BBS-Net: 基于二分支主干策略的 RGB-D 显著目标检测网络

范登平[1,*], 翟英杰[2,*], Ali Borji[3], 杨巨峰[2,#], 邵岭[1,4]

[1] 起源人工智能研究院，阿布扎比，阿联酋
[2] 南开大学, 天津，中国
[3] 美国 HCL 公司，纽约，美国
[4] 穆罕默德 · 本 · 扎耶德人工智能大学, 阿布扎比，阿联酋
dengpfan@gmail.com, zhaiyingjie@mail.nankai.edu.cn
https://github.com/zyjwuyan/BBS-Net

摘要 多尺度特征融合在计算机视觉领域是一个经典问题, 在检测、分割和分类不同尺度的物体时尤为重要。然而与之相比，如何最优化融合多尺度及多模态的特征是一个更具挑战的难题。在这篇文章中, 我们利用 RGB-D 固有的多尺度和多模态特点, 首次提出了一个新颖的级联改良网络。具体做法是, 我们 1) 提出了一种二分支主干策略 (bifurcated backbone strategy, BBS [5]) 把多尺度特征分为了教师特征和学生特征，2）利用深度增强模块 (depth-enhanced module, DEM) 从通道和空间的角度提取有用的深度特征信息，从而以互补的方式融合 RGB 和深度模态。我们把这种简单有效的架构命名为 *BBS-Net*, 它能够应用于各种主干网络，其性能在 7 个数据集的 4 个评价指标上超过 18 种当前最先进的方法。

关键词: RGB-D 显著目标检测 · 二分支主干策略网络

1 简介

多模态多尺度特征融合 [38] 在很多计算机视觉任务（如，目标检测 [8, 21, 27, 41, 66]，语义分割 [30, 31, 33, 63]，共同显著目标检测 [19, 68]，物体分类 [2, 71]）中起到关键作用。在这篇文章中，我们尝试利用多模态多尺度特征融合进行 RGB-D 显著目标检测 [4, 70]，其中，RGB-D 显著目标检测任务旨在从 RGB 和深度图像中发现并分割出视觉上最显著的物体 [2, 71]。为了充分结合 RGB 和深度模态进行显著目标检测，研究者们已经探索了几种

[5] 本文为 ECCV2020 论文 [22] 的中文翻译版

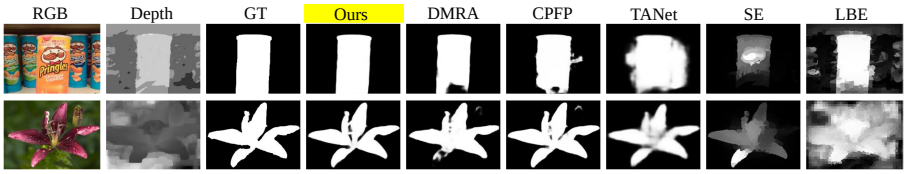

图 1: 用最先进的基于深度网络的模型（DMRA [50], CPFP [70], TANet [4], 和本文的 *BBS-Net*) 和基于手工特征的方法（SE [28] 和 LBE [25]）产生的显著物体检测结果图。本文的方法能够生成更高质量的显著图并且在极具挑战的场景下还能够抑制背景的干扰（第一行：复杂的背景；第二行：具有噪声的深度图像）。

多模态策略 [3, 5] 并且取得了一些鼓舞人心的结果。然而，现有的显著目标检方法仍然存在着两方面挑战：

(1) **如何有效整合多模态特征**。就像 [43, 60] 讨论的一样，教师特征[6]包含了具有判别力的语义信息从而能够为定位显著物体提供强力的指导，而学生特征携带着丰富的细节信息，从而对优化边缘很有用处。因此，以往的 RGB-D 显著目标检测方法往往致力于通过逐步融合策略 [46, 72] 或者专门的聚合策略 [50, 70] 来整合多级特征。然而，这些操作都是直接对不同等级的特征进行融合而没有考虑到不同等级特征的特点，因此在低级特征容易引入噪声 [4, 62]。所以，这些方法往往会容易被背景干扰（如图1的第一行）。

(2) **如何从深度模态中提取有用信息**。以前的方法在结合 RGB 和深度信息的时候通常将深度图当做第四通道的输入 [13, 49]，通过简单的相加操作 [23, 24] 或者相乘操作 [9, 74] 进行混合。这些方法把深度信息和 RGB 信息同等对待，忽略了一个事实——深度图主要关注物体间的空间距离，而 RGB 负责捕获颜色和纹理信息。因此，由于这些模态的差异，这种简单的操作不是很奏效。除此之外，很多情况下深度图的质量往往很低，这可能会向网络中引入一些冗余的特征和噪音。比如，图1中第二行的深度图既模糊又含有噪声，这就是很多方法（如 DMRA-iccv19 [50]）不能检测出完整显著物体的原因。

为了解决这些问题，我们提出了一个新颖的二分支主干策略网络 (***BBS-Net***) 用于 RGB-D 显著目标检测。如图2 (b) 所示，*BBS-Net* 包含了两个级联解码阶段。在第一个阶段，用一个标准的解码器 $\mathbf{F}_{CD1}$ 聚合教师特征生成一张初始的显著图 S_1。在第二个阶段，用初始的显著图 S_1 与学生特征相乘改善学生特征，然后另外一个解码器 $\mathbf{F}_{CD2}$ 聚合改良后的特征生成最终的显

[6] 我们在这篇文章中交替使用 '高级特征 & 低级特征' 和 '教师特征 & 学生特征'。

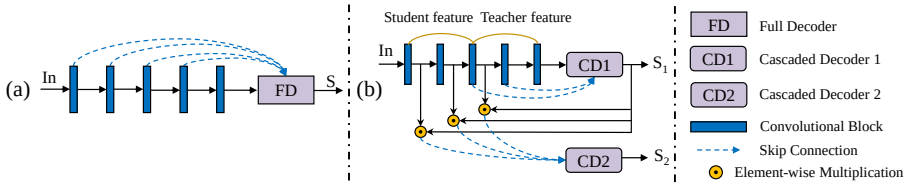

图 2: (a) 现有的 RGB-D 显著目标检测领域的不同等级特征的聚合方法 [3, 4, 46, 50, 59, 70, 72]。(b) 在本文中，我们建议利用一种二分支主干策略把不同层级的特征分为学生特征和教师特征。初始的显著图 S_1 被用于改良学生特征，从而有效地抑制噪声信息，然后改善后的特征通过一个级联解码器就能够生成最终的预测图 S_2。

