属性特征保留的 RGB-D 显著性检测模型

周涛 [1]，付华柱 [2]，陈耿 [2]，周毅 [3]，范登平 [2]*，邵岭 [2]
[1] 南京理工大学计算机科学与工程学院 [2] 起源人工智能研究院 [3] 东南大学计算机科学与工程学院

摘要

RGB-D 显著性检测得益于它的有效性以及方便地捕获深度线索而引起了越来越多的关注。现有的工作多侧重于通过各种融合策略来学习一个共享的特征，很少有工作明确考虑如何保留模态的特定特征。本文从一个新角度提出了一种用于 RGB-D 显著性检测的属性特征保留网络（SP-Net），该网络通过探索共享信息和模态的特定属性（如特定性）来提高显著性检测的性能。具体而言，SP-Net 网络利用两个模态特定网络和一个共享学习网络来生成特定及共享显著性图。本文提出了交叉增强融合模块（CIM）来融合共享学习网络中的交叉模态特征，然后将这些特征传播到下一层来融合跨层次信息。本文还提出了多模态特征聚合模块（MFA），该模块将单个解码器生成的模态特征融合到共享解码器中，这丰富了多模态的互补信息，从而提高了显著性检测的性能。此外，本文使用跳跃连接将编码层和解码层之间的分层特征进行组合。在六个基准数据集上得到的实验结果表明，本文的 SP-Net 优于其它最前沿的方法。代码：https://github.com/taozh2017/SPNet。

1 引言

显著性检测的目标是在给定场景中定位视觉上最突出的单个或者多个目标 [46]。它已经被广泛应用于各项与视觉有关的任务，例如图像理解 [76]，视频分割/语义分割 [58, 55]，动作识别 [51, 55]，以及行人重识别 [67]。尽管在显著性检测已经取得了重大进展，但是在复杂的场景中，比如背景杂乱或者低对比度照明条件下的场景，准确定位显著性物体仍然是一个挑战。近来，随着智能设备中深度传感器的大量使用，深度图被引入来提供几何和空间信息，从而提高显著性检测的性能。因此，融合 RGB 图和深度图在显著性检测领域获得了越来越多的关注 [64, 39, 69]。

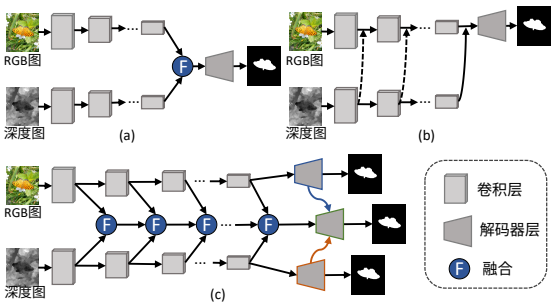

图 1: 现有的 RGB-D 显著性检测框架和本文的模型之间的比较。(a) RGB 图和深度图分别被送入两个独立的网络流，然后将其生成的特征融合成高层特征并送入一个解码器中（例如，[4, 37, 5, 25]）。(b) 使用分割-整合子网络将深度特征融合到 RGB 网络中（例如，[6, 66, 73]）。(c) 本文的方法明确探讨了共享信息和模态的特定特征。然后将模态特定解码器中学到的特征融合到共享解码器中，从而提高显著性检测的性能。

对于 RGB-D 显著性检测而言，有效地融合 RGB 图和深度图是至关重要的。一些方法通过简单串联的方式来实现早期融合策略。例如，这些模型 [46, 52, 56, 41]直接将 RGB 图和深度图级联成为一个四通道的输入。但是，这种类型的融合没有考虑到两种模式之间分布的差异，因此可能导致特征融合不准确。此外，基于后期融合策略的模型大多使用两个并行的网络来产生 RGB 数据和深度数据的独立显著性图，然后再将这两个显著性图进行融合，得到最终的预测图 [24, 57, 15]。但是这种后期融合的策略很难捕捉到两种模态之间复杂的相互关联。

*本文为 ICCV2021 [71]中译版。通讯：dengpfan@gmail.com。

目前，许多**中间融合**策略方法利用两个独立的网络分别学习两种模态的中间特征，然后将融合后的特征输入后续网络或解码器（如图 1（a）所示）。此外，其它方法在多个尺度上进行跨模态融合 [7, 4, 37, 5, 25, 27]。因此，RGB 图和深度图之间复杂的相关性能够在这两种模态之间得到有效地利用。此外，一些方法通过分割-整合子网络 [6, 66, 73] 来利用深度信息从而增强 RGB 图的特征（如图 1（b）所示）。例如，Zhao 等人 [66] 在基于 CNN 的框架中通过引入对比度先验的方式来增强深度信息，然后利用流体金字塔融合模块将增强的深度信息与 RGB 图的特征进行融合。Zhu 等人 [73] 先利用一个独立的子网络来提取深度特征，然后再将其融入 RGB 网络中。值得注意的是，上述方法主要侧重于通过融合特征的方式来学习共享特征，然后使用解码器来生成最终的显著性图。更重要的是，如果基于深度信息的特征学习缺少具有监督功能的解码器来指导 [66, 73]，网络可能就无法获得最佳深度特征。从多模态学习的视角来看，许多工作表明 [26, 42, 72, 70]，探索共享信息和模态的特定属性可以提高模型的性能。然而，很少 RGB-D 显著性检测模型能够明确的利用模态的特定属性。

为此，本文提出了一种新型的用于 RGB-D 显著性检测的属性特征保留网络（称为 SP-Net），它不仅可以探索共享信息，还可以利用模态的特定属性来提高显著性检测的性能。具体而言，SP-Net 使用两个编码器子网络来提取两种模态的多尺度特征，同时提出一个交叉增强融合模块（CIM）来融合跨模态的特征。然后，本文使用 U-Net [53] 结构来构建一个特定模态的解码器，它使用跳跃连接来融合编码器层和解码器层之间的分层特征。通过这种方式，就可以在每个独立的解码器中学习强大的模态特定特征。此外，本文还构建了一个共享解码器，它利用跳跃连接融合之前多个交叉增强融合模块的分层特征。为了充分利用模态的特定特征，本文提出了一个多模态特征聚合模块（MFA）将这些模态特征融合到共享解码器中。最后，本文形成一个统一的、端到端的可训练的框架来实现 RGB-D 显著性检测。

