基于注意机制和多尺度跨模态融合的 RGB-D 显著性检测

崔志强[1)], 冯正勇[1)], 王峰[2)*], 刘强[1)]

[1)] (西华师范大学电子信息工程学院 南充 637000)
[2)] (渭南师范学院物理与电气工程学院 渭南 714099)
(wangfeng81113@163.com)

摘 要: 针对基于深度卷积神经网络的 RGB-D 显著性检测性能差等问题, 提出利用注意机制和多尺度跨模态融合进行 RGB-D 显著性检测的方法. 首先采用多尺度残差注意模块对骨干网络提取的特征进行预处理; 然后提出多尺度跨模态融合策略, 对高层 RGB 特征和深度特征进行融合, 获得初始显著图; 最后采用边界细化模块细化初始显著图中目标的边界, 使最终显著图包含敏锐的边界和完整的突出目标. 在 5 个基准数据集上与 10 种先进方法进行实验的结果表明, 所提方法在 4 个评价指标上均处于前 3 名; 尤其是在 NJUD 和 SIP 数据集上, 该方法在 4 个指标上提升了 0.5%~1.5%.

关键词: 图像处理; 显著性检测; 注意机制; 多尺度跨模态融合
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19479

RGB-D Saliency Detection Based on Attention Mechanism and Multi-Scale Cross-Modal Fusion

Cui Zhiqiang[1)], Feng Zhengyong[1)], Wang Feng[2)*], and Liu Qiang[1)]

[1)] (*College of Electronic and Information Engineering, China West Normal University, Nanchong* 637000)
[2)] (*College of Physics and Electrical Engineering, Weinan Normal University, Weinan* 714099)

Abstract: Aiming at the poor performance of RGB-D saliency detection based on deep convolution neural network, a method of RGB-D saliency detection using attention mechanism and multi-scale cross-modal fusion is proposed. Firstly, the multi-scale residual attention module is used to preprocess the features extracted from the backbone network. Secondly, a multi-scale cross-modal fusion strategy is proposed to fuse the high-level RGB and depth features to obtain the initial saliency map. Finally, the boundary refinement module is utilized to refine the boundary of the object in the initial saliency map, so that the final saliency map contains sharp boundaries and complete salient objects. The results of experiments with ten advanced methods on five benchmark datasets show that the proposed method ranks in the top three in four evaluation metrics, especially on the NJUD and SIP datasets, where the method improves by 0.5%—1.5% in four metrics.

Key words: image processing; saliency detection; attention mechanism; multi-scale cross-modal fusion

显著性检测通过模拟人类视觉感知系统自主地检测出给定图像上最吸引人的目标, 已在图像分割[1]、视觉追踪[2]、目标识别[3]以及图像检索[4]等各种计算机视觉任务中得到了广泛应用.

收稿日期: 2021-11-19; 修回日期: 2022-04-24. **基金项目**: 渭南市科技局重点项目(ZDYF-JCYJ-196); 渭南师范学院人才项目(2020RC11); 西华师范大学英才项目(17YC046). **崔志强**(1997—), 男, 硕士研究生, 主要研究方向为计算机视觉、显著性检测; **冯正勇**(1978—), 男, 博士, 教授, 硕士生导师, 主要研究方向为深度学习、人工智能; **王峰**(1981—), 男, 博士, 讲师, 论文通信作者, 主要研究方向为深度学习、图像融合; **刘强**(1995—), 男, 硕士研究生, 主要研究方向为计算机视觉、变化检测.

传统的 RGB-D 显著性检测通过探索 RGB 和深度线索中的有用信息来设计手工特征，由于手工特征表达能力较低，检测效果往往不佳. 随着深度学习的快速发展，相对于传统的检测方法，基于卷积神经网络(convolutional neural network, CNN)的 RGB-D 显著性检测取得了较大的进步. Liu 等[5]提出一种自交互的注意力模块以融合 RGB 和深度图，并且集成自注意力和交互注意力机制准确地传播上下文信息; Zhang 等[6]提出一个互补交互模块，探索模态间的互补性和融合跨模态特征; Piao 等[7]设计深度感应多尺度加权模块和循环注意力模块探索 RGB 特征和深度特征之间的互补性，提升显著性检测的性能. 然而，当处理具有挑战的场景(如背景杂乱、前景背景相似等)时，上述方法不能完整地检测图像中的突出目标，造成此问题的原因如下: (1) 未对骨干网络提取的特征信息进行处理而直接进行跨模态融合，以及个别低质量深度图的负面影响; (2) RGB 图像和深度图像未得到充分的互补融合，致使遗漏一些重要特征信息.

近年来，由于注意力机制能够筛选出重要的特征区域，大多数方法将其引入 RGB-D 显著性检测中. Zhou 等[8]引入注意力图区分显著物体和背景区域，以减少低质量深度图的负面影响; Li 等[9]通过深度监督注意力机制对显著区域进行加权; 凌艳等[10]提出一个多尺度上下文增强模块处理骨干网络提取的特征信息. 不同于以上方法，本文在 Liu 等[11]提出的感受野块(receptive field block, RFB)的基础上加入空间注意(spatial attention, SA)模块和通道注意(channel attention, CA)模块，设计一个多尺度残差注意模块(multi-scale residual attention module, MSRA); 将此模块应用到骨干网络提取的分层特征上，分别从 RGB 图像和深度图像中提取丰富且重要的特征信息，也可以缓解低质量深度图的负面影响.

