflenetV2：Practical guidelines for efficient cnn architecture design//Proceedings of the European Conference on Computer Vision. Munich，Germany，2018：116-131

[112] Chen L，Zhu Y，Papandreou G，et al. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the European Conference on Computer Vision. Munich，Germany，2018：801-818

[113] Zhao X，Zhang L，Pang Y，et al. A single stream network for robust and real-time RGB-D salient object detection//Proceedings of the European Conference on Computer Vision. Online，2020：646-662

[114] Ji W，Li J，Zhang M，et al. Accurate RGB-Dsalient object detection via collaborative learning//Proceedings of the European Conference on Computer Vision. Online，2020：52-69

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

[116] Ren G，Yu Y，Liu H，et al. Dynamic knowledge distillation with noise elimination for rgb-dsalient object detection. Sensors，2022，22(16)：6188

[117] Fu K，Fan D P，Ji G P，et al. Siamese network for RGB-D salient object detection and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，44(9)：5541-5559

[118] Wu Y H，Liu Y，Xu J，et al. MobileSal：Extremely efficient RGB-D salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，44(12)：10261-10269

[119] Zhang M，Fei S X，Liu J，et al. Asymmetric two-stream architecture for accurate RGB-D saliency detection//Proceedings of the European Conference on Computer Vision. Online，2020：374-390

[120] Wang J，Chen S，Lv X，et al. Guided residual network for RGB-D salient object detection with efficient depth feature learning. The Visual Computer，2022，38：1803-1814

[121] Chen S，Fu Y. Progressively guided alternate refinement network for RGB-D salient object detection//Proceedings of the European Conference on Computer Vision. Online，2020：520-538

[122] Jin X，Yi K，Xu J. MoADNet：Mobile asymmetric dual-stream networks for real-time and lightweight RGB-D salient object detection. IEEE Transactions on Circuits and Systems for Video Technology，2022，32(11)：7632-7645

[123] Huang N，Jiao Q，Zhang Q，et al. Middle-level feature fusion for lightweight RGB-D salient object detection. IEEE Transactions on Image Processing，2022，31：6621-6634

[124] Niu Y，Geng Y，Li X，et al. Leveraging stereopsis for saliency analysis//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence，USA，2012：454-461

[125] Ju R，Ge L，Geng W，et al. Depth saliency based on anisotropic center-surround difference//Proceedings of the IEEE International Conference on Image Processing. France，2014：1115-1119

[126] Liu N，Zhang N，Shao L，et al. Learning selective mutual attention and contrast for RGB-D saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，44(12)：9026-9042

[127] Ciptadi A，Hermans T，Rehg J M. An indepth view of saliency//Proceedings of the British Machine Vision Conference. Rome，Italy，2013：9-13

[128] Li N，Ye J，Ji Y，et al. Saliency detection on light field//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus，USA，2014：2806-2813

[129] Cheng Y，Fu H，Wei X，et al. Depth enhanced saliency detection method//Proceedings of International Conference on Internet Multimedia Computing and Service. Xiamen，China，2014：23-27

[130] Zhu C，Li G. A three-pathway psychobiological framework of salient object detection using stereoscopic technology//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice，Italy，2017：3008-3014

[131] Borji A，Cheng M，Jiang H，et al. Salient object detection：A benchmark. IEEE Transactions on Image Processing，2015，24(12)：5706-5722

[132] Radhakrishna A，Sheila S H，Francisco J E，et al. Frequency-tuned salient region detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami，USA，2009：1597-1604

[133] Fan D，Cheng M，Liu Y，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

[134] Fan D，Cheng G，Yang C，et al. Enhanced-alignment measure for binary foreground map evaluation//International Joint Conference on Artificial Intelligence. Stockholm，Sweden，2018：698-704

[135] Zhang Y，Gong K，Zhang K，et al. Meta-transformer：A unified framework for multimodal learning. arXiv preprintarXiv：2307.10802，2023

[136] De Boer P T，Kroese D P，Mannor S，et al. Atutorial on the cross-entropy method. Annals of Operations Research，2005，134(1)：19-67