本文主要的**贡献**归纳如下：

- 本文为 RGB-D 显著性检测提出了一种保留属性特征的新型网络（SP-Net），它可以探索共享信息，并保留模态的特定属性。
- 本文提出了一个交叉增强融合模块（CIM）来融合多模态的特征并学习两种模式的共享特征。然后，每个 CIM 的输出被传播到下一层，以获取跨层信息。
- 本文提出了一个简单有效的多模态特征聚合模块（MFA）来融合学到的模态特定特征。它能充分利用在特定模态解码器中学到的特征，来提高显著性检测的性能。
- 在六个公共数据集上进行的实验表明，本文的模型比三十个基准方法更有优势。此外，本文还通过进行属性评价来研究许多最前沿的 RGB-D 显著性检测方法在不同挑战因素下的性能（例如，显著性目标的数量，室内或者室外环境，以及光照条件），这种评价在以前的研究中是没有的。

2 相关工作

2.1 RGB-D 显著性检测

早期基于 RGB-D 的显著性检测模型通常从输入的 RGB-D 数据中提取手工制作的特征。例如 Lang 等人 [31] 提出了第一个 RGB-D 显著性检测工作，该工作利用高斯混合模型对深度信息引导的显著性目标的分布进行建模。之后，几种基于不同准则的方法被探索出来，例如中心-周围差异法 [30, 24]、对比度法 [13, 46, 52]、中心/边界先验法 [75, 36] 以及背景包围法 [19]。然而，由于手工制作的特征表达能力有限，所以这些方法的性能都不尽如人意。得益于深度卷积神经网络（CNN）的快速发展，最近一些基于深度学习的工作 [50, 66, 48, 63, 18] 取得了满意的结果。例如，Qu 等人 [50] 开发了一个 CNN 模型，这个模型将不同的低层显著性线索融合到分层特征中，提高了显著性检测的性能。Chen 等人 [4] 提出了一个互补感知的融合模块来有效地融合 RGB 信息和深度图之间跨模态、跨层次的特征。Piao 等人 [48]提出了一个深度信息引导的多尺度循环注意力网络来增强跨模态特征的融合。Fan 等人 [18] 设计了一个深度净化单元来过滤掉那些低质量的深度图。其它的大多数模型 [7, 37, 5, 25, 33, 35] 采用不同的融合策略，在多个尺度上进行跨模态融合。

2.2 多模态学习

最近，多模态（或多视图）学习引起了越来越多的关注，因为大多数数据可以从多个来源获得或者用

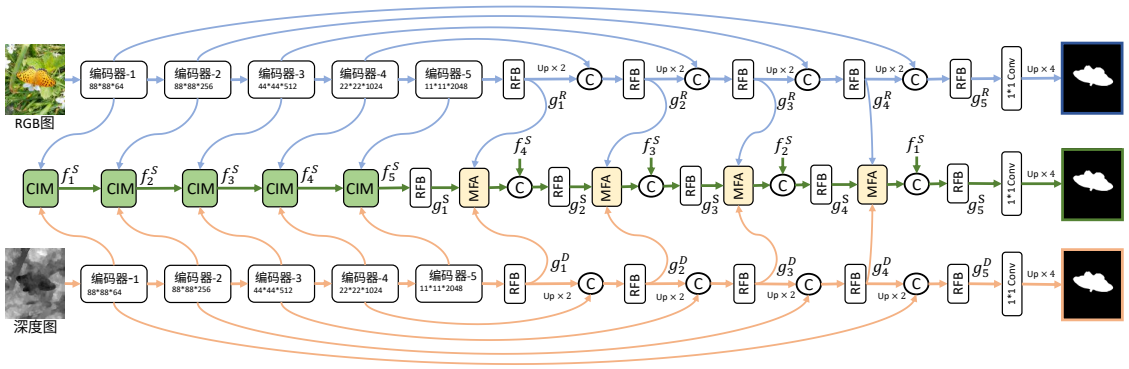

图 2: 本文的 SP-Net 的整体架构图。本文的模型由两个模态的特定学习网络和一个共享学习网络组成。模态的特定学习网络用于保留每个模态（即 RGB 图或深度图）的私有属性，而共享网络用于融合跨模态特征并探索它们的补充信息。采用跳跃连接来融合编码器层和解码器层之间的分层特征。把从特定模态解码器中学习到的特征融合到共享解码器中，来提供丰富的多模态互补信息，从而提高显著性检测的性能。这里的"C"表示特征串联。

不同类型的特征来表示。一种常见的策略是直接将这种多模态数据的特征向量串联成一个长向量。然而，这种简单的串联策略可能无法挖掘多模态数据之间复杂的相关性。因此，开发了一些多模态的学习方法来有效地融合不同模态的互补信息，以达到提高模型性能的效果。这些流行的方法可以分为以下三种类型。1）联合训练法 [3, 14] 试图最大程度地减少不同模态之间的分歧；2）多核学习法 [23] 利用一组来自多模态的预定义核，并利用学到的核权重整合这些模态；3）子空间学习法 [60, 62] 假设存在一个由不同模态共享的潜在子空间，并且多种模态可以源自一个潜在隐表示。此外，为了有效地融合多模态数据，研究者们还探索了几种基于深度学习的模型。例如，Ngiam 等人 [44] 本文的从音频和视频输入的信息中学习共享特征的方法。Eitel 等人 [16] 分别使用两个独立的 CNN 网络来处理 RGB 信息和深度信息，然后使用后期融合网络将它们组合起来以实现 RGB-D 显著性检测。此外，Hu 等人 [26] 提出了一种可共享可独享的多视图学习算法，以探索多模态数据的更多特性。Lu 等人 [42] 为跨模态的行人重识别任务开发了一个共享的特定特征迁移框架。

3 本文方法

图 2展示了用于 RGB-D 显著性检测的属性特征保留网络的框架。首先将 RGB 图和深度图输入模态的特定双流学习网络，来获得它们的多级特征表示，其次提出 CIM 模块来学习它们的共享特征。最后分别采用特定解码器子网络和共享解码器子网络生成显著性预测图。此外，来自编码器网络的原始特征通过跳跃连接被融合到解码器中。为了充分利用从模态的特定解码器中学到的特征，本文提出了 MFA 模块将这些特征融合到共享解码器中，以提高显著性检测的性能。下面给出每个关键部分的详细信息。

3.1 模态的特定学习网络

如图 2 所示，使用在 ImageNet 数据集 [54] 上预训练过的 Res2Net-50 [22] 建立模态的特定网络。因此，在 RGB 图和深度图的模态特定编码器子网络中分别有五个多层级特征，即 $F^R = [f_m^R, m = 1, 2, \ldots, 5]$ 和 $F^D = [f_m^D, m = 1, 2, \ldots, 5]$。在本文的研究中，用 $W \times H$ 来表示模态特定编码器子网络的输入分辨率。所以，第一层的特征分辨率为 $\frac{H}{8} * \frac{W}{8}$，其它分辨率为 $\frac{H}{2^m} * \frac{W}{2^m}$（当 $m > 1$ 时）。此外，如果将第 m 层的特征通道数标定为 C_m（$m = 1, 2, \ldots$），那么就能得到 $C = [64, 256, 512, 1024, 2048]$。