通常，显著性检测中 RGB 特征和深度特征的融合方法划分为 3 种: (1) 早期融合. Liu 等[12]直接把 RGB 图像和深度图像进行连接操作形成四通道输入，然后将其输入到网络中实现显著图的预测; (2) 后期融合. Wang 等[13]使用 2 个并行的网络结构分别学习 RGB 和深度图的特征信息，然后将其级联起来生成最终的显著图; (3) 中期融合，也称多尺度融合. 陈正等[14]通过点乘和相加操作融合 2 个模态的特征信息; Li 等[15]设计了一个跨模态深度加权结合块，利用深度特征来增强 RGB 特征; Chen 等[16]提出一个密集的跨层反馈拓扑结构，其中每层进行跨模态交互，然后密集地反馈到浅层; Liu 等[17]提出一种跨模态自适应门控融合生成对抗网络来融合 2 个模态的特征信息; Li 等[18]提出一个跨模态加权策略，通过 3 个 RGB-D 交互模块分别进行低层、中层和高层的跨模态融合. 不同于上述方法，本文提出一个多尺度跨模态融合策略. 首先，把经过 MSRA 获取的高层特征送到 2 种类型的密集融合模块(dense fusion module, DFM)，充分地融合 2 个模态间的信息，产生一个初始显著图; 然后提出边界细化模块(boundary refinement module, BRM)挖掘低层中的边界细节信息，细化初始显著图中目标的边界，进而得出包含敏锐边界和完整目标的最终显著图.

1 本文方法

1.1 整体结构

本文方法的整体结构如图 1 所示，包括特征提取模块、MSRA、相同尺度密集融合模块(same scale dense fusion module, S-DFM)、不同尺度密集融合模块(different scale dense fusion module, D-DFM)和 BRM 共 5 个模块. 其中，后 3 个模块构成本文的多尺度跨模态融合策略. 首先采用 ResNet50[19]骨干网络分别从输入的 RGB 和深度图像中提取 5 层特征图 $F_a^b \in \mathbb{R}^{H\times W\times C}$. 其中，$a\in\{r,d\}$，$b\in\{1,2,3,4,5\}$，$H,W$ 和 C 分别表示特征图的高度、宽度和通道维度数; 然后对提取的 5 层特征图 $F_a^b \in \mathbb{R}^{H\times W\times C}$，分别应用 1×1 的卷积层得到降维后的特征图 $F_a^b \in \mathbb{R}^{H\times W\times 64}$; 再将降维后的特征图输入到 MSRA 中，获取丰富且重要的 RGB 特征 $\{R_1,R_2,R_3,R_4,R_5\}$ 和深度特征 $\{D_1,D_2,D_3,D_4,D_5\}$; 最后将高层的 RGB 特征 $\{R_3,R_4,R_5\}$ 和深度特征 $\{D_3,D_4,D_5\}$ 送入 S-DFM 和 D-DFM 进行密集的融合，产生初始显著图 S_{ini}，使用低层的 RGB 特征 $\{R_1,R_2\}$ 和深度特征 $\{D_1,D_2\}$ 细化初始显著图中目标的边界，产生最终的显著图 S_{fin}.

1.2 MSRA

现实生活中，给定图像中的突出目标尺度是多变的，因此需要获取不同尺度大小的特征信息，增加显著性检测的可靠性. 虽然主干网络有足够大的理论感受野覆盖图像中大部分的突出目标，但是实验已经证明，有效的感受野比理论感受野要小. 经过 Liu 等[11]提出的 RFB 后，虽然可以捕获丰富的全局上下文信息，也解决了显著性检测的

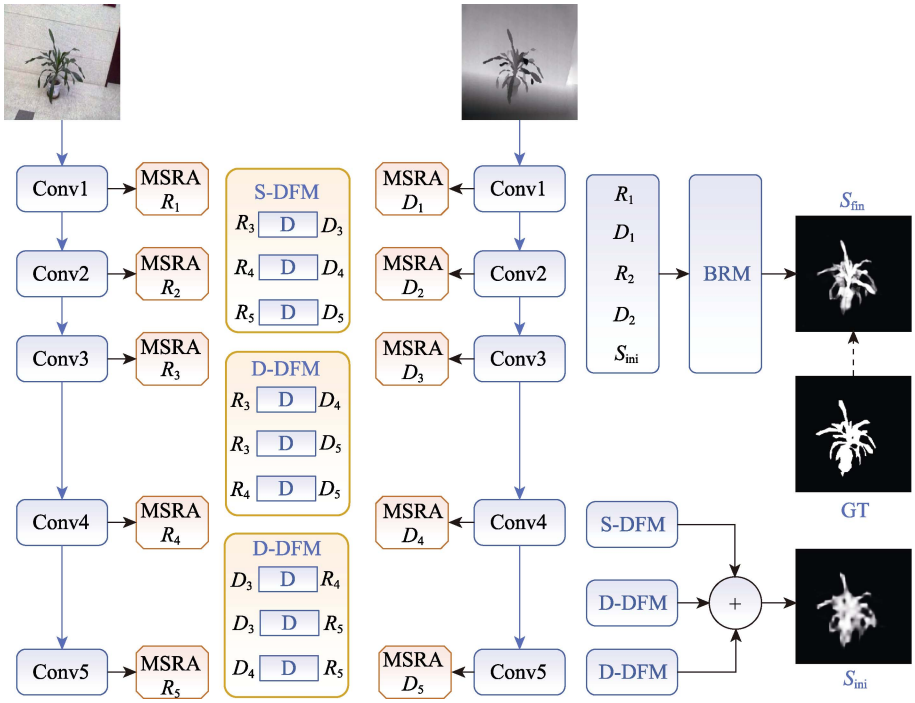


图 1　本文方法整体结构

尺度问题，但是所提取的特征信息量大且没有信息的着重点. 考虑 CA 机制和 SA 机制可以有效地筛选出通道和空间上哪一部分区域更值得关注，本文巧妙地把 RFB 与 SA 机制和 CA 机制结合起来形成 MSRA，在通过 RFB 覆盖更大范围区域捕获丰富特征信息的同时，也通过注意机制从丰富的特征信息中筛选出重要的特征信息.

MSRA 结构如图 2 所示，由 5 个并行分支构成：第 1 个分支由一个 1×1 和一个 3×3 卷积层构成，在无膨胀率的情况下对输入特征图提取特征信息；中间 3 个分支由上至下膨胀率依此为 3, 5 和 7，并且这 3 个分支添加 CA 模块和 SA 模块，旨在获取丰富特征信息的同时更加关注重要的信息；最后一个分支由一个 1×1 卷积层构成，用于降低通道数. 最终将前 4 个分支获得的特征图进行拼接操作，再与包含输入特征原始信息的第 5 个分支相加形成残差结构.