著图 S_2。据我们所知，*BBS-Net* 是 RGB-D 任务上第一个探索级联改良策略的网络，我们的贡献点包括：

1 为了充分有效利用不同等级的特征，我们提出了一种二分支主干策略（BBS）。这种策略的依据是高级特征能够提供具有判别力的语义信息，这有助于消除低级特征的冗余。
2 为了充分地从深度图中获取有用的信息并且改善 RGB 和深度特征的匹配能力，我们引入了一个深度增强模块（depth-enhanced module，DEM），它包含了两个连续的注意力机制（通道注意力和空间注意力），通道注意力利用了深度特征通道间的联系而空间注意力能够发现哪些深度区域更具判别性。
3 我们的 *BBS-Net* 在 7 个公共数据集上性能大幅度超过了 18 个 SOTA 方法。此外，实验显示我们的框架适用于多种主干网络，这意味着，我们的基于级联改良策略的二分支主干网络对于不同等级和模态特征的融合具有普遍意义。

2 相关工作

尽管近年来基于 RGB 图片的显著目标检测 [7, 39, 57, 65, 67] 已经被广泛地研究，但是大部分算法还是很难处理复杂的场景（比如，凌乱的背景 [16]，低对比度的场景，光照变化的情况）[4, 50]。作为 RGB 模态的互补模态，深度模态包含丰富的空间距离信息 [50]，并且能够有助于理解复杂的场景。因此，研究者们开始结合 RGB 和深度信息来解决显著目标检测问题 [15]。

传统模型. 早期的 RGB-D 算法主要利用手工特征 [9, 74]。一些算法依靠基于对比的知识，通过计算颜色、边缘、纹理和区域的对比来计算局部区域的显著性。比如 [15] 利用区域对比计算分割区域的对比强度。在 [10] 中，每个像素的显著性值依靠颜色对比和表面法线计算。然而这些局部对比方法主要关注显著目标的边界，并且容易被高频信息所干扰 [51]。因此改良算法（比如全局对比 [11]，空间先验 [9]，背景先验 [53]）通过结合局部和全局信息进行显著性计算。为了有效地结合 RGB 模态和深度模态中的显著性信息，研究者们研究了不同的融合策略。[13, 49] 把深度图像当做是除 RGB 以外的第四个通道的输入一起处理他们（早期融合）。这种操作看起来很简单但是忽略了 RGB 模态和深度模态的不同之处，因此往往不能够得到可靠的结果。为了有效地从两种模态中独立地提取显著性信息，一些算法 [23, 74] 先分别利用两个主干网络预测显著图，然后再结合两个显著图的结果（后期融合）。除此之外，考虑到 RGB 模态和深度模态可能会相互影响，其他一些方法将两个分支网络提取到的 RGB 特征和深度特征融合，然后利用融合后的特征进行显著性预测（中期融合）。实际上，这三种融合策略在深度方法中也很常见，我们提出的模型可以看做是一种中期融合策略。

深度模型. 早期的深度方法 [51, 53] 首先提取手工特征，然后把手工特征输入深度网络中计算显著性得分。然而这些方法需要提前设计低级的手工特征，并且不能以端对端的形式训练。最近，研究者们以一种自底向上的方式 [29] 利用深度网络提取 RGB 和深度特征。相对于手工设计的特征，深度特征包含更多的语义和上下文信息，因此能够更好地捕获 RGB 和深度模态的表示并得到令人鼓舞的结果。这些深度模型 [5, 50] 的成功源于两方面的融合，一方面是从不同网络层次中提取多尺度特征，然后有效地融合他们；另一方面是融合了来自两个不同模态的特征。为了有效地聚合多尺度特征，研究者们设计了不同的网络结构。比如 [46] 把一个四通道的 RGB-D 图像输入到主干网络中，然后从每个小分支得到显著性输出（单流网络）。Chen 等人 [3] 利用两个主干网络分别提取 RGB 和深度特征，然后用级联互补的策略进行融合（双流网络）。后来，为了以自底向上的形式进行多模态互补，Chen 等人 [4] 提出了一种包含两个独立模态的主干网络和一个并行的交叉模态蒸馏分支的三流网络用于学习互补信息（三流网络）。然而，深度图像经常是比较低质量的并且包含噪声和干扰信息，这极大地影响了显著性模型的性能。为了解决这个问题，Zhao 等人 [70] 设计了一个对比增强的网络通过改善深度图像的质量来提高性能。Fan 等人 [20] 提出了一个深度过滤单元

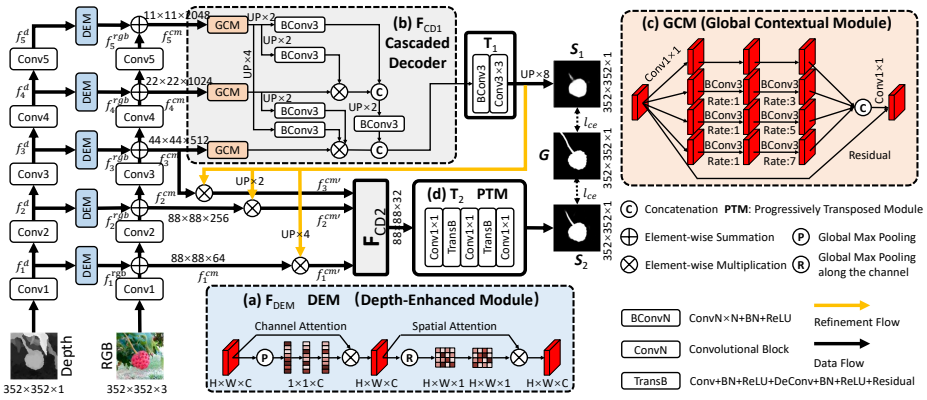

图 3: ***BBS-Net* 的网络架构**。特征提取阶段: '*Conv1*'∼'*Conv5*' 代表了来自 ResNet-50 [32] 的不同中间层。来自深度分支的多尺度特征 $f_1^d \sim f_5^d$ 首先被深度增强单元 DEM 增强，然后和来自 RGB 分支的特征 $f_1^{rgb} \sim f_5^{rgb}$ 融合。阶段 1: 混合模态教师特征 $f_3^{cm} \sim f_5^{cm}$ 首先通过级联解码器 (b) 生成初始显著图 S_1。阶段 2: 学生特征 $f_1^{cm} \sim f_3^{cm}$ 被初始显著图 S_1 校正，然后用另外一个级联解码器预测最终的显著图 S_2。

评估深度图像的质量，并且能够过滤掉那些质量低影响性能的深度图像。最近的著名工作也探索了不确定性 [64]、双边注意力机制 [69]、图网络 [47] 和共同学习策略 [26]，都取得了较好的性能。