一旦得到高层级的特征 f_5^R 和 f_5^D，就将它们送入模态的特定解码器子网络，产生单独的显著性图。另外，本文利用 U-Net [53]结构来构建模态的特定解码器，其中编码器层和解码器层之间的跳跃连接用于融合分层级特征。串联的特征（只有解码器子网络

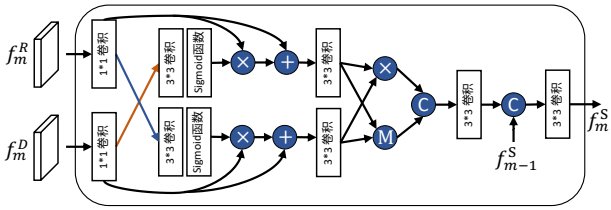

图 3: 交叉增强型融合模块（CIM）的示意图。图中‘C ’ 表示特征连接，“+”, “×”, “M” 分别表示元素相加、相乘和最大化。

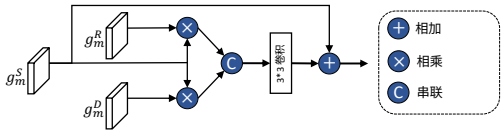

图 4: 多模态特征聚合模块（MFA）示意图。

第一阶段的 f_5^R 或 f_5^D）被输入给感受野模块（RFB）[61]来捕获全局上下文信息。值得注意的是，模态的特定学习网络能够通过保留特定属性来学习每种模态有效且强大的私有特征，然后将这些特征融合到共享解码器子网络中从而提高显著性检测的性能。

3.2 共享学习网络

如图 2 所示，在共享学习网络中，本文方法通过融合来自 RGB 图和深度图的跨模态特征来学习它们的共享特征，并将这个共享特征输入共享解码器来生成最终的显著性图。另外，本文在编码器层和解码器层之间采用跳跃连接，以融合分层级特征。本文还充分利用了模态特定解码器学到的特征，并将这些特征融合到共享解码器中，以此提高显著性检测性能。

3.2.1 交叉增强融合模块

本文提出了一个 CIM 模块来有效地融合跨模态的特征。以 $f_m^R \in \mathbb{R}^{W_m * H_m * C_m}$ 和 $f_m^D \in \mathbb{R}^{W_m * H_m * C_m}$ 为例（方便起见，第 m 层的宽度、高度和通道数分别用 W_m、H_m 和 C_m 来表示），本文使用卷积核大小为 1×1 的卷积层，将通道数减少到 $C_m/2$ 来进行加速。CIM 模块包括两部分，跨模态特征增强部分和自适应特征融合部分。本文首先使用交叉增强策略，通过学习两种模态的增强特征来挖掘二者间的相关性。

具体而言，如图 3 所示，把这两个特征送入一个具有 Sigmoid 激活函数的 3×3 的卷积层，随后就得到归一化的特征图，例如，$w_m^R = \sigma(Conv_3(f_m^R)) \in [0,1]$ 和 $w_m^D = \sigma(Conv_3(f_m^R)) \in [0,1]$，其中 σ 表示 Sigmoid 激活函数。将归一化的特征图看作特征级的注意力图从而自适应地增强特征表示以有效利用两种模态之间的相关性。通过这种方式，可以使用一种模态的特征图来增强另一种模态的特征图。此外，为了保留每种模态的原始信息，本文采用残差连接来融合增强的特征和原始特征。因此，本文将两种模态的交叉增强特征表示如下：

$$\begin{cases} f_m^{R'} = f_m^R + f_m^R \otimes w_m^D, \\ f_m^{D'} = f_m^D + f_m^D \otimes w_m^R, \end{cases} \tag{1}$$

其中 $\otimes$ 表示元素相乘。

本文获得交叉增强的特征（即 $f_m^{R'}$ 和 $f_m^{D'}$）之后的关键任务就是有效地融合它们。许多方法可以用于融合不同模态的特征，包括元素相乘和最大化。然而，目前还不清楚哪种方法最适合的对于某个特定任务。为了利用不同策略的优势，本文应用了元素相乘和最大化，然后将结果串联起来。具体来说，$f_m^{R'}$ 和 $f_m^{D'}$ 两个特征首先被送入一个 3×3 卷积层以获得它们的平滑表示，然后进行元素乘法和最大化。因此，可以得到：

$$\begin{cases} p_{mul} = Bconv_3(f_m^{R'}) \otimes Bconv_3(f_m^{D'}) \\ p_{max} = Max(Bconv_3(f_m^{R'}), Bconv_3(f_m^{D'})) \end{cases} \tag{2}$$

其中 $Bconv(\cdot)$ 表示一连串顺序运算，它包括一个 3×3 的卷积紧接着是归一化和 ReLU 函数。然后本文将结果串联为 $p_{cat} = [p_{mul}, p_{max}] \in \mathbb{R}^{W_m * H_m * C_m}$，并且通过 $Bconv_3$ 操作得到 $p_{cat}^1 = Bconv_3(p_{cat})$，最后给这两部分进行自适应加权。此外，将得到的 p_{cat}^1 和第 $(m-1)$ 个 CIM 模块的输出 f_{m-1}^S 进行串联，然后将结果输入第二个 $Bconv_3$ 进行运算。最后，本文得到第 m 个 CIM 的输出 f_m^S。注意，当 $m = 1$ 时，不需要使用 1×1 的卷积层来减少通道数。特别是，在没有先前的输出 f_{m-1}^S（也就是当 $m = 1$）时，就只需要将串联的特征输入到一个 $Bconv_3$ 中进行运算。

值得注意的是，本文的 CIM 模块可通过交叉增强的特征学习有效地利用两种模态之间的相关性，并通过自适应加权方法来融合它们。把融合后的特征表示 f_m^S 传播到下一层，来捕捉和融合跨层级信息。

3.2.2 多模态特征聚合

为了充分利用在模态特定解码器中学到的特征，本文提出了一个简单有效的模块（MFA），该模块将那些在模态特定解码器中学到的特征融合到共享解码器中。具体来说，本文使用 g_m^S 表示共享解码器第 m 层的共享特征，并且使用 g_m^R 和 g_m^D 表示在模态特定解码器中学习的特定特征. 如图 4 所示，将 g_m^R 和 g_m^D 两个特征与当前层的共享特征相乘，即 $g_m^{RS} = g_m^S \otimes g_m^R$ 和 $g_m^{DS} = g_m^S \otimes g_m^D$。这两个特征被进一步串联（$[g_m^{DR}, g_m^{DS}]$），然后将串联的结果输入 $Bconv(\cdot)$ 进行运算就会得到 g_m^{Sc}。最后，本文通过加法运算将卷积特征 g_m^{Sc} 与原始特征 g_m^S 相结合就能得到 MFA 模块的输出。