1.3　多尺度跨模态融合

RGB 和深度图像的融合方法大致分为早期融合、中期融合和后期融合. 本文提出一种多尺度跨

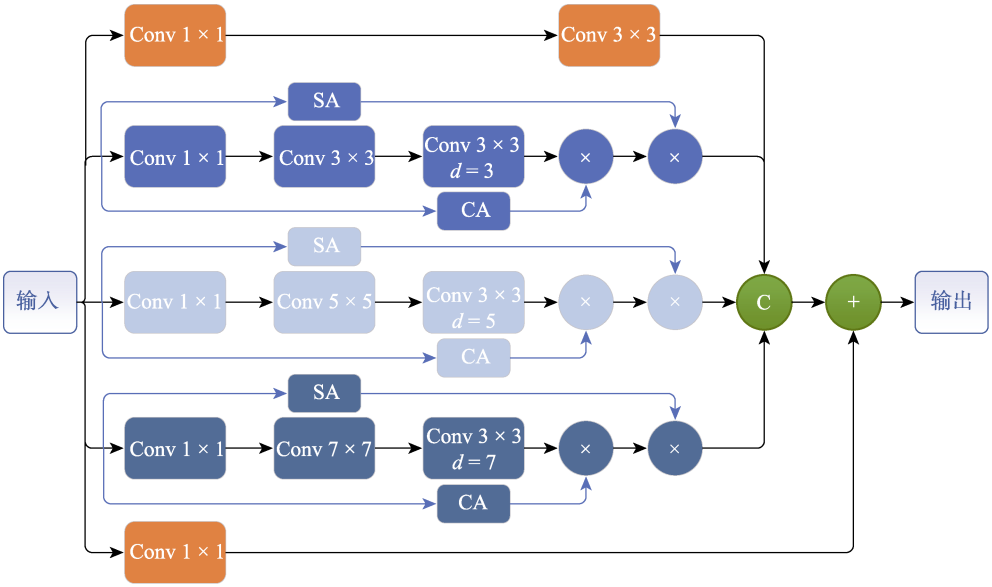


图 2　MSRA 结构

模态融合策略，其充分地探索模态间的互补性，有效地融合 2 个模态提取的特征，分为 2 步：首先对高层特征跨模态密集融合产生初始显著图；其次使用包含丰富边界细节信息的低层特征细化初始显著图，使得最终的显著图包含敏锐的边界和完整的突出目标.

(1) 密集融合. 由于高层特征包含丰富的语义信息有助于定位突出目标，因此有效地融合 2 个模态的高层特征是一个至关重要的问题. 从相同尺度和不同尺度的跨模态密集融合出发，本文设计 S-DFM 和 D-DFM 这 2 种类型的 DFM 充分地挖掘模态间像素层的互补信息，同时通过 2 种类型的密集融合捕获更全面的突出目标信息.

S-DFM 如图 3a 所示. 对经过 MSRA 处理的高 3 层 RGB 特征 $\{R_3,R_4,R_5\}$ 和深度特征 $\{D_3,D_4,D_5\}$，首先通过 3 个经典的密集块进行相同尺度的跨模态融合，产生 3 个具备位置信息和空间结构信息的强有力的混合特征图；然后通过上采样、卷积和连接操作整合 3 个特征图，产生一个特征图. 第 i 个 RGB 特征和第 j 个深度特征融合产生的特征图为

$$F_{ij}=D\left(R_i,D_j\right) \tag{1}$$

其中，$i=j=\{3,4,5\}$，R_i 和 D_j 分别表示第 i 个 RGB 特征和第 j 个深度特征，$D(\cdot)$ 表示一个密集融合操作. 通过式(1)可以得到 3 个包含复杂信息的特征图 $\{F_{33},F_{44},F_{55}\}$，进行下述操作

$$\begin{cases}F'_{55}=\mathrm{Conv}\left[U_4\left(F_{55}\right)\right]\\ F'_{44}=\mathrm{Conv}\left\{U_2\left\{\mathrm{Conv}\left[U_2\left(F_{55}\right)\right]\oplus F_{44}\right\}\right\}\\ F'_{33}=F'_{55}\oplus F'_{44}\oplus F_{33}\\ F_{\mathrm{fuse1}}=\mathrm{Conv}\left[C\left(F'_{33},F'_{44},F'_{55}\right)\right]\end{cases}.$$

其中，$U_2(\cdot)$ 和 $U_4(\cdot)$ 分别表示 2 倍和 4 倍上采样操作；$\mathrm{Conv}(\cdot)$ 表示一个 3×3 的卷积操作；$\oplus$ 表示元素相加操作；$C(\cdot)$ 表示连接操作；F'_{ij} 表示特征 F_{ij} 跨尺度融合后的增强特征；F_{fuse1} 表示相同尺度跨模态融合得到的特征图.

考虑 RGB 图像包含丰富的位置细节信息，而深度图像包含丰富的空间结构信息，仅仅通过 3 个相同尺度的跨模态融合无法充分地挖掘它们之间的互补性，也会丢失一些重要的特征信息. 因此，本文使用 2 个如图 3b 所示的 D-DFM 进行不同尺度的跨模态融合. 经过 MSRA 处理得到高层 RGB 特征和深度特征后，分别将它们送进密集块中产生具有复杂信息的特征图.

