基于对比度先验和流式金字塔的 RGBD 显著性目标检测

赵嘉星 [1,*] 曹洋 [1,*] 范登平 [1,*] 程明明 [1] 李炫毅 [1] 张乐 [2]
[1] 南开大学计算机学院 [2] 新加坡 A*STAR
https://mmcheng.net/rgbdsalpyr/

摘要

大量深度传感器为 RGBD 图像的显著性目标检测提供了有价值的补充信息。然而，由于 RGB 和深度信息之间的内在差异，使用 ImageNet 预先训练的骨干模型从深度通道中提取特征并直接与 RGB 特征融合，这种方法效果并不是最好的。在本文中，我们利用对比度先验这一过去在非深度学习的 SOD 方法中占主导地位的方法，将其引入到基于 CNNs 的体系结构中，以增强深度信息。增强的深度信息进一步调整用于 SOD 的 RGB 特征，使用一种新型的流式金字塔结构，可以更好地利用多尺度跨模态特征。在 5 个具有挑战性基准数据集上的综合实验表明，CPFP[1]结构比 9 种最先进的方法具有更好的性能。

1. 介绍

显著性目标检测 (SOD) 旨在区分场景中视觉上最独特的对象或区域。它有着广泛的应用，包括视频/图像分割 [17, 40]，目标识别 [46]，视觉跟踪 [3]，前景图评估 [14, 15]，图像检索 [6, 16, 22, 38]，内容感知图像编辑 [8]，信息检索 [59]，照片合成 [5, 29]，以及弱监督语义分割 [52]。最近，基于卷积神经网络 (CNNs) 的方法 [28, 36, 39] 已经成为 SOD 任务的主流方法，在具有挑战性的基准方面 [13] 取得了令人振奋的结果。然而，现有的基于 CNNs 的 SOD 方法主要处理 RGB 图像，当图像中的目标与背景相似时，可能会产生令人不满意的结果。

如图 1 所示，来自流行设备 (如 Kinect 和 iPhone X) 的深度信息为识别显著性物体提供了重要的补充信息。尽管在过去的几年已经提出了几种基于 RGBD 的 SOD 基准 [32, 42] 和方法 [4, 18, 20, 49]，但是如何有效地利用深度信息，特别是在深度神经网络 [4] 的背景下，仍然没有被充分探索。

图 1. RGBD 显著性数据集中的样例：NJU2000 [32], NLPR [42] 和 SSB [41]。深度信息对发现显著目标起着重要的补充作用。

如图 2 所示，现有的基于 RGBD 的 SOD 方法通常通过简单的拼接或通过早期融合 [42, 49]，后期融合 [18]，或中期融合 [20] 来融合 RGB 和深度输入/特征。我们认为，由于两个主要原因，通过简单的拼接来直接跨模态融合可能并不是最好的：

1) **缺乏高质量的深度图**

从最先进的探测设备获得的深度图比 RGB 图像噪声更大，纹理更少，这给深度特征的提取带来了挑战。我们缺乏一个具有良好预训练的骨干网络从深度图（像 ImageNet [10] 级别的大规模深度数据集）中提取有力的特征。

2) **次优多尺度跨模态融合**

深度图和 RGB 具有非常不同的特性，使得这两种模态的有效多尺度融合变得困难。例如，与其他颜色相比，“绿色”与“植物”类的相关性要强得多。但是，没有深度值具有这样的相关性。当采用诸如线性组合或

*共同一作. 程明明 (cmm@nankai.edu.cn) 是本文的通讯作者。
[1]本文为 CVPR2019 [55] 的中文翻译版。

拼接这样的简单融合策略时，两种模态之间的固有差异可能会导致不兼容问题。

我们不再使用 ImageNet 预先训练的骨干网络从深度图中提取特征，然后用现有方法 [4, 18, 20, 49] 融合 RGB 和深度信息，而是使用对比度先验来增强深度信息。然后，增强的深度图被用作注意力图来处理 RGB 特征以获得高质量的 SOD 结果。在 CNNs 普及之前，对比度先验一直是发现显著物体的主要方法，不仅在计算机视觉领域 [2, 7, 30, 43]，在神经科学 [11] 和认知心理学 [50] 中也是如此。通过重新使用具有我们对比度增强网络的对比度先验，我们桥接了来自 RGB 通道的 CNN 代表性特征和来自深度通道有力的显著性先验。具体地，我们通过测量显著区域和非显著区域之间的对比度以及它们的一致性，提出了对比度增强网络的对比度损失函数。以完全可微的方式设计的对比度增强网络可以很容易地通过反向传播进行训练，并与其他 CNN 模块一起工作。

高质量的基于 RGBD 的 SOD 需要有效的多尺度跨模态特征融合。与现有的基于 CNN [4, 27, 28, 56] 的多尺度特征融合方法不同，我们需要额外考虑特征兼容性问题。我们设计了流式金字塔结构，以分层的方式融合跨模态 (RGB 和深度) 信息。灵感来自 Hou 等人 [28] 和 Zhao 等人 [56]，我们的融合方案包含从较高 CNN 层到较低 CNN 层丰富的短连接，同时以金字塔的形式融合特征。在融合过程中，来自两个模态的特征通过几个非线性层，使得反向传播机制能够调整它们的特征表达以获得更好的兼容性。

我们通过广泛的消融实验和对比，从实验上验证了我们模型设计的有效性。即使用简单的骨干网络 (VGG-16 [48])，我们的方法与基于 RGBD 的最先进的 SOD 方法相比表现出显著的性能。总而言之，我们的主要贡献有三个方面。

- 我们设计了一种对比度损失，来利用非深度学习中广泛使用的对比度先验来增强深度图。我们基于 RGBD 的 SOD 模型成功地利用了传统对比度先验和深层 CNN 特征的优点。
- 为了更好地利用多尺度跨模态特征，我们提出了一种流式金字塔集成策略，并通过实验验证了该策略的有效性。
- 没有花里胡哨的操作，例如 HHA [24]、超像素 [54] 或者 CRF [33]，我们的模型在 5 个广泛使用的基准数据集上很大程度上超过了 9 个最先进的可替代方案。