3 方法

3.1 概览

流行的 RGB-D 显著目标检测模型倾向于直接聚合不同等级的特征 (图2(a))。如图3所示，*BBS-Net* 的网络流不同于上述介绍的模型。我们首先在 § 3.2介绍基于级联改良策略的二分支主干策略网络，然后在 § 3.3介绍我们引入的用于挖掘有用深度信息的深度增强单元。

3.2 二分支主干策略 (BBS)

我们建议充分利用高层多模态特征中挖掘的语义信息并且以一种级联改良的方式来抑制背景的干扰。我们采用了一种二分支主干策略把不同等级的多模态特征分为了两组，即 $\mathbf{Q}_1 = \{Conv1,\ Conv2,\ Conv3\}$ 和 $\mathbf{Q}_2 = \{Conv3,\ Conv4,\ Conv5\}$，其中 *Conv3* 是分割点。这样的话，每个小组内的特征仍然保留着原来的多尺度信息。

级联改良策略. 为了有效地利用来自两个分组的特征，整个网络采用了一种级联改良策略。这个策略首先用三种混合模态的教师特征 $\mathbf{Q}_2$ 生成初始显著图，然后用初始显著图 S_1 改良学生特征。利用这种机制，我们的模型能够迭代地改良低级特征的细节信息。其依据是：高级特征包含丰富的有利于定位显著性物体的全局环境信息，低级特征携带了大量的细节信息，有助于改善边缘。换句话说，这个策略通过利用不同层级的特征的特性，能够有效地消除低级多模态特征的噪声信息，并且以一种级联改良的方式生成最终的预测图像。

具体来说，我们首先通过融合 RGB 特征和被 DEM (图3(a)) 增强过的深度特征计算混合模态特征 $\{f_i^{cm}; i = 1, 2, ..., 5\}$。在第一个阶段，三种混合模态的教师特征 $f_3^{cm}, f_4^{cm}, f_5^{cm}$ 被第一个级联解码器进行聚合，这个过程定义为：

$$S_1 = \mathbf{T}_1\big(\mathbf{F}_{CD1}(f_3^{cm}, f_4^{cm}, f_5^{cm})\big), \tag{1}$$

其中 S_1 是初始的显著图，$\mathbf{F}_{CD1}$ 是第一个级联解码器，$\mathbf{T}_1$ 是两个简单的卷积层，把通道数从 32 变为 1。在阶段 2，初始显著图 S_1 被用于改良三级混合模态学生特征，这个过程定义为：

$$f_i^{cm'} = f_i^{cm} \odot S_1, \tag{2}$$

其中 $f_i^{cm'}$ $(i \in \{1, 2, 3\})$ 代表改良后的特征，$\odot$ 表示元素级别上的相乘操作。然后，三种改良后的学生特征被另外一个级联解码器聚合，紧接着后面是一个级联上采样模块，这个过程定义为：

$$S_2 = \mathbf{T}_2\Big(\mathbf{F}_{CD2}(f_1^{cm'}, f_2^{cm'}, f_3^{cm'})\Big), \tag{3}$$

其中，S_2 是最终的显著图。$\mathbf{T}_2$ 代表级联上采样模块，$\mathbf{F}_{CD2}$ 表示第二个级联解码器。最后，我们用下面的损失函数来共同优化这两个阶段：

$$\mathcal{L} = \alpha \ell_{ce}(S_1, G) + (1 - \alpha)\ell_{ce}(S_2, G), \tag{4}$$

其中，ℓ_{ce} 为广泛使用的二元交叉熵损失函数，$\alpha \in [0, 1]$ 控制两部分损失的权值大小。ℓ_{ce} 的计算过程如下：

$$\ell_{ce}(S, G) = G \log S + (1 - G) \log(1 - S), \tag{5}$$

其中，S 是预测的显著图，G 表示真实的二值显著图。

级联解码器. 对于由来自网络中不同层的 RGB 特征和深度特征融合生成的两组多等级多模态的特征 $\{f_i^{cm}, f_{i+1}^{cm}, f_{i+2}^{cm}\}, i \in \{1,3\}$，我们需要高效地利用每一个组内特征的多尺度和多等级信息来实现我们的级联改良策略。因此，我们引入了一个轻量的级联解码器来聚合两组多等级多模态特征。如图3(b)所示，级联解码器包含了三个全局环境单元（GCM）和一个简单的特征聚合策略。GCM 是在 RFB [45] 基础上改进的，它包含了额外的一个扩大感受野的分支和一个残差连接以便保存原有信息。具体而言，如图3(c) 所示，GCM 模块包含了四个平行分支。对于所有分支，首先利用一个 1×1 卷积操作把特征图的通道数降为 32。然后对于第 k ($k \in \{2,3,4\}$) 个分支，我们先进行一个卷积核为 $2k-1$，膨胀率为 1 的卷积操作。然后紧接着，我们进行一个卷积核为 3，膨胀率为 $2k-1$ 的卷积操作。这里的目标是从多模态特征中提取全局环境信息。接下来，四个分支的输出在通道层面上拼接起来，并且再用一个 1×1 卷积把输出的通道数降为 32。最终，拼接后的特征和输入特征形成一个残差连接。两个级联解码器的 GCM 模块的输出定义为：

$$f_i^{gcm} = \mathbf{F}_{GCM}(f_i), \tag{6}$$

为了进一步改善多模态特征的表示能力，我们利用一个金字塔式的相乘和拼接操作来聚合 GCM 输出的特征 $\{f_i^{gcm}, f_{i+1}^{gcm}, f_{i+2}^{gcm}\}, i \in \{1,3\}$。如图3(b) 所示，首先，对于每个改良后的特征 f_i^{gcm} 都会通过与比它更高级别的所有特征相乘进行更新：

$$f_i^{gcm'} = f_i^{gcm} \odot \Pi_{k=i+1}^{k_{max}} Conv\Big(\mathbf{F}_{UP}\left(f_k^{gcm}\right)\Big), \tag{7}$$

其中，$i \in \{1,2,3\}$, $k_{max} = 3$ 或者 $i \in \{3,4,5\}$, $k_{max} = 5$。$Conv(\cdot)$ 表示标准的 3×3 卷积，$\mathbf{F}_{UP}$ 表示如果特征不是相同大小时就通过上采样变为相同大小。$\odot$ 代表元素级别的相乘操作。其次，更新后的特征通过一个级联拼接操作来生成输出：

$$S = \mathbf{T}\left(\left[f_k^{gcm'}; Conv\Big(\mathbf{F}_{UP}\left[f_{k+1}^{gcm'}; Conv\left(\mathbf{F}_{UP}(f_{k+2}^{gcm'})\right)\right]\Big)\right]\right), \tag{8}$$

