Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx
Research Article

属性特征保留的 RGB-D 显著性检测模型

周涛 [1,2], 范登平 [3](✉), 陈耿 [4], 周毅 [5], 付华柱 [6]

Abstract 基于 RGB 图和深度图的显著性目标检测 (SOD) 吸引了越来越多的研究兴趣。现有的 RGB-D 显著性模型通常采用融合策略，从 RGB 图和深度图中学习共享表征，但是很少有方法明确考虑如何保留模态的特定特征。本文的研究提出了一个新的框架，即属性特征保留网络 (Specificity-preserving Network, SPNet)，它通过探索共享信息和模态的特定属性来提高显著性检测的性能。具体来说，本文使用了两个特定模态网络和一个共享学习网络来生成特定和共享的显著性预测图。为了有效融合共享学习网络中的跨模态特征，本文提出了一个交叉增强融合模块 (Cross-enhanced Integration Module，CIM)，它将融合后的特征传播到下一层来融合跨层级信息。此外，为了捕获互补的丰富多模态信息，本文使用了多模态特征聚合模块 (Multi-modal Feature Aggregation, MFA) 将单个解码器生成的特定模态特征集成到共享解码器中，进而提高显著性检测的性能。本文使用跳跃连接将编码层和解码层之间的分层特征充分地联合。在六个公开的 RGB-D 显著性检测基准和三个伪装目标检测基准数据集上的大量实验表明，本文的 SPNet 优于前沿方法。该项目可在此处获得: https://github.com/taozh2017/SPNet。

Keywords 显著性目标检测、RGB-D、交叉增强融合模块、多模态特征聚合.

1 引言

显著性目标检测（Salient Object Detection, SOD, 也称为显著性检测）旨在模拟人类视觉注意力机制并在给定场景中定位最具视觉区分性的目标 [61]。显著性目标检测已经广泛应用于多个视觉相关的任务，例如图像理解 [109]、动作识别 [67, 71]、视频/语义分割 [71, 77] 以及行人重识别 [95]。尽管已经取得了重大进展，但是在许多具有挑战性的场景中准确定位显著性目标依旧具有挑战性，比如背景杂乱、低对比度照明条件下的场景，以及显著性目标与背景相似的场景。最近，随着智能设备中深度传感器的广泛使用，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目标检测领域 [5, 21, 24, 42, 52, 86, 87, 99, 102] 获得了越来越多的关注，而自适应融合 RGB 信息和深度图是一项具有挑战性的任务。

在过去的几年中，许多 RGB-D 显著性检测方法被提出。这些方法通常关注如何有效地融合 RGB 图和深度图。现有的融合策略可以分为早期融合、后期融合和中间融合。早期融合策略通常采用简单的串联方式来整合两种模态。例如 [53, 61, 68, 72, 79] 直接将 RGB 图和深度图整合在一起，形成四通道的输入。然而，该类型的融合策略没有考虑两种模态之间的

1 南京理工大学计算机科学与工程学院
2 系统控制与信息处理教育部重点实验室
3 苏黎世联邦理工学院 (dengpfan@gmail.com)
4 西北工业大学计算机科学与工程学院
5 东南大学计算机科学与工程学院
6 起源人工智能研究院

*** 本文为 CVMJ22 [101] 中译版，由刘静怡译，周涛、范登平校对。

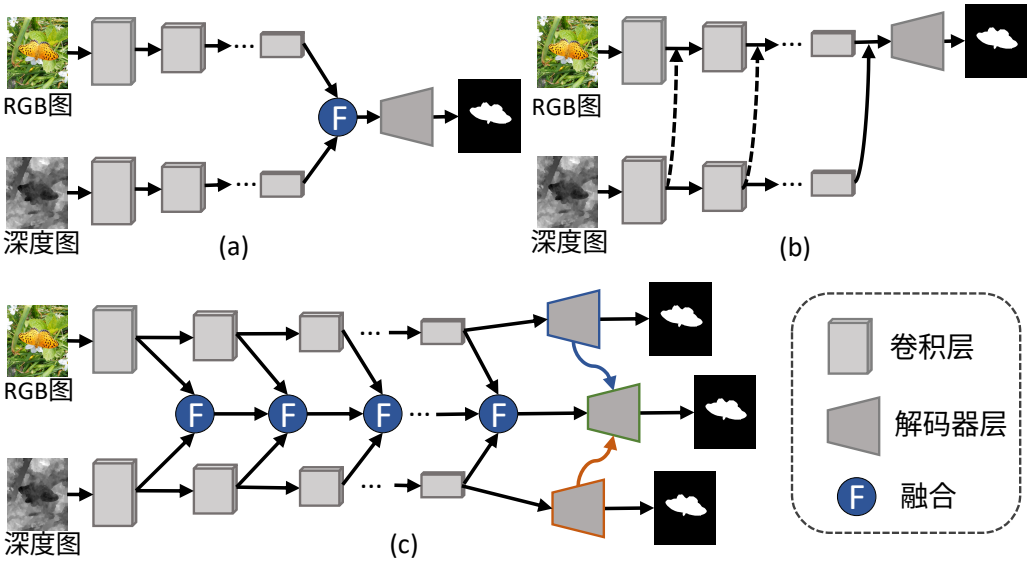

Fig. 1 现有的 RGB-D 显著性检测框架和本文的模型之间的比较。(a) RGB 图和深度图分别输入到两个独立的网络流，然后融合高层级特征并输入到一个解码器中，来预测显著性图（例如 [3, 4, 28, 48]）。(b) 使用辅助子网络将深度特征融合到 RGB 网络中（例如 [6, 22, 84, 94, 106]）。(c) 本文采用两个特定模态网络和一个共享学习网络来探索特定模态特征和共享信息。然后将特定模态解码器中学到的特征融合到共享解码器中，从而提高显著性检测的性能。