值得注意的是，学习到的特定模态的特征被用来增强共享特征，并提供丰富而又互补的跨模态信息。更重要的是，为了保存特定模态的特征，特定模态解码器被赋予了一个监督信号来进行学习指导，这有利于将特征融合到共享解码器并生成最终的预测图。

3.3 损失函数

最后，本文形成了一个统一、端到端的可训练框架。总体损失函数由 $\mathcal{L}_{sp}$ 和 $\mathcal{L}_{sh}$ 两部分组成，它们分别用于模态特定解码器和共享解码器。方便起见，分别用 S_R 和 S_D 表示使用 RGB 图和深度图时生成的预测图，使用 S_{sh} 表示使用其共享表示时的预测图，G 表示真值图。因此，总的损失函数可以表述如下：

$$\mathcal{L}_{total} = \mathcal{L}_{sh}(S_{sh}, G) + \mathcal{L}_{sp}(S_R, G) + \mathcal{L}_{sp}(S_D, G) \tag{3}$$

在公式 (3) 中，本文利用了 $\mathcal{L}_{sp}$ 和 $\mathcal{L}_{sh}$ 的像素位置感知损失 [59]，这可以给予难识别和易识别的像素不同的关注，从而提高显著性检测的性能。

4 实验

4.1 实验设置

数据集：为了验证所提出模型的有效性，本文在六个公开 RGB-D SOD 数据集上对其进行了评估，包括 NJU2K [30], NLPR [46], DES [10], SSD [74], STERE [45]和 SIP [18]。本文根据 [48, 18]的设置，从 NJU2K 数据集 [30]中选取 1485 个样本，从 NLPR [46]中选取 700 个样本，总共 2195 个样本用于训练。其余用于测试的样本来自 NJU2K（500）和 NLPR（300）以及整个 DES（135）、SSD（80）、STERE（1,000）和 SIP（929）。

评估指标：本文采用了四个广泛使用的指标进行定量评价。1）S-measure（S_α）[8]用于评估区域感知（S_r）和目标感知（S_o）之间的结构相似性。它被定义为 $S_\alpha = \alpha * S_o + (1-\alpha) * S_r$，其中 $\alpha \in [0,1]$，它是一个权衡参数，默认设置为 0.5。2）E_ϕ [17]它被用来捕获像素级别的统计数据及其局部像素匹配信息，它被定义为 $E_\phi = \frac{1}{W*H}\sum_{i=1}^{W}\sum_{i=1}^{H}\phi_{FM}(i,j)$，其中 ϕ_{FM} 表示增强对齐矩阵 [17]。3）F-measure [1]（F_β）用于综合考虑精度和召回率，可以通过 $F_\beta = \left(1+\beta^2\right)\frac{P*R}{\beta^2 P+R}$ 得到加权调和平均值。为了加强精度，β^2 的值被设置为 0.3 [1]。本文使用 $[0, 255]$ 区间不同的阈值来计算 F 方法，这就产生了一组 F 测量值，本文报告 F_β 的最大值。4）平均绝对误差 (MAE) [47]是用于通过计算差值的平均值来评估真实值和归一化预测之间像素级的平均相对误差。

RGB-D **显著性检测对比模型**：将本文的模型与 30 种基准的 RGB 显著性检测方法进行对比，其中包括 8 种基于手工特征的传统模型（即 LHM [46], ACSD [30], LBE [19], DCMC [12], SE [24], MDSF [56], CDCP [75] 和 DTM [11]）和 22 个深度模型（即 DF [50], CTMF [25], PCF [4], AFNet [57], CPFP [66], MMCI [6], TANet [5], DMRA [48], cmSalGAN [29], ASIFNet [32], ICNet [34], A2dele [49], JLDCF [20], S^2MA [38], UCNet [63], SSF [65], Cas-GNN [43], CMMS [33], D^3Net [18], CoNet [28], DANet [68], PGAR [9]）方法细节在此省略，请参考相关论文。

实验细节：本文的模型是用 PyTorch 实现的，并在一个 32GB 内存的 NVIDIA Tesla V100 GPU 上进行训练。使用在 ImageNet [54]上进行预训练的骨干网络 (Res2Net-50 [22])。因为 RGB 图和深度图的通道数不同，所以深度编码器的输入通道修改为 1。本文采用 Adam 算法来优化本文的模型。初始学习率设置为 $1e-4$，每 60 次迭代除以 10。RGB 图和深度图的输入分辨率被调整为 352×352。使用各种策略增强训练图像，包括随机翻转、旋转和边缘裁剪。批量大小被设置为 20，模型训练超过 200 轮。在测试阶段，RGB 图和深度图被调整为 352×352 的大小，然后将它们输入模型来获得预测图。将预测图重新缩放到原始大小来进行最终的评估。最后，共享解码器输出的结果作为本文模型的最终预测图。

表 1: 在 6 个公开的 RGB-D 显著性数据集上使用 4 个广泛使用的评价指标（即 S_α [8], max E_ϕ[17], max F_β [1]和 M [47]）与 8 个具有代表性的传统模型和 22 个深度模型进行对比产生的基准测试结果。'↑ '和 '↓" 表示越大或越小越好。每个模型的下标表示出版年份。最佳结果以下划线 突出显示。

