基于跨模态特征融合的 RGB-D 显著性目标检测

陈正, 赵晓丽*, 张佳颖, 尹明臣, 叶翰辰, 周浩军

(上海工程技术大学电子电气工程学院 上海 201600)
(evawhy@163.com)

摘要: 为了解决基于彩色图像的显著性检测在多目标或小目标等场景下无法准确检测出显著目标的问题, 提出了一种基于 RGB-D 跨模态特征融合的显著性检测网络模型, 该网络模型以改进的全卷积神经网络(FCN)为双流主干网络, 分别提取彩色与深度特征并作出预测, 最后利用 Inception 结构融合生成最终显著图. 针对原 FCN 实际感受野远低于理论感受野, 没有真正利用图像全局信息的问题, 设计了双分支结构的全局与局部特征提取块, 利用全局特征分支提取全局信息并指导局部特征提取, 并以此构建了改进的 FCN. 此外, 考虑到不同层级上彩色与深度特征之间的差异性, 提出了跨模态特征融合模块, 采用点积有选择性地融合彩色和深度特征, 与加法和级联相比, 采用点乘可以有效减少噪声与冗余信息. 通过在 3 个公开基准数据集上与 21 种主流网络相比的综合实验表明, 所提模型在 S 值、 F 值和 MAE 这 3 个指标上基本处于前 3 水平, 同时对模型大小进行了比较, 其大小仅为 MMCI 的 4.7%, 与现有最小模型 A2dele 相比减少了 22.8%.

关键词: RGB-D 显著性目标检测; 双分支特征; 跨模态融合

中图法分类号: TP391.41

DOI: 10.3724/SP.J.1089.2021.18710

RGB-D Image Saliency Detection Based on Cross-Model Feature Fusion

Chen Zheng, Zhao Xiaoli*, Zhang Jiaying, Yin Mingchen, Ye Hanchen, and Zhou Haojun

(School of Electric and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201600)

Abstract: In order to solve the problem that saliency detection based on RGB images cannot accurately detect salient object in scenes such as multiple targets or small targets, a novel saliency detection network based on RGB-D cross-modal feature fusion is proposed. The network takes the improved fully convolutional network (FCN) as the dual-stream backbone network, extracts the color and depth features and makes predictions respectively, and finally uses the inception structure fusion to generate the final saliency map. Aiming at the problem that the actual receptive field of the original FCN is far lower than the theoretical receptive field, and the global image information is not really used, a dual-branch structure global and local feature extraction block is designed. The global feature branch is used to extract global information and guide local feature extraction. This builds an improved FCN. In addition, considering the differences between color and depth features at different levels, a cross-modal feature fusion module is proposed, which uses dot product to selectively fuse color and depth features. Compared with addition and cascade, it uses dot multiplication. It can effectively reduce noise and redundant information. Comprehensive experiments on three benchmark datasets demonstrate that compared with 21 mainstream networks, this model is basically in the top three levels in terms

收稿日期: 2020-10-12; 修回日期: 2021-05-13. 基金项目: 国家自然科学基金(61772328). 陈正(1995—), 男, 硕士, 主要研究方向为计算机视觉、显著性检测、深度学习; 赵晓丽(1975—), 女, 博士, 副教授, 硕士生导师, CCF 会员, 论文通讯作者, 主要研究方向为视频与图像处理、模式识别、计算机视觉、智能计算; 张佳颖(1995—), 女, 硕士, 主要研究方向为三维点云、计算机视觉; 尹明臣(1994—), 男, 硕士, 主要研究方向为视频处理、计算机视觉; 叶翰辰(1993—), 男, 硕士研究生, 主要研究方向为对抗攻击、计算机视觉; 周浩军(1997—), 男, 硕士研究生, 主要研究方向为深度学习、计算机视觉.

of S value, F value and MAE. At the same time, the size of the model is analyzed. In comparison, its size is only 4.7% of MMCI, which has a decrease of 22.8% compared with the existing smallest model A2dele.