分布差距，因而会导致特征融合的不准确。后期融合策略使用两个并行的网络流为 RGB 信息和深度数据生成独立的显著性图，并将其融合来获得最终的预测图 [16, 27, 75]。但是，要捕捉这两种模态间复杂的相互关系仍然具有挑战性。

最近的研究主要集中在中间融合策略上，它利用两个独立的网络分别学习两种模态的中间特征，然后将融合之后的特征输入后续的网络或解码器中（如图 1 (a) 所示）。其它方法则在多个尺度上进行跨模态融合 [3, 4, 7, 28, 32, 33, 48]。因此，可以有效地利用两种模态间复杂的关联性。进一步的方法是利用深度信息，通过一个辅助子网络来增强 RGB 特征 [6, 94, 106]（如图 1(b) 所示）。例如，Zhao 等人 [94] 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 [106] 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。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。此外，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 [94, 106]，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 [31, 54, 103, 105]，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，很少有 RGB-D 显著性检测模型能够明确地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 SPNet），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。SPNet 使用两个编码器子网络来提取两种模态的多尺度特征（例如 RGB 图和深度图），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 [69] 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特定模态特征，它能捕获特定模态属性来提供跨模态的互补性。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，这个框架用共享信息和特定模态信息来提高显著性检测的性能。本文的主要贡献概括如下：

- 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（称为 SPNet），它可以从 RGB 图和深度图中探索共享信息，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态的共享特征。然后，每个 CIM 的输出被传播到下一层，以获取丰富的跨层信息。
- 本文提出了一个有效的多模态特征聚合模块（MFA）来融合学到的特定模态特征。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。
- 在六个公共的 RGB-D 显著性目标检测数据集和三个伪装目标检测（Camouflaged Object Detection，COD）数据集上进行的实验结果表明，本文的模型比其它前沿方法更有优势。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，光照条件以及目标尺寸），先前的工作并没有进行这种评价研究。

本文扩展了先前在 [104] 中的工作，具体包括如下几个方面：

- 本文提供更多的讨论，包括（i）讨论了本文提出的 CIM 模块和现有融合策略之间的差异，(ii) 讨论了本文提出的 CIM 模块和 MFA 模块。
- 本文提供更多的细节，包括（i）现有的 RGB 显著性目标检测方法；(ii) 对整合多层次/多尺度特征的重要性的讨论；(iii) 对本文的评估指标进行更好的描述。
- 本文通过消融实验和基于属性的评价，来验证共享解码器的有效性，并且研究不同数量的 CIM 模块对性能的影响，同时还展示了本文的模型可以有效地处理目标尺寸的变化。
- 本文将 SPNet 应用于一个新的 RGB-D 任务：伪装目标检测，并且证明了本文的模型比现有方法的优越性。

2 相关工作

本节回顾了与提出的模型相关性最高的三种类型的工作，即 RGB 显著性目标检测、RGB-D 显著性目标检测和多模态学习。

2.1 RGB 显著性目标检测

早期的显著性目标检测方法基于手工制作的特征和各种显著性先验方法，例如背景先验 [110]、颜色对比先验 [1]、紧凑性先验 [100] 以及中心先验 [36]。然而，这些传统方法的普遍性和有效性是有限的。随着深度学习在计算机视觉领域的突破，各种基于深度学习的显著性目标检测方法已经被开发出来，并取得了满意的成果。例如，Hou 等人 [30] 提出了一种新颖的显著性目标检测法，这个方法将短连接引入跳跃层结构，并将它们融入到整体嵌套的边缘检测器中。Wang 等人 [74] 提出了一种用于显著性目标检测的循环全卷积网络框架，并取得了满意的结果。Liu 等人 [50] 提出将全局上下文和局部上下文模块分层嵌入到自上而下的路径中，这个方法可以在每个像素的上下文区域里产生注意力。Deng 等人 [13] 提出了一种带有残差细化模块的循环残差细化网络，用来准确检测显著性目标。更多的方法参见综述 [76]。尺度变化是显著性目标检测任务中的一个关键挑战。因此，许多方法 [60, 78, 89, 91] 提出来整合多层次多尺度特征来改善显著性目标检测的结果。本文考虑了如何有效地结合跨模态（RGB 和深度）特征，以及如何通过交叉增强融合模块利用多层次信息。

2.2 RGB-D 显著性目标检测

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 [38] 率先在 RGB-D 显著性检测工作中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。此后，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 [27, 37]、对比度法 [14, 61, 68]、中心/边界先验法 [46, 108] 以及背景包围法 [23]。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（CNNs）的快速发展，最近一些基于深度学习的工作 [21, 63, 66, 86, 94] 取得了满意的成果。例如，Qu