	NJU2K [30]				STERE [45]				DES [10]				NLPR [46]				SSD [74]				SIP [18]			
模型	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$
LHM_{14} [46]	.514	.632	.724	.205	.562	.683	.771	.172	.562	.511	.653	.114	.630	.622	.766	.108	.566	.568	.717	.195	.511	.574	.716	.184
ACSD_{14} [30]	.699	.711	.803	.202	.692	.669	.806	.200	.728	.756	.850	.169	.673	.607	.780	.179	.675	.682	.785	.203	.732	.763	.838	.172
LBE_{16} [19]	.695	.748	.803	.153	.660	.633	.787	.250	.703	.788	.890	.208	.762	.745	.855	.081	.621	.619	.736	.278	.727	.751	.853	.200
DCMC_{16} [12]	.686	.715	.799	.172	.731	.740	.819	.148	.707	.666	.773	.111	.724	.648	.793	.117	.704	.711	.786	.169	.683	.618	.743	.186
SE_{16} [24]	.664	.748	.813	.169	.708	.755	.846	.143	.741	.741	.856	.090	.756	.713	.847	.091	.675	.710	.800	.165	.628	.661	.771	.164
MDSF_{17} [56]	.748	.775	.838	.157	.728	.719	.809	.176	.741	.746	.851	.122	.805	.793	.885	.095	.673	.703	.779	.192	.717	.698	.798	.167
CDCP_{17} [75]	.669	.621	.741	.180	.713	.664	.786	.149	.709	.631	.811	.115	.669	.621	.741	.180	.603	.535	.700	.214	.595	.505	.721	.224
DTM_{20} [11]	.706	.716	.799	.190	.747	.743	.837	.168	.752	.697	.858	.123	.733	.677	.833	.145	.677	.651	.773	.199	.690	.659	.778	.203
DF_{17} [50]	.763	.804	.864	.141	.757	.757	.847	.141	.752	.766	.870	.093	.802	.778	.880	.085	.747	.735	.828	.142	.653	.657	.759	.185
CTMF_{18} [25]	.849	.845	.913	.085	.848	.831	.912	.086	.863	.844	.932	.055	.860	.825	.929	.056	.776	.729	.865	.099	.716	.694	.829	.139
PCF_{18} [4]	.877	.872	.924	.059	.875	.860	.925	.064	.842	.804	.893	.049	.874	.841	.925	.044	.841	.807	.894	.062	.842	.838	.901	.071
AFNet_{19} [57]	.772	.775	.853	.100	.825	.823	.887	.075	.770	.729	.881	.068	.799	.771	.879	.058	.714	.687	.807	.118	.720	.712	.819	.118
CPFP_{19} [66]	.878	.877	.923	.053	.879	.874	.925	.051	.872	.846	.923	.038	.888	.867	.932	.036	.807	.766	.852	.082	.850	.851	.903	.064
MMCI_{19} [6]	.859	.853	.915	.079	.873	.863	.927	.068	.848	.822	.928	.065	.856	.815	.913	.059	.813	.781	.882	.082	.833	.818	.897	.086
TANet_{19} [5]	.878	.874	.925	.060	.871	.861	.923	.060	.858	.827	.910	.046	.886	.863	.941	.041	.839	.810	.897	.063	.835	.830	.895	.075
DMRA_{19} [48]	.886	.886	.927	.051	.886	.886	.938	.047	.900	.888	.943	.030	.899	.879	.947	.031	.857	.844	.906	.058	.806	.821	.875	.085
cmSalGAN_{20} [29]	.903	.896	.940	.046	.900	.894	.936	.050	.913	.899	.943	.028	.922	.907	.957	.027	.791	.735	.867	.086	.865	.864	.906	.064
ASIFNet_{20} [32]	.889	.888	.927	.047	.878	.878	.927	.049	.934	.935	.974	.019	.906	.888	.944	.030	.857	.834	.884	.056	.857	.859	.896	.061
ICNet_{20} [34]	.894	.891	.926	.052	.903	.898	.942	.045	.920	.913	.960	.027	.923	.908	.952	.028	.848	.841	.902	.064	.854	.857	.903	.069
A2dele_{20} [49]	.871	.874	.916	.051	.878	.879	.928	.044	.886	.872	.920	.029	.898	.882	.944	.029	.802	.776	.861	.070	.828	.833	.889	.070
JLDCF_{20} [20]	.903	.903	.944	.043	.905	.901	.946	.042	.929	.919	.968	.022	.925	.916	.962	.022	.830	.795	.885	.068	.879	.885	.923	.051
S^2MA_{20} [38]	.894	.889	.930	.053	.890	.882	.932	.051	.941	.935	.973	.021	.915	.902	.953	.030	.868	.848	.909	.052	.872	.877	.919	.057
UCNet_{20} [63]	.897	.895	.936	.043	.903	.899	.944	.039	.933	.930	.976	.018	.920	.903	.956	.025	.865	.854	.907	.049	.875	.879	.919	.051
SSF_{20} [65]	.899	.896	.935	.043	.893	.890	.936	.044	.904	.884	.941	.026	.914	.896	.953	.026	.845	.824	.897	.058	.876	.882	.922	.052
Cas-GNN_{20} [43]	.911	.903	.933	.035	.899	.901	.930	.039	.905	.906	.947	.028	.919	.904	.947	.028	.872	.862	.915	.047	.875	.879	.919	.051
CMMS_{20} [33]	.900	.897	.936	.044	.895	.893	.939	.043	.937	.930	.976	.018	.915	.896	.949	.027	.874	.864	.922	.046	.872	.877	.911	.058
CoNet_{20} [28]	.895	.893	.937	.046	.908	.905	.949	.040	.909	.896	.945	.028	.908	.887	.945	.031	.853	.840	.915	.059	.858	.867	.913	.063
DANet_{20} [68]	.899	.910	.935	.045	.901	.892	.937	.043	.924	.928	.968	.023	.915	.916	.953	.028	.864	.866	.914	.050	.875	.892	.918	.054
PGAR_{20} [9]	.909	.907	.940	.042	.907	.898	.939	.041	.913	.902	.945	.026	.930	.916	.961	.024	.865	.838	.898	.057	.876	.876	.915	.055
$\text{D}^3\text{Net}_{21}$ [18]	.900	.900	.950	.041	.899	.891	.938	.046	.898	.885	.946	.031	.912	.897	.953	.030	.857	.834	.910	.058	.860	.861	.909	.063
SP-Net (本文)	.925	.935	.954	.028	.907	.915	.944	.037	.945	.950	.980	.014	.927	.925	.959	.021	.871	.883	.915	.044	.894	.916	.930	.043

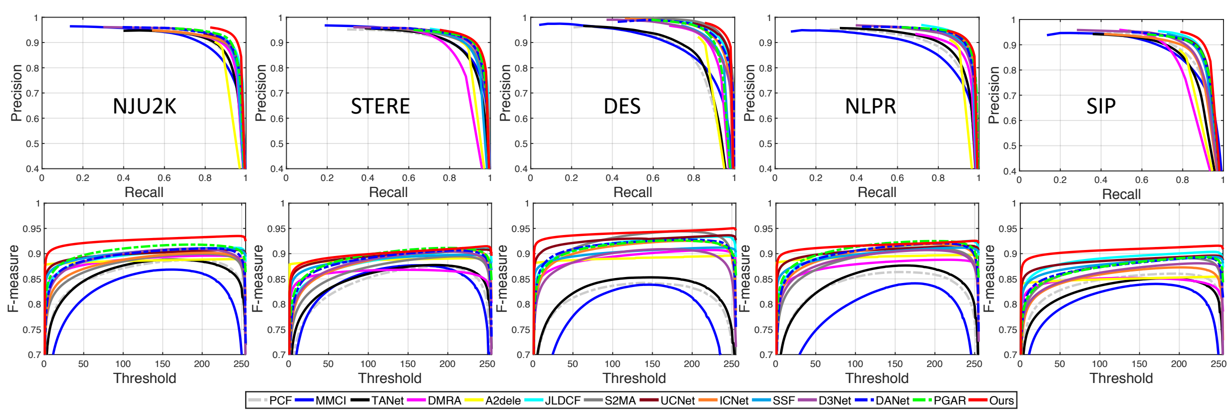

图 5: 在 NJU2K [30], STERE [45], DES [10], NLPR [46], and SIP [18]数据集上的 PR 曲线（如图第 1 行所示），F-measure 曲线（如图第 2 行所示）