其中，S 是生成的显著图，$k \in \{1,3\}$，$[x;y]$ 表示 x 和 y 的拼接操作。在第一个阶段，$\mathbf{T}$ 表示两个连续的卷积层 ($\mathbf{T}_1$)，在第二个阶段，则表示 PTM 模块 ($\mathbf{T}_2$)。第二个解码器的输出 (88×88) 是真实图像分辨率 (352×352) 的四分之一，因此，直接上采样到真实图像的大小会丢失一些细节信息。为了解决这个问题，我们设计了一个简单但是有效的级联上采样模块 PTM，以一

种级联上采样的方式生成最终的显著图像 S_2。这个模块有两个基于残差连接的反卷积模块 [34] 和三个连续的 1×1 卷积组成。每个基于残差连接的上采样模块包含一个 3×3 卷积和一个基于残差连接的反卷积。

注意到我们的级联改良策略与最近的工作 R3Net [14], CRN [6] 和 RFCN [58] 是不同的。不同点主要在于在不同等级特征和初始显著图的利用上。我们的模型最显著的不同和优点在于，我们只需要一轮的显著性改良就能生成较好的结果，而同类工作往往需要更多次迭代，这会增加模型的预测时间和计算资源的消耗。另外我们的模型也不同于 CPD [62]，我们同时考虑了低级特征的细节信息以及高级特征的语义信息，并且能够同时抑制住低级特征的噪声信息向高级特征扩散。

3.3 深度增强单元 (DEM)

当融合 RGB 和深度特征的时候有两个问题需要解决。一个是由于两部分特征的模态差异要解决的匹配性问题，另一个是要解决在低质量深度图中的冗余和噪声信息。受 [61] 启发，我们引入了一个深度增强单元来改善多模态特征的匹配性并且从深度特征中提取有用信息。

具体而言，f_i^{rgb}, f_i^d 表示来自 RGB 和深度分支第 i $(i \in 1, 2, ..., 5)$ 个旁分支的输出。如图3所示，每个深度增强单元被放到每个深度旁分支的后面来增强深度特征的匹配性。这种旁侧处理过程能够增强深度特征的显著性表示能力，并且很好得保留了多等级和多尺度信息。两个模态特征的融合过程表示如下：

$$f_i^{cm} = f_i^{rgb} + \mathbf{F}_{DEM}(f_i^d), \tag{9}$$

其中，f_i^{cm} 表示第 i 层的多模态特征。如图3(a) 所示，深度增强单元包含了一个连续的通道注意力操作和一个空间注意力操作，深度增强单元的操作定义为：

$$\mathbf{F}_{DEM}(f_i^d) = \mathbf{S}_{att}\Big(\mathbf{C}_{att}(f_i^d)\Big), \tag{10}$$

其中，$\mathbf{C}_{att}(\cdot)$ 和 $\mathbf{S}_{att}(\cdot)$ 分别表示空间注意力和通道注意力操作。更具体而言：

$$\mathbf{C}_{att}(f) = \mathbf{M}\Big(\mathbf{P}_{max}(f)\Big) \otimes f, \tag{11}$$

$\mathbf{P}_{max}(\cdot)$ 表示对于每个特征图的全局最大池化操作，f 表示输入的特征图，$\mathbf{M}(\cdot)$ 是一个两层感知机，$\otimes$ 表示有维度扩展的相乘操作。空间注意力操作

定义为:

$$\mathbf{S}_{att}(f) = Conv\Big(\mathbf{R}_{max}(f)\Big) \odot f, \tag{12}$$

其中，$\mathbf{R}_{max}(\cdot)$ 表示对于特征图上沿着通道轴上的每个点的全局最大池化操作。我们的深度增强单元不同于以往的 RGB-D 模型。以前的模型融合 RGB 和深度特征的时候是通过直接的拼接 [3, 4, 72]，通过对比先验增强深度图像 [70] 或者是用一个简单的卷积层处理不同等级的深度特征 [50]。据我们所知，我们是首次在深度分支的多层输出后面引入注意力机制来获取有用信息。我们的实验（表4和图5）验证了本文方法在提高多模态特征匹配性上的有效性。而且本文用的空间和通道注意力机制与 [61] 中的操作不同，我们仅仅利用了一个单独的全局最大池化操作从深度分支中提取有用信息，并且能够同时降低模型的复杂度，这主要是基于显著目标检测在图像中发现最重要的区域的目的。

4 实验

4.1 实验设置

数据集. 我们在七个数据集上进行了实验，包括 NJU2K [35], NLPR [49], STERE [48], SIP [20], DES [9], LFSD [40], 和 SSD [73]。

训练测试集划分. 按照 [50, 70] 的训练策略，我们用了 1485 张 NJU2K 的图像和 700 张 NLPR 的图像进行于训练。NJU2K 和 NLPR 剩下的图像以及整个 STERE，DES，LFSD，SSD，SIP 数据集用于测试。

评价指标. 我们用了四个广泛使用的评价指标，包括 S-measure (S_α) [17], maximum E-measure (E_ξ) [18], maximum F-measure (F_β) [1], 和 mean absolute error (MAE)。评价代码见http://dpfan.net/d3netbenchmark/。

对比方法. 我们对比了十个手工特征模型 [9, 12, 25, 28, 35, 42, 49, 52, 55, 74]，以及八个深度模型 [3–5, 29, 50, 51, 59, 70]。采用他们的默认设置重新训练了上述模型，对于那些没有开源代码的模型，则使用他们公布的结果作为对比。

模型速度. 当批次大小设置为 1 和 10 的时候，BBSNet 模型的速度分别为 14fps 和 18fps。

实现细节. 模型实现采用了 PyTorch [56] 框架，并且在一块 1080Ti GPU 上做了实验。ImageNet [37] 上的预训练模型被用来初始化我们的主干模型 ResNet-50 [32] 的参数。同时去掉了 ResNet50 最后的池化层和全连接层，并且把中间 5 个卷积块的输出作为旁支输出。RGB 分支和深度分支是不共享

表 1: 七个数据集上四个评价指标 S-measure (S_α), max F-measure (F_β), max E-measure (E_ξ) 和 MAE (M) 上的对比结果，↑ (↓) 表示越高 (低) 越好。每行最好的性能用粗体表示。