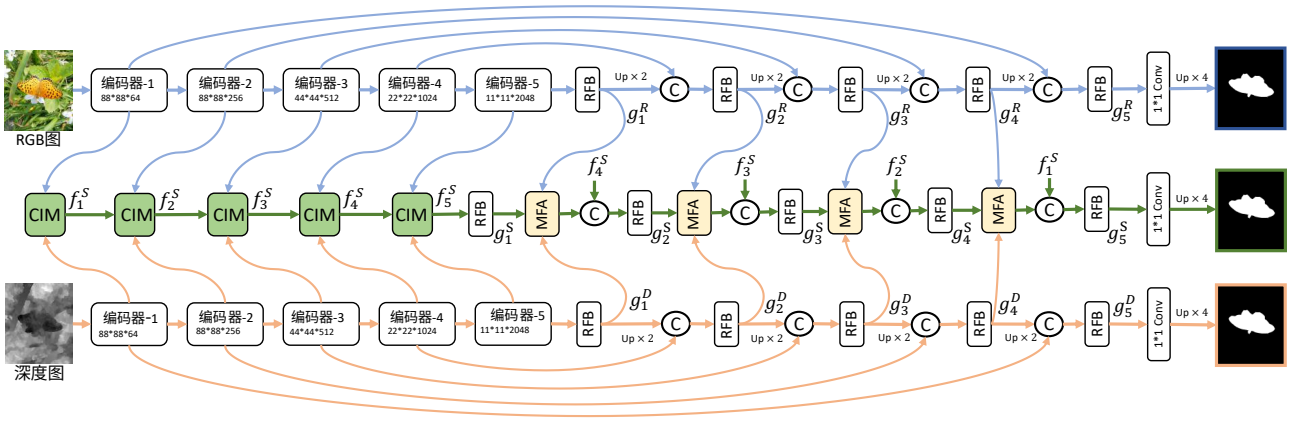

Fig. 2 本文的 SPNet 架构图，本文的模型由两个模态的特定学习网络和一个共享学习网络组成。模态的特定学习网络用于保留了 RGB 图或深度图的私有属性，而共享网络用于融合跨模态特征并探索它们的补充信息。采用跳跃连接来融合编码器层和解码器层之间的分层特征。把从特定模态解码器中学习到的特征融合到共享解码器中，来提供丰富的多模态互补信息，从而提高显著性检测的性能。这里的“C”表示特征串联。

等人 [66] 使用 CNN 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 [3] 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 [63] 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 [21] 设计了一个深度净化单元来去除那些低质量的深度图。其它的大多数模型 [4, 7, 28, 41, 44, 48] 采用不同的融合策略，在多个尺度上进行跨模态融合。

2.3 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，因为数据通常来自多个源域或者用不同类型的特征来表示。一种传统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三种类型（i）联合训练法 [2, 15] 试图最大程度地减少不同模态之间的分歧，(ii) 多核学习法 [26] 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，以及 (iii) 子空间学习法 [81, 85] 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 [56] 提出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 [17] 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 [31] 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 [54] 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

本节首先整体介绍 SPNet，然后描述模型中的两个关键组件，即特定模态学习网络和共享学习网络，最后提供整体损失函数。

3.1 概述

图 2 展示了基于属性特征保留网络的 RGB-D 显著性检测框架。首先，将 RGB 图和深度图模态输入特定双流学习网络，来获得它们的多级特征表示，然后利用提出的 CIM 模块来学习它们的共享特征。其次，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。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。最后，为了充分利用从特定模态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。下面给出每个关键部分的详细信息。

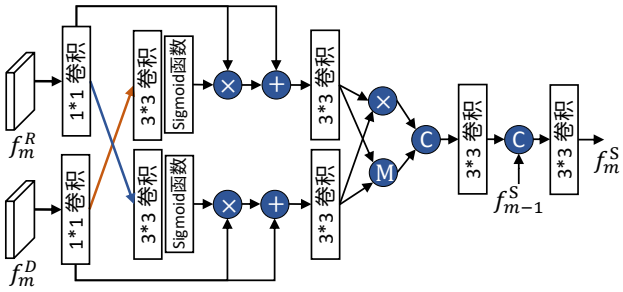

Fig. 3 交叉增强型融合模块（CIM）的示意图。图中“C”、“+”、“×”和“M”分别表示特征连接、元素相加、相乘和最大化。

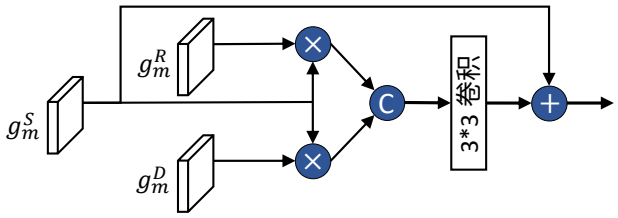

Fig. 4 多模态特征聚合模块（MFA）的示意图。其中“C”、“+”和“×”分别表示特征连接、元素相加和元素相乘。

3.2 模态特定特征学习网络

如图 2 所示，使用在 ImageNet 数据集 [70] 上预训练过的 Res2Net-50 [25] 建立模态的特定网络。因此，在 RGB 图和深度图的特定模态编码器子网络中分别有五个多层级特征，即 $F^R = [f_m^R$ 以及 $F^D = [f_m^D, m = 1, \ldots, 5]$。特定模态编码器子网络的输入分辨率为 $W \times H$。从而，第一层的特征分辨率为 $(H/8) \times (W/8)$，其他分辨率为 $(H/2^m) \times (W/2^m)$（对于 $m > 1$）。第 m 层的通道特征数记为 C_m，其中 $C_m = [64, 256, 512, 1024, 2048]$。得到高层级的特征 f_5^R 和 f_5^D 后，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。本文进一步利用 U-Net [69] 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（只有解码器子网络第一阶段的 f_5^R 或 f_5^D）被输入给感受野模块（RFB） [82] 来捕获全局上下文信息。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。

3.3 共享学习网络

3.3.1 结构

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。本文在编码器层和解码器层之间再次采用跳跃连接，以融合分层级特征。本文还充分利用了特定模态解码器学到的特征，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。

3.3.2 交叉增强融合模块

本文的 CIM 模块有效地融合跨模态的特征。将第 m 层的宽度、高度和通道数分别用 W_m、H_m 和 C_m 来表示。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，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。$w_m^R = \sigma(\mathrm{Conv}_3(f_m^R)) \in [0,1]$ 以及 $w_m^D = \sigma(\mathrm{Conv}_3(f_m^R)) \in [0,1]$，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 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。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：

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

其中，$\otimes$ 表示元素相乘。

本文获得交叉增强的特征 $f_m^{R'}$ 和 $f_m^{D'}$ 之后的关键任务就是有效地融合它们。许多策略可以用于融合不同模态的特征，包括元素相乘和最大化。但是，对于一些特定任务，尚不清楚哪种策略是最适合的。为了利用不同策略的优势，本文采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：

$$\begin{aligned} p_{\text{mul}} &= \text{BConv}_3(f_m^{R'}) \otimes \text{BConv}_3(f_m^{D'}), \\ p_{\text{max}} &= \max(\text{BConv}_3(f_m^{R'}), \text{BConv}_3(f_m^{D'})), \end{aligned} \tag{2}$$

