控制与决策
Control and Decision

基于非对称跨模态融合的RGB-D显著目标检测

于明, 邢章浩, 刘依

引用本文:
于明,邢章浩,刘依. 基于非对称跨模态融合的RGB-D显著目标检测[J]. 控制与决策, 2023, 38(9): 2487-2495.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2084

基于非对称跨模态融合的RGB-D显著目标检测

于 明[1,2], 邢章浩[1], 刘 依[2†]

(1. 河北工业大学 电子信息工程学院，天津 300401；2. 河北工业大学 人工智能与数据科学学院，天津 300401)

摘 要: 目前大多数RGB-D显著目标检测方法在RGB特征和Depth特征的融合过程中采用对称结构,对两种特征进行相同的操作,忽视了RGB图像和Depth图像的差异性,易造成错误的检测结果. 针对该问题,提出一种基于非对称结构的跨模态融合RGB-D显著目标检测方法,利用全局感知模块提取RGB图像的全局特征,并设计了深度去噪模块滤除低质量Depth图像中的大量噪声;再通过所提出的非对称融合模块,充分利用两种特征间的差异性,使用Depth特征定位显著目标,用于指导RGB特征融合,补足显著目标的细节信息,利用两种特征各自的优势形成互补. 通过在4个公开的RGB-D显著目标检测数据集上进行大量实验,验证所提出的方法优于当前的主流方法.

关键词: RGB-D图像；显著目标检测；非对称融合；全局感知模块；深度去噪模块

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2084

引用格式: 于明,邢章浩,刘依. 基于非对称跨模态融合的RGB-D显著目标检测[J]. 控制与决策, 2023, 38(9): 2487-2495.

RGB-D salient object detection with asymmetric cross-modal fusion

YU Ming[1,2], *XING Zhang-hao*[1], *LIU Yi*[2†]

(1. School of Electronic and Information Engineering，Hebei University of Technology，Tianjin 300401，China；2. School of Artificial Intelligence，Hebei University of Technology，Tianjin 300401，China)

Abstract: Most RGB-D salient object detection methods use a symmetric structure during the fusion process to perform the same operation on the RGB features and Depth features. This fusion method ignores the difference between the RGB image and the Depth image, which is likely to cause false detection results. In order to solve it, this paper proposes a cross-modal fusion RGB-D salient object detection method based on an asymmetric structure. In this paper, a global perception module (GPM) is designed to extract the global features of RGB images, and a deep denoising module (DDM) is designed to filter out a large amount of noise in low-quality depth images. Then through the asymmetric fusion module designed, we make full use of the difference between the two features differences, use the depth feature to locate salient objects, so as to guide RGB feature fusion and complement the detailed information of salient objects, and use the respective advantages of the two features to form a complement. A large number of experiments are carried out on four publicly available RGB-D salient object detection datasets, and the experimental results verify that the proposed method outperforms the state-of-the-art methods.

Keywords: RGB-D image；salient object detection；asymmetric fusion；global perception module；depth denoising module

0 引 言

图像显著目标检测旨在模拟人的视觉注意力机制,在图像中标注出最能吸引人注意的显著目标. 显著目标检测可以将计算资源优先分配给显著区域,因此常常作为一种预处理手段广泛应用在图像处理的各研究领域,例如目标跟踪[1]、语义分割[2]和医学图像处理[3]等.

早期的显著目标检测方法[4]大多仅使用RGB图像完成检测,尽管近几年的卷积神经网络在RGB显著目标检测算法中展现出强大的特征提取能力,但在复杂背景、光照不均匀或者纹理相似的场景中,检测结果仍然不够精确. Depth图像中含有丰富的空间信

收稿日期: 2021-11-29；**录用日期:** 2022-04-27.

基金项目: 国家自然科学基金青年项目 (61806071, 62102129)；河北省自然科学基金面上项目 (F2019202381, F2019202464).

责任编委: 张文安.

†通讯作者. E-mail: liuyi@scse.hebut.edu.cn.

息,可以作为一种补充特征帮助RGB特征定位显著目标,RGB-D显著目标检测渐渐取代RGB显著目标检测,成为研究的热点问题.

现有的RGB-D显著目标检测方法主要分为3类:前期融合、中期融合和后期融合. 前期融合在提取特征阶段就将RGB图像和Depth图像级联成为四通道图像,然后使用卷积神经网络(CNN)提取特征,得到最终的预测显著图. 这种方式完全没有考虑RGB图像和Depth图像的差异性,往往达不到预期的效果. 后期融合的方式采用多流网络模型,每个网络分别生成预测显著图,通过卷积等操作融合预测显著图. 后期融合进行图像级别融合,交互有限,难以获得高精度的预测结果. 相较于前期融合和后期融合,中期融合使用双流网络,对提取的特征进行融合,交互能力大大增强,得到的检测结果准确性更高. 文献[5]使用门控融合模块控制同一尺度的RGB特征与Depth特征在融合特征中所占比重,得到融合特征之后再生成预测显著图. 虽然中期融合带来了检测精度的提升,但是大多数方法没有考虑RGB特征和Depth特征两者之间的差异性,且融合前的特征为卷积神经网络提取的初始特征,没有充分发挥出两种模态的优势,限制了检测结果的进一步提升.

为了解决上述问题,提升RGB-D显著目标检测的精度,本文提出一种基于非对称跨模态融合的RGB-D显著目标检测算法. 本文算法的主要思路基于RGB特征与Depth特征的差异性:RGB图像主要包含光照、纹理和颜色等复杂信息;Depth图像包含丰富的空间信息,特征中包含的信息复杂度不同,对卷积神经网络提取的初始特征处理方式也应不同,在跨模态融合时发挥不同的作用,利用两种特征各自的优势形成互补.

本文主要贡献如下:

1) 针对大多数RGB-D显著目标检测方法忽视了RGB图像与Depth图像的差异性,容易造成错误检测结果的问题,提出一种基于非对称结构的跨模态融合RGB-D显著目标检测方法,并设计了全局感知模块提取RGB特征中的全局特征,以及深度去噪模块来滤除低质量Depth图中的大量噪声. 在跨模态融合之前,对初始RGB特征和Depth特征进行不同的处理,双流网络呈现非对称性.