数据集		指标	手工方法										深度方法								
			LHM	CDB	DESM	GP	CDCP	ACSD	LBE	DCMC	MDSF	SE	DF	AFNet	CTMF	MMCI	PCF	TANet	CPFP	DMRA	BBS-Net
			[49]	[42]	[9]	[52]	[74]	[35]	[25]	[12]	[55]	[28]	[51]	[59]	[29]	[5]	[3]	[4]	[70]	[50]	Ours
NJU2K	[35]	$S_\alpha\uparrow$	.514	.624	.665	.527	.669	.699	.695	.686	.748	.664	.763	.772	.849	.858	.877	.878	.879	.886	**.921**
		$F_\beta\uparrow$	.632	.648	.717	.647	.621	.711	.748	.715	.775	.748	.804	.775	.845	.852	.872	.874	.877	.886	**.920**
		$E_\xi\uparrow$	.724	.742	.791	.703	.741	.803	.803	.799	.838	.813	.864	.853	.913	.915	.924	.925	.926	.927	**.949**
		$M\downarrow$	.205	.203	.283	.211	.180	.202	.153	.172	.157	.169	.141	.100	.085	.079	.059	.060	.053	.051	**.035**
NLPR	[49]	$S_\alpha\uparrow$	.630	.629	.572	.654	.727	.673	.762	.724	.805	.756	.802	.799	.860	.856	.874	.886	.888	.899	**.930**
		$F_\beta\uparrow$	.622	.618	.640	.611	.645	.607	.745	.648	.793	.713	.778	.771	.825	.815	.841	.863	.867	.879	**.918**
		$E_\xi\uparrow$	.766	.791	.805	.723	.820	.780	.855	.793	.885	.847	.880	.879	.929	.913	.925	.941	.932	.947	**.961**
		$M\downarrow$	.108	.114	.312	.146	.112	.179	.081	.117	.095	.091	.085	.058	.056	.059	.044	.041	.036	.031	**.023**
STERE	[48]	$S_\alpha\uparrow$	.562	.615	.642	.588	.713	.692	.660	.731	.728	.708	.757	.825	.848	.873	.875	.871	.879	.835	**.908**
		$F_\beta\uparrow$	.683	.717	.700	.671	.664	.669	.633	.740	.719	.755	.757	.823	.831	.863	.860	.861	.874	.847	**.903**
		$E_\xi\uparrow$	.771	.823	.811	.743	.786	.806	.787	.819	.809	.846	.847	.887	.912	.927	.925	.923	.925	.911	**.942**
		$M\downarrow$	.172	.166	.295	.182	.149	.200	.250	.148	.176	.143	.141	.075	.086	.068	.064	.060	.051	.066	**.041**
DES	[9]	$S_\alpha\uparrow$	.562	.645	.622	.636	.709	.728	.703	.707	.741	.741	.752	.770	.863	.848	.842	.858	.872	.900	**.933**
		$F_\beta\uparrow$	.511	.723	.765	.597	.631	.756	.788	.666	.746	.741	.766	.728	.844	.822	.804	.827	.846	.888	**.927**
		$E_\xi\uparrow$	.653	.830	.868	.670	.811	.850	.890	.773	.851	.856	.870	.881	.932	.928	.893	.910	.923	.943	**.966**
		$M\downarrow$	.114	.100	.299	.168	.115	.169	.208	.111	.122	.090	.093	.068	.055	.065	.049	.046	.038	.030	**.021**
LFSD	[40]	$S_\alpha\uparrow$	.553	.515	.716	.635	.712	.727	.729	.753	.694	.692	.783	.738	.788	.787	.786	.801	.828	.839	**.864**
		$F_\beta\uparrow$	.708	.677	.762	.783	.702	.763	.722	.817	.779	.786	.817	.744	.787	.771	.775	.796	.826	.852	**.859**
		$E_\xi\uparrow$	.763	.766	.811	.824	.780	.829	.797	.856	.819	.832	.857	.815	.857	.839	.827	.847	.863	.893	**.901**
		$M\downarrow$	.218	.225	.253	.190	.172	.195	.214	.155	.197	.174	.145	.133	.127	.132	.119	.111	.088	.083	**.072**
SSD	[73]	$S_\alpha\uparrow$	.566	.562	.602	.615	.603	.675	.621	.704	.673	.675	.747	.714	.776	.813	.841	.839	.807	.857	**.882**
		$F_\beta\uparrow$	.568	.592	.680	.740	.535	.682	.619	.711	.703	.710	.735	.687	.729	.781	.807	.810	.766	.844	**.859**
		$E_\xi\uparrow$	.717	.698	.769	.782	.700	.785	.736	.786	.779	.800	.828	.807	.865	.882	.894	.897	.852	.906	**.919**
		$M\downarrow$	.195	.196	.038	.180	.214	.203	.278	.169	.192	.165	.142	.118	.099	.082	.062	.063	.082	.058	**.044**
SIP	[20]	$S_\alpha\uparrow$	.511	.557	.616	.588	.595	.732	.727	.683	.717	.628	.653	.729	.716	.833	.842	.835	.850	.806	**.879**
		$F_\beta\uparrow$	.574	.620	.669	.687	.505	.763	.751	.618	.698	.661	.657	.712	.694	.818	.838	.830	.851	.821	**.883**
		$E_\xi\uparrow$	.716	.737	.770	.768	.721	.838	.853	.743	.798	.771	.759	.819	.829	.897	.901	.895	.903	.875	**.922**
		$M\downarrow$	.184	.192	.298	.173	.224	.172	.200	.186	.167	.164	.185	.118	.139	.086	.071	.075	.064	.085	**.055**

权重的，这两个分支的不同在于深度分支的输入是 1 个通道。其他的参数我们按照 PyTorch 的默认设置。这里使用 Adam [36] 优化模型。初始的学习率设置为 1e-4 并且每隔 60 轮下降 10 倍。输入图像都统一调整为 352×352 的尺寸。所有的训练图像用随机翻转、旋转和边界裁减进行数据增强。批次大小设置为 10 的时候，训练模型 200 轮大约需要 10 个小时。

4.2 与最先进方法的对比

结果对比. 如表1所示，我们的方法在四个评价指标 7 个数据集上超过所有的对比方法。对于最好的对比方法（ICCV'19 DMRA [50] 和 CVPR'19 CPFP [70]）我们在四个评价指标 S_α, $maxF_\beta$, $maxE_\xi$, M 上有 2.5% ∼ 3.5%, 0.7% ∼ 3.9%, 0.8% ∼ 2.3%, 0.009 ∼ 0.016 的提升幅度。