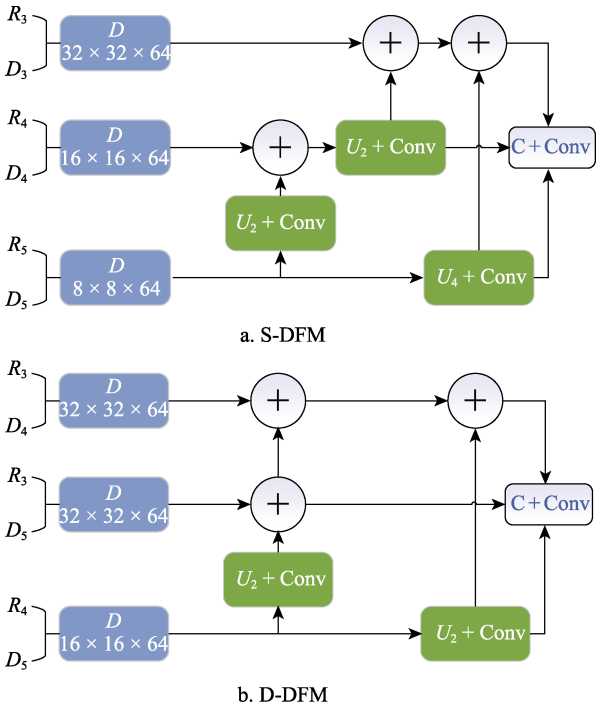


图 3 2 种类型的 DFM

首先，如同式(1)那样进行密集融合操作，但此处 $i=\{3,4,5\}$，$j=\{3,4,5\}$ 且 $i\neq j$，因此将得到 6 个特征图 $\{F_{34},F_{35},F_{45},F_{43},F_{53},F_{54}\}$；然后，分别对前 3 个以及后 3 个特征图通过一系列卷积、上采样、相加和连接操作进行整合. 前 3 个特征图的详细整合过程为

$$\begin{cases}F'_{45}=\mathrm{Conv}\left[U_2\left(F_{45}\right)\right]\\ F'_{35}=F_{35}\oplus F'_{45}\\ F'_{34}=F_{34}\oplus F'_{35}\oplus F'_{45}\\ F_{\mathrm{fuse2}}=\mathrm{Conv}\left[C\left(F'_{34},F'_{35},F'_{45}\right)\right]\end{cases}.$$

其中，F_{fuse2} 表示不同尺度跨模态融合的特征图. 类似地，后 3 个特征图通过上述操作也可以得到一个不同尺度跨模态融合的特征图 F_{fuse3}. 最终的初始显著图为 $S_{\mathrm{ini}}=F_{\mathrm{fuse1}}\oplus F_{\mathrm{fuse2}}\oplus F_{\mathrm{fuse3}}$. 其中，$S_{\mathrm{ini}}$ 表示初始显著图.

(2) 边界细化. 由于低层特征包含丰富的边界细节信息，有助于细化突出目标边界，因此本文使用低层特征细化初始显著图，使最终显著图具备敏锐的边界细节和完整的突出目标. 如图 4 所示，边界细化部分使用经 MSRA 处理后低层的 RGB 特征和深度特征，公式为

$$\begin{cases} F_1 = U_2\left(R_1 \oplus D_1\right) \oplus \left[\left(R_1 \oplus D_1\right) \otimes S_{\text{ini}}\right] \\ F_2 = U_4\left(R_2 \oplus D_2\right) \oplus \left[\left(R_2 \oplus D_2\right) \otimes S_{\text{ini}}\right] \\ S_{\text{fin}} = \text{Conv}\left[C\left(F_1, F_2\right)\right] \end{cases}.$$

其中，$\otimes$表示元素相乘操作；F_1和F_2表示细化后的特征图；S_{fin}表示最终的显著图.

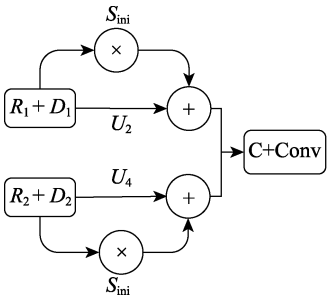


图 4　BRM 结构

1.4　损失函数

交叉熵可以表示真实值与预测值之间的差异，其值越小，表明模型预测效果越好. 本文使用二值交叉熵损失评估模型，公式为 $L = L_{\text{bce}}\left(S_{\text{fin}}, \text{GT}\right)$. 其中，$L$表示模型的总损失；$L_{\text{bce}}$表示二值交叉熵损失函数；$S_{\text{fin}}$表示最终的预测显著图; GT 表示真值(ground truth)显著图.

2　实验及结果分析

2.1　数据集

本文在5个具有挑战的数据集上进行实验，包括 NJUD[20], NLPR[21], LFSD[22], SSD[23]和 SIP[24]. 其中, NJUD 由 1985 幅立体图像对组成，这些图像均来自 Internet, 3D 电影和 FujiW3 立体相机拍摄的照片; NLPR 包含 1000 幅 RGB 图像及其对应的深度图，这些图像都是通过标准的 Microsoft Kinect 收集得到的; LFSD 由 Lytro 光场相机收集的 100 幅光场图像组成，包含 60 张室内场景和 40 张室外场景; SSD 通过 3 部立体电影制作而成，包含室内和室外场景共 80 个样本; SIP 由 929 幅带标注的高分辨率图像组成，每幅图像都包含多个显著人物. 为了公平对比，训练测试集划分按照文献[25-26]相同的设置，训练集包含 NJUD 中 1 485 个样本和 NLPR 数据集中 700 个样本；测试集包含 NJUD 和 NLPR 数据集中的剩余图像，以及整个 LFSD, SSD 和 SIP 数据集.

2.2　评价指标

本文采用 5 个广泛使用的指标评价提出的方法，包括 *S*-measure(S_α)，*E*-measure(E_ξ)，*F*-measure (F_β)、平均绝对值误差(mean absolute error, MAE)和准确率-召回率(precision-recall, P-R)曲线.

S_α用于评估区域感知S_{r}和目标感知S_{o}之间的结构相似性，定义为

$$S_\alpha = \alpha \times S_{\text{o}} + \left(1-\alpha\right) \times S_{\text{r}}.$$

其中，$\alpha \in [0,1]$表示平衡参数，默认值设为 0.5.

E_ξ用于获取图像水平统计信息及其局部像素匹配信息，定义为 $E_\xi = \dfrac{1}{W \times H}\sum_{I=1}^{W}\sum_{J=1}^{H}\boldsymbol{\xi}\left(i,j\right)$. 其中，$W$和$H$分别表示显著图的宽和高；$\boldsymbol{\xi}(\cdot)$表示增强对角矩阵.