2) 根据两种特征之间的差异性设计了非对称跨模态融合模块,利用Depth特征包含的空间、轮廓信息快速定位显著目标位置,指导RGB特征融合,补足显著目标的细节信息.

3) 在4个公开数据集上进行大量实验,实验结果验证了本文方法优于对比的16个RGB-D显著目标检测算法,可以在复杂环境中较好地检测到显著目标,且边缘较清晰,显著目标较为均匀高亮.

1 相关工作

传统方法:早期的RGB-D显著目标检测方法主要依赖设计具体的公式提取图像的颜色、边缘和对比度等特征. 文献[6]提取中心-周围对比度特征,根据像素点在周围环境中的突出程度衡量显著性. 文献[7]通过区域增长算法粗略估计显著目标区域,归一化得到粗显著图,最后使用最短距离作为后处理手段得到细化的预测显著图. 传统方法提取的特征属于低层次语义特征,难以生成更抽象的高层次语义特征,限制了基于传统方法的RGB-D显著目标检测算法的上限.

深度学习方法:卷积神经网络可以提取低层次和高层次的语义特征,弥补了传统方法只能提取低层次语义特征的不足,深度学习在RGB-D显著目标检测领域取得了良好的成绩. 基于深度学习的RGB-D显著目标检测算法大多为双流结构,采用中期融合的方式可以取得更好的效果. 文献[8]使用迁移学习,将在RGB数据库上训练好的模型作为双流网络的基础提取特征,再将RGB特征和Depth特征跨模态融合,处理RGB特征和Depth特征的方式相同,呈对称性. 文献[9]将特征分为高层次语义特征和低层次语义特征两类,分别设计了对称的融合模块.

上述的方法都具有以下特点:1)跨模态融合的输入特征均为卷积神经网络提取的初始特征;2)跨模态融合模块中,对两种模态的特征采用相同的处理方式,没有根据其特性做差别处理. 这些方法在融合过程中追求两种模态特征的相似性,忽略了两者之间的差异性. 文献[10]首次将对称网络概念引入RGB-D显著目标检测领域,将RGB图像和Depth图像分别输入ResNet-50和VGG16网络中,仅在提取初始特征时使用非对称结构,后续的双流网络中使用相同的模块,难以有效利用两种模态特征之间的差异性. 文献[11]使用不同的结构处理两种模态特征,没有提出明确的融合模块,融合方式较为简单. 不同于上述两种非对称模型,本文在特征提取阶段使用特征提取能力更强的ResNet-50网络提取初始特征,分别在RGB数据流和Depth数据流中设计相应的模块针对性地处理初始特征. 此外,不同于文献[11]在Depth数据流中简单进行跨模态融合,本文设计了明确的非对称融合模块,利用其差异性形成优势互补,从而实现更准

确的显著目标检测.

2　本文算法

2.1　整体框架

如图1所示,本文方法基于双流网络,将Depth图像和RGB图像分别输入ResNet-50网络中提取特征,提取到的Depth特征记为$f_{d_i}(i=1,2,3,4,5)$,RGB特征记为$f_{r_i}(i=1,2,3,4,5)$,共5个特征块,以下公式中多次出现下标i,表示具有相同宽和高的特征块,其中$i=1,2,3,4,5$. 使用深度去噪模块(depth denoising module, DDM)去除f_{d_i}中的噪声,与此同时,使用全局感知模块(global perception module, GPM)提取f_{r_i}中的全局信息,再将两种模态特征输入非对称融合模块(asymmetric fusion module, AFM),得到跨模态融合特征cm_i,解码器部分采用文献[12]中的BBS策略(二分支主干策略),融合特征cm_i经过解码器解码后生成预测显著图S_1和S_2,使用真值图G对预测显著图进行监督训练,选择S_2作为最终的预测显著图.

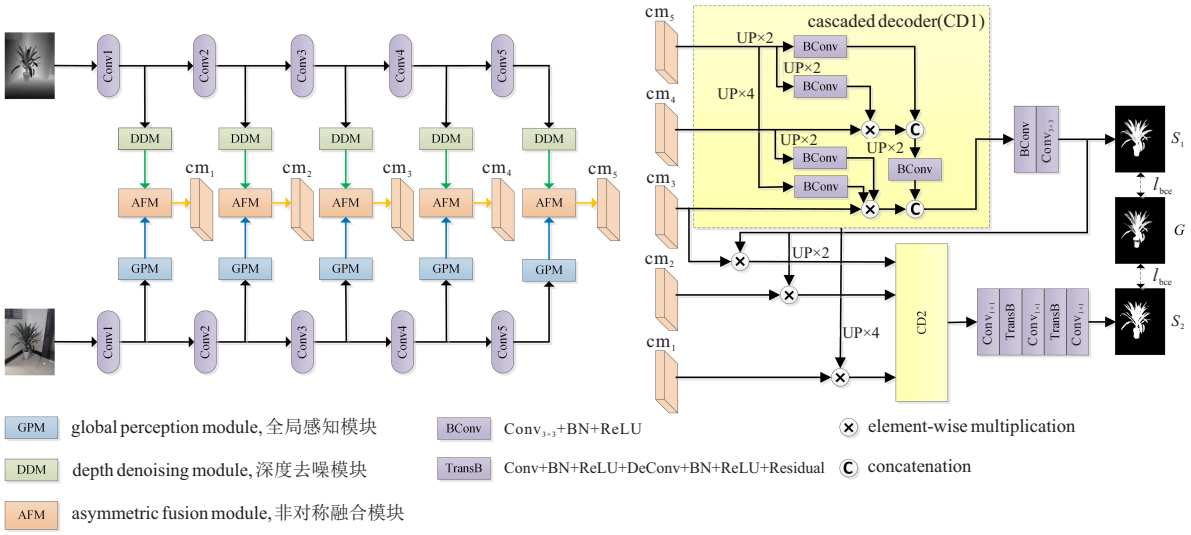


图 1　基于非对称跨模态融合RGB-D显著目标检测框架

2.2　深度去噪模块(DDM)

深度图由红外结构光或者飞行时间(time of flight, TOF)原理测得,因此在快速移动或者前景、背景深度相似的情况下拍摄到的深度图质量较差,低质量的深度图通常会造成误判. 本文将造成误判的深度图视为含有较多噪声的深度信息,ResNet-50网络提取的初始化Depth特征中含有较多噪声,不能直接进行跨模态融合. 深度去噪模块利用注意力机制去除深度信息中冗余的噪声,将初始化特征重构为对显著目标判断有利的特征.