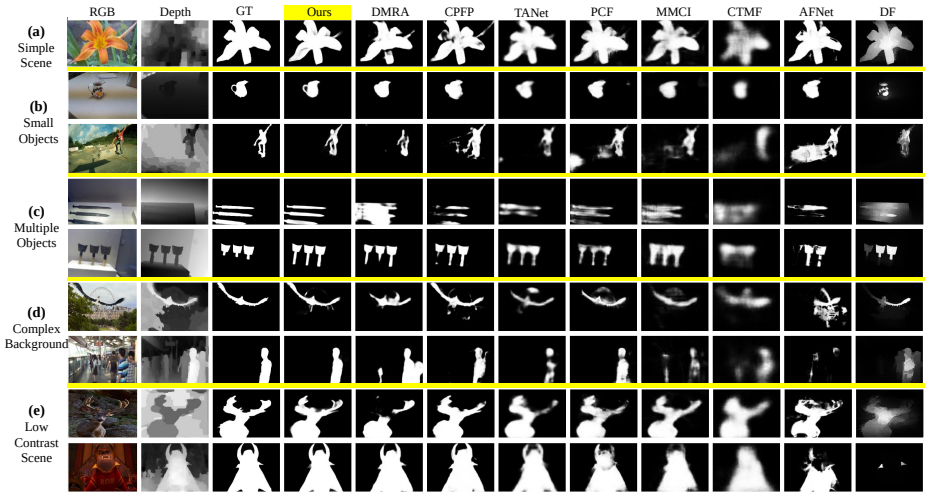

图 4: 我们的模型和八个先进方法在显著图结果上的对比。

表 2: 不同主干网络的实验。

模型	NJU2K [35]		NLPR [49]		STERE [48]		DES [9]		LFSD [40]		SSD [73]		SIP [20]	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
CPFP [70]	.879	.053	.888	.036	.879	.051	.872	.038	.828	.088	.807	.082	.850	.064
DMRA [50]	.886	.051	.899	.031	.835	.066	.900	.030	.839	.083	.857	.058	.806	.085
BBS-Net(VGG-16)	.916	.039	.923	.026	.896	.046	.908	.028	.845	.080	.858	.055	.874	.056
BBS-Net(VGG-19)	.918	.037	.925	.025	.901	.043	.915	.026	.852	.074	.855	.056	.878	**.054**
BBS-Net(ResNet-50)	**.921**	**.035**	**.930**	**.023**	**.908**	**.041**	**.933**	**.021**	**.864**	**.072**	**.882**	**.044**	**.879**	.055

视觉上的对比. 图4展示了我们的方法和几个对比方法预测的显著图。这些图像分为了简单的场景 (a) 和各种复杂度的场景，包括小物体 (b)、多个物体 (c)、复杂的背景 (d) 和低对比度场景 (e)。首先，(a) 是个简单的例子。前景的花在 RGB 图像中很明显，但是深度图质量很差且包含了一些干扰信息。因此性能最好的两个模型 DMRA 和 CPFP 都不能完整地将显著目标检测出来。而我们的方法能够更有效地利用深度图像的互补信息并且消除深度图像的副作用。第二，(b) 展示了两个小物体图像例子。尽管第一行的茶壶的把子很细小，我们的方法仍然能够很准确的检测到它。第三，(c) 展示两个包含多个物体的图像。我们的方法能够检测出所有的显著目标，并且能够很好地把它们分割出来。尽管 (c) 中第一行的深度图像没有清晰的信息，我们的算法也能够正确的预测出显著图。第四，(d) 是具有复杂背景的两个例子。我们的方法能够生成可靠的结果，而其他方法错误地把背景当成了显著性物

表 3: 不同特征融合策略的对比研究。

#	设置	NJU2K [35]		NLPR [49]		STERE [48]		DES [9]		LFSD [40]		SSD [73]		SIP [20]	
		$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
1	Low3	.881	.051	.882	.038	.832	.070	.853	.044	.779	.110	.805	.080	.760	.108
2	High3	.902	.042	.911	.029	.886	.048	.912	.026	.845	.080	.850	.058	.833	.073
3	All5	.905	.042	.915	.027	.891	.045	.901	.028	.845	.082	.848	.060	.839	.071
4	BBS-NoRF	.893	.050	.904	.035	.843	.072	.886	.039	.804.	.105	.839	.069	.843	.076
5	BBS-RH	.913	.040	.922	.028	.881	.054	.919	.027	.833	.085	.872	.053	.866	.063
6	BBS-RL (ours)	**.921**	**.035**	**.930**	**.023**	**.908**	**.041**	**.933**	**.021**	**.864**	**.072**	**.882**	**.044**	**.879**	**.055**

体。最后，(e) 展示了两个低对比度的例子，我们的方法能够通过抑制背景的干扰和从深度图中捕获有用的信息从而产生可靠的结果。

5 讨论

扩展性. 现有的深度 RGB-D 模型主要使用三种主干网络架构 VGG-16 [54], VGG-19 [54] 和 ResNet-50 [32]。为了进一步验证我们方法的扩展性，我们在表2提供了使用不同主干网络的对比。数据显示，我们的模型在这三种架构上均能够超过最先进方法 CPFP [70] 和 DMRA [50]。

融合策略. 我们进行了几方面的实验来验证我们的级联改良策略的有效性。实验结果在表3和图5(a) 中。‘Low3’ 表示我们只用一个解码器聚合低三层的特征，并且没有来自初始显著图的改良。学生特征包含丰富的有利于改善边缘的细节信息但是同时也引入了很多背景的干扰。仅仅聚合低等级的特征并不能产生较好的结果并且会产生许多模糊块（图5(a) 的第一行和第二行）或者不能定位到显著性物体（图5(a) 的第三行）。‘High3’ 仅仅聚合高等级的教师特征 *Conv3∼5* 生成显著图，与学生特征相比，教师特征是更‘老练’的，因此包含更多的语义信息，能够帮助定位显著性物体，同时也保存了一些边缘信息，因此聚合教师特征能够产生更好的结果。‘All5’ 同时用一个单独的解码器聚合 5 层特征，它能够产生与 ‘High3’ 差不多的结果，但是会产生一些由学生特征引入的噪音。‘BBS-NoRF’ 直接去掉我们的改良流，这会导致很差的结果。‘BBS-RH’ 是一种与我们的方法相反的改良策略，也就是说，先用解码器聚合低级的学生特征 *Conv1∼3* 生成初始显著图，然后用初始显著图改良高级教师特征 *Conv3∼5*，最后用解码器聚合高级教师特征生成最终的显著图。这种改良策略比我们的方法效果差，因为用这种改良策略并不能消除掉学生特征中的噪声信息。除此之外，与 ‘All5’ 相比，我们