为全面考虑精度和召回率，加权的调和平均值定义为 $F_\beta = \left(1+\beta^2\right)\dfrac{P \times R}{\beta^2 P + R}$. 其中，设置 β^2 =0.3 以增强精度. 本文使用不同的阈值$[0,255]$计算 F_β.

2.3　实现细节

本文方法采用 Pytorch 框架，在一块 3080GPU 上做实验；主干网络采用 ResNet50[19]，去掉最后的池化层和全连接层，并且采用 ImageNet 上的预训练模型进行初始化. 为了防止过拟合，本文对所有训练图像实施随机翻转、旋转和边界裁剪；输入图像的分辨率统一调整为 256×256 像素；采用 Adam 优化模型，初始学习率设为 0.000 1，每隔 60 轮下降 10 倍；实验中批次大小设为 8，训练模型 200 轮大约需 6 h.

2.4　结果对比与分析

本文方法在 5 个广泛使用的公开基准数据集上与 10 种方法进行比较，包括 TANet[27], CPFP[28], DMRA[7], D3Net[26], ICNet[15], CMMS[29], CoNet[30], SSF[6], S2MA[5], ASIFNet[9]. 为了公平对比，直接使用由作者提供的显著图或预计算的评价结果.

(1) 定量评估. 表 1 和图 5 所示为在 5 个数据集和 5 个评价指标下，本文方法与 10 种方法的定量比较结果. 表 1 中，结果最好的用粗体标注(下文同)，可以明显地观察到，在 NJUD, NLPR 和 SIP 数据集上，本文方法优于所有对比方法. 从图 5 可以看出，在 NJUD, NLPR 和 SIP 数据集上，本文方法(红线所示)在不同的阈值下优于所有对比方法.

(2) 定性评估. 图 6 所示为本文方法与 10 种方法的视觉对比结果. 为了证实本文方法的鲁棒性，选取 7 幅典型的场景图像进行对比，其中，前 5 行图像来自 NJUD 数据集，后 2 行图像来自 NLPR 数据集. 可以看出，本文方法检测的显著图在不同的

表 1　在 5 个基准数据集上的定量评估对比

数据集	评价指标	方法										
		TANet[27]	CPFP[28]	DMRA[7]	D3Net[26]	ICNet[15]	CMMS[29]	CoNet[30]	SSF[6]	S2MA[5]	ASIFNet[9]	本文
NJUD	S	0.878	0.883	0.886	0.900	0.894	0.904	0.895	0.899	0.894	0.889	**0.915**
	$F_{\max}$	0.874	0.881	0.886	0.900	0.891	0.897	0.892	0.896	0.889	0.888	**0.909**
	$E_{\max}$	0.925	0.930	0.927	0.938	0.924	0.936	0.937	0.934	0.930	0.927	**0.947**
	MAE	0.060	0.051	0.051	0.047	0.052	0.044	0.047	0.043	0.054	0.047	**0.039**
NLPR	S	0.886	0.889	0.899	0.912	0.923	0.919	0.908	0.914	0.915	0.907	**0.925**
	$F_{\max}$	0.863	0.868	0.880	0.897	0.908	0.904	0.887	0.896	0.902	0.888	**0.911**
	$E_{\max}$	0.941	0.932	0.947	0.953	0.952	0.955	0.945	0.954	0.953	0.949	**0.959**
	MAE	0.041	0.036	0.031	0.030	0.029	0.028	0.031	**0.026**	0.030	0.030	0.027
LFSD	S	0.801	0.828	0.847	0.825	**0.868**	0.845	0.862	0.859	0.837	0.822	0.867
	$F_{\max}$	0.796	0.825	0.856	0.810	**0.871**	0.858	0.859	0.861	0.835	0.824	0.863
	$E_{\max}$	0.847	0.872	0.901	0.862	0.903	0.886	0.907	0.900	0.873	0.861	**0.909**
	MAE	0.111	0.088	0.076	0.095	0.071	0.082	0.071	0.067	0.095	0.089	0.075
SIP	S	0.835	0.850	0.806	0.860	0.854		0.858				**0.876**
	$F_{\max}$	0.830	0.851	0.821	0.861	0.857		0.867				**0.881**
	$E_{\max}$	0.895	0.903	0.874	0.909	0.903		0.913				**0.921**
	MAE	0.075	0.064	0.086	0.063	0.069		0.063				**0.058**
SSD	S	0.839	0.807	0.857	0.857	0.848	0.858	0.853		**0.868**	0.857	**0.868**
	$F_{\max}$	0.810	0.766	0.844	0.834	0.841	0.839	0.840		**0.848**	0.834	0.846
	$E_{\max}$	0.897	0.852	0.906	0.908	0.903	0.900	**0.915**		0.909	0.884	0.909
	MAE	0.063	0.082	0.059	0.058	0.064	0.053	0.060		0.052	0.056	**0.051**

注. 加粗数值表示最优值.