其中，$\text{BConv}(\cdot)$ 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 $p_{\text{cat}} = [p_{\text{mul}}, p_{\text{max}}] \in \mathbb{R}^{W_m \times H_m \times C_m}$，并且通过 BConv_3 操作得到 $p_{\text{cat}}^1 = \text{BConv}_3(p_{\text{cat}})$，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 BConv_3 中进行运算。

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。一些工作 [53, 61, 72] 通过整合 RGB 图和深度图形成四通道的输入（例如使用级联操作），其它方法使用跨模态融合策略，例如使用基于注意力的融合模块 [4, 7]，融合细化模块 (例如使用求和模块) [48] 等。与这些方法不同，本文的 CIM 模块主要利用 RGB 图和深度图之间的相关性，自适应地增强跨模态特征，得到一个融合的特征表示。

3.3.3 多模态特征聚合

为了充分利用在特定模态解码器中学到的特征，本文提出了一个简单有效的 MFA 模块，该模块将那些在特定模态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 表示在特定模态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 $g_m^{RS} = g_m^S \otimes g_m^R$ 以及 $g_m^{DS} = g_m^S \otimes g_m^D$。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 $Bconv(\cdot)$ 进行运算就得到 g_m^{Sc}。最后，通过加法运算将卷积特征 g_m^{Sc} 与原始特征 g_m^S 融合得到 MFA 模块的输出。

在 MFA 模块中，学习到的特定模态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。具体来说，本文使用两个模态的特定特征 g_m^R 和 g_m^D 来增强 g_m^S。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。本文还注意到 CIM 模块和 MFA 模块之间的区别：CIM 模块用于学习融合的多模态特征表示（即 RGB 图和深度图的特征），而 MFA 模块则利用学到的模态的特定特征在共享解码器中形成一个融合的特征表示。

3.4 损失函数

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，它们分别用于特定模态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：

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

在公式中，本文利用了 $\mathcal{L}_{sp}$ 和 $\mathcal{L}_{sh}$ 的像素位置感知损失 [80]，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 实验结果与分析

本节首先给出实验设置，包括数据集、评估指标和实现细节。然后进行定量和定性评估，并通过消融实验来验证每个部分的有效性。最后，进行基于属性的评价，来展示本文模型在面对不同挑战时的有效性。

4.1 实验设置

4.1.1 数据集

为了验证模型的有效性，本文在六个公开 RGB-D SOD 数据集上对其进行了评估，包括 NJU2K [37]、NLPR [61]、DES [10]、SSD [107]、STERE [57] 和 SIP [21]。每个数据集的详细信息参见 https://github.com/taozh2017/RGBD-SODsurvey。公平起见，本文根据 [21, 63] 的设置生成训练和测试集。训练集总共包含 2195 个样本，其中，1485 个样本来自 NJU2K 数据集 [37]，700 个样本来自 NLPR 数据集 [61]。其

Tab. 1 对比 8 个具有代表性的传统模型和 23 个深度模型在六个公共 RGB-D 数据集上的结果。使用四种广泛使用的评估指标对结果进行基准测试：S_α [8], 最大 E_ϕ[18]、最大 F_β [1] 和 $\mathcal{M}$ [62])。“↑ / ↓” 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**粗体**突出显示。

	NJU2K [37]				STERE [57]				DES [10]				NLPR [61]				SSD [107]				SIP [21]			
模型	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$\mathcal{M}\downarrow$
LHM_{14} [61]	.514	.632	.724	.205	.562	.683	.771	.172	.562	.511	.653	.114	.630	.622	.766	.108	.566	.568	.717	.195	.511	.574	.716	.184
$ACSD_{14}$ [37]	.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} [23]	.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} [27]	.664	.748	.813	.169	.708	.755	.846	.143	.741	.741	.856	.090	.756	.713	.847	.091	.675	.710	.800	.165	.628	.661	.771	.164
$MDSF_{17}$ [72]	.748	.775	.838	.157	.728	.719	.809	.176	.741	.746	.851	.122	.805	.793	.885	.095	.673	.703	.779	.192	.717	.698	.798	.167
$CDCP_{17}$ [108]	.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} [66]	.763	.804	.864	.141	.757	.757	.847	.141	.752	.766	.870	.093	.802	.778	.880	.085	.747	.735	.828	.142	.653	.657	.759	.185
$CTMF_{18}$ [28]	.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} [3]	.877	.872	.924	.059	.875	.860	.925	.064	.842	.804	.893	.049	.874	.841	.925	.044	.841	.807	.894	.062	.842	.838	.901	.071
$AFNet_{19}$ [75]	.772	.775	.853	.100	.825	.823	.887	.075	.770	.729	.881	.068	.799	.771	.879	.058	.714	.687	.807	.118	.720	.712	.819	.118
$CPFP_{19}$ [94]	.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
$TANet_{19}$ [4]	.878	.874	.925	.060	.871	.861	.923	.060	.858	.827	.910	.046	.886	.863	.941	.041	.839	.810	.897	.063	.835	.830	.895	.075
$DMRA_{19}$ [63]	.886	.886	.927	.051	.886	.886	.938	.047	.900	.888	.943	.030	.899	.879	.947	.031	.857	.844	.906	.058	.806	.821	.875	.085
$cmSalGAN_{20}$ [35]	.903	.896	.940	.046	.900	.894	.936	.050	.913	.899	.943	.028	.922	.907	.957	.027	.791	.735	.867	.086	.865	.864	.906	.064
$ASIFNet_{20}$ [40]	.889	.888	.927	.047	.878	.878	.927	.049	.934	.935	.974	.019	.906	.888	.944	.030	.857	.834	.884	.056	.857	.859	.896	.061
$ICNet_{20}$ [43]	.894	.891	.926	.052	.903	.898	.942	.045	.920	.913	.960	.027	.923	.908	.952	.028	.848	.841	.902	.064	.854	.857	.903	.069
$A2dele_{20}$ [64]	.871	.874	.916	.051	.878	.879	.928	.044	.886	.872	.920	.029	.898	.882	.944	.029	.802	.776	.861	.070	.828	.833	.889	.070
$JL\text{-}DCF_{20}$ [24]	.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} [51]	.894	.889	.930	.053	.890	.882	.932	.051	.941	.935	.973	.021	.915	.902	.953	.030	.868	.848	.909	.052	.872	.877	.919	.057
$UCNet_{20}$ [86]	.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} [90]	.899	.896	.935	.043	.893	.890	.936	.044	.904	.884	.941	.026	.914	.896	.953	.026	.845	.824	.897	.058	.876	.882	.922	.052
$HDFNet_{20}$ [58]	.908	.911	.944	.038	.900	.900	.943	.041	.926	.921	.970	.021	.923	.917	**.963**	.023	**.879**	.870	**.925**	.045	.886	.894	.930	.047
$Cas\text{-}GNN_{20}$ [55]	.911	.903	.933	.035	.899	.901	.930	.039	.905	.906	.947	.028	.919	.904	.947	.028	.872	.862	.915	.047	.875	.879	.919	.051
$CMMS_{20}$ [41]	.900	.897	.936	.044	.895	.893	.939	.043	.937	.930	.976	.018	.915	.896	.949	.027	.874	.864	.922	.046	.872	.877	.911	.058
$CoNet_{20}$ [34]	.895	.893	.937	.046	**.908**	.905	**.949**	.040	.909	.896	.945	.028	.908	.887	.945	.031	.853	.840	.915	.059	.858	.867	.913	.063
$DANet_{20}$ [97]	.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
D^3Net_{21} [21]	.900	.900	.950	.041	.899	.891	.938	.046	.898	.885	.946	.031	.912	.897	.953	.030	.857	.834	.910	.058	.860	.861	.909	.063
SPNet（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**