4.2 性能对比

定量比较: 如表 1 所示，本文的方法在六个数据集上都以较大的优势优于八个传统方法（LHM [46], ACSD [30], LBE [19], DCMC [12], SE [24], MDSF [56], CDCP [75]以及 DTM [11]）。此外，本文的方法在 NJU2K、DES 和 SIP 数据集上的表现优于所有比较的方法，而且在四个评估指标上获得了最佳表现。此外，值得注意的是，相较于大多数 RGB-D 显著性检测方法，本文的模型在 STERE 和 NLPR 数据集上获得了更好的性能。本文的模型在 STERE 数据集

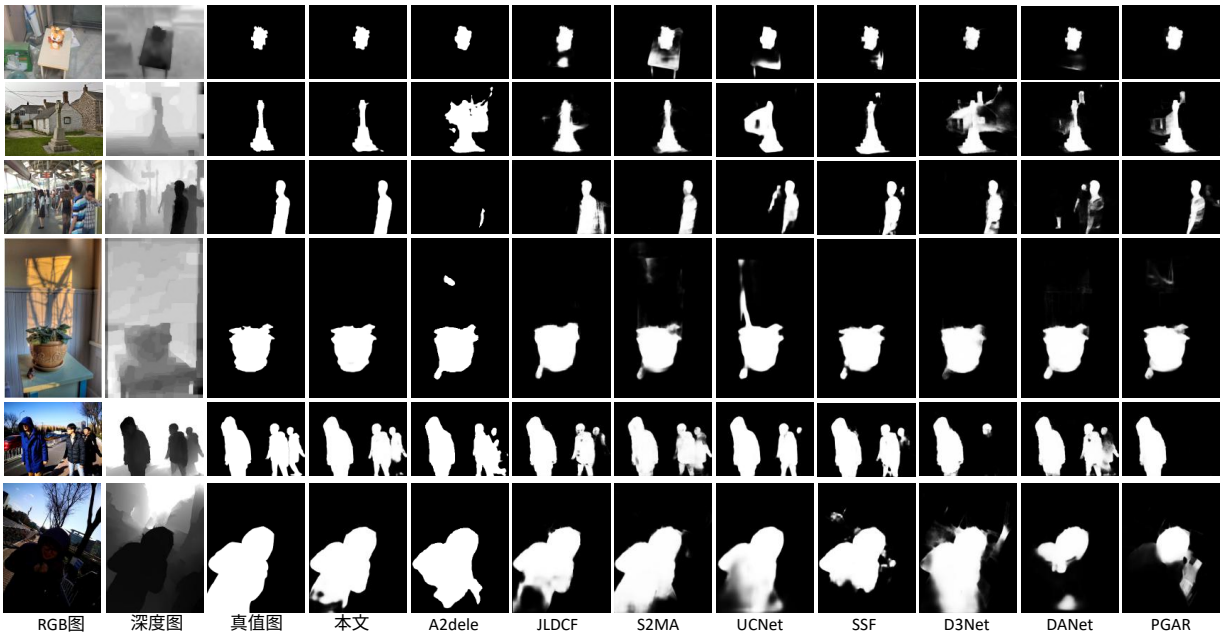

图 6: 本文的方法和 8 个最新的方法的视觉对比（其中包括 A2dele [49]，JLDCF [20]，S2MA [38]，UCNet [63]，SSF [65]，D3Net [18]，DANet [68] ，以及 PGAR [9]）

上与 CoNet 相当，在 NLPR 数据集上与 JLDCF 和 PGAR 相当。总体而言，在给定场景的情况下本文的 SP-Net 在定位显著性目标时表现出了很好的性能。此外，本文在图 5 中显示了 PR 曲线 [2]和 F-measure 曲线。为清楚起见，本文提供了 14 种 RGB-D 显著性检测方法的结果，这其中包含 13 种具有完整显著性图的 SOTA 模型。正如图所示，在这些呈现的数据集上本文模型的优越性更加明显。

定性比较: 将本文的模型与 8 个最前沿方法进行比较，图 6 显示了其中几个具有代表性的结果。第一行展示的是小目标情况。本文的方法以及 A2dele、PGAR 和 D3Net 可以准确地检测到显著性目标，而 JLDCF、S2MA、SSF 和 UCNet 则预测到一些非目标区域。在第 2 和第 3 行中，本文背景复杂的例子。从对比结果可以看出，本文的方法和 S2MA 方法产生可靠的结果，而其它 RGB-D 显著性检测模型不能正确定位显著性目标或者混淆了背景和目标。在第 4 行中，除了 D3Net 方法之外，对比的其它方法都检测到了一个很小的非目标区域。本文在第 5 行中展示了具有多个显著性目标的例子，在此情况下准确定位所有显著性目标是极具挑战性的。本文的方法能够定位所有显著性目标，相较于其它方法，能够更准确的分割它们并且产生更清晰的边缘。本文在最后一行展示了弱光条件下的情况。有些方法不能将显著性目标的全部区域完整的检测出来。本文的模型能够通过抑制背景干扰来提高显著性检测的性能，从而产生满意的结果。

表 2: 消融实验的定量评估

	NJU2K [30]		STERE [45]		DES [10]		NLPR [46]		SSD [74]		SIP [18]	
	$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$
本文	.925	.028	.907	.037	.945	.014	.927	.021	.871	.044	.894	.043
A1	.916	.034	.898	.042	.939	.016	.926	.022	.869	.047	.892	.044
A2	.921	.031	.895	.042	.938	.016	.925	.022	.865	.051	.896	.042
A3	.919	.032	.895	.043	.938	.016	.929	.020	.864	.049	.887	.048
A4	.924	.029	.903	.038	.930	.019	.927	.023	.867	.049	.888	.046
B1	.918	.034	.901	.041	.939	.017	.922	.024	.858	.050	.885	.048
B2	.924	.029	.900	.041	.941	.015	.926	.022	.864	.049	.893	.044
B3	.921	.031	.903	.039	.938	.016	.925	.022	.863	.050	.891	.045
C	.913	.037	.900	.047	.935	.019	.922	.025	.861	.055	.880	.051