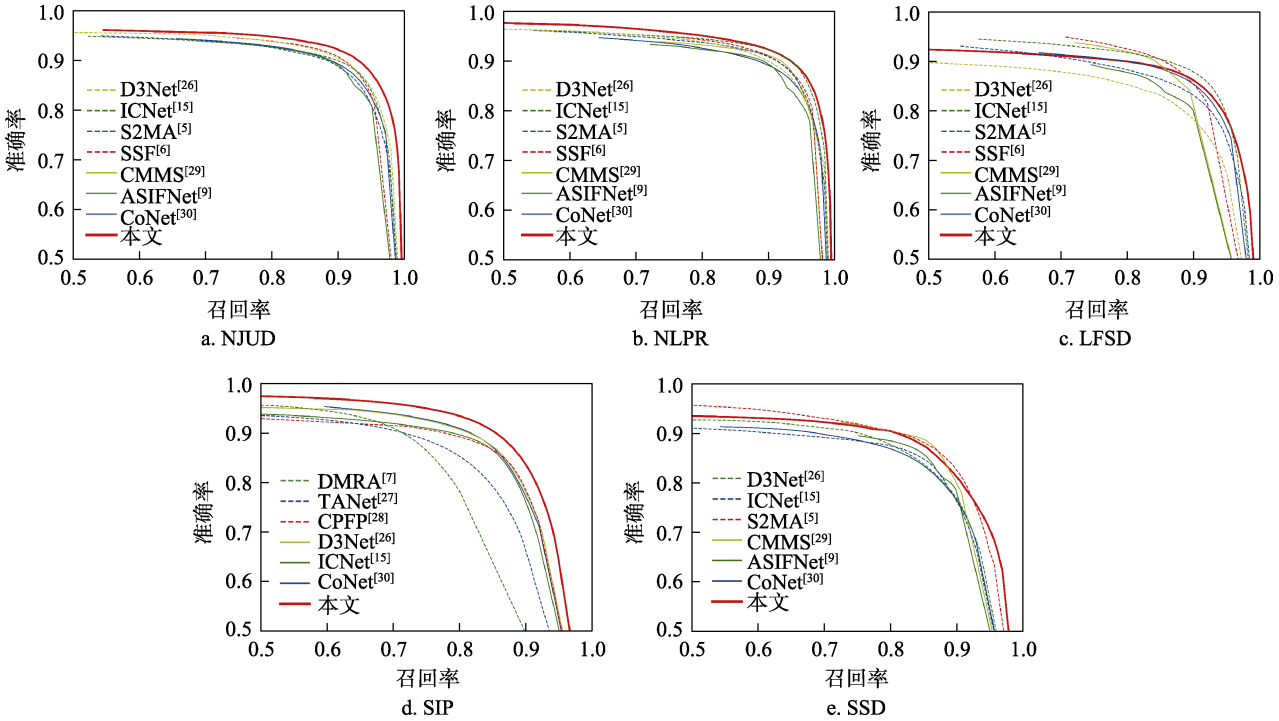


图 5　5 个数据集上 P-R 曲线对比

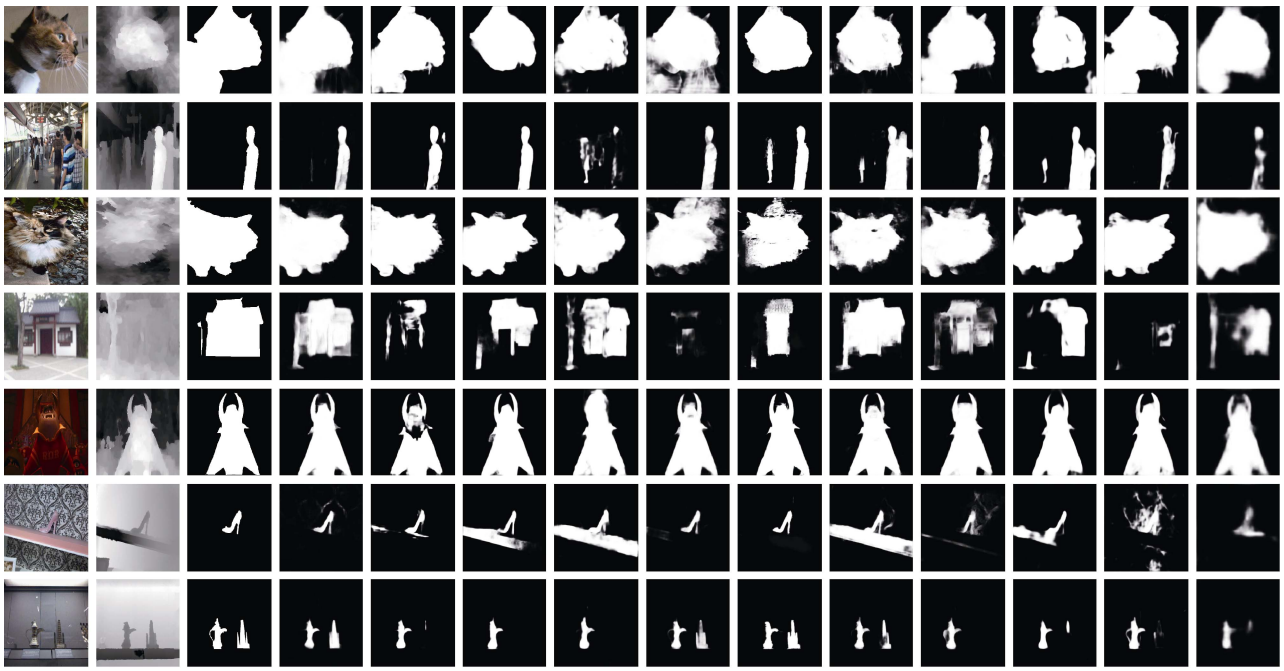


图 6　11 种方法在典型复杂场景上的视觉对比

场景下与真值图更相似. 在简单场景(第 1 行)下, RGB 图像中的突出目标猫是非常明显的, 但是深度图是低质量的; 与其他方法相比, 本文方法可以检测出更为完整的突出目标. 在复杂背景(第 2 行)下, 本文方法可以过滤背景信息的干扰, 产生令人满意的结果; 在前景背景相似(第 3 行)下, 本文方法可以关注到空间和通道维度上重要的特征区域, 比其他方法检测结果要好; 在低质量深度图(第 1 行和第 4 行)下, MSRA 可以在一定程度上缓解低质量深度图的消极影响, 因此与其他方法相比, 本文方法能检测出更完整的目标; 在低对比度场景(第 5 行)下, 本文方法通过密集融合探索模态间的互补信息, 产生完整的显著目标; 在小目标场景(第 6 行)下, 本文方法关注重要的特征区域, 过滤周围不重要的区域, 可以较好地检测出小目标; 在多目标场景(第 7 行)下, 与其他方法相比, 本文方法能检测出图像上所有的突出目标.

2.5　消融实验

本文在 LFSD 和 SIP 数据集上进行消融实验, 在验证不同模块有效性的同时, 也可以证实本文方法有较好的泛化能力.