如图2所示,深度去噪模块主要由通道注意力机制、空间注意力机制和批标准化等操作组成.

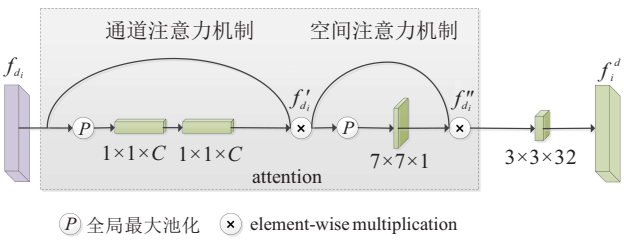


图 2　深度去噪模块

f_{d_i}表示ResNet-50网络提取的初始化Depth特征,使用两个连续大小为$1\times1\times C$的卷积核完成通道间注意力操作,通道注意力机制可以区分哪些通道中含有重要特征,哪些通道中含有较多噪声. 具体计算方法如下:

$$f'_{d_i}=\mathrm{Conv}_{1\times1}(\mathrm{Conv}_{1\times1}(P(f_{d_i})))\otimes f_{d_i}. \tag{1}$$

其中:$\mathrm{Conv}_{1\times1}(\cdot)$是卷积核大小为$1\times1$的卷积操作,$P$表示全局最大池化,$\otimes$为逐元素相乘.

通道注意力机制去除了含有噪声的特征通道,空间注意力机制则负责提取重要通道内具体的特征,是对通道注意力机制的重要补充. 实现方式如下:

$$f''_{d_i}=\mathrm{Conv}_{7\times7}(P(f'_{d_i}))\otimes f'_{d_i}. \tag{2}$$

其中:$\mathrm{Conv}_{7\times7}(\cdot)$是卷积核大小为$7\times7$的卷积操作,$P$表示全局最大池化,$\otimes$为逐元素相乘.

将上述经过注意力机制去噪的特征f''_{d_i}进行标准化,特征通道数降为32,有利于后续的融合过程,即

$$f_i^d=\delta(\mathrm{BN}(\mathrm{Conv}_{3\times3}(f''_{d_i}))). \tag{3}$$

其中:δ为ReLU激活函数,$\mathrm{BN}(\cdot)$为批标准化操作,$\mathrm{Conv}_{3\times3}(\cdot)$代表卷积核大小为$3\times3$的卷积操作.

2.3 全局感知模块

为了获得更加丰富的语义特征,全局信息对于显著目标检测任务必不可少. 相较于Depth图像,RGB图像是三通道图像,含有较为丰富的颜色、亮度、纹理等复杂信息,可用于提取全局特征. 本文在RGB特征流中加入全局感知模块,提取全局特征. 如图3所示,全局感知模块主要由注意力机制和空洞卷积组成,首先,使用同深度去噪模块相同结构的注意力机制,提取重要特征的同时寻找RGB特征与Depth特征之间的相似性.

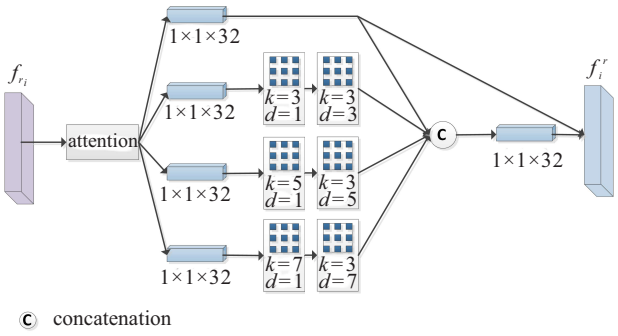


图 3 全局感知模块

然后,受到RFB结构[13](感受野模块)的启发,使用多分支空洞卷积扩大感受野,有利于提取全局特征. 此处共有4个分支,每个分支使用$1\times1\times32$的卷积操作将通道数降为32,后3个分支使用卷积核大小为$k\times k$、膨胀率为d的连续空洞卷积,将4个分支的特征级联后使用$1\times1\times32$的卷积操作,再与第1分支的特征残差连接,得到通道数为32的特征f_i^r. 具体实现方式如下:

$$\begin{cases} f_{r_i}^1 = \mathrm{Conv}_{1\times1}(\mathrm{Att}(\mathrm{fr}_i));\\ f_{r_i}^2 = \mathrm{Conv}_{3\times3}^3(\mathrm{Conv}_{3\times3}(\mathrm{Conv}_{1\times1}(\mathrm{Att}(f_{r_i}))));\\ f_{r_i}^3 = \mathrm{Conv}_{3\times3}^5(\mathrm{Conv}_{5\times5}(\mathrm{Conv}_{1\times1}(\mathrm{Att}(f_{r_i}))));\\ f_{r_i}^4 = \mathrm{Conv}_{3\times3}^7(\mathrm{Conv}_{7\times7}(\mathrm{Conv}_{1\times1}(\mathrm{Att}(f_{r_i})))). \end{cases} \tag{4}$$

$$f_i^r = \mathrm{Conv}_{1\times1}(\mathrm{Cat}(f_{r_i}^1, f_{r_i}^2, f_{r_i}^3, f_{r_i}^4)) \otimes f_{r_i}^1. \tag{5}$$

其中:$\mathrm{Att}(\cdot)$表示注意力机制,同图2的Attention结构;$\mathrm{Conv}_{k\times k}^d(\cdot)$表示卷积核大小为$k\times k$、膨胀率为$d$的卷积操作,上标为空时默认$d=1$,即普通卷积操作;$\mathrm{Cat}(\cdot)$代表级联操作;$\oplus$表示逐元素相加.