Key words: RGB-D saliency object detection; dual branch feature; cross-model fusion

显著性目标检测(salient object detection, SOD)是以认知科学、计算机视觉等学科为理论基础,通过模拟人类视觉系统检测图像或视频中的重要信息,目前已广泛应用于图像检索^[1]、目标检测^[2-3]、图像分割^[4]、视频压缩^[5]与视频跟踪^[6]等任务。

近年来,随着深度卷积神经网络的成功应用,显著性检测迅猛发展,大量基于颜色、亮度等视觉特征的显著性模型被相继提出。Li 等^[3]首次利用深度神经网络搭建了一个基于多尺度特征的显著性模型。Hou 等^[7]提出了深度监督的显著性检测模型(deeply supervised salient object detection, DSS),该模型利用全卷积神经网络提取多级多尺度特征,然后通过引入跳层结构融合提取到的多级多尺度特征。Wu 等^[8]提出级联部分解码器模型(cascaded partial decoder, CPD),将主干网络中较深的特征进行整合,得到初始显著性图,进而通过整体注意力模块细化特征,获得最终的显著性图。Liu 等^[9]以 U-Net 为主干网络,利用其提出的像素级上下文注意力网络(pixel-wise contextual attention network, PiCANet)整合全局上下文信息以及多尺度局部上下文信息,从而进一步提升检测性能。Feng 等^[10]利用全局感知器模块从整体上细化最显著的特征,并利用注意力反馈模块在相应的编解码器之间传递信息。Qin 等^[11]先用一个深度监督的自底向上/自顶向下结构进行粗显著性预测,然后用另一个自底向上/自顶向下模块对显著性映射的残差进行细化,并采用混合损失提升边缘质量。

然而,RGB 显著性检测面临着 2 大挑战:(1)当显著目标和背景具有相似的外观时,仅依靠 RGB 信息难以有效地区分目标与背景;(2)当同一物体包含不同色彩时,容易被误判为不同对象。而深度图包含丰富的空间结构和三维布局信息,能保持检测区域的完整性并提供大量额外的显著性线索,以区分目标与背景,因此,利用深度信息可以有效地提高显著性检测效果。Ciptadi 等^[12]首次在 RGB 基础上引入了深度信息,提出了基于 RGB-D 的显著性分割模型。Peng 等^[13]提出了一个多阶段 RGB-D 模型,该模型同时考虑了低层特征对比度、中层区域分组和高级先验增强所产生的

深度和外观线索。由于手工特征的不稳定性, Han 等^[14]将基于 RGB 的深度神经网络转换为适用于深度视图的网络,并自动融合 2 个视图的深度表示以获得最终显著性图。考虑 RGB 与深度信息之间的差异性, Zhu 等^[15]分别构造了处理 RGB 信息的主网络与处理深度信息的子网络,并将子网络整合到主网络中。Chen 等^[16]在 RGB-D 融合网络中设计了互补感知融合(complementarity-aware fusion, CA-Fusion)模块学习互补信息,通过级联 CA-Fuse 模块,从深到浅密集地增加逐层监督来逐步融合多级信息。Piao 等^[17]提出了深度诱导多尺度循环注意力网络,该网络使用包含残差结构的深度细化块融合 RGB 与深度互补信息,将多尺度上下文特征与深度信息相结合以精准定位显著性目标,同时利用循环注意力模块进一步提升模型性能。Fan 等^[18]在特征金字塔网络(feature pyramid networks, FPN)的基础上构建了深度净化单元进行初步的深度图滤波,同时搭建了三流特征学习模块学习跨模态特征。综上所述,现有 RGB-D 显著性检测方法主要是在主干网络的基础上提出一些子网络学习 RGB-D 互补信息以及进行特征细化、特征融合,但其网络结构大多过于复杂,参数量大,训练较为困难。