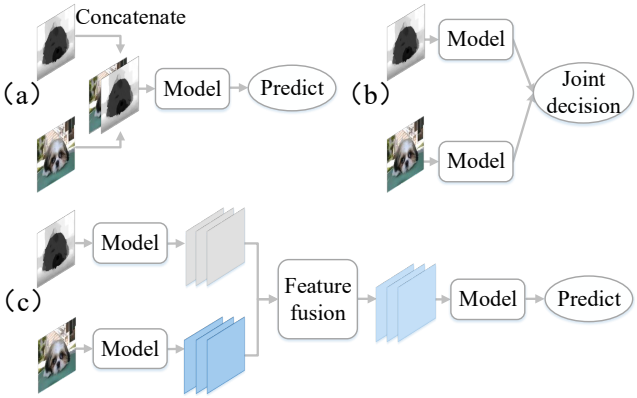

图 2. 使用深度图的三种方法。(a) 早期融合 (例如 [42, 49]) (b) 后期融合 (例如 [18]) (c) 中期融合 (例如 [20]) 详细情况在 2.2 节中介绍。

2. 相关工作

2.1. 显著性物体检测

SOD 早期的方法依赖于各种人工设计的特征 [7, 26, 37, 41]。近年来，表征学习正在成为实际的标准并且取得了很大的提高。Li 等人 [35] 通过预先训练的深度卷积网络提取每个超像素的多尺度特征来得到显著图。将每个超像素周围三个不同尺度包围盒的特征组合成一个特征向量，以整合多尺度信息。在 [57] 中，Zhao 等人提出了一种用于显著性目标检测的多语境深度学习框架，该框架分别使用两个不同的 CNN 来提取全局和局部信息。Lee 等人 [34] 综合考虑了从 CNN 中提取的高层特征和人工设计的特征。使用多个的 1×1 卷积和 RELU 层将高级特征和人工特征融合成一个特征向量。在上述方法中，输入都是超像素，模型需要多次运行才能得到显著目标的预测结果。Liu 等人 [39] 设计了一个两阶段网络，由另一个网络以分层递进的方式生成一个粗略的降维预测图，并对其进行细化。Li 等人 [36] 提出了一种深对比度网络，该网络不仅考虑了像素级信息，而且将分割级别融合到网络中。在 [28] 中介绍了一种具有短连接的深层结构，它基于 HED 结构 [53] 添加了高级特征和低级特征的连接。

2.2. 基于 RGBD 的显著性物体检测

如图 2 所示，现有的 RGBD 显著目标检测方法可以分为三类。第一种方法如图 2(a) 所示，在最早阶段融合输入，并将深度图直接视为输入的一个通道 [42, 49]。

图 2（b）代表的第二种方法采用“晚融合”策略，RGB 和深度图分别产生预测，并且结果被集成到一个单独的后处理步骤中，例如像素级加法和乘法。例如，Fan 等人 [18] 使用深度对比度和深度加权颜色对比度来测量区域的显著性值。Fang 等人 [19] 利用从 DC-T 系数中提取的深度来表示图像斑块的能量。Cheng 等人 [9] 根据视觉显著刺激在颜色空间和深度空间中的规律计算显著性。此外，Desingh 等人 [12] 利用非线性支持向量回归来融合这些预测图。第三种方法如图 2(c) 所示，它结合了从不同网络提取的深度特征和 RGB 特征。例如，Feng 等人 [20] 提出了一种新型的 RGBD 显著特征来捕捉角度方向的扩散。类似地，R.Shigematsu 等人 [47] 建议提取背景线索，以及低层深度信息。

目前，在 RGBD 显著性检测中采用 CNNs 的方法未能获得更具辨别力的基于学习的特征。如上所述，基于 CNNs 的方法几乎属于第三种方案。在 [44] 中，Qu 等人首先为每个超像素/斑块生成 RGB 和深度特征向量，然后将这些特征向量送入 CNN 得到显著性置信度值，最后用拉普拉斯传播得到最终的显著图。Han 等人 [25] 提出了一种两视 (RGB 和 Depth) CNN 从 RGB 图像和相应的深度图中提取特征，并将这些特征与新的全连接层同时连接起来，得到最终的显著图。Chen 等人 [4] 设计了一种渐进式融合方法。为了融合多尺度信息，它将所有较深层的预测跳跃连接到较浅层。而不同尺度的信息在融合前被预测为预测图，即在多尺度融合之前已经完成了特征的跨模态互补。

3. 提出的方法

CPFP 的整体结构如图 3 所示。VGG-16 采用特征增强模块 (FEM) 和流式金字塔集成。基于对比度先验，FEM 在 VGG-16 的五个阶段增强了 RGB 特征。详细细节在 3.1 节中介绍。然后利用流式金字塔对多尺度跨模态特征进行融合。细节在 3.2 节中描述。

3.1. 特征增强模块 (FEM)

我们提出通过使用来自深度图的信息增强 RGB 输入的特征。然而，简单地使用深度图进行调制可能会降低最终的性能，因为深度图通常是有噪声的。相反，我们提出了一种新的特征增强模块，该模块由用于学习增强深度图的对比度增强网络和用于特征调制的跨模态融合策略组成。特征增强模块独立于 RGB 流的网络主干。为了公平比较，这里我们使用 [4] 中建议的 VGG-16 并且删除了最后三层。VGG-16 网络包括 5 个卷积模块并且每个模块的输出分别进行 [2，4，8，16，32] 倍的下采样。如图 3 所示，我们在每个模块后边都添加一个特征增强模块 (FEM)，以获得增强的特征。FEM 包括对比度增强网络和跨模态融合，这将在 3.1.1 节和 3.1.2 节中介绍。