2.4 非对称融合模块(AFM)

在基于双流网络的RGB-D显著目标检测任务中,跨模态融合模块经常采用对称的结构,融合过程中对两种模态的特征进行相同的处理[8-9],其理论基础为寻找RGB特征与Depth特征之间的相似性. 但是Depth特征与RGB特征之间存在固有的差异性,Depth特征含有的空间特征有利于定位显著目标在图像中的位置,RGB特征含有更加丰富的语义,结合全局感知模块得到上下文信息,可以补充显著目标的细节信息. 非对称融合模块主要利用两种特征之间的差异性,对两种特征进行不同的操作,使用Depth特征快速定位,勾勒出显著目标的轮廓,指导RGB特征补足显著目标的边缘等细节信息,形成非对称的融合模块,充分发挥两种模态各自的特点,实现优势互补. 非对称融合模块的结构如图4所示.

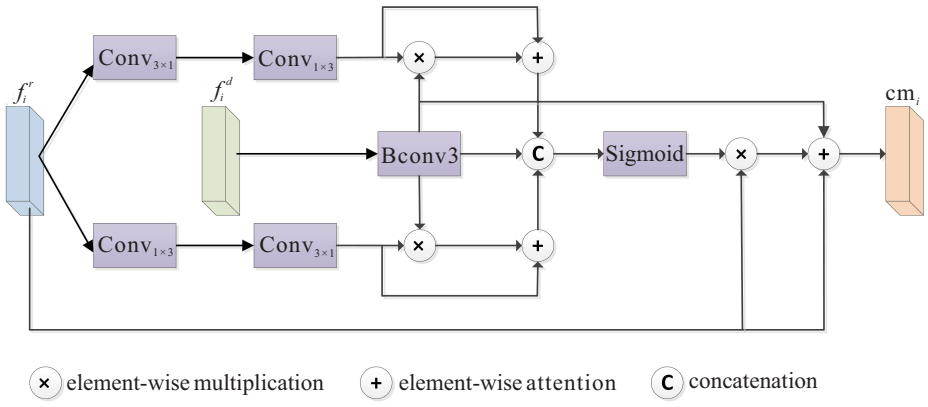


图 4 非对称融合模块

非对称模块输入的特征为f_i^d、f_i^r,首先,通过批标准化、卷积操作和ReLU激活函数处理f_i^d,得到特征f_g用于指导分别经过3×1和1×3卷积操作的f_i^r,补充显著目标的上下文信息,通过上下分支调换卷积操作顺序得到不同的特征块,进行两次融合;再将两个特征块级联,避免单次融合丢失较多重要信息;接着,使用残差连接和级联操作进一步融合特征,通过Sigmoid激活函数得到初步的融合特征;最后,为了最大程度地保留输入特征的原有重要信息,使用逐元素乘积和残差连接的方式,结合初步融合特征得到最终

的融合特征cm_i. 以上过程可以表示为:

$$f_g = \delta(\mathrm{BN}(\mathrm{Conv}_{3\times3}(f_i^d))), \tag{6}$$

$$\mathrm{cm}_i^1 = \mathrm{Conv}_{1\times3}(\mathrm{Conv}_{3\times1}(f_i^r)) \otimes f_g \oplus \mathrm{Conv}_{1\times3}(\mathrm{Conv}_{3\times1}(f_i^r)), \tag{7}$$

$$\mathrm{cm}_i^2 = \mathrm{Conv}_{3\times1}(\mathrm{Conv}_{1\times3}(f_i^r)) \otimes f_g \oplus \mathrm{Conv}_{3\times1}(\mathrm{Conv}_{1\times3}(f_i^r)), \tag{8}$$

$$\mathrm{cm}_i = \delta(\mathrm{Cat}(\mathrm{cm}_i^1, \mathrm{cm}_i^2, f_g)) \otimes f_i^r \oplus f_g \oplus f_i^r. \tag{9}$$

其中: δ、σ分别代表ReLU激活函数和Sigmoid激活函数; cm_i^1和cm_i^2为两个处理RGB特征的分支; $\mathrm{BN}(\cdot)$代表批标准化; $\mathrm{Conv}_{i\times j}(\cdot)$代表卷积操作, 下标为卷积核大小; $\mathrm{Cat}(\cdot)$表示级联操作; $\otimes$表示逐元素相乘, $\oplus$表示逐元素相加.

表1展示了本文设计的3个模块的输入输出特征尺寸, 深度去噪模块和全局感知模块的输出特征通道数均标准化为32.

表1　模块输入输出特征尺寸

卷积块	深度去噪模块		全局感知模块		非对称融合模块	
	输入	输出	输入	输出	输入	输出
1	88 × 88 × 64	88 × 88 × 32	88 × 88 × 64	88 × 88 × 32	88 × 88 × 32	88 × 88 × 32
2	88 × 88 × 256	88 × 88 × 32	88 × 88 × 256	88 × 88 × 32	88 × 88 × 32	88 × 88 × 32
3	44 × 44 × 512	44 × 44 × 32	44 × 44 × 512	44 × 44 × 32	44 × 44 × 32	44 × 44 × 32
4	22 × 22 × 1 024	22 × 22 × 32	22 × 22 × 1 024	22 × 22 × 32	22 × 22 × 32	22 × 22 × 32
5	11 × 11 × 2 048	11 × 11 × 32	11 × 11 × 2 048	11 × 11 × 32	11 × 11 × 32	11 × 11 × 32

2.5　损失函数

假设输入图像的高和宽分别为H、W, 给定输入RGB图像$X_r \in \boldsymbol{R}^{H\times W\times 3}$, Depth图像$X_d \in \boldsymbol{R}^{H\times W\times 1}$和对应的真值图$G \in \{0,1\}^{H\times W\times 1}$, 通过解码器可以得到两个预测的显著图$(S_1, S_2) \in \{0,1\}^{H\times W\times 1}$. 使用二元交叉熵损失函数分别计算$S_1, S_2$与$G$之间的误差, 即

$$l_{\mathrm{bce}}(S_j, G) = G\log(S_j) + (1-G)\log(1-S_j), \tag{10}$$

其中$j = 1, 2$.

最终的损失函数由上述两个损失函数加权得到, 即

$$l = \alpha l_{\mathrm{bce}}(S_1, G) + (1-\alpha) l_{\mathrm{bce}}(S_2, G). \tag{11}$$

其中权值$\alpha \in [0,1]$, 可以用于控制两个损失函数在最终的损失函数中所占的比重.

3　实验结果

本文在4个广泛使用的标准数据集上进行评估, 包括NJU2K[6]、NLPR[14]、SIP[15]和STERE[16].