针对已有 RGB-D 显著性检测方法存在的上述问题,本文提出了一种交互式双流显著性检测模型。为了解决全卷积神经网络(fully convolutional network, FCN)^[19]理论感受野与实际感受野不一致,无法获取全局信息的问题,设计了全局与局部特征提取模块(global and local feature extraction block, GL Block)以获取全局信息并指导局部特征提取,并以此搭建了改进的 FCN(improved FCN, IFCN);受注意力机制启发并考虑 RGB 与深度信息之间的差异性,采用点乘获取彩色与深度特征中共同的显著性特征,搭建了跨模态特征融合模块(cross-model feature fusion module, CFFM)交叉融合 RGB 和深度信息以获得更好的显著性检测结果。在 NJU2000^[20], NLPR^[13]和 STEREO^[21]这 3 个公开数据集上的综合实验表明,本文模型在主流评价指标上具有较高的精度。

1 本文模型

本文网络模型采用了如图 1 所示的双流网络. 由于 FCN 理论感受野远低于实际感受野, 本文利用所提出的 GL Block 获取并融合全局与局部特征, 同时代替 FCN 原有卷积块搭建主干网络——IFCN,

以生成初始显著性图. 为了获取彩色与深度信息的共显著性特征, 设计了 CFFM, 考虑到浅层特征的噪声较多, 本文在网络较深层次利用 CFFM 交叉融合彩色及深度特征, 减少冗余特征. 最后通过 Inception^[22]结构融合初始显著性图, 提升网络的尺度适应性.

图 1 本文网络模型

1.1 IFCN

众所周知, 浅层次的卷积层感受野较小, 保留了大量的细节信息, 有助于精细化分割图像; 深层次的卷积层感受野相对较大, 可以学习到一些抽象特征, 有助于分类性能的提高. FCN 采用跳级结构利用浅层信息辅助逐步上采样, 获得精细化分割图像. 对于分割网络, 感受野越大, 网络所捕捉到的范围越大, 能够用于分析的信息越多, 分割效果越好. 在 FCN 中, 其 FC7 层理论感受野为 404×404 像素, 足以获取全局上下文信息. 但是, Liu 等^[23]的实验却证明, 其实际感受野只有全图的 $1/4$. 为了获取更大的感受野, 通常可以采用增加网络深度和使用大卷积核的方法. 然而, 如果一味地通过增加网络深度捕捉全局上下文信息, 不仅会大幅度增加网络负担, 而且容易造成梯度爆炸或梯度消失; 此外, 使用大的卷积核会导致计算量的暴增, 不利于网络深度的增加, 计算性能也会降低.

受 SE^[24]结构的启发, 本文设计了一个 GL Block, 该模块采用双分支结构分别提取局部与全局特征, 其结构如图 2 所示. 局部特征提取分支先用 stride 为 2, 卷积核为 3×3 , 激活函数为 ReLU 的卷积层将输入特征图降为原特征图的 $1/4$, 然后利用 2 个相同的 stride 为 1 的卷积层进行局部特征提

取. 全局特征提取分支先使用全局平均池化显式提取全局特征, 以整合全局空间信息, 为降低分支网络计算量, 引入了瓶颈结构, 然后利用 Softmax 学习全局特征分布, 最后采用点乘的方式融合全局与局部特征.

本文提出的 IFCN(图 1)延续了 FCN 的整体架构, 包含 5 个卷积块, 其中第 1 个卷积块为标准卷积块(考虑到原 FCN 在第 1 个卷积上对图像进行 100 的边缘填充操作会引入大量额外噪声, 本网络舍弃了该操作), 其余为 GL Block 用于捕捉全局与局部特征, 并通过全局特征指导局部特征提取. 然后利用反卷积进行上采样, 并通过跳级连接融合浅层信息逐步辅助上采样.