3.1.1 对比度增强网络 (CEN)

在前人工作 [14] 的启发，我们发现 SOD 中前景和背景的对比度以及前景的均匀分布是至关重要的特性。为了有效地利用这些特点，我们在对比度增强网络中设计了对比度损失函数。对比度增强网络的结构如图 3 所示。为了科学地测量对比度损失的影响，对于 CEN 中的其他部分，我们选择了几个共同的层和简单的结构，这不会影响性能。参数详情在 4.1 节中介绍。对比度损失包含三项：前景目标分布损失 l_f、背景分布损失 l_b 和整个深度图像分布损失 l_w。在我们的例子中，我们简单地将图像中的显著目标视为前景目标。

首先，对于前景和背景目标，增强后的深度图应该与原始深度图一致。因此，对于生成的增强深度图，前景目标分布损失 l_f 和背景分布损失 l_b 可以表示为：

$$\begin{aligned} l_f &= -\log(1-4*\sum_{(i,j)\in F}\frac{(p_{i,j}-\hat{p_f})^2}{N_f}), \\ l_b &= -\log(1-4*\sum_{(i,j)\in B}\frac{(p_{i,j}-\hat{p_b})^2}{N_b}), \end{aligned} \tag{1}$$

F 和 B 分别是真实显著图中的显著目标区域和背景。N_f 和 N_b 分别表示显著目标和背景的像素数。类似地，$\hat{p_f}$ 和 $\hat{p_b}$ 分别表示增强的深度图中的前景和背景中像素值的平均值。

$$\hat{p_f} = \sum_{(i,j)\in F}\frac{p_{i,j}}{N_f}, \hat{p_b} = \sum_{(i,j)\in B}\frac{p_{i,j}}{N_b}. \tag{2}$$

如等式（1）所定义，我们对显著目标和背景的内部方差进行建模，以提高与原始深度图的一致性。使用 Sigmoid 层将对比度增强网络的输出压缩到 [0，1] 的范围。在本例中，内部方差的最大值为 0.25，因此我们将方差乘以 4，以确保 log 函数的输入范围是从 0 到 1。

其次，应增强前景和背景目标之间的对比度。因此，我们定义整个深度图像分布损失函数 l_w 为：

$$l_w = -\log(\hat{p_f}-\hat{p_b})^2. \tag{3}$$

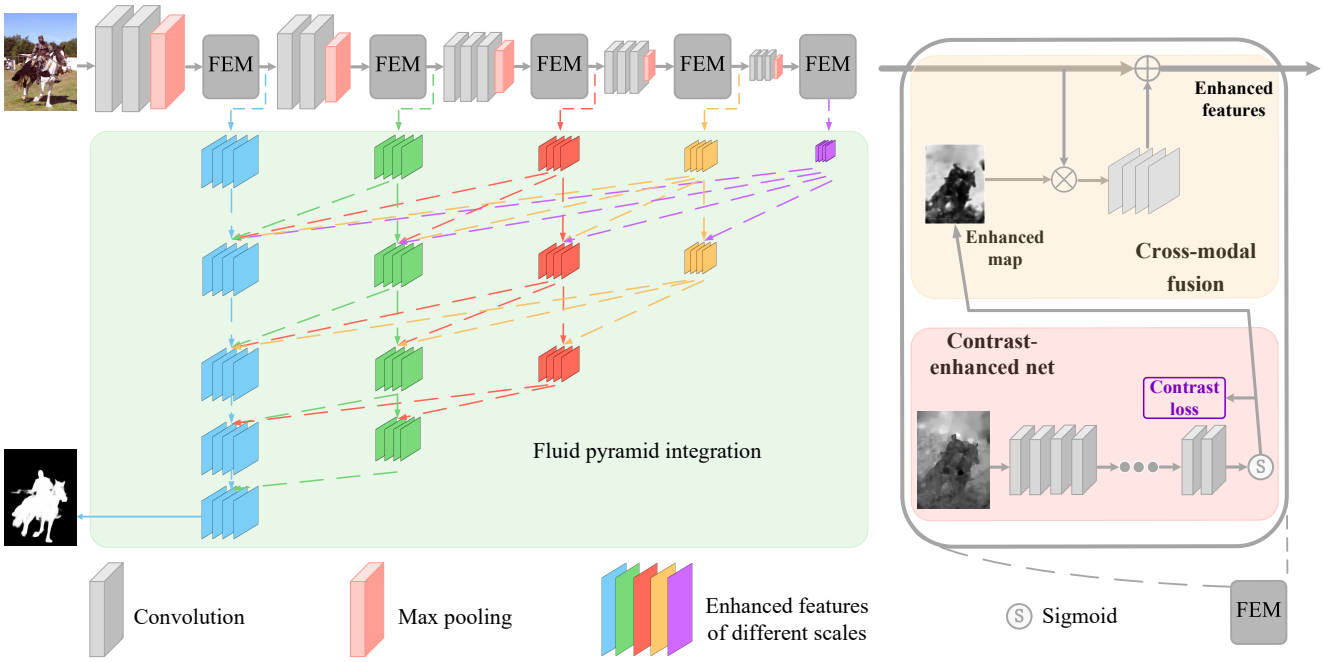

图 3. CPFP 结构。该结构包含两个模块：特征增强模块 (FEM) 和流式金字塔集成模块。FEM 包含两个子模块：对比度增强网络和跨模态融合。在对比度增强网络中，我们利用一种新的对比度损失来利用深层网络中的对比度先验来生成增强图，然后在 VGG-16 的所有 5 个阶段通过跨模式融合得到增强的特征。为了融合多尺度的跨模态特征，设计了流式金字塔结构。结构的详细信息将在第 3 节中介绍。