3.1　评价指标

评价指标用于对比本文算法和目前主流算法的检测性能, 使用5种评价指标评估本文算法.

F分数[17]是一种评估区域相似度的常见评价指标, 计算公式为

$$F_\beta^j = \frac{(1+\beta^2)(P^j)^2 R^j}{\beta^2 P^j + R^j}. \tag{12}$$

其中: P^j和R^j分别是阈值为j时的查准率和查全率; β^2设为0.3.

S分数[18]是一种结合了区域感知结构相似度(S_r)和对象感知结构相似度(S_o)的评价指标. 计算公式为

$$S_\alpha = \alpha S_o + (1-\alpha) S_r, \tag{13}$$

其中$\alpha \in [0,1]$, 用于平衡S_r和S_o, 本文将α设为0.5.

E分数[19]是用于评估二进制显著性图像的评估指标, 同时使用了图像级别和像素级别的统计数据. 其计算公式为

$$E_\varphi = \frac{1}{w\times h}\sum_{x=1}^{w}\sum_{y=1}^{h}\varphi(x,y). \tag{14}$$

其中: w和h分别是显著图的宽和高, $\varphi(\cdot)$是增强对齐矩阵.

MAE评估预测显著图与真值图的平均绝对误差, 计算公式为

$$\mathrm{MAE} = \frac{1}{N}|S-G|. \tag{15}$$

其中: S和G分别代表预测显著图与真值图, N为图像的总体像素数.

P-R曲线: 阈值在[0, 255]之间取值, 对显著图进行二值化, 计算出成对出现的准确率和召回率, 不同的阈值对应不同的准确率和召回率对, 将准确率作为横坐标, 召回率作为纵坐标绘制P-R曲线图.

3.2　实验细节

本文模型使用PyTorch框架实现, 使用的硬件设备为GTX 1080Ti. ResNet-50网络在ImageNet数据集上进行预训练, 动量、学习率和批处理大小分别设置为0.99, 1e-4和10, 优化方式选用Adam算法. 除此之外, 使用NJU2K中的1 478张图片和NLPR中的710张图片进行训练, 分辨率大小统一为352 × 352. 为解

决训练数据不足的问题,使用数据增强的方式,例如随机翻转、随机裁剪、旋转等. 训练周期为250个周期,训练时长约为12 h. 实验结果显示,损失函数权重$\alpha = 0.5$时本文方法可以取得最好效果,得到的模型大小为204 MB,平均检测一张图片需要0.040 s,约为25 FPS,可以满足实时检测需求.

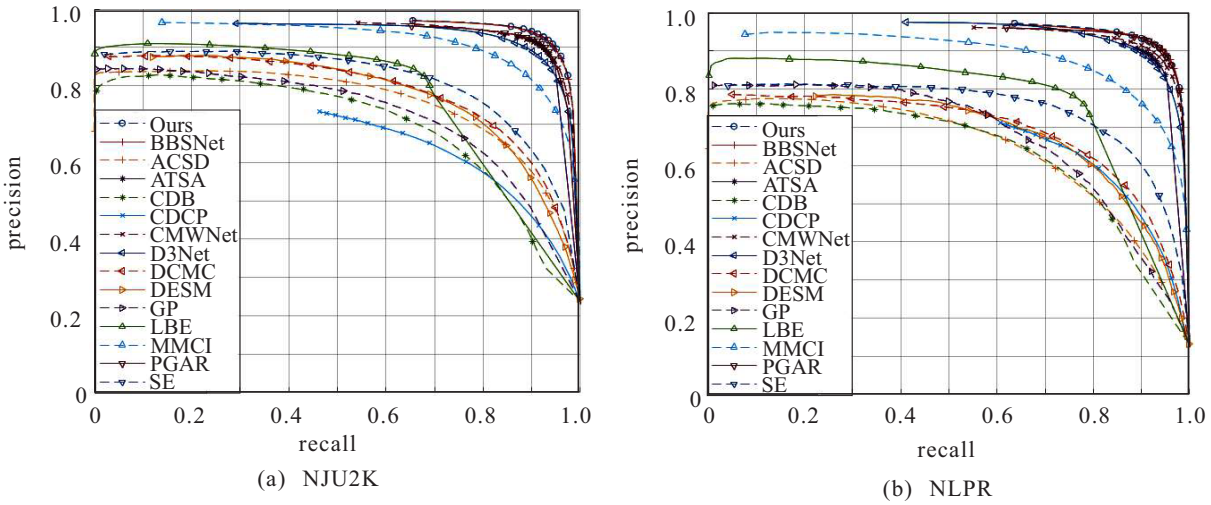
3.3 结果分析

将本文算法与8个传统算法(CDB[20], DESM[21], GP[22], ACSD[6], LBE[23], DCMC[24], CDCP[7], SE[25]),8个深度学习算法(MMCI[26], CMWNet[27], BBSNet[12], PGAP[28], ATSA[11], D3Net[15], CCAFNet[29], DQAS[30])进行对比. 为了公平比较,对比算法的预测显著图由相应论文提供,或者由论文提供的代码得出. 表2为本文算法与对比算法在4个数据集上进行的F分数、S分数、E分数和MAE的评估,其中,F分数、S分数和E分数越高越好,MAE越小越好. 为了更全面比较实验结果,将F分数细分为最大F分数(maxFm)、自适应F分数(adpFm)、平均F分数(meanFm). 同理,将E分数细分为最大E分数(maxEm)、自适应E分数(adpEm)、平均E分数(meanEm). 如表2所示,每行最好的结果加粗表示,次好的结果用斜体表示,在排名第3的数据下增加了下划线,本文算法在NJU2K, NLPR和SIP数据集上均取得了最好的结果,优于其他模型,在STERE数据集上3项指标为最优,3项指标是次优,S分数和MAE排在第3.