图 2 GL Block

1.2 CFFM

考虑已有跨模态特征融合方式大多在加法或

级联的基础上设计, 容易引入冗余噪声且结构复杂加算量较大, 本文受注意力机制以及 Wei 等^[25]研究启发, 采用点乘搭建了 CFFM 融合彩色特征 $f_r \in \mathbb{R}^{H \times W \times C}$ 和深度特征 $f_d \in \mathbb{R}^{H \times W \times C}$. 相对而言, 彩色图具有鲜明的外观与纹理信息, 而深度图提供了清晰的目标形状、轮廓和空间结构, 这些对于生成精确的显著图非常重要. 考虑到浅层深度特征含有大量噪声, 本文在较深层次利用 CFFM 进行特征融合.

CFFM 通过特征交叉融合减少特征之间的差异. 该算法首先通过点乘提取 f_r 和 f_d 的共性特征, 然后通过元素加法分别与 f_r 和 f_d 进行合并. 与现有研究中采用的直接加法或级联方法相比, CFFM 避免了在 f_r 和 f_d 中引入过多的冗余与噪声信息, 这些信息会对显著性图的生成产生不利影响. 通过多次特征融合, f_r 和 f_d 将逐渐吸收彼此的有用信息, 使其优势互补, 即减少 f_r 的冗余信息, 锐化 f_d 的边界.

CFFM 包含 f_r 和 f_d 这 2 个分支, 如图 3 所示. 首先对 f_r 和 f_d 施加一个 3×3 的卷积, 在压缩通道的同时降低模块计算量, 以便后续处理; 然后通过点乘对特征进行融合变换, 使融合特征具有清晰的边界和语义一致性; 最后仍采用一个 3×3 的卷积恢复原有通道并添加到原 f_r 和 f_d , 以精细化特征表示. 其过程可以表示为

$$f_r = f_r + W_2(W_r(f_r) \times W_d(f_d)) \quad (1)$$

$$f_d = f_d + W_2(W_r(f_r) \times W_d(f_d)) \quad (2)$$

其中, W_r, W_d, W_2 均为 3×3 卷积, 用于压缩及恢复通道. 整个模块采用对称式结构, f_r 将细节信息传递给 f_d 以精细化边缘, f_d 将显著性语义传递给 f_r 以摒弃冗余信息.

图 3 CFFM

1.3 损失函数

在显著性检测中, 二元交叉熵(binary cross entropy, BCE)是最常用的损失函数, 其计算方法为

$$L(\hat{y}, y) = \begin{cases} -\lg \hat{y}, & y = 1 \\ -\lg(1 - \hat{y}), & y = 0 \end{cases} \quad (3)$$

然而, BCE 损失函数的前提条件是样本均衡, 而在现实场景中通常存在着正负样本不平衡的问题, 即在背景占主导地位的图片中前景像素的损失会被稀释, BCE 损失无法解决背景-前景不平衡的问题. 因此, 本文引入了 Focal Loss^[26]解决此问题, 即

$$L(\hat{y}, y) = \begin{cases} -\alpha(1 - \hat{y})^\gamma \lg \hat{y}, & y = 1 \\ -(1 - \alpha)\hat{y}^\gamma \lg(1 - \hat{y}), & y = 0 \end{cases} \quad (4)$$

Focal Loss 在 BCE 的基础上加入 γ 参数, 降低易分类样本的损失, 使网络更关注于困难样本, 同时加入平衡因子 α , 增大前景对于损失函数的贡献, 以平衡正负样本.

2 实验及结果分析

本文模型基于 PyTorch 实现, 机器配置 2 块 GTX 1080Ti (11 GB)显卡. 使用 Adam 优化器^[27]进行训练, 训练冲量、学习率、权值衰减率和批大小分别设置为(0.9, 0.999), 0.0005, 10^{-5} , 16. 本文模型为端到端的模型, 不需要进行预训练或其他操作.