(1) MSRA 的有效性. 本文进行不同策略的实验: a. w/o MSRA. 对骨干网络提取的不同尺度分层特征不经过 MSRA 预处理, 直接用于后续的跨模态融合; b. with RFB. 去掉 MSRA 模块中的 CA 和 SA, 使用 RFB 对骨干网络提取的分层特征进行处理, 然后执行后续操作. 从表 2 可以看出, 使用 MSRA 处理后, $S, E_{\max}, F_{\max}$ 和 MAE 4 个评价指标均有较大提升, 如在 LFSD 数据集上, 4 个指标分别提升 3%, 3.8%, 3.7%和 1.9%; 在 SIP 数据集上, 4 个指标分别提升 3.1%, 3.6%, 1.5%和 1.8%, 证实了 MSRA 的有效性. 图 7 所示为不同策略下的检测结果, 可以看出, 不经过 MSRA 处理的特征图中突出目标不完整且边界模糊, 去掉 CA 和 SA 后检测的图包含一些背景区域, 证实 MSRA 能够提取丰富且重要的特征信息, 检测的结果令人满意. 另外, 当处理低质量深度图时(第 3 行), MSRA 能够有效地缓解低质量深度图的负面影响, 而其他策略检测效果较差.

(2) 多尺度跨模态融合的有效性. 本文进行不同网络结构设计的实验: a. w/o S-DFM. 去掉多尺度跨模态融合策略中的 S-DFM 部分; b. w/o D-DFM. 去掉多尺度跨模态融合策略中的 D-DFM 部分; c. w/o BRM. 去掉多尺度跨模态融合策略中的 BRM 部分. 从表 2 可以看出, 丢掉多尺度跨模态融合策略中的任何一个模块都会降低模型的检测效果, 证实多尺度跨模态融合策略的有效性. 图 8 所示为多尺度跨模态融合策略不同设置的视觉对比结果, 可以看出, 在不同的场景下去掉 BRM 后, 检测结果图丢失一些细节信息, 并且突出目标边界模糊; 当去掉 S-DFM 部分时, 在不同场景下检测图中的突出目标信息不完整, 这是因为去掉相同尺度的

跨模态融合会丢失一些目标信息；当去掉 D-DFM 部分时，不同场景下检测的突出目标会多检或漏检一些重要特征信息，这是因为去掉不同尺度的跨模态融合不能捕获一些复杂重要的特征信息.

表 2　LFSD 和 SIP 数据集上的消融实验

方法	LFSD				SIP			
	S	F_{max}	E_{max}	MAE	S	F_{max}	E_{max}	MAE
本文	**0.867**	**0.863**	**0.909**	**0.075**	**0.876**	**0.881**	**0.921**	**0.058**
w/o MSRA	0.837	0.825	0.872	0.094	0.845	0.845	0.906	0.076
with RFB	0.852	0.851	0.894	0.089	0.852	0.854	0.908	0.073
w/o BRM	0.857	0.851	0.896	0.080	0.866	0.867	0.914	0.064
w/o S-DFM	0.859	0.857	0.895	0.080	0.870	0.874	0.917	0.061
w/o D-DFM	0.861	0.860	0.898	0.079	0.865	0.869	0.910	0.064

注. 加粗数值表示最优值.

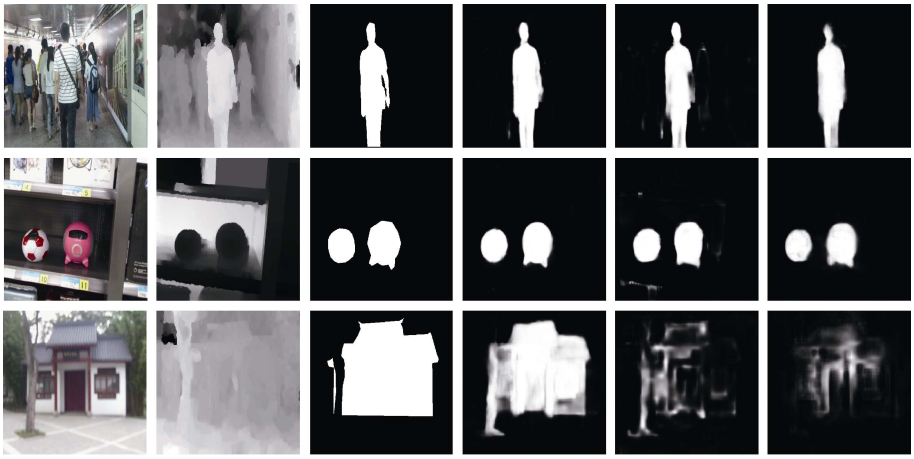


图 7　MSRA 不同策略的视觉对比

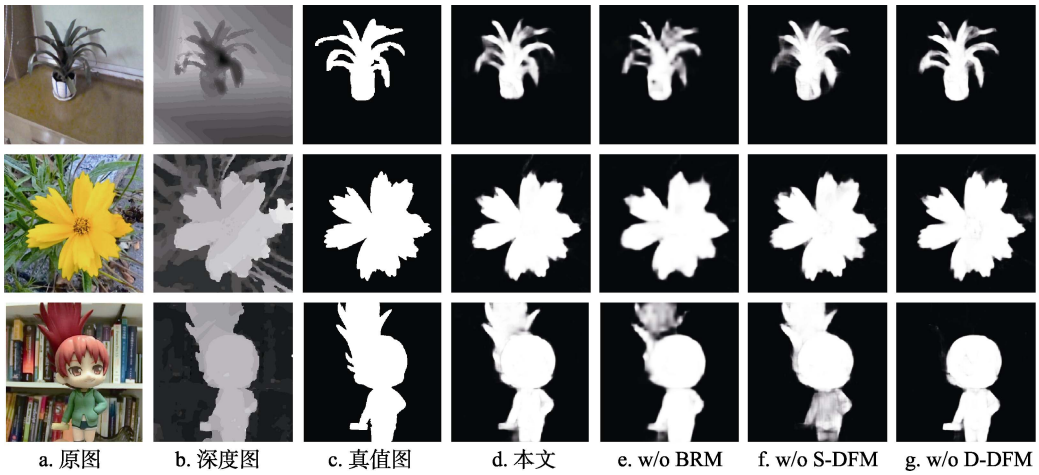
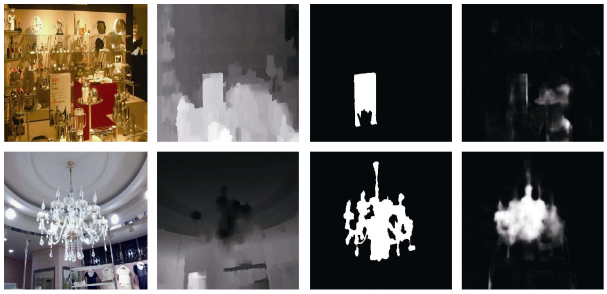