通过对平均差值的建模，保证了前景目标和背景目标之间的对比度尽可能的大。$\hat{p_f}$ 和 $\hat{p_b}$ 介于 0 到 1 之间，因此 log 函数中的参数取值范围为 0 到 1。

最后，对比度损失 l_c 可以表示为：

$$l_c = \alpha_1 l_f + \alpha_2 l_b + \alpha_3 l_w, \tag{4}$$

其中 α_1 和 α_2 以及 α_3 是预定义的参数。我们建议将它们分别设置为 5、5 和 1。

如图 3 所示，增强后的深度图与原始深度图相比对比度更高。此外，前景和背景的分布更加均匀。

3.1.2 跨模态融合

跨模式融合是特征增强模块的一个子模块，目的是用增强后的深度图调制 RGB 特征。单通道增强图的作用类似于注意力图 [21,51]。具体而言，我们将每个块的 RGB 特征图乘以增强的深度图，以增强显著区域和非显著区域之间的特征对比度。进一步添加残差连接以保留原始 RGB 特征。我们将这些特征图称为**增强特征** $\tilde{F}$，其计算公式为：

$$\tilde{F} = F + F \otimes D_E, \tag{5}$$

F 是原始的 RGB 特征，D_E 表示由所提出的对比度增强网络生成的增强图。$\otimes$ 表示像素相乘。

如图 3 所示，通过在每个块的末端插入特征增强模块，我们分别获得了 $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4, \tilde{F}_5$ 五种不同尺度的增强特征。

3.2. 流式金字塔融合 (FPI)

在处理跨模态信息时，特征兼容性是关键。基于最近在多尺度特征融合方面的成功，我们设计了一个如图 3 所示的流式金字塔结构。流式金字塔可以在多尺度层面更充分地利用跨模态特征，有助于确保特征兼容性。

具体地说，我们的金字塔有 5 层。第一层由 5 个节点组成，每个节点是一组不同尺度的增强特征。然后，我们通过将 $\tilde{F}_2, \tilde{F}_3, \tilde{F}_4, \tilde{F}_5$ 上采样到与 $\tilde{F}_1$ 相同的大小并添加这些上采样特征来构造第二层的第一个节点。同样，我们将 $\tilde{F}_3, \tilde{F}_4, \tilde{F}_5$ 向上采样到与 $\tilde{F}_2$ 相同的大小，并将它们相加以构建第二层的第二个节点。这样，对于金字塔的第 n($n \in \{1,2,3,4,5\}$) 层，总共有 n 个节点，并且每个节点与来自金字塔的第 (n−1) 层 (在这种情况下，第 0 层是修改的 VGG-16 主干) 的所有高层信息

融合。然后经过 transition 卷积层和 Sigmoid 层得到最终的显著图 P。与 [4] 相比，它将预测的显著图拼接起来，所提出的融合方法作用在特征图上。而特征在多尺度融合前保留了更丰富的跨模态信息。也就是说，流式金字塔融合了多尺度层次和跨模态层次的信息。与 [56] 相比它融合了传统金字塔方法中的特征，而 FPI 通过更丰富的连接将所有高层特征引入到金字塔每层节点的低层特征中，称为*流式连接*。流式连接为不同尺度的跨模态特征提供了更多的交互，有助于多尺度层次的特征兼容。

受 [53] 的启发，我们在每个尺度的增强深度图上增加了深度监督。因此，总的损失函数 L 可以表示为：

$$L = l_s + \sum_{i=1}^{5} l_{c_i}, \tag{6}$$

其中 l_s 表示预测图和真实显著图之间的交叉熵损失。l_{c_i} 表示第 i 个特征增强模块中的对比度损失。对比度损失如上所述，交叉熵损失可以计算为：

$$l_s = Y \log P + (1-Y)\log(1-P), \tag{7}$$

其中 P 和 Y 分别表示预测图和真实显著图。

4. 实验

4.1. 实施细节

本文的方法不依赖于特定的主干网络。在这项工作中，为了公平比较，我们选择了 VGG-16 [48] 。本文的网络用 Caffe 平台 [31] 实现。参考 [4] 的设置，我们从 NJU2000 [32] 中随机选取 1400 个样本以及 NLPR 中的 650 个样本 [42] 用于训练。我们还从 NJU2000 中抽取了 100 幅图像并从 NLPR 中抽取了 50 幅图像作为验证集。其余的图像用作测试集。我们随机翻转训练集中的图像来进行数据增强。

对比度增强网络中的参数细节。我们简单地使用两个卷积层加一个 RELU 层，以确保增强图具有与原始特征图相同的尺寸。在第一层卷积中，核大小、通道数量和步长被设置为 (4，32，2)。在第二层卷积中，核大小、通道数目和步长被设置为 (3，32，1)。之后重复这种两层模块，直到特征图在融合的位置保持与 RGB 特征相同的大小。然后，接着是另外两层卷积层。它们的核大小、通道数和步长分别为 (3，32，1) 和 (3，1，1)。之后，将输出放入 Sigmoid 层以生成最终的增强图。采用 Sigmoid 层以确保增强的特征图的值落在 [0，1] 范围内。

训练阶段。在训练阶段，我们对网络进行 10，000 次迭代。初始学习速率设置为 1e-7，并在 7000 次迭代后学习率除以 10。权重衰减和动量分别设置为 0.0005 和 0.9。

权重衰减和动量分别设置为 0.0005 和 0.9。我们在单个 NVIDIA Titan X GPU 上训练我们的网络。batch size 和 iter size 大小分别设置为 1 和 10。新增加的卷积层参数均采用高斯核进行初始化。对于长度或宽度大于 400 的图像，我们将其调整为新的长度和宽度，在保持长宽比不变的情况下，最大值为 400。