2.1 数据集及评价指标

在 NJU2000, NLPR 和 STEREO 这 3 个公开 RGB-D 基准数据集上进行了充分实验以验证本文算法. NJU2000 数据集包含 1985 幅从互联网上收集的 RGB 图像, 并采用光流技术获得深度图像, 以及对应的真值图. NLPR 数据集包含 1000 幅 RGB 图像、深度图以及对应的真值图, 其中包含了超过 400 种物体. STEREO 数据集包含了 797 幅 RGB 图像、深度图以及对应的真值图.

为了与其他方法公平对比, 本文在 NJU2000 数据集中选取 1585 幅图像作为训练集, 400 幅图像作为测试集; 在 NLPR 数据集中选取 800 幅图像作为训练集, 200 幅图像作为测试集; 在 STEREO 数据集中选取 637 幅图像作为训练集, 160 幅图像作为测试集.

本文采用平均精确度 P 、平均召回率 R 、 F 值和平均绝对误差(mean absolute error, MAE)等作为评价指标, 并绘制了 P - R 曲线对模型的性能进行评估.

F 值是 P 和 R 的加权调和平均, 计算公式为

$$F = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R} \quad (5)$$

其中, 设 $\beta^2=0.3$. 本文以最大 F 值 ($F_{\max}$) 作为评价指标. P 和 R 的计算公式分别为

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (6)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (7)$$

其中, N_{TP} , N_{FP} , N_{FN} 分别代表真阳性、假阳性、假阴性的数量.

MAE 定义为显著性输出与真值之间的平均像素绝对差, 即

$$MAE = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - G(x, y)| \quad (8)$$

其中, H 和 W 分别为图像的长和宽; $S(x, y)$ 和 $G(x, y)$ 分别为 (x, y) 处的显著性预测值和真值.

通过结合区域感知结构相似性 S_r 和对象感知结构相似性 S_o 来测量图像质量, 计算公式为

$$S = \alpha S_o + (1 - \alpha) S_r \quad (9)$$

其中, 通常设 $\alpha=0.5$.

2.2 实验对比

为了证明本文模型的有效性, 将其与 21 种显著性模型进行了比较, 如表 1 所示, 其中包括 LHM^[13], ACSD^[20], GP^[28], LBE^[29], DCMC^[30], CDCP^[31], DSS^[7], CTMF^[14], DF^[32], PiCANet^[9], PAGRN^[33], CDB^[34], PCA^[16], MMCI^[35], PDNet^[15], PoolNet^[36], CPD^[8], CPFP^[37], DMRA^[17], SSF^[38] 和 A2dele^[39].

如表 1 所示, 与其他同类模型相比, 本文模型在 3 个数据集上 MAE, S 和 $F_{\max}$ 这 3 个指标值基本居于前三的水平, 定量证明了本文模型的优越性.

为公平比较, 本文采用原文献给出的显著性图, 采用同一代码绘制了 P - R 曲线和 F 值曲线对比图(由于无法获取部分开源算法在 STEREO 数据集上的显著性图, 本文仅对 NJU2000 和 NLPR 数据集进行对比), 如图 4 和图 5 所示. 可以看出, 在不同阈值下, 本文模型始终具有一定的优势, 这意味着本文模型具有很好的检测显著区域和生成精确显著图的能力.

为了进一步体现本文模型的优越性, 就模型大小与其他模型进行比较, 以评估网络复杂度, 结果如图 6 所示. 可以非常直观地看出本文模型大小比综合性能最佳的 DMRA 小 81.6%, 比现有最小模型 A2dele 小 22.8%.