图5是在NJU2K和NLPR两个数据集上的P-R曲线,本文方法更靠近右上角,显示了本文方法相对其他对比方法的优越性. 图6是各模型的可视化预测显著图,本文提出的算法得到了更准确的预测图. 具体来说,图6中的第1行、第2行、第7行的Depth图像为低质量图像,从视觉上看显著目标与背景融为一体,显著目标轮廓难以看清,但是本文算法仍然准确标注出显著目标. 相较于对比算法,本文算法成功保留了自行车辐条的细节信息;第5行、第6行聚焦在检测显著人物上,相较于对比算法,本文算法较少将影子信息检测为显著目标,更接近真值图,减少了冗余信息. 第8行显示,本文算法充分利用了Depth图像的深度、空间信息,根据深度的不同,没有将后排的雕像标注出来,在特征融合过程中发挥了Depth特征的定位功能. 第3行和第4行Depth图像为高质量图像,ATSA算法错误地将部分背景区域标注成显著区域,本文算法没有误判显著目标,虽然本文算法与ATSA算法均使用非对称的结构分别处理RGB特征和Depth特征,但是ATSA算法在Depth数据流一侧进行简单跨模态融合,本文则设计了非对称融合模块进行跨模态融合,融合过程不易丢失重要信息,因此显著区域误判情况较少.

3.4 消融实验

本文为了验证所使用的结构和模块的有效性,进行了消融对比实验. 表3显示了在SIP和STERE两个数据集上的5组消融对比实验结果,其中4个评价指标,↑表示值越大越好,↓表示值越小越好. 实验设置中的BL表示不添加本文使用的3个模块,仅使用ResNet-50提取初始特征后通过简单的级联融合,结合解码器组成的基础模型;BL+AFM表示在BL的基础模型上添加非对称融合模块的实验结果;实验设置的第1列、第2列为Depth特征流和RGB特征流使用的预处理模型,‘–’代表该过程没有使用文中提到的模型;第5行对应的是本文使用模型. 从第1行和第2行的数据对比可以看出,在SIP和STERE两个数据集上,S分数的平均提升为1.4 %,在MAE指标上取得的平均提升为1.1 %,证实了本文所提出的非对称模