4.3 消融实验

为了验证模型的不同部分的贡献，本文通过从完整模型中删除或替换它们来进行消融实验。

(A) CIM **模块的有效性**. 由于 CIM 是用来融合跨模态特征并学习它们的共享特征的，所以本文利用直接串联方法来代替 CIM 模块。具体来说，将两个特征 f_m^R 和 f_m^D（如图 3 所示）直接串联，然后输入

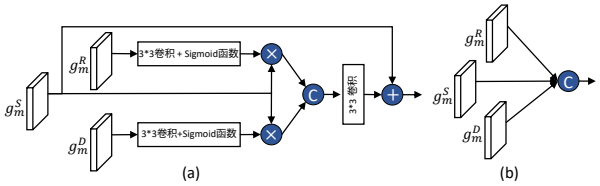

图 7: MFA 模块与其它融合策略的比较。

一个 3×3 的卷积层从而获得每一层的融合结果。本文在表 2 中把这个评估结果标记为“A1”。从比较结果可以看出，本文的模型在使用 CIM 模块时比简单的使用特征串联方法表现的更好。这也说明了 CIM 模块在提高显著性检测性能方面的贡献。此外，CIM 模块还有两个部分，即跨模态特征增强部分和自适应特征融合部分。因此，为了评估每个部分的贡献，本文将仅具有跨模态特征增强部分和仅具有自适应特征融合部分的 CIM 模块分别标记为“A2”和“A3”。当把这两个独立的部分与完整的 CIM 模块进行比较时，可以看出本文的完整的 CIM 的有效性。此外，在 CIM 模块中，通过将上一层的特征传播到下一层的方式来捕捉跨层的关联性。为了验证传播策略的有效性，本文在 CIM 模块中删除了这个传播，并标记为“A4”。“A4”和 CIM 模块的比较结果表明，这种传播策略提高了显著性检测的性能。

(B) MFA **的有效性**. 在本文的框架中，MFA 模块充分利用了在特定模态解码器中学到的特征，然后将这些特征融合到共享解码器中来提供更多的多模式互补信息，为了证明其有效性，本文删除了这个模块并标记为“B1”。此外，本文将其他两种特征融合策略与本文的 MFA 模块进行比较。如图 7 所示，一种是跨模态特征增强融合策略；另一种是简单的串联策略。这两种策略的对比实验分别标记为“B2”和“B3”如表 2 所示，将“B1”和本文的完整模型进行对比，对比结果表明了将特征融合到共享解码器中的有效性。将“B2”、“B3”和本文的全部模型相比较，可以看出 MFA 模块优于其他两种融合策略。

(C) **特定模态解码器的有效性**. 本文删除了两个特定模态解码器，评估结果如表 2 的“C”所示。可以看出，如果不使用这两个部分，显著性检测的性能将会下降。这证明了模态特定解码器的有效性，它能提供监督信号来确保学习得到模态的特定属性。

4.4 属性评价

许多挑战性的因素会影响 RGB-D 显著性检测模型的性能，例如显著性目标的数量、室内或室外的环境、光照条件等。因为可以展示前沿模型在处理这些挑战因素的优势和劣势，评估这些条件下的显著性检测性能是很有意义的。

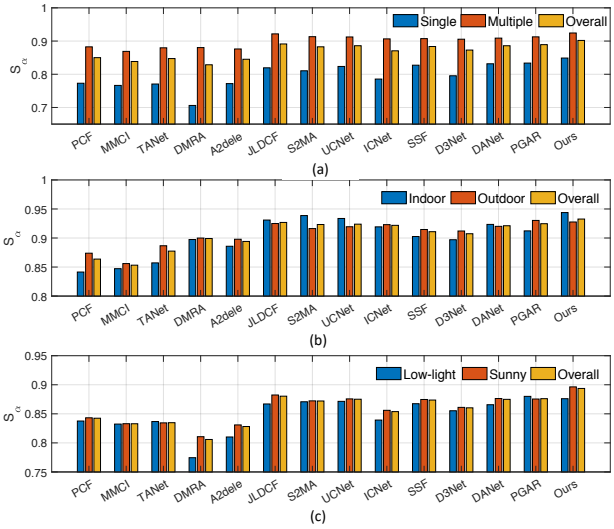

图 8: 基于属性的评价。(a) 显著目标的数量（即单个或多个），(b) 室内或室外环境，以及 (c) 光照条件（弱光或者阳光）。

(1) 单一目标与多目标的比较。在这个评估中，本文构建了一个混合数据集，其中包含 1,229 张从 NLPR [46]和 SIP [18]数据集收集的图像。使用 S_α 的对比结果如图 8 (a) 所示。可以看出识别单个显著性目标比识别多个显著性目标要容易。此外，本文的模型在定位单个目标和多个目标方面都优于其它最前沿的方法。(2) 室内与室外的比较。DES [10]和 NLPR [46]数据集包括室内和室外场景，因此本文构建了从这两个数据集收集的混合数据集。比较结果如图 8 (b) 所示。可以看出，与室外场景相比，许多模型在检测室内场景中的显著性目标上更加困难，而 JLDCF、S2MA、UCNet、ICNet、SSF、DANet 和本文的模型在室外场景中的表现要更好一点。(3) 光照条件。本文在 SIP 数据集 [18]上进行评估并将数据分为两类，即光照充足和光照不足。对比结果如图 8 (c) 所示。可以看出，所有的模型在低光照条件下检测显著性目标时都遭受到挑战，这证明了低光照条件对显著性目标检测性能的负面影响。

5 结论

本文提出了一种用于 RGB-D 显著性检测的新型属性特征保留网络 SP-Net。现有大多数模型主要侧重于学习共享特征，和这些模型不同，本文的模型不仅探索了共享的跨模态信息，还捕捉了模态的特定特征来提高显著性检测的性能。此外，本文的 CIM 模块可以跨模态和跨层级传播信息，而 MFA 模块则可以为共享解码器提供模态的特定属性，从而增强多模态信息的互补性。在六个具有挑战性的基准数据集上进行的定量和定性评估表明，本文的 SP-Net 比现有其它的 RGB-D 显著性检测的方法更具优势。本文的模型也计划应用于光场显著性检测任务 [21, 40]。

参考文献

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In CVPR, pages 1597–1604. IEEE, 2009. 5, 6

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. IEEE TIP, 24(12):5706–5722, 2015. 7

[3] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In ICML, pages 129–136, 2009. 3

[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, 5, 6

[5] Hao Chen and Youfu Li. Three-stream attention-aware network for RGB-D salient object detection. IEEE TIP, 28(6):2825–2835, 2019. 1, 2, 5, 6

[6] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognition, 86:376–385, 2019. 1, 2, 5, 6