表 4: 消融实验分析。‘BM’ 表示基础模型。‘CA’ 表示通道注意力机制。‘SA’ 是空间注意力机制。‘PTM’ 代表级联上采样模块。

#	设置				NJU2K [35]		NLPR [49]		STERE [48]		DES [9]		LFSD [40]		SSD [73]		SIP [20]	
	BM	CA	SA	PTM	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
1	✓				.908	.045	.918	.029	.882	.055	.917	.027	.842	.083	.862	.057	.864	.066
2	✓	✓			.913	.042	.922	.027	.896	.048	.923	.025	.840	.086	.855	.057	.868	.063
3	✓		✓		.912	.045	.918	.029	.891	.054	.914	.029	.855	.083	.872	.054	.869	.063
4	✓	✓	✓		.919	.037	.928	.026	.900	.045	.924	.024	.861	.074	.873	.052	.869	.061
5	✓	✓	✓	✓	**.921**	**.035**	**.930**	**.023**	**.908**	**.041**	**.933**	**.021**	**.864**	**.072**	**.882**	**.044**	**.879**	**.055**

表 5: 级联解码器的有效性分析。

设置	NJU2K [35]		NLPR [49]		STERE [48]		DES [9]		LFSD [40]		SSD [73]		SIP [20]	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
元素级相乘	.915	.037	.925	.025	.897	.045	.925	.022	.856	.073	.868	.050	**.880**	**.052**
级联解码器	**.921**	**.035**	**.930**	**.023**	**.908**	**.041**	**.933**	**.021**	**.864**	**.072**	**.882**	**.044**	.879	.055

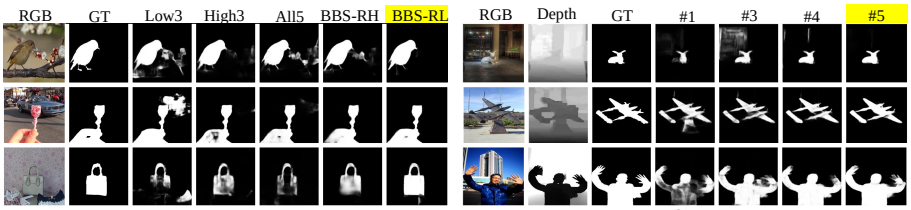

图 5: (a): 不同聚合策略的显著性结果图, (b): 不断增加不同模块的显著经结果图变化。‘#’ 表示表4的对应行。

的方法能够充分利用不同层级特征的信息，因此能够得到很好的性能，也就是有更少的背景上的干扰和更锋利的边缘（如图5(a) 的 ‘BBS-RL’ 所示）。

不同模块的有效性分析。如表4和图5(b) 所示，是多个模块有效性的消融分析。基础模型 BM 表示 *BBS-Net* 模型没有添加任何额外模块（CA，SA 和 PTM）。如表1和表4所示，注意到仅仅 BM 就几乎能在所有数据集上超过现有方法。增加通道注意力和空间注意力操作能够在多数数据集上提升模型的性能。这个结论可以从表4第二行和第三行中看出来。同时增加它们的时候（表4的第四行），相对于 BM，所有数据集上的性能都会被提升。能够很容易从图5(b) 中 ‘#3’ 和 ‘#4’ 推断出，通道注意力操作和空间注意力操作使得模型更关注深度特征中有信息量的部分，从而能够对背景干扰进行抑制。最后我们在第二个级联解码器后面加了一个级联上采样模块逐步把特征图输出

表 6: 在三个数据集上与最先进的 RGB SOD 方法的对比。‘w/o D 和 ‘w/ D’ 表示训练和测试的时候是否利用深度图信息。

方法	CPD [62]		PoolNet [43]		PiCANet [44]		PAGRN [67]		R3Net [14]		Ours (w/o D)		Ours (w/ D)	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
NJU2K [35]	.894	.046	.887	.045	.847	.071	.829	.081	.837	.092	.914	.038	**.921**	**.035**
NLPR [49]	.915	.025	.900	.029	.834	.053	.844	.051	.798	.101	.925	.026	**.930**	**.023**
DES [9]	.897	.028	.873	.034	.854	.042	.858	.044	.847	.066	.912	.025	**.933**	**.021**

为真实图像的分辨率大小。表4的第五行和图5(b) 的’#5’ 列显示‘PTM’能够实现明显的性能提升，并且能够生成更细节的边缘。

为了进一步分析级联解码器的有效性，我们把级联解码器换为一个简单的元素级别的相乘操作来进行对比。也就是说，我们首先用 1×1 卷积和上采样操作把来自不同层的特征变为相同的尺度，然后再通过元素级别上的相乘操作进行融合。表5中的实验结果证明了级联解码器的有效性。

深度图的有效性. 为了研究深度信息是否真的能够提升 SOD 的性能，我们进行了表6所示的两方面的实验。(i) 我们比较了 5 个最先进的 RGB 显著目标检测方法（CPD [62], PoolNet [43], PiCANet [44], PAGRN [67] 和 R3Net [14])。对于这些方法我们不输入深度信息，用我们的训练测试集划分重新对这些方法进行训练。实验结果显示，由于利用了深度信息，‘Ours (w/ D)’能够显著地超过对比方法。(ii) 不利用深度信息训练我们的模型‘Ours (w/o D)’，也就是把深度分支的输入设置为 0。通过比较‘Ours (w/ D)’和‘Ours (w/o D)’，深度信息能够显著地提升模型的性能。这两方面的实验共同证明了深度图带来的增益，这是因为深度图可以看做是一种先验知识，能够为显著目标检测提供空间距离信息和边缘上的指导。

6 结论

本文提出了一个新颖的多等级多模态学习的框架 *BBS-Net* 在七个 RGB-D 数据集上达到最先进的效果。本文的模型是基于级联改良的新颖的二分支主干策略。重要的是，我们提出的模型不依赖主干网络，未来也能够用于其他相关领域的研究，比如语义分割、目标检测和分类等。

参考文献

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604 (2009)

2. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE TIP **24**(12), 5706–5722 (2015)
3. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for RGB-D salient object detection. In: CVPR. pp. 3051–3060 (2018)
4. Chen, H., Li, Y.: Three-stream attention-aware network for RGB-D salient object detection. IEEE TIP **28**(6), 2825–2835 (2019)
5. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. IEEE TOC **86**, 376–385 (2019)
6. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: CVPR. pp. 1511–1520 (2017)
7. Chen, S., Tan, X., Wang, B., Lu, H., Hu, X., Fu, Y.: Reverse attention-based residual network for salient object detection. IEEE TIP **29**, 3763–3776 (2020)
8. Cheng, G., Han, J., Zhou, P., Xu, D.: Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE TIP **28**(1) (2018)
9. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: ICIMCS. pp. 23–27 (2014)
10. Ciptadi, A., Hermans, T., M. Rehg, J.: An in depth view of saliency. In: BMVC (2013)
11. Cong, R., Lei, J., Fu, H., Hou, J., Huang, Q., Kwong, S.: Going from RGB to RGBD saliency: A depth-guided transformation model. IEEE TOC pp. 1–13 (2019)
12. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE SPL **23**(6), 819–823 (2016)
13. Cong, R., Lei, J., Fu, H., Huang, Q., Cao, X., Ling, N.: HSCS: Hierarchical sparsity based co-saliency detection for RGBD images. IEEE TMM **21**(7), 1660–1671 (2019)
14. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: IJCAI. pp. 684–690 (2018)
15. Desingh, K., Krishna, K., Rajanand, D., Jawahar, C.: Depth really matters: Improving visual salient region detection with depth. In: BMVC. pp. 1–11 (2013)
16. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV. pp. 186–202 (2018)

17. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. pp. 4548–4557 (2017)
18. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: IJCAI. pp. 698–704 (2018)
19. Fan, D.P., Lin, Z., Ji, G.P., Zhang, D., Fu, H., Cheng, M.M.: Taking a deeper look at co-salient object detection. In: CVPR. pp. 2919–2929 (2020)
20. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. IEEE TNNLS (2020)
21. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: CVPR. pp. 8554–8564 (2019)
22. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: ECCV (2020)
23. Fan, X., Liu, Z., Sun, G.: Salient region detection for stereoscopic images. In: DSP. pp. 454–458 (2014)
24. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W.: Saliency detection for stereoscopic images. IEEE TIP **23**(6), 2625–2636 (2014)
25. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for RGB-D salient object detection. In: CVPR. pp. 2343–2350 (2016)
26. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In: CVPR. pp. 3052–3062 (2020)
27. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: ECCV (2020)
28. Guo, J., Ren, T., Bei, J.: Salient object detection for RGB-D image via saliency evolution. In: ICME. pp. 1–6 (2016)
29. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: CNNs-Based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE TOC **48**(11), 3171–3183 (2018)
30. Han, J., Yang, L., Zhang, D., Chang, X., Liang, X.: Reinforcement cutting-agent learning for video object segmentation. In: CVPR. pp. 9080–9089 (2018)
31. Han, Q., Zhao, K., Xu, J., Cheng, M.M.: Deep hough transform for semantic line detection. In: ECCV) (2020)
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

33. He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y.: Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In: AAAI019. pp. 8417–8424 (2019)
34. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for RGBD semantic segmentation. In: ICIP. pp. 1440–1444 (2019)
35. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: ICIP. pp. 1115–1119 (2014)
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
37. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
38. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR. pp. 5455–5463 (2015)
39. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: ICCV. pp. 7274–7283 (2019)
40. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: CVPR. pp. 2806–2813 (2014)
41. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: ECCV. pp. 355–370 (2018)
42. Liang, F., Duan, L., Ma, W., Qiao, Y., Cai, Z., Qing, L.: Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing **275**, 2227–2238 (2018)
43. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR. pp. 3917–3926 (2019)
44. Liu, N., Han, J., Yang, M.H.: PiCANet: Learning pixel-wise contextual attention for saliency detection. In: CVPR. pp. 3089–3098 (2018)
45. Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In: ECCV. pp. 404–419 (2018)
46. Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P.: Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing **363**, 46–57 (2019)
47. Luo, A., Li, X., Yang, F., Jiao, Z., Cheng, H., Lyu, S.: Cascade graph neural networks for rgb-d salient object detection. In: ECCV (2020)
48. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR. pp. 454–461 (2012)

49. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: a benchmark and algorithms. In: ECCV. pp. 92–109. Springer (2014)

50. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV. pp. 7254–7263 (2019)

51. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: RGBD salient object detection via deep fusion. IEEE TIP **26**(5), 2274–2285 (2017)

52. Ren, J., Gong, X., Yu, L., Zhou, W., Ying Yang, M.: Exploiting global priors for RGB-D saliency detection. In: CVPRW. pp. 25–32 (2015)

53. Shigematsu, R., Feng, D., You, S., Barnes, N.: Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features. In: ICCVW. pp. 2749–2757 (2017)

54. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

55. Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., Ren, T.: Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE TIP **26**(9), 4204–4216 (2017)

56. Steiner, B., DeVito, Z., Chintala, S., Gross, S., Paszke, A., Massa, F., Lerer, A., Chanan, G., Lin, Z., Yang, E., et al.: PyTorch: An imperative style, high-performance deep learning library. In: NIPS. pp. 8024–8035 (2019)

57. Su, J., Li, J., Zhang, Y., Xia, C., Tian, Y.: Selectivity or invariance: Boundary-aware salient object detection. In: ICCV. pp. 3798–3807 (2019)

58. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Salient object detection with recurrent fully convolutional networks. IEEE TPAMI **41**(7), 1734–1746 (2018)

59. Wang, N., Gong, X.: Adaptive fusion for RGB-D salient object detection. IEEE Access **7**, 55277–55284 (2019)

60. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: CVPR. pp. 3127–3135 (2018)

61. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: CBAM: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)

62. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR. pp. 3907–3916 (2019)

63. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. In: ICCV. pp. 7223–7233 (2019)

64. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: CVPR. pp. 8582–8591 (2020)

65. Zhang, L., Wu, J., Wang, T., Borji, A., Wei, G., Lu, H.: A multistage refinement network for salient object detection. IEEE TIP **29**, 3534–3545 (2020)

66. Zhang, Q., Huang, N., Yao, L., Zhang, D., Shan, C., Han, J.: Rgb-t salient object detection via fusing multi-level cnn features. IEEE TIP **29**, 3321–3335 (2020)

67. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: CVPR. pp. 714–722 (2018)

68. Zhang, Z., Jin, W., Xu, J., Cheng, M.M.: Gradient-induced co-saliency detection. In: ECCV (2020)

69. Zhang, Z., Lin, Z., Xu, J., Jin, W., Lu, S.P., Fan, D.P.: Bilateral attention network for RGB-D salient object detection. arXiv preprint arXiv:2004.14582 (2020)

70. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for RGBD salient object detection. In: CVPR. pp. 3927–3936 (2019)

71. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: EGNet: Edge guidance network for salient object detection. In: CVPR. pp. 8779–8788 (2019)

72. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: Pdnet: Prior-model guided depth-enhanced network for salient object detection. In: ICME. pp. 199–204 (2019)

73. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: ICCVW. pp. 3008–3014 (2017)

74. Zhu, C., Li, G., Wang, W., Wang, R.: An innovative salient object detection using center-dark channel prior. In: ICCVW. pp. 1509–1515 (2017)