[137] Mattyus G，Luo W，Urtasun R. Deep road mapper：extracting road topology from aerial images//Proceedings of the IEEE International Conference on Computer Vision. Venice，Italy，2017：3438-3446

[138] Xiao F，Pu Z，Chen J，et al. DGFNet：Depth-guided cross-modality fusion network for RGB-D salient object detection. IEEE Transactions on Multimedia，2024，26：2648-2658

[139] Zhao Y，Zhao J，Li J，et al. RGB-D salient object detection with ubiquitous target awareness. IEEE Transactions on Image

Processing, 2021,30: 7717-7731

[140] Li A, Mao Y, Zhang J, et al. Mutual information regularization for weakly-supervised RGB-Dsalient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(1):397-410

[141] Xu Y, Yu X, Zhang J, et al. Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting. IEEE Transactions on Image Processing, 2022, 31: 2148-2161

[142] Liu Z, Hayat M, Yang H, et al. Deep hypersphere feature regularization for weakly supervised RGB-D salient object detection. IEEE Transactions on Image Processing, 2023, 32: 5423-5437

[143] Zhao X, Pang Y, Zhang L, et al. Self-supervised pretraining for RGB-D salient object detection//Proceedings of the AAAI Conference on Artificial Intelligence, 36: 3. Online, 2022: 3463-3471

[144] Luo Z, Liu N, Zhao W, et al. VSCode: General visual salient and camouflaged object detection with 2D Prompt Learning// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA, 2024: 17169-17180

[145] Huang N, Yang Y, Xi R, et al. Salient object detection from arbitrary modalities. IEEE Transactions on Image Processing, 2024,32(3):235-252

HUANG Nian-Chang, Ph. D., assistant researcher. His research interests include multi-modal image processing and deep learning.

YANG Yang, Ph. D. candidate. His current research interests include multi-modal image processing and deep learning.

ZHANG Qiang, Ph. D., professor, Ph. D. supervisor. His current research interests include computer vision and intelligent image processing.

HAN Jun-Gong, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision, artificial intelligence and machine learning.

Background

Salient object detection imitates the human visual system to identify the most visually appealing objects or regions in an image. As an important prepossessing step, it has been widely applied in many computer vision tasks. Recently, with the rapid development of deep learning, deep learning based RGB SOD algorithms have received widespread attention and achieved significant advancements. Towards complex scenarios, depth information has been introduced into SOD, achieving encouraging results. For researchers and engineers in related fields to better understand the research progress of RGB-D SOD task, this paper conducts a comprehensive investigation from the key issues and the critical technologies of RGB-D SOD.

Existing reviews related to RGB-D SOD mainly categorize those deep learning based RGB-D SOD algorithms according to the model structures or the task objectives. Such classification approaches enable readers to gain a good understanding of the network architecture of existing models. However, they fail to provide a comprehensive overview for readers to analyze those key issues addressed by existing methods in RGB-D SOD together with their corresponding solutions. Differently, this paper first analyzes and summarizes some major issues in RGB-D SOD, accordingly, organizing and analyzing the recent deep learning based RGB-D SOD algorithms. Based on the analysis of existing methods, some challenges and future research trends are described. This paper aims to review the latest methods for researchers from the perspective of key issues and further promote the development of RGB-D SOD technologies. We hope that this paper inspires subsequent research works.

Our research group has been devoted to the research of multi-modal image processing and computer vision, including multi-sensor image fusion, unimodal/multi-modal image salient object detection, multi-modal image semantic segmentation and so on. We have published more than 50 papers in the international journals and CCF A conferences. Especially, we have published more than 10 papers that focus on salient object detection.

This work was supported by the China Postdoctoral Science Foundation under Grant No. 2023M742745, the Postdoctoral Fellowship Program of CPSF under Grant No. GZB20230559 and Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515110165. It was also supported by the National Natural Science Foundation of China under Grant No. 61773301, the Shaanxi Innovation Team Project under Grant No. 2018TD-012 and the State Key Laboratory of Reliability and Intelligence of Electrical Equipment under Grant No. EERI_KF2022005, He-bei University of Technology.