推理阶段。在推断阶段，我们调整预测显著图的大小以保持与原始 RGB 图像相同的分辨率。

4.2. 数据集和评估指标

数据集。我们在 5 个广泛使用的 RGBD 数据集上进行了实验。NJU2000 [32] 包含 2003 个立体图像对，具有不同的对象和复杂的、具有挑战性的场景，以及真实显著图。立体图像是从 3D 电影、互联网和富士 W3 立体相机拍摄的照片中收集的。NLPR [42] 也称为 RGBD1000 数据集，包含 1000 幅图像。在每个图像中可能存在多个显著对象。利用 Microsoft Kinect 获得了不同光照条件下的结构光深度图像。SSB [41] 也称为立体数据集，由 1000 对双目图像组成。LFSD [37] 是一个小数据集，其中包含 100 张带有深度信息和人类标记的真实显著图像。深度信息通过 Lytro 光场相机获得。RGBD135 [9] 也称为 DES，它由 7 个室内场景组成，包含由 Microsoft Kinect 收集的 135 幅室内图像。

评估指标。我们采用了 4 个常用指标，即 S-measure、平均 F-measure、最大 F-measure 和平均绝对误差 (MAE)，以及最近提出的结构度量 (S-measure [14]) 来评价不同方法 [2] 的性能。

F-measure 是平均精确度和平均召回率的调和平均值，公式如下：

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{8}$$

按照 [1] 中的建议，精度比召回率更重要，我们也设置 $\beta^2 = 0.3$。在 [2] 的基础上，给出了不同阈值 (0-255) 下的平均 F-measure、最大 F-measure。

设 P 和 Y 表示显著图和归一化为 [0，1] 的真实显著图。为对非显著区域 [2] 公平比较，我们通过以下方

Dataset	Metric	LHM [42]	GP [45]	LBE [20]	SE [23]	CDCP [58]	DF [44]	MDSF [49]	CTMF [25]	PCF [4]	Our CPFP
SSB1000	S-measure ↑	0.562	0.588	0.660	0.708	0.713	0.757	0.728	0.848	0.875	0.879
	meanF ↑	0.378	0.405	0.501	0.610	0.643	0.616	0.527	0.758	0.818	0.842
	maxF ↑	0.683	0.671	0.633	0.755	0.668	0.756	0.719	0.831	0.860	0.873
[41]	MAE ↓	0.172	0.182	0.250	0.143	0.149	0.141	0.176	0.086	0.064	0.051
NJU2000	S-measure ↑	0.514	0.527	0.695	0.664	0.669	0.763	0.748	0.849	0.877	0.878
	meanF ↑	0.328	0.357	0.606	0.583	0.594	0.663	0.628	0.779	0.840	0.850
	maxF ↑	0.632	0.647	0.748	0.747	0.621	0.815	0.775	0.845	0.872	0.877
[32]	MAE ↓	0.205	0.211	0.153	0.169	0.180	0.136	0.157	0.085	0.059	0.053
LFSD	S-measure ↑	0.557	0.640	0.736	0.698	0.717	0.791	0.700	0.796	0.794	0.828
	meanF ↑	0.396	0.519	0.611	0.640	0.680	0.679	0.521	0.756	0.761	0.811
	maxF ↑	0.712	0.787	0.726	0.791	0.703	0.817	0.783	0.791	0.779	0.826
[37]	MAE ↓	0.211	0.183	0.208	0.167	0.167	0.138	0.190	0.119	0.112	0.088
RGBD135	S-measure ↑	0.578	0.636	0.703	0.741	0.709	0.752	0.741	0.863	0.842	0.872
	meanF ↑	0.345	0.411	0.576	0.619	0.585	0.604	0.523	0.756	0.765	0.815
	maxF ↑	0.511	0.600	0.788	0.745	0.631	0.766	0.746	0.844	0.804	0.838
[9]	MAE ↓	0.114	0.168	0.208	0.089	0.115	0.093	0.122	0.055	0.049	0.037
NLPR	S-measure↑	0.630	0.654	0.762	0.756	0.727	0.802	0.805	0.860	0.874	0.888
	meanF ↑	0.427	0.443	0.626	0.624	0.621	0.684	0.649	0.753	0.809	0.840
	maxF ↑	0.622	0.603	0.745	0.720	0.655	0.792	0.793	0.834	0.847	0.869
[42]	MAE ↓	0.108	0.155	0.081	0.099	0.117	0.078	0.095	0.063	0.052	0.036

表 1. 定量比较结果包括 S-measure、平均 F-measure、最大 F-measure 和 MAE 在 5 个流行的数据集上的比较结果。↑&↓ 分别表示越大越好和越小越好。每行前三名的分数分别标为红色、蓝色和绿色。

式计算 MAE 得分：

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - Y(x,y)|, \tag{9}$$

其中 W 和 H 分别是显著图的宽度和高度。

MAE 和 F-measure 指标都忽略了结构相似性评估，然而行为视觉研究表明人类视觉系统对场景 [14] 中的结构高度敏感。因此，我们另外引入了 S-measure [14] 以便进行更全面的评估。S-measure 将区域感知 (S_r) 和目标感知 (S_o) 结构相似性组合为最终的结构指标：

$$S - measure = \alpha * S_o + (1 - \alpha) * S_r, \tag{10}$$

其中 $\alpha \in [0,1]$ 是平衡参数，设置为 0.5。

4.3. 消融实验与分析

在这一部分中，我们将探讨所提出的方法中不同模块对 NJU2000 数据集的影响。

特征增强模块。为了证明对比度增强网络的有效性。我们将主干网络的结果 (表示为 B) 与在主干中添加 FEM 的结果 (表示为 B+C) 进行比较。如表 2 所示，比较第 1 行和第 3 行，可以看出所提出的 FEM 对结果带来了显著提升。此外，在图 4 中我们还展示了深度图与其对应增强图的一些视觉比较。显然，与原始深度图像相比，显著区域和非显著区域之间的对比度增强，同时两个区域内的值变得更加一致。此外，我们还直接使用原始深度图作为增强图 (表示为表 2 中第 2 行 B+D) 对结果进行评估，结果表明 B+D 有负面效应。这是情理之中的。从图 4 中所示的原始深度图我们可以看出，显著区域和非显著区域之间的对比度不够明显并且显著区域和背景中的噪声较多。虽然 B+C 有所不同，但可视实例如图 6 所示。比较主干网络 (B,