图 8　多尺度跨模态融合不同策略的视觉对比

2.6　失败案例

利用 MSRA 和多尺度跨模态融合策略，本文方法可以清晰地检测出大多图像中的显著区域，但面对如图 9 中更复杂的情况时，本文方法检测效果不佳. 从图 9 可以看出，第 1 行中目标多样且背景复杂，本文方法不能区分前景背景，检测效果较差；第 2 行中目标透明且复杂，本文方法不能明确地学习到细节信息，检测结果较模糊.



a. 原图　b. 深度图　c. 真值图　d. 本文

图 9　本文方法的失败案例

3　结　语

本文提出一种基于注意机制和多尺度跨模态融合的显著性检测方法，包含 MSRA，DFM 和 BRM 这 3 个重要模块. 其中，MSRA 在 RFB 的基础上加入 SA 和 CA，从骨干网络的输出特征中提取丰富且重要的特征信息；对于高层的特征，设计了 2 种类型的 DFM，一种用于相同尺度的融合，另一种用于不同尺度的融合，可以有效地探索模态间的互补信息，并充分地融合这些信息产生包含复杂信息的初始显著图；BRM 利用低层特征包含的边界细节信息细化高层特征产生的初始显著图，使产生的最终显著图包含完整的突出目标. 定量和定性的实验结果证实，本文方法在 5 个公开基准数据集上的实验结果表现优秀.

参考文献(References):

[1] Ye L W, Rochan M, Liu Z, *et al*. Cross-modal self-attention network for referring image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 10494-10503

[2] Jia X, Lu H C, Yang M H. Visual tracking via coarse and fine structural local sparse appearance models[J]. IEEE Transactions on Image Processing, 2016, 25(10): 4555-4564

[3] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[4] Gao Y, Wang M, Tao D C, *et al*. 3-D object retrieval and recognition with hypergraph analysis[J]. IEEE Transactions on Image Processing, 2012, 21(9): 4290-4303

[5] Liu N, Zhang N, Han J W. Learning selective self-mutual attention for RGB-D saliency detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020:13753-13762

[6] Zhang M, Ren W S, Piao Y R, *et al*. Select, supplement and focus for RGB-D saliency detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 3469-3478

[7] Piao Y R, Ji W, Li J J, *et al*. Depth-induced multi-scale recurrent attention network for saliency detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7253-7262

[8] Zhou X F, Li G Y, Gong C, *et al*. Attention-guided RGB-D saliency detection using appearance information[J]. Image and Vision Computing, 2020, 95: 103888

[9] Li C Y, Cong R M, Kwong S, *et al*. ASIF-net: attention steered interweave fusion network for RGB-D salient object detection[J]. IEEE Transactions on Cybernetics, 2021, 51(1): 88-100

[10] Ling Yan, Chen Ying. Salient object detection with multiscale context enhanced fully convolutional network[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(11): 2007-2016(in Chinese)
(凌艳，陈莹. 多尺度上下文信息增强的显著目标检测全卷积网络[J]. 计算机辅助设计与图形学学报，2019, 31(11): 2007-2016)

[11] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 404-419

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

[13] Wang N N, Gong X J. Adaptive fusion for RGB-D salient object detection[J]. IEEE Access, 2019, 7: 55277-55284

[14] Chen Zheng, Zhao Xiaoli, Zhang Jiaying, *et al*. RGB-D image saliency detection based on cross-model feature fusion[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(11): 1688-1697(in Chinese)
(陈正，赵晓丽，张佳颖，等. 基于跨模态特征融合的 RGB-D 显著性目标检测[J]. 计算机辅助设计与图形学学报, 2021, 33(11): 1688-1697)

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

[16] Chen H, Li Y F, Su D. Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection[J]. IEEE Transactions on Cybernetics, 2020, 50(11): 4808-4820

[17] Liu Z Y, Zhang W, Zhao P. A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection[J]. Neurocomputing, 2020, 387: 210-220

[18] Li G Y, Liu Z, Ye L W, *et al*. Cross-modal weighting network for RGB-D salient object detection[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 665-681

[19] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[20] Ju R, Ge L, Geng W J, *et al*. Depth saliency based on anisot-

ropic center-surround difference[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2014: 1115-1119

[21] Peng H W, Li B, Xiong W H, *et al*. RGB-D salient object detection: a benchmark and algorithms[C] //Proceedings of the 13th European Conference on Computer Vision. Heidelberg: Springer, 2014: 92-109

[22] Li N Y, Ye J W, Ji Y, *et al*. Saliency detection on light field[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 2806-2813

[23] Li G, Zhu C B. A three-pathway psychobiological framework of salient object detection using stereoscopic technology[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 3008-3014

[24] Fan D P, Cheng M M, Liu Y, *et al*. Structure-measure: a new way to evaluate foreground maps[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4558-4567

[25] Chen S H, Fu Y. Progressively guided alternate refinement network for RGB-D salient object detection[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 520-538

[26] Fan D P, Lin Z, Zhao Z Z, *et al*. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(5): 2075-2089

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

[28] Zhao J X, Cao Y, Fan D P, *et al*. Contrast prior and fluid pyramid integration for RGB-D salient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3922-3931

[29] Li C Y, Cong R M, Piao Y R, *et al*. RGB-D salient object detection with cross-modality modulation and selection[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 225-241

[30] Ji W, Li J J, Zhang M, *et al*. Accurate RGB-D salient object detection via collaborative learning[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 52-69