表 1 本文与 21 种 RGB-D 显著性检测模型在 3 个数据集上的表现

模型	NJU2000			NLPR			STEREO		
	S	$F_{\max}$	MAE	S	$F_{\max}$	MAE	S	$F_{\max}$	MAE
LHM ^[13]	0.514	0.632	0.205	0.630	0.622	0.108	0.562	0.683	0.172
ACSD ^[20]	0.699	0.711	0.202	0.673	0.607	0.179	0.692	0.669	0.200
GP ^[28]	0.527	0.647	0.211	0.654	0.611	0.146	0.588	0.671	0.182
LBE ^[29]	0.695	0.748	0.153	0.762	0.745	0.081	0.660	0.633	0.250
DCMC ^[30]	0.686	0.715	0.172	0.724	0.648	0.117	0.731	0.740	0.148
CDCP ^[31]	0.669	0.612	0.180	0.727	0.645	0.112	0.713	0.664	0.149
DSS ^[7]	0.807	0.776	0.108	0.816	0.755	0.076	0.841	0.814	0.087
CTMF ^[14]	0.849	0.845	0.085	0.860	0.825	0.056	0.848	0.786	0.087
DF ^[32]	0.763	0.804	0.141	0.802	0.682	0.099	0.757	0.761	0.142
PiCANet ^[9]	0.847	0.806	0.071	0.834	0.778	0.053	0.868	0.835	0.062
PAGRN ^[33]	0.829	0.827	0.081	0.844	0.795	0.051	0.851	0.856	0.067
CDB ^[34]	0.624	0.700	0.202	0.629	0.659	0.113	0.615	0.717	0.166
PCA ^[16]	0.877	0.844	0.059	0.873	0.794	0.044	0.880	0.845	0.061
MMCI ^[35]	0.877	0.852	0.079	0.856	0.815	0.059	0.873	0.812	0.080
PDNet ^[15]	0.883	0.832	0.062	0.835	0.740	0.064	0.874	0.833	0.064
PoolNet ^[36]	0.872	0.850	0.057	0.867	0.791	0.046	0.898	0.877	0.045
CPD ^[8]	0.862	0.853	0.059	0.862	0.840	0.037	0.885	0.880	0.046
CPFP ^[37]	0.878	0.877	0.053	0.888	0.867	0.036	0.879	0.874	0.051
DMRA ^[17]	0.886	0.886	0.051	0.899	0.879	0.030	0.886	0.886	0.047
SSF ^[38]	0.899	0.886	0.043	0.914	0.875	0.026	0.893	0.880	0.044
A2dele ^[39]	0.851	0.874	0.051	0.867	0.878	0.028	0.868	0.884	0.043
本文	0.888	0.892	0.047	0.873	0.890	0.043	0.893	0.904	0.044

图 4 不同模型 P - R 曲线对比 图 5 不同模型 F 值曲线对比

图 6 模型大小比较

为了更加直观地表现出本文模型的优越性, 对上表 1 中部分开源模型进行了可视化, 如图 7 所示, 前 5 行为分割效果较好的图像, 通过对比可知, 本文模型不仅能够突出目标区域, 而且在小对象(第 1, 3, 4, 5 行)场景下也能够有效地抑制背景噪声, 同时处理前景与背景不平衡的问题; 由最后一行分割效果较差的图像可知, 本文模型在处理细节信息上还有待加强, 但其效果也明显优于其他模型。

2.3 消融实验

本节进行了消融实验, 以更好地了解不同模

块对模型的影响, 考虑了以下 4 种模型:

- (1) 骨干网络, 既不使用 GL Block, 也不使用 CFFM;
- (2) 仅使用 GL Block;
- (3) 仅使用 CFFM;
- (4) 同时使用 GL Block 和 CFFM。

如表 2 所示, 从 MAE, $F_{\max}$ 等指标定量分析, 本文提出的 CFFM 和 GL Block 在 3 个数据集上各项指标均具有不同程度的提升并且拥有较强的泛化能力; 与骨干网络相比, 当同时使用 CFFM 和 GL Block 时, 在 NJU2000, NLPR 和 STEREO 这 3 个数据集上 MAE 分别降低了 0.021, 0.032, 0.036, $F_{\max}$ 值分别提升了 0.057, 0.145, 0.075, 提升明显, 且具有较强的鲁棒性。