图 4. 深度图及其增强图之间的视觉比较。显著区域和非显著区域之间的对比度增强，同时这些区域的数值更加一致。

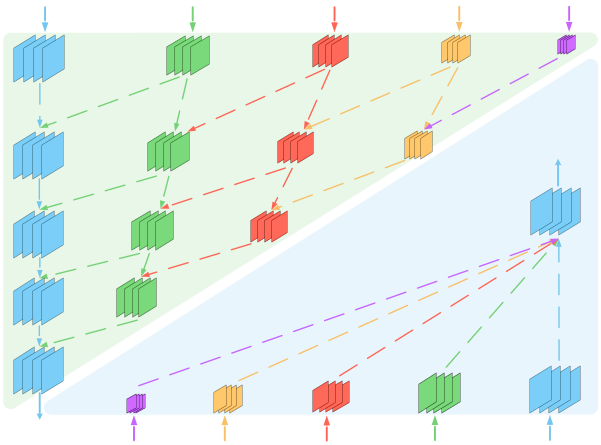

图 5. 不同的融合方法。左上角的是金字塔融合结构 [56] (表 2 中的 P)。右下角是简单的多尺度融合结构 [36](表 2 中的 M)。

图 3 中的第 3 列) 以及主干融合原始深度图像 (B+D，图 6 中第 4 列) 的结果，我们可以看到原始深度图像作用不是很好。当我们在主干中加入特征增强模块来融合跨模态信息，结果如图 6 中第 5 列 (B+C) 所示。在深度信息的帮助下，骨干网络中被误认为显著目标的区域被成功去除。结果表明，在对深度图进行对比度优先增强后，当从 RGB 特征中检测遇到困难时，深度信息有很大的帮助。例如，RGB 图像中的某些区域有噪波 (因为颜色、纹理、亮度等) 在深度层面上的分布微不足道。

流式金字塔。与一些传统的多尺度方法 [36, 56] 相比，该融合方法可以更充分地利用信息，有助于多尺度层次上的跨模态特征兼容。在表 2 中，第 3 行和最后一行表示添加 FPI 之前 (B+C) 和添加 FPI 之后 (B+C+FP) 模型的性能。从数值上看，金字塔融合策略非常有效，贡献了近十个百分点。为了说明金字塔结构的作用，我们首先采用简单的融合方法，如图 5 右下角所示，将多尺度特征上采样到相同的大小，然后直接将它们连接在一起 [36]。我们将此方法表示为 B+C+M，并在表 6 中的第 4 行显示了其性能。结果表明，这种多尺度融合方法的作用非常有限。然后使用图 5 左上角所示方法，使用金字塔结构对这些特征进行分层融合 [56] ，在表 2 中第 5 行表示为 B+C+P。从数值上看，金字塔融合方法比直接融合方法更有效，并贡献了近 4 个点的提升。然后我们在金字塔上添加了流式连接，结果进一步改进，如第 6 行所示。在视觉上，如图 6 所示，比较第 5(B+C) 列和第 6(B+C+M)

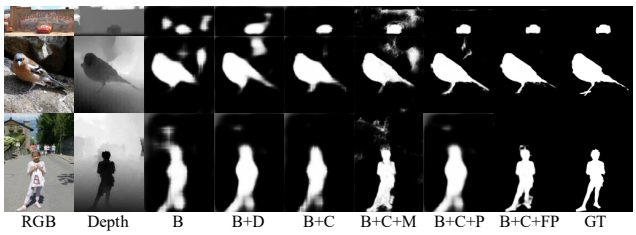

图 6. 不同模型的视觉比较。索引的含义可以从表 2 的标题中看到。

Model	meanF↑	maxF↑	MAE↓
B [40]	0.714	0.791	0.115
B + D	0.708	0.788	0.121
B + C	0.756	0.806	0.094
B + FP	0.758	0.814	0.092
B + C + M	0.748	0.824	0.105
B + C + P	0.789	0.844	0.078
B + D + FP	0.783	0.842	0.081
B + C + FP	0.851	0.877	0.053

表 2. 不同模型的消融研究。B 表示基础模型 (VGG)。D 表示深度图。B+D 表示我们直接使用原始深度图作为增强图。C 表示对比度增强的网络，M 表示简单的多尺度融合，如图 5 右下角所示。P 表示金字塔融合，如图 5 左上角所示。FP 表示流式金字塔融合方法。详细情况在 4.3 节中介绍。

列的结果。可以看出，融合多尺度信息后，边缘细节得到改善。但之前被对比度屏蔽的非显著区域 (第 5 列) 再次出现。造成这一现象的原因是跨模态信息融合在多尺度上遇到了特征兼容问题。然后利用金字塔结构 (B+C+P) 对多尺度信息进行更充分的融合。因为特征更好地互补，所以非显著区域变得更小。之后我们使用流式连接 (B+C+FP)，在金字塔的每一层将高层特征融合到低级特征中，显著目标的位置变得更好。特征互补实现了最佳性能。

4.4. 与最先进的模型进行对比

我们将该模型与 9 种基于 RGBD 的显著目标检测模型进行了比较，这些模型包括 LHM [42]、GP [45]、LBE [20]、SE [23]、CTMF [25]、DF [44]、MDSF [49]、CDCP [58] 和 PCF [4]。注意，上述方法的所有显著图都是通过运行源代码或由作者预先计算生成的。对于所有比较的方法，我们都使用本文建议的默认设置。对于目前没有发布代码的模型，我们感谢作者帮助运行结果。