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)。

4.1.2 评价指标

本文采用了四个广泛使用的指标对模型的有效性进行评估。它们的定义如下：

- **结构衡量标准**。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-measure S_α [8]

$$S_\alpha = \alpha S_o + (1-\alpha)\, S_r, \tag{4}$$

  其中，$\alpha \in [0,1]$，它是一个权衡参数，默认设置为 0.5 [8]。
- **F-测量方法**。给定一个显著性映射 S，将它转换为二进制映射 M，然后计算 Precision 和 Recall [1]

$$\text{Precision} = \frac{|M \cap G|}{|M|}, \qquad \text{Recall} = \frac{|M \cap G|}{|G|}, \tag{5}$$

其中，G 表示真实图。一种广泛使用的策略是使用一组从 0 到 255 的阈值来划分 S。对于每个阈值，本文计算一对召回率和精度值，然后将所有值组合起来得到一条 PR 曲线。F-measure F_β [1] 通过加权调和平均值将精度和召回率结合起来：

$$F_\beta = \left(1+\beta^2\right) \frac{\text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}, \tag{6}$$

其中，β^2 的值被设置为 0.3 以强调精度 [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。

- **增强对齐测量标准**。E_ϕ [18] 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

$$E_\phi = \frac{1}{WH}\sum_{i=1}^{W}\sum_{i=1}^{H}\phi_{FM}\left(i,j\right), \tag{7}$$

其中，ϕ_{FM} 表示增强对齐矩阵 [18]。

- **平均绝对误差** ($\mathcal{M}$)。它用于通过计算差值的平均值来评估真实值 (即，G) 和归一化预测 (即，S) 之间像素级的平均相对误差。它的定义为

$$\mathcal{M} = \frac{1}{W*H}\sum_{i=1}^{W}\sum_{i=1}^{H}\left|S\left(i,j\right)-G\left(i,j\right)\right|, \tag{8}$$

其中，W 和 H 分别表示宽度和高度。$\mathcal{M}$ 评估显著性预测图和真实图之间的相似性，并将其归一化为 $[0,1]$。

4.1.3 实现细节

本文的模型采用 PyTorch 实现，并在一个 32GB 内存的 Nvidia Tesla V100 GPU 上进行训练。使用在 ImageNet [70] 上进行预训练的骨干网络 Res2Net-50 [25]。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，然后将它们作为输入获得显著性预测图。将预测图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。

4.2 结果对比

4.2.1 对比的模型

将本文提出的 SPNet 模型与 31 个 RGB-D 显著性检测方法进行比较，其中包括 8 种人工设计的传统模型：LHM [61]、ACSD [37]、LBE [23]、DCMC [12]、SE [27]、MDSF [72]、CDCP [108]、DTM [11] 以及 23 种深度学习模型：DF [66]、CTMF [28]、PCF [3]、AFNet [75]、CPFP [94]、MMCI [6]、TANet [4] 、DMRA [63]、cmSalGAN [35]、ASIFNet [40]、ICNet [43]、A2dele [64]、JL-DCF [24]、S^2MA [51]、UCNet [86]、SSF [90]、HDFNet [58]、Cas-GNN [55]、CMMS [41]、D^3Net [21]、CoNet [34]、DANet [97] 和 PGAR [9]。更多细节可参见 [102]。

4.2.2 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 [61]、ACSD [37]、LBE [23]、DCMC [12]、SE [27]、MDSF [72]、CDCP [108] 和 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 SPNet 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，在图 6 中展示了 F-measure 曲线，提供了 29 种 RGB-D 显著性检测方法的结果，这其中包含 28 种具有完整显著性图的 SOTA 模型。在这些呈现的数据集上，本文模型的优越性显而易见。

此外，本文在 ReDWeb-S 数据集上将 SPNet 与最近的 13 个最先进的模型进行比较。其它模型的结果来自https://github.com/nnizhang/SMAC，而本文（使用 NJU2K 数据集 [37] 和 NLPR 数据集 [61] 进行训练）的结果在 ReDWeb-S 数据集上测试得到。结果如表 2 所示。本文方法的效果比大多数其它方法好，在 ReDWeb-S 数据集上与 UCNet 和 JL-DCF 相当。