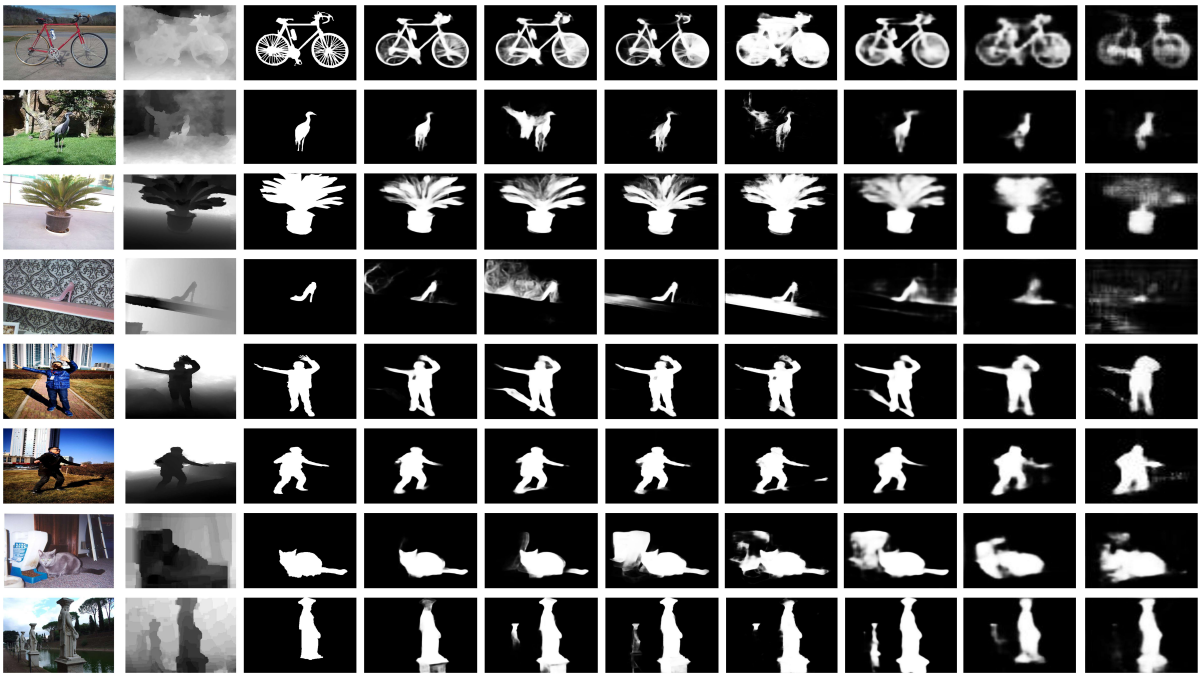


(a) NJU2K (b) NLPR

图5 在NJU2K和NLPR数据集上的P-R曲线

表 2　本文算法与经典算法在4个数据集上的对比结果

数据集	算法	评估指标							
		Sm↑	MAE↓	adpEm↑	meanEm↑	maxEm↑	adpFm↑	meanFm↑	maxFm↑
NJU2K	CDB	0.624	0.203	0.745	0.565	0.742	0.648	0.482	0.648
	DESM	0.665	0.283	0.682	0.590	0.791	0.632	0.550	0.717
	GP	0.527	0.211	0.716	0.466	0.703	0.655	0.357	0.647
	ACSD	0.699	0.202	0.786	0.593	0.803	0.696	0.512	0.711
	LBE	0.695	0.153	0.791	0.655	0.803	0.740	0.606	0.748
	DCMC	0.686	0.172	0.791	0.619	0.799	0.717	0.556	0.715
	CDCP	0.669	0.180	0.747	0.706	0.741	0.624	0.595	0.621
	SE	0.664	0.169	0.772	0.624	0.813	0.734	0.583	0.748
	MMCI	0.858	0.079	0.878	0.851	0.915	0.812	0.793	0.852
	ATSA	0.902	0.041	*0.920*	0.935	0.940	0.891	0.894	0.906
	D3Net	0.895	0.051	0.892	0.912	0.932	0.840	0.860	0.889
	CMW	0.903	0.046	0.911	0.923	0.936	0.880	0.881	0.902
	PGAR	0.909	0.042	0.916	0.929	0.940	0.893	0.892	0.907
	CCAF	0.910	*0.037*	*0.920*	–	–	0.898	–	0.911
	DQAS	0.892	0.051	0.910	–	0.928	0.856	0.867	0.891
	BBSNet	*0.921*	**0.035**	**0.924**	*0.938*	*0.949*	*0.902*	*0.902*	*0.920*
	Ours	**0.923**	**0.035**	**0.924**	**0.941**	**0.953**	**0.904**	**0.907**	**0.924**
NLPR	CDB	0.629	0.114	0.809	0.565	0.791	0.613	0.422	0.618
	DESM	0.572	0.312	0.698	0.541	0.805	0.563	0.430	0.640
	GP	0.654	0.146	0.804	0.571	0.723	0.659	0.451	0.611
	ACSD	0.673	0.179	0.742	0.578	0.780	0.535	0.429	0.607
	LBE	0.762	0.081	0.855	0.719	0.855	0.736	0.736	0.745
	DCMC	0.724	0.117	0.786	0.684	0.793	0.614	0.543	0.648
	CDCP	0.727	0.112	0.800	0.781	0.820	0.608	0.609	0.645
	SE	0.756	0.091	0.839	0.742	0.847	0.692	0.624	0.713
	MMCI	0.856	0.059	0.872	0.841	0.913	0.730	0.737	0.815
	ATSA	0.908	0.028	0.944	0.944	0.949	0.872	0.879	0.896
	D3Net	0.906	0.034	0.932	0.923	0.946	0.834	0.853	0.885
	CMW	0.917	0.029	0.940	0.939	0.951	0.859	0.877	0.903
	PGAR	**0.930**	*0.024*	**0.955**	0.949	*0.961*	**0.885**	*0.896*	0.916
	CCAF	*0.922*	0.026	0.952	–	–	*0.882*	–	0.909
	DQAS	0.900	0.034	0.938	–	0.938	0.858	0.855	0.884
	BBSNet	**0.930**	**0.023**	0.952	*0.950*	*0.961*	*0.882*	*0.896*	*0.918*
	Ours	**0.930**	**0.023**	*0.954*	**0.952**	**0.964**	**0.885**	**0.899**	**0.919**
SIP	CDB	0.557	0.192	0.771	0.455	0.737	0.624	0.341	0.620
	DESM	0.616	0.298	0.742	0.564	0.770	0.664	0.496	0.669
	GP	0.588	0.173	0.774	0.511	0.768	0.699	0.411	0.687
	ACSD	0.732	0.172	0.827	0.614	0.838	0.727	0.542	0.763
	LBE	0.727	0.200	0.841	0.651	0.853	0.733	0.571	0.751
	DCMC	0.683	0.186	0.786	0.598	0.743	0.645	0.499	0.618
	CDCP	0.595	0.224	0.722	0.683	0.721	0.495	0.482	0.505
	SE	0.628	0.164	0.756	0.592	0.771	0.662	0.515	0.661
	MMCI	0.833	0.086	0.886	0.845	0.897	0.795	0.771	0.818
	ATSA	0.845	0.071	0.890	0.887	0.896	0.862	0.861	0.865
	D3Net	0.864	0.063	0.902	0.894	0.910	0.831	0.832	0.862
	CMW	0.867	0.062	0.906	0.900	0.913	0.851	0.851	0.874
	PGAR	0.876	0.055	0.908	*0.906*	0.915	0.854	0.854	0.876
	CCAF	0.877	*0.054*	0.915	–	–	0.866	–	0.881
	DQAS	–	–	–	–	–	–	–	–
	BBSNet	*0.879*	0.055	*0.916*	*0.906*	*0.922*	*0.872*	*0.868*	*0.883*
	Ours	**0.883**	**0.052**	**0.918**	**0.912**	**0.925**	**0.873**	**0.872**	**0.890**
STERE	CDB	0.615	0.166	0.808	0.561	0.823	0.713	0.489	0.717
	DESM	0.642	0.295	0.675	0.579	0.811	0.594	0.519	0.700
	GP	0.588	0.182	0.784	0.509	0.743	0.711	0.405	0.671
	ACSD	0.692	0.200	0.793	0.592	0.806	0.661	0.478	0.669
	LBE	0.660	0.250	0.749	0.601	0.787	0.595	0.501	0.633
	DCMC	0.731	0.148	0.831	0.655	0.819	0.742	0.590	0.740
	CDCP	0.713	0.149	0.796	0.751	0.786	0.666	0.638	0.664
	SE	0.708	0.143	0.825	0.665	0.846	0.748	0.610	0.755
	MMCI	0.873	0.068	0.901	0.873	0.927	0.829	0.813	0.863
	ATSA	0.897	**0.039**	*0.921*	**0.939**	**0.944**	*0.884*	**0.883**	*0.901*
	D3Net	0.891	0.054	0.904	0.908	0.930	0.883	0.844	0.881
	CMW	0.905	0.043	0.917	0.928	**0.944**	0.869	0.872	*0.901*
	PGAR	*0.907*	*0.041*	0.919	0.927	0.939	0.880	*0.878*	0.898
	CCAF	0.891	0.044	*0.921*	–	–	0.870	–	0.887
	DQAS	0.897	0.048	0.919	–	0.932	0.857	0.861	0.888
	BBSNet	**0.908**	*0.041*	**0.925**	0.928	*0.942*	**0.885**	**0.883**	**0.903**
	Ours	0.906	0.042	**0.925**	*0.929*	*0.942*	**0.885**	**0.883**	*0.901*



(a) RGB (b) Depth (c) GT (d) Ours (e) BBSNet (f) PGAR (g) CMW (h) D3Net (i) ATSA (j) MMCI

图 6 与经典算法的可视化对比

表 3 消融实验结果对比

模型	实验设置		SIP				STERE			
	Depth	RGB	Sm↑	MAE↓	adpEm↑	adpFm↑	Sm↑	MAE↓	adpEm↑	adpFm↑
BL	–	–	0.862	0.066	0.904	0.857	0.886	0.056	0.914	0.865
BL+AFM	–	–	0.874	0.056	0.912	0.865	0.902	0.044	0.922	0.880
BL+AFM	DDM	DDM	0.881	0.055	0.916	0.865	0.903	0.045	0.919	0.876
BL+AFM	GPM	GPM	0.882	0.053	0.919	0.870	0.903	0.044	0.922	0.880
BL+AFM	DDM	GPM	0.883	0.052	0.918	0.873	0.906	0.042	0.925	0.885

块的有效性. 同理,从第2行和第5行的数据对比可以看出,融合前使用深度去噪模块和全局感知模块能有效提升检测的精度. 为了验证非对称结构的有效性,第3行和第4行采用对称的融合前处理方式,通过与第5行的比较可以看出,对称结构得到的结果略低于非对称结构的实验结果,证实采用非对称的结构更有利于两种模态特征在融合时发挥各自的优势.