如表 1 所示，在包含最大 F-measure、平均 F-measure 和 MAE 的大多数评价指标上，我们的方法都

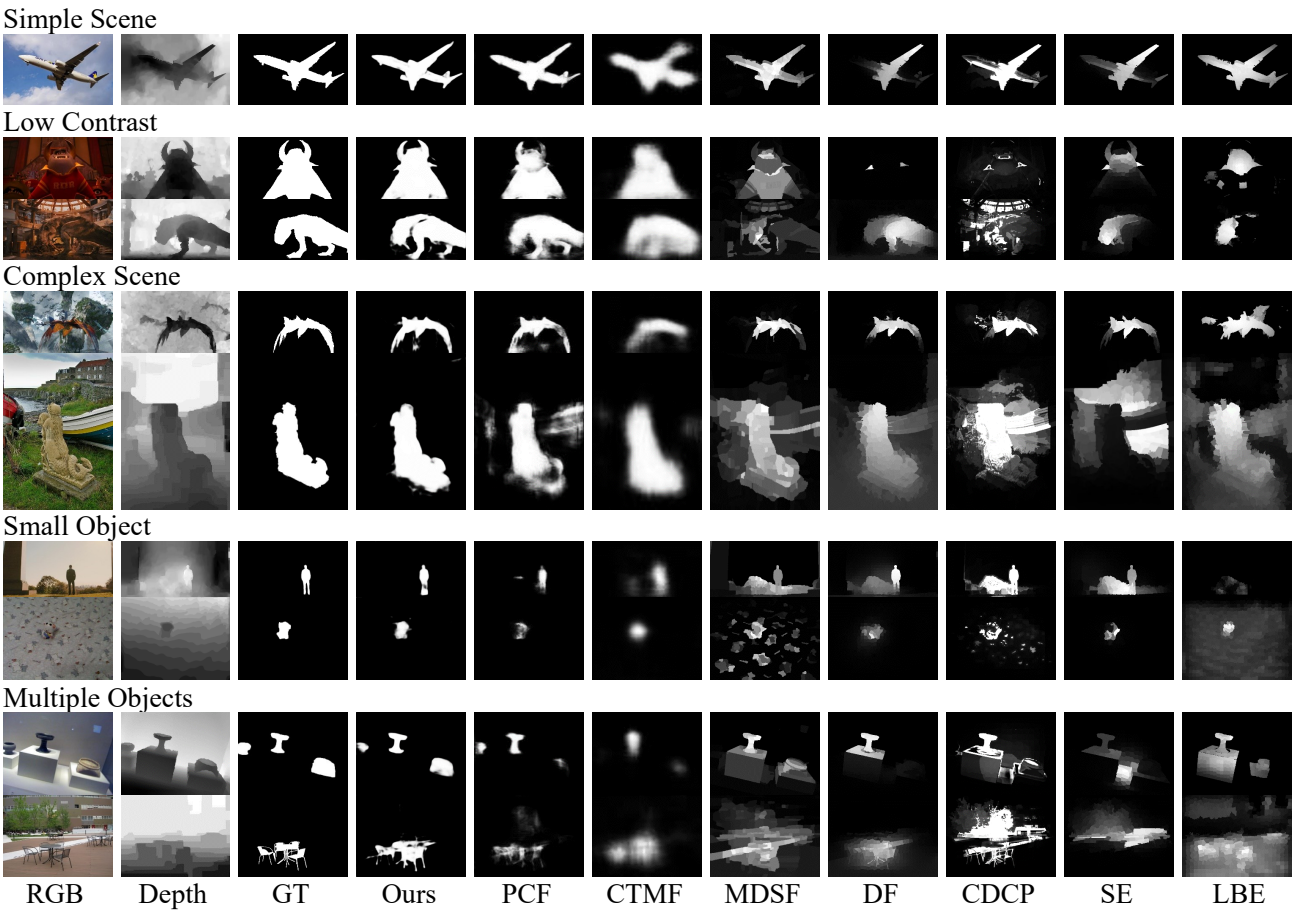

图 7. SSB1000，NJU2000，LFSDRGBD135 和 NLPR 的可视化结果。

优于最新的方法。与最近提出的基于 CNNs 的方法相比，我们的方法在常用的数据集上具有明显的优势。在图 7 中，我们给出了一些可视化结果。我们总结了显著目标检测中的几个具有挑战性的情况：低对比度、复杂场景、小目标和多目标。如图 7 所示，我们在第一行给出了一个简单的例子，几乎所有的方法都表现得很好。在第 2-3 行，我们显示了一些低对比度图像，其中显著目标和背景之间的颜色差异不明显。然而，如果它们的深度差异像显示的样本那样明显，我们可以利用这些深度信息来帮助模型检测显著目标。与早期的方法 (右) 相比，我们的结果更加完整。与 PCF [4] 和 CTMF [25] 等基于学习的方法相比，我们的细节要更好。此外，我们还抽样了一些场景复杂的图像 (第 4-5 行)。在这些图像中，由于场景的复杂性，其他方法将背景误认为是显著目标。然而，我们的模型表现得非常好。这两种类型的图像进一步说明了用所提出的方法使用深度信息是合理的。然后，我们展示了另外两种具有挑战性的情况，小目标和多目标。在这些具有挑战性的情况下，可以看出，我们的模型不仅能通过高层信息很好地定位显著目标，而且通过低层信息能很好地分割目标。

5. 总结

在本文中，我们提出了一种新型的对比度损失函数来监督对比度增强网络。该网络基于对比度先验信息，对深度图进行了明显增强。增强图和 RGB 特征共同作用，增强显著区域和非显著区域之间的对比度，同时保证这些区域内部的一致性。此外，我们设计了一种流式金字塔融合方法来利用多尺度的跨模态特征。与单模态特征的多尺度融合策略相比，流式金字塔结构更适合多尺度跨模态融合，以更好地处理特征兼容性问题。我们的方法极大提高了广泛使用的数据集的技术水平，并且能够在具有挑战性的情况下捕获显著区域。