[7] Hao Chen, You-Fu Li, and Dan Su. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In IEEE IROS, pages 6821–6826. IEEE, 2018. 2

[8] Ming-Ming Chen and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. IJCV, 129:2622–2638, 2021. 5, 6

[9] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In ECCV. Springer, 2020. 5, 6, 7

[10] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In ICIMCS, pages 23–27, 2014. 5, 6, 7, 8

[11] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. IEEE TCYB, 2019. 5, 6

[12] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. SPL, 23(6):819–823, 2016. 5, 6

[13] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In BMVC, 2013. 2

[14] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. IEEE TMM, 17(11):2049–2058, 2015. 3

[15] Yu Ding, Zhi Liu, Mengke Huang, Ran Shi, and Xiangyang Wang. Depth-aware saliency detection using convolutional neural networks. Journal of Visual Communication and Image Representation, 61:1–9, 2019. 1

[16] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In IROS, pages 681–687. IEEE, 2015. 3

[17] 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. 5, 6

[18] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. IEEE TNNLS, 32(5):2075–2089, 2021. 2, 5, 6, 7, 8

[19] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for RGB-D salient object detection. In CVPR, pages 2343–2350, 2016. 2, 5, 6

[20] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for RGB-D salient object detection and beyond. IEEE TPAMI, 2021. 5, 6, 7

[21] Keren Fu, Yao Jiang, Ge-Peng Ji, Tao Zhou, Qijun Zhao, and Deng-Ping Fan. Light field salient object detection: A review and benchmark. arXiv preprint arXiv:2010.04968, 2020. 9

[22] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. IEEE TPAMI, 2020. 3, 5

[23] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. JMLR, 12:2211–2268, 2011. 3

[24] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for RGB-D image via saliency evolution. In ICME, pages 1–6. IEEE, 2016. 1, 2, 5, 6

[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 TCYB, 48(11):3171–3183, 2017. 1, 2, 5, 6

[26] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Sharable and individual multi-view metric learning. IEEE TPAMI, 40(9):2281–2288, 2017. 2, 3

[27] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated RGB-D salient object detection. In CVPR, pages 9471–9481, 2021. 2

[28] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In ECCV, 2020. 5, 6

[29] Bo Jiang, Zitai Zhou, Xiao Wang, Jin Tang, and Bin Luo. cmsalgan: RGB-D salient object detection with cross-view generative adversarial networks. IEEE TMM, 2020. 5, 6

[30] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In ICIP, pages 1115–1119. IEEE, 2014. 2, 5, 6, 7

[31] Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In ECCV, pages 101–115. Springer, 2012. 2

[32] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. IEEE TCYB, 2020. 5, 6

[33] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In ECCV. Springer, 2020. 2, 5, 6

[34] Gongyang Li, Zhi Liu, and Haibin Ling. Icnet: Information conversion network for RGB-D based salient object detection. IEEE TIP, 29:4873–4884, 2020. 5, 6

[35] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for RGB-D salient object detection. In ECCV. Springer, 2020. 2

[36] Fangfang Liang, Lijuan Duan, Wei Ma, Yuanhua Qiao, Zhi Cai, and Laiyun Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing, 275:2227–2238, 2018. 2

[37] Di Liu, Yaosi Hu, Kao Zhang, and Zhenzhong Chen. Two-stream refinement network for RGB-D saliency detection. In ICIP, pages 3925–3929. IEEE, 2019. 1, 2

[38] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for RGB-D saliency detection. In CVPR, 2020. 5, 6, 7

[39] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In ICCV, 2021. 1

[40] Nian Liu, Wangbo Zhao, Dingwen Zhang, Junwei Han, and Shao Ling. Light field saliency detection with dual local graph learning and reciprocative guidance. In ICCV, 2021. 9

[41] Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing, 363:46–57, 2019. 1

[42] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In CVPR, pages 13379–13389, 2020. 2, 3

[43] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural net-

works for RGB-D salient object detection. In ECCV. Springer, 2020. 5, 6

[44] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In ICML, 2011. 3

[45] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In CVPR, pages 454–461. IEEE, 2012. 5, 6, 7

[46] 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, 8

[47] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In CVPR, pages 733–740. IEEE, 2012. 5, 6

[48] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In ICCV, pages 7254–7263, 2019. 2, 5, 6

[49] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In CVPR, 2020. 5, 6, 7

[50] 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. 2, 5, 6

[51] Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In CVPR, pages 1454–1461, 2009. 1

[52] 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. 1, 2

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, pages 234–241. Springer, 2015. 2, 3

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, et al. Imagenet large scale visual recognition challenge. IJCV, 115(3):211–252, 2015. 3, 5

[55] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In ECCV, pages 218–234. Springer, 2016. 1

[56] 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, 5, 6

[57] Ningning Wang and Xiaojin Gong. Adaptive fusion for RGB-D salient object detection. IEEE Access, 7:55277–55284, 2019. 1, 5, 6

[58] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. IEEE TPAMI, 40(1):20–33, 2017. 1

[59] Jun Wei, Shuhui Wang, and Qingming Huang. F3Net: Fusion, feedback and focus for salient object detection. AAAI, 2019. 5

[60] Martha White, Xinhua Zhang, Dale Schuurmans, and Yao-liang Yu. Convex multi-view subspace learning. In NIPS, pages 1673–1681, 2012. 3

[61] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In CVPR, pages 3907–3916, 2019. 4

[62] Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In CVPR, pages 4279–4287, 2017. 3

[63] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. Uncertainty inspired RGB-D saliency detection. IEEE TPAMI, 2021. 2, 5, 6, 7

[64] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. RGB-D saliency detection via cascaded mutual information minimization. In ICCV, 2021. 1

[65] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for RGB-D saliency detection. In CVPR, 2020. 5, 6, 7

[66] 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, 2, 5, 6

[67] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. IEEE TPAMI, 39(2):356–370, 2016. 1

[68] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for

robust and real-time RGB-D salient object detection. In ECCV. Springer, 2020. 5, 6, 7

[69] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. Computational Visual Media, pages 1–33, 2021. 1

[70] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. IEEE TMI, 39(9):2772–2781, 2020. 2

[71] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In IEEE ICCV, pages 4681–4691, 2021. 1

[72] Tao Zhou, Changqing Zhang, Xi Peng, Harish Bhaskar, and Jie Yang. Dual shared-specific multiview subspace clustering. IEEE TCYB, 50(8):3517–3530, 2019. 2

[73] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In ICME, pages 199–204, 2019. 1, 2

[74] Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In ICCVW, pages 3008–3014, 2017. 5, 6, 7

[75] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In ICCVW, pages 1509–1515, 2017. 2, 5, 6

[76] 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, 2014. 1