本文进一步比较了在提出的模型中使用不同的骨干网络的效果，结果见表 3。当使用 Res2Net-50 作为骨干网络时，本文提出的模型效果更好，即使是用 ResNet-50 作为骨干网络，本文的模型仍比其它前沿方法的表现更好（对比表 1和表 3中结果）。

4.2.3 定性比较

将本文的模型与 8 个最前沿方法进行比较，图 7 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。第 2 和第

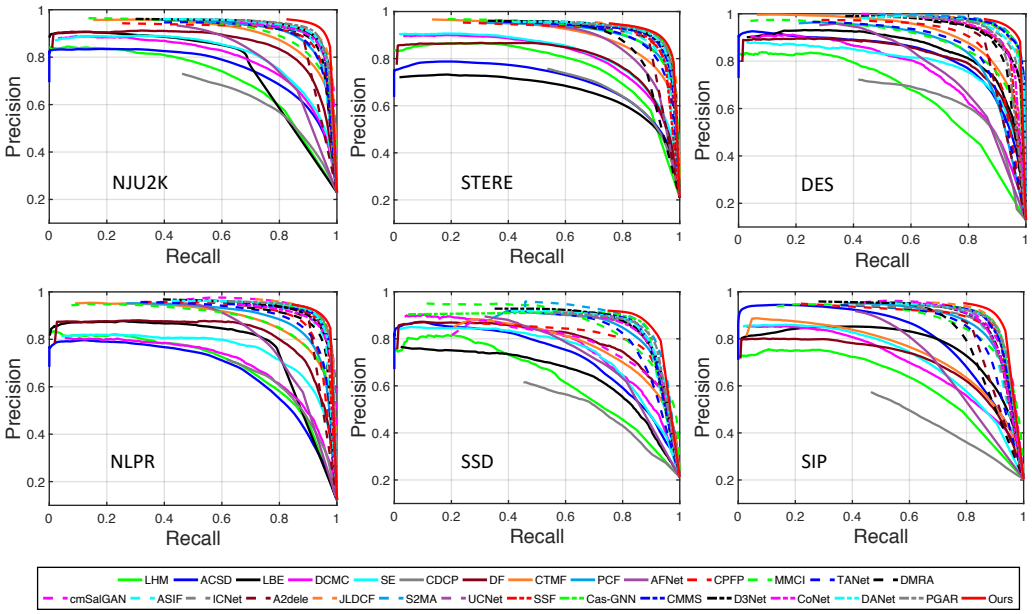

Fig. 5 六个数据集（NJU2K [37]、STERE [57]、DES [10]、NLPR [61]、SSD [107] 和 SIP [21]）上的 PR 曲线。

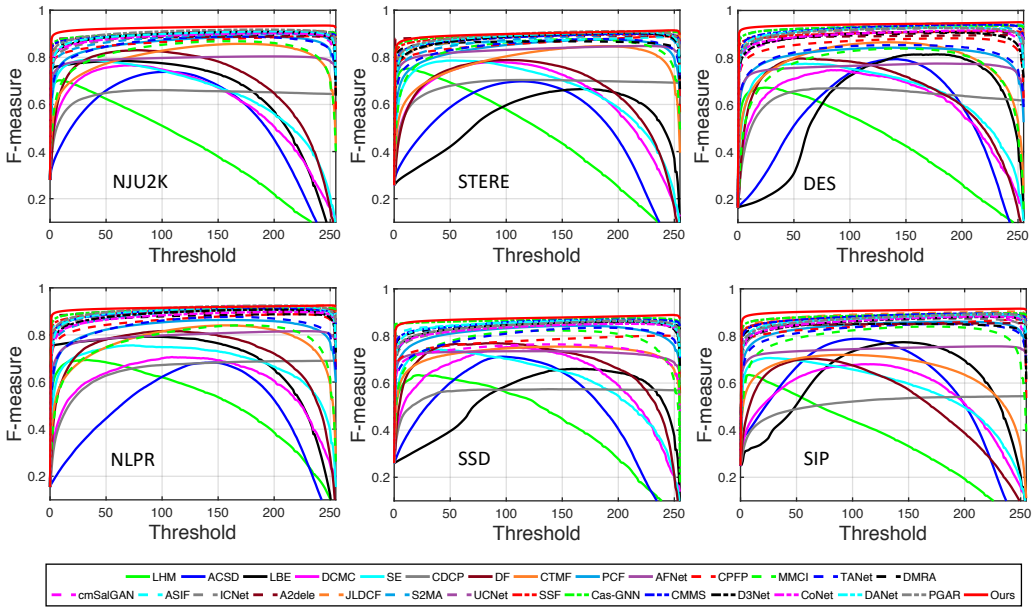

Fig. 6 在 6 个数据集（NJU2K [37]、STERE [57]、DES [10]、NLPR [61]、SSD [107] 和 SIP [21]）上不同阈值的 F 测量曲线。