参考文献

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In CVPR, 2009. 5

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient Object Detection: A Benchmark. IEEE TIP, 24(12):5706–5722, 2015. 2, 5

[3] Ali Borji, Simone Frintrop, Dicky N Sihite, and Laurent Itti. Adaptive object tracking by learning background context. In CVPRW, pages 23–30. IEEE, 2012. 1

[4] Hao Chen and Youfu Li. Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection. In CVPR, pages 3051–3060, 2018. 1, 2, 3, 5, 6, 7, 8

[5] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. ACM TOG, 28(5):124, 2009. 1

[6] Ming-Ming Cheng, Qi-Bin Hou, Song-Hai Zhang, and Paul L Rosin. Intelligent visual media processing: When graphics meets vision. JCST, 32(1):110–121, 2017. 1

[7] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. IEEE TPAMI, 37(3):569–582, 2015. 2

[8] Ming-Ming Cheng, Fang-Lue Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Repfinder: finding approximately repeated scene elements for image editing. ACM TOG, 29(4):83, 2010. 1

[9] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In ICIMCS, page 23. ACM, 2014. 3, 5, 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255. Ieee, 2009. 1

[11] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. Annual review of neuroscience, 18(1):193–222, 1995. 2

[12] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In BMVC, 2013. 3

[13] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In ECCV. Springer, 2018. 1

[14] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In ICCV, pages 4548–4557, 2017. 1, 3, 5, 6

[15] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In IJCAI, pages 698–704, 2018. 1

[16] Deng Ping Fan, Juan Wang, and Xue Mei Liang. Improving image retrieval using the context-aware saliency areas. In Applied Mechanics and Materials, volume 734, pages 596–599. Trans Tech Publ, 2015. 1

[17] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In CVPR, 2019. 1

[18] Xingxing Fan, Zhi Liu, and Guangling Sun. Salient region detection for stereoscopic images. In DSP, pages 454–458, 2014. 1, 2, 3

[19] Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin. Saliency detection for stereoscopic images. IEEE TIP, 23(6):2625–2636, 2014. 3

[20] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In CVPR, pages 2343–2350, 2016. 1, 2, 3, 6, 7

[21] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In CVPR, volume 2, page 3, 2017. 4

[22] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. IEEE TIP, 21(9):4290–4303, 2012. 1

[23] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for rgb-d image via saliency evolution. In IEEE ICME, pages 1–6. IEEE, 2016. 6, 7

[24] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d im-

ages for object detection and segmentation. In ECCV, pages 345–360. Springer, 2014. 2

[25] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics, 2017. 3, 6, 7, 8

[26] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In NIPS, pages 545–552, 2007. 2

[27] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In CVPR, pages 447–456, 2015. 2

[28] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In CVPR, pages 5300–5309. IEEE, 2017. 1, 2

[29] Shi-Min Hu, Tao Chen, Kun Xu, Ming-Ming Cheng, and Ralph R Martin. Internet visual media processing: a survey with graphics and vision applications. The Visual Computer, 29(5):393–405, 2013. 1

[30] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI, 20(11):1254–1259, 1998. 2

[31] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In ACM Multimedia, pages 675–678. ACM, 2014. 5

[32] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In IEEE ICIP, pages 1115–1119, 2014. 1, 5, 6

[33] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In NIPS, pages 109–117, 2011. 2

[34] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In ICCV, pages 660–668, 2016. 2

[35] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In ICCV, pages 5455–5463, 2015. 2

[36] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In ICCV, pages 478–487, 2016. 1, 2, 7

[37] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In CVPR, pages 2806–2813, 2014. 2, 5, 6

[38] Guanghai Liu and Dengping Fan. A model of visual attention for natural image retrieval. In IEEE ISCC-C, pages 728–733, 2013. 1

[39] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In ICCV, pages 678–686, 2016. 1, 2

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In ICCV, pages 3431–3440, 2015. 1, 7

[41] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In CVPR, pages 454–461, 2012. 1, 2, 5, 6

[42] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: a benchmark and algorithms. In ECCV, pages 92–109. Springer, 2014. 1, 2, 5, 6, 7

[43] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In CVPR. IEEE, 2012. 2

[44] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. IEEE TIP, 26(5):2274–2285, 2017. 3, 6, 7

[45] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting Global Priors for RGB-D Saliency Detection. In CVPRW, pages 25–32, 2015. 6, 7

[46] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. IEEE TCSVT, 24(5):769–779, 2014. 1

[47] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning RGB-D Salient Object Detection using background enclosure, depth contrast, and top-down features. In ICCVW, 2017. 3

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 2, 5

[49] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware

salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE TIP, 26(9):4204–4216, 2017. 1, 2, 6, 7

[50] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. Cognitive psychology, 12(1):97–136, 1980. 2

[51] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In CVPR, pages 3156–3164, 2017. 4

[52] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI, 39(11):2314–2320, 2017. 1

[53] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In ICCV, pages 1395–1403, 2015. 2, 5

[54] Jiaxing Zhao, Bo Ren, Qibin Hou, and Ming-Ming Cheng. Flic: Fast linear iterative clustering with active search. In AAAI, 2018. 2

[55] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In CVPR, pages 3927–3936, 2019. 1

[56] Kai Zhao, Wei Shen, Shanghua Gao, Dandan Li, and Ming-Ming Cheng. Hi-fi: Hierarchical feature integration for skeleton detection. In IJCAI, 2018. 2, 5, 7

[57] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In CVPR, pages 1265–1274, 2015. 2

[58] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In ICCVW, 2017. 6, 7

[59] Jun-Yan Zhu, Jiajun Wu, Yan Xu, Eric Chang, and Zhuowen Tu. Unsupervised object class discovery via saliency-guided multiple class learning. IEEE TPAMI, 37(4):862–875, 2015. 1