4 结 论

本文从利用RGB特征与Depth特征的差异性角度出发,提出一种基于非对称跨模态融合的RGB-D显著目标检测方法. 为了充分利用两种特征之间的差异性,使用不同的模块处理两种特征之后再进行跨模态融合. 本文利用深度去噪模块滤除噪声,有利于后续的特征融合;RGB特征中则含有丰富的语义信息,因此设计了全局感知模块更准确地提取全局特征;在跨模态融合过程中,使用不同的结构处理两种差异性的特征,使用Depth特征快速定位显著目标,提取显著目标的轮廓,用于指导RGB特征融合,补足显著目标的细节信息. 在4个公开的RGB-D显著目标检测数据集上进行了大量的实验,结果显示本文方法优于当前的主流方法,能够在复杂环境中较准确地检测到显著目标,且边缘较清晰,显著目标内侧较为均匀高亮.

参考文献(References)

[1] Wang J, Huang W C. Object tracking based on saliency and adaptive background constraint[C]. The 39th Chinese Control Conference (CCC). Shenyang, 2020: 6533-6538.

[2] LuoW, YangM, ZhengW. Weakly-supervised semantic segmentation with saliency and incremental supervision updating[J]. Pattern Recognition, 2021, 115: 107858.

[3] 高源, 于晓升, 吴成东, 等. 基于显著性检测和改进局部高斯分布拟合模型的眼底图像视盘边界自动提取[J]. 控制与决策, 2019, 34(1): 151-156.
(Gao Y, Yu X S, Wu C D, et al. Automatic optic disc boundary extraction based on saliency object detection and modified local Gaussian distribution fitting model in retinal images[J]. Control and Decision, 2019, 34(1): 151-156.)

[4] 周静波, 黄伟. 基于低秩矩阵恢复的视觉显著性目标检测与细化[J]. 控制与决策, 2021, 36(7): 1707-1713.
(Zhou J B, Huang W. Saliency object detection and

refinement based on low rank matrix recovery[J]. Control and Decision, 2021, 36(7): 1707-1713.)

[5] Zhou W J, Chen Y Z, Liu C, et al. GFNet: Gate fusion network with Res2Net for detecting salient objects in RGB-D images[J]. IEEE Signal Processing Letters, 2020, 27: 800-804.

[6] Ju R, Ge L, Geng W J, et al. Depth saliency based on anisotropic center-surround difference[C]. IEEE International Conference on Image Processing. Paris, 2014 : 1115-1119.

[7] Zhu C B, Li G, Wang W M, et al. An innovative salient object detection using center-dark channel prior[C]. IEEE International Conference on Computer Vision Workshops. Venice, 2017: 1509-1515.

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

[9] 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[C]. Proceedings of the 28th ACM International Conference on Multimedia. NewYork, 2020: 4107-4115.

[10] Liu C, Zhou W J, Chen Y Z, et al. Asymmetric deeply fused network for detecting salient objects in RGB-D images[J]. IEEE Signal Processing Letters, 2020, 27: 1620-1624.

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

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

[13] Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]. Proceedings of the European Conference on Computer Vision (ECCV). Piscataway: IEEE, 2018: 385-400.

[14] Peng H W, Li B, Xiong W H, et al. RGBD salient object detection: A benchmark and algorithms[C]. European Conference on Computer Vision. Berlin: Springer International Publishing, 2014: 92-109.

[15] Fan D P, Lin Z, Zhang 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.

[16] Niu Y Z, Geng Y J, Li X Q, et al. Leveraging stereopsis for saliency analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition. Providence, 2012: 454-461.

[17] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009: 1597-1604.

[18] 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. Piscataway: IEEE, 2017: 4548-4557.

[19] Fan D P, Gong C, Cao Y, et al. Enhanced-alignment measure for binary foreground map evaluation[J/OL]. 2018, arXiv: 1805.10421.

[20] Liang F F, Duan L J, Ma W, et al. Stereoscopic saliency model using contrast and depth-guided-background prior[J]. Neurocomputing, 2018, 275: 2227-2238.

[21] Cheng Y P, Fu H Z, Wei X X, et al. Depth enhanced saliency detection method[C]. Proceedings of International Conference on Internet Multimedia Computing and Service. Berlin: Springer, 2014: 23-27.

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

[23] Feng D, Barnes N, You S D, et al. Local background enclosure for RGB-D salient object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 2343-2350.

[24] Cong R M, Lei J J, Zhang C Q, et al. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion[J]. IEEE Signal Processing Letters, 2016, 23(6): 819-823.

[25] Guo J F, Ren T W, Bei J. Salient object detection for RGB-D image via saliency evolution[C]. IEEE International Conference on Multimedia and Expo. Seattle, 2016: 1-6.

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

[27] Li G Y, Liu Z, Ye L W, et al. Cross-modal weighting network for RGB-D salient object detection[C]. European Conference on Computer Vision. Berlin: Springer, 2020: 665-681.

[28] Chen S H, Fu Y. Progressively guided alternate refinement network for RGB-D salient object detection[C]. European Conference on Computer Vision. Berlin: Springer, 2020: 520-538.

[29] Zhou W J, Zhu Y, Lei J S, et al. CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images[J]. IEEE Transactions on Multimedia, 2022, 24: 2192-2204.

[30] Wang X H, Li S, Chen C, et al. Depth quality-aware selective saliency fusion for RGB-D image salient object detection[J]. Neurocomputing, 2021, 432: 44-56.

作者简介

于明 (1964–), 男, 教授, 博士生, 从事图像处理、模式识别等研究, E-mail: yuming@scse.hebut.edu.cn;

邢章浩 (1997–), 男, 硕士生, 从事图像处理的研究, E-mail: xingzhanghao_ai@163.com;

刘依 (1977–), 女, 讲师, 硕士, 从事图像处理、模式识别等研究, E-mail: liuyi@scse.hebut.edu.cn.