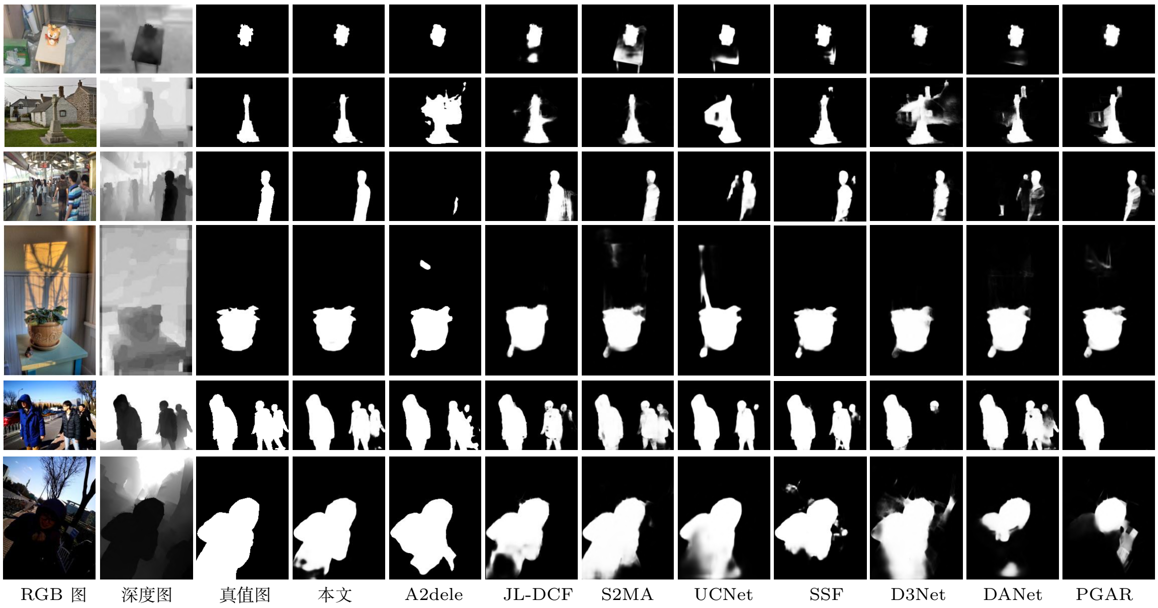

Fig. 7 本文的方法和八个最前沿方法的视觉比较，A2dele [64]、JL-DCF [24]、S2MA [51]、UCNet [86]、SSF [90]、D3Net [21]、DANet [97] 和 PGAR [9]。

Tab. 2 本文模型和 13 种最先进方法在 ReDWeb-S 数据集上的结果：CTMFF [28]、PCF [3]、AFNet [75]、MMCI [6]、CPFP [94]、DMRA [63]、TANet [4]、A2dele [64]、UCNet [86]、JL-DCF [24]、S^2MA [51]、SSF [90] 和 D^3Net [21]。

模型	CTMF	PCF	AFNet	MMCI	CPFP	DMRA	TANet	A2dele	UCNet	JL-DCF	S^2MA	SSF	D^3Net	Ours
$S_\alpha \uparrow$	0.641	0.655	0.546	0.660	0.685	0.592	0.656	0.641	0.713	0.734	0.711	0.595	0.689	0.710
$F_\beta \uparrow$	0.607	0.627	0.549	0.641	0.645	0.579	0.623	0.603	0.710	0.727	0.696	0.558	0.673	0.715
$E_\phi \uparrow$	0.739	0.743	0.693	0.754	0.744	0.721	0.741	0.672	0.794	0.805	0.781	0.710	0.768	0.800
$\mathcal{M} \downarrow$	0.204	0.166	0.213	0.176	0.142	0.188	0.165	0.160	0.130	0.128	0.139	0.189	0.149	0.129

Tab. 3 本文的模型使用不同骨干网络的结果。

骨干网络	NJU2K [37]				STERE [57]				DES [10]				NLPR [61]				SSD [107]				SIP [21]			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
ResNet-50	.922	.934	.952	.030	.904	.914	.942	.037	.936	.944	.974	.016	.930	.931	.965	.020	.869	.876	.906	.044	.896	.916	.934	.041
Res2Net-50	.925	.935	.954	.028	.907	.915	.944	.037	.945	.950	.980	.014	.927	.925	.959	.021	.871	.883	.915	.044	.894	.916	.930	.043

3 行展示了 2 个背景复杂的例子。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。

Tab. 4 不同方法的推理时间（ms）和模型规（MB）的比较。

方法	本文	JL-DCF	S2MA
模型规模	175.3	124.5	82.7
推理时间	91.7	21.8	22.1
方法	UCNet	SSF	HDFNet
模型规模	31.3	32.9	153.2
推理时间	31.8	45.7	57.1

4.2.4 推理时间和模型规模

本文在内存为 24G 的 NVIDIA TESLA P40 GPU 上测试了不同方法的推理时间。不同方法的推理时

Tab. 5 消融实验的定量评估。

	NJU2K [37]		STERE [57]		DES [10]		NLPR [61]		SSD [107]		SIP [21]	
	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$
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
C1	.913	.037	.900	.047	.935	.019	.922	.025	.861	.055	.880	.051
C2	.916	.034	.906	.040	.923	.021	.924	.022	.866	.049	.882	.051

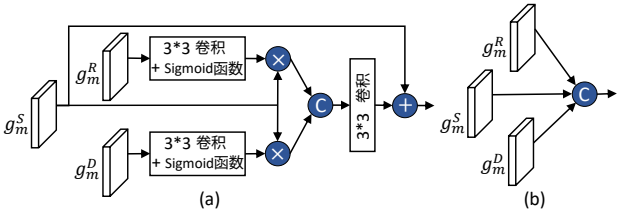

Fig. 8 MFA 模块与其它融合策略的比较。

间和模型规模如表 4 所示，这些方法包括 SPNet、JL-DCF [24]、S2MA [51]、UCNet [86]、SSF [90] 和 HDFNet [58]。由于本文的模型采用了两个模态的特定网络和一个共享学习网络来生成单独以及共享的显着性预测图。因此，与其它方法相比，本文的模型规模相对较大，在显著性预测方面也需要更多的推理时间。为此，本文希望在未来的工作中设计轻量级网络来提升效率。

4.3 消融实验

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

4.3.1 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，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，然后输入一个 3×3 的卷积层从而获得每一层的融合结果。在表 5 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。

4.3.2 MFA 模块的有效性

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态互补信息。为了证明其有效性，本文删除了这个模块并标记为“B1”。此外，本文将其它两种特征融合策略与本文的 MFA 模块进行比较。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 5 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”和本文模型相比较，可以看出 MFA 模块优于其它两种融合策略。