如图 8 和图 9 所示, 本文绘制了 4 种模型在 3 个数据集上的 P - R 曲线和 F 值曲线。由图 8 可知, 同时使用 CFFM 和 GL Block 的模型 P - R 曲线可以将其他曲线完全包住, 表明其性能优于仅使用 CFFM 或 GL Block 的模型。由图 9 可知, 同时使用 CFFM 和 GL Block 的模型在不同阈值下的 F 值整体表现更加优秀。从视觉上看, 如图 10 所示, 仅使

图 7 本文模型与其他模型可视化对比

表 2 不同模块对模型性能的影响

模块	NJU2000			NLPR			STEREO		
	S	MAE	$F_{\max}$	S	MAE	$F_{\max}$	S	MAE	$F_{\max}$
骨干网络	0.833	0.068	0.835	0.742	0.075	0.745	0.789	0.080	0.829
CFFM	0.847	0.063	0.851	0.844	0.051	0.839	0.841	0.061	0.857
GL Block	0.851	0.060	0.856	0.828	0.049	0.832	0.867	0.061	0.859
GL Block +CFFM	0.888	0.047	0.892	0.873	0.043	0.890	0.893	0.044	0.904

图 8 不同模型 P - R 曲线 图 9 不同模型 F 值曲线

图 10 不同模型可视化对比

用 CFFM 和仅使用 GL Block 的模型所生成的显著性图能够有效地检测出显著性目标,但同时使用 CFFM 和 GL Block 的模型所生成的显著性图细节信息更丰富,轮廓更分明。

3 结 语

本文提出了一种基于 RGB-D 跨模态特征融合模块的显著性目标检测模型,首先利用 GL Block 提取全局与局部特征并搭建了改进的 FCN,同时通过 CFFM 融合较高层 RGB-D 特征,减少特征之间的差异性,最后利用 Inception 结构对初始显著

性图充分融合,增强网络的尺度适应性。在 NJU2000, NLPR, STEREO 这 3 个公开的 RGB-D 基准数据集上进行充分实验与比较,表明本文模型具有良好的检测效果,具有较强的鲁棒性,并且网络模型较小,可移植性较高。进一步的研究可以着眼于如何更好地融合 RGB 和深度信息,充分利用深度和 RGB 互补信息,丰富图像细节,以得到更好的显著性检测结果。

参考文献(References):

- [1] Zheng L, Wang S J, Liu Z Q, *et al.* Fast image retrieval: query

- pruning and early termination[J]. *IEEE Transactions on Multimedia*, 2015, 17(5): 648-659
- [2] Liu Feng, Shen Tongsheng, Lou Shuli, *et al.* Deep network saliency detection based on global model and local optimization[J]. *Acta Optica Sinica*, 2017, 37(12): 264-272(in Chinese) (刘峰, 沈同圣, 娄树理, 等. 全局模型和局部优化的深度网络显著性检测[J]. *光学学报*, 2017, 37(12): 264-272)
 - [3] Li G B, Yu Y Z. Visual saliency based on multiscale deep features[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 5455-5463
 - [4] 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
 - [5] Hadizadeh H, Bajić I V. Saliency-aware video compression[J]. *IEEE Transactions on Image Processing*, 2014, 23(1): 19-33
 - [6] Mahadevan V, Vasconcelos N. Saliency-based discriminant tracking[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2009: 1007-1013
 - [7] Hou Q B, Cheng M M, Hu X W, *et al.* Deeply supervised salient object detection with short connections[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 3203-3212
 - [8] Wu Z, Su L, Huang Q M. Cascaded partial decoder for fast and accurate salient object detection[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2019: 3907-3916
 - [9] Liu N, Han J W, Yang M H. PiCANet: learning pixel-wise contextual attention for saliency detection[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 3089-3098
 - [10] Feng M Y, Lu H C, Ding E. Attentive feedback network for boundary-aware salient object detection[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2019: 1623-1632
 - [11] Qin X B, Zhang Z C, Huang C Y, *et al.* BASNet: boundary-aware salient object detection[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2019: 7479-7489
 - [12] Ciptadi A, Hermans T, Rehg J M. An in depth view of saliency[C] // *Proceedings of British Machine Vision Conference*. Durham: BMVA Press, 2013: 112.1-112.11
 - [13] Peng H W, Li B, Xiong W H, *et al.* RGBD salient object detection: a benchmark and algorithms[C] // *Proceedings of European Conference on Computer Vision*. Heidelberg: Springer, 2014: 92-109
 - [14] Han J W, Chen H, Liu N, *et al.* CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion[J]. *IEEE Transactions on Cybernetics*, 2018, 48(11): 3171-3183
 - [15] Zhu C B, Cai X, Huang K, *et al.* PDNet: prior-model guided depth-enhanced network for salient object detection[C] // *Proceedings of the IEEE International Conference on Multimedia and Expo*. Los Alamitos: IEEE Computer Society Press, 2019: 199-204
 - [16] Chen H, Li Y F. Progressively complementarity-aware fusion network for RGB-D salient object detection[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 3051-3060
 - [17] 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 International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2019: 7254-7263
 - [18] 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
 - [19] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440
 - [20] Ju R, Ge L, Geng W J, *et al.* Depth saliency based on anisotropic center-surround difference[C] // *Proceedings of the IEEE International Conference on Image Processing*. Los Alamitos: IEEE Computer Society Press, 2014: 1115-1119
 - [21] Niu Y Z, Geng Y J, Li X Q, *et al.* Leveraging stereopsis for saliency analysis[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2012: 454-461
 - [22] Szegedy C, Liu W, Jia Y Q, *et al.* Going deeper with convolutions[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 1-9
 - [23] Liu W, Rabinovich A, Berg A C. ParseNet: looking wider to see better[OL]. [2020-10-12]. <https://arxiv.org/abs/1506.04579>
 - [24] Hu J, Shen L, Albanie S, *et al.* Squeeze-and-excitation networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(8): 2011-2023
 - [25] Wei J, Wang S H, Huang Q M. F³Net: fusion, feedback and focus for salient object detection[OL]. [2020-10-12]. <https://arxiv.org/abs/1911.11445>
 - [26] Lin T Y, Goyal P, Girshick R, *et al.* Focal loss for dense object detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(2): 318-327
 - [27] Kingma D P, Ba J. Adam: a method for stochastic optimization[OL]. [2020-10-12]. <https://arxiv.org/abs/1412.6980v4>
 - [28] Ren J Q, Gong X J, 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*. Los Alamitos: IEEE Computer Society Press, 2015: 25-32
 - [29] Feng D, Barnes N, You S D, *et al.* Local background enclosure for RGB-D salient object detection[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 2343-2350
 - [30] 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
- [31] Zhu C B, Li G, Wang W M, *et al.* An innovative salient object detection using center-dark channel prior[C] //Proceedings of the IEEE International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2017: 1509-1515
- [32] Qu L Q, He S F, Zhang J W, *et al.* RGBD salient object detection via deep fusion[J]. IEEE Transactions on Image Processing, 2017, 26(5): 2274-2285
- [33] Zhang X N, Wang T T, Qi J Q, *et al.* Progressive attention guided recurrent network for salient object detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 714-722
- [34] 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
- [35] 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
- [36] Liu J J, Hou Q, Cheng M M, *et al.* A simple pooling-based design for real-time salient object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3917-3926
- [37] Zhao J X, Cao Y, Fan D P, *et al.* Contrast prior and fluid pyramid integration for RGBD salient object detection[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 3927-3936
- [38] 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: 3472-3481
- [39] Piao Y R, Rong Z K, Zhang M, *et al.* A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9060-9069
- [40] Zhao J X, Liu J J, Fan D P, *et al.* EGNet: edge guidance network for salient object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 8779-8788