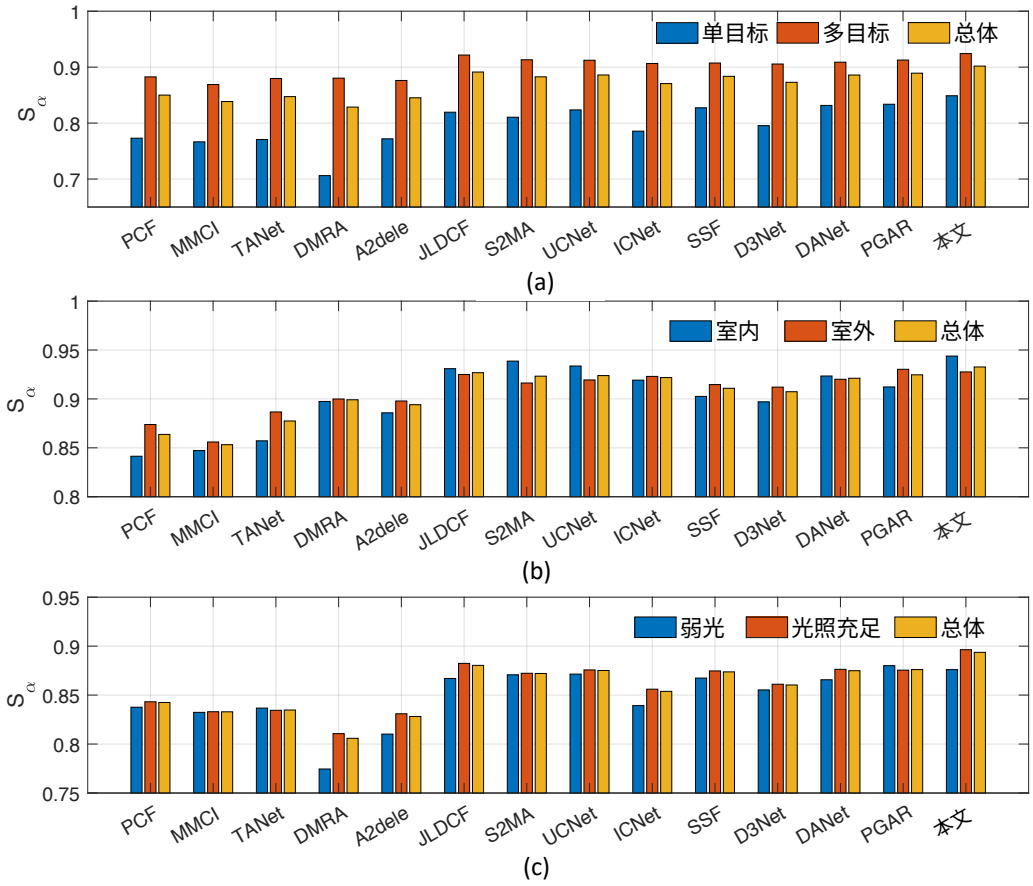

Fig. 9 基于属性的评价，涉及（a）显著性目标的数量（一个或多个），（b）室内与室外环境以及（c）光照条件（弱光与光照充足）。

4.3.3 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 5 中“C1”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。这证明了特定模态解码器的有效性，它能提供监督信号来确保学习得到模态的特定属性。为了进一步评估这两个特定模态解码器的有效性，本文增加了一个实验来对比两种情况下的显著性检测结果。一种是使用共享解码器生成的结果，另一种是使用两个特定模态解码器生成的结果。结果如表 5 中“C2”所示，可以看出，共享解码器的性能优于两个特定模态解码器的组合，这表明共享解码器可以结合多模态共享信息和特定模态特性来改善显著性检测结果。

4.3.4 不同数量的 CIM 的影响

为了研究 CIM 数量改变的影响，本文使用五个完整的 CIM 模型与两个降级的版本进行比较。“CIM_1”代表仅将 CIM 应用于编码器网络中最后一层的特。“CIM_3”代表将 CIM 应用于编码器网络中最后三层特征。结果如表 6 所示，对于大多数的数据集，使用五个 CIM 模型的效果更好。

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。因此，本文评估了不同条件下的显著性检测性能，展示了前沿算法在处理这些挑战因素时的优势和劣势。

Tab. 6 使用不同数量 CIM 模块的结果。

	NJU2K		STERE		DES		NLPR		SSD		SIP	
	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$M\downarrow$
CIM_1	.918	.034	.908	.039	.929	.019	.928	.022	.865	.047	.889	.046
CIM_3	.920	.032	.900	.041	.935	.017	.928	.021	.857	.049	.891	.045
本文	**.925**	**.028**	**.907**	**.037**	**.945**	**.014**	**.927**	**.021**	**.871**	**.044**	**.894**	**.043**

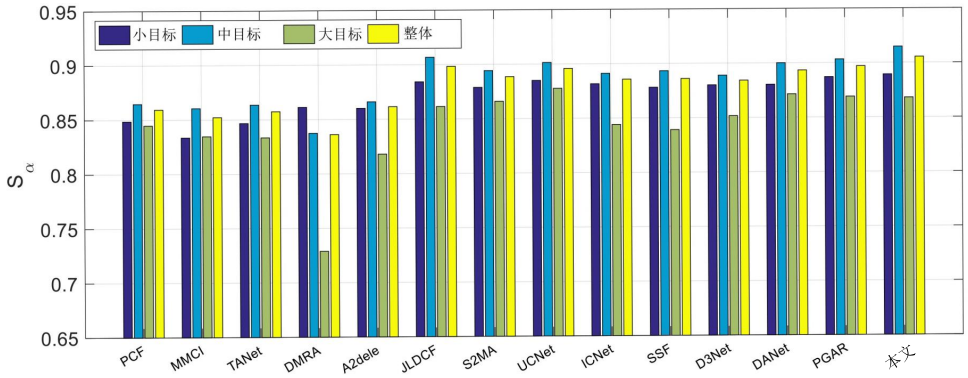

Fig. 10 基于显著性目标尺寸的属性评价。

4.4.1 单一目标与多目标的比较

在这项评估中，本文构建了一个混合数据集，由 1,229 张从 NLPR [61] 和 SIP [21] 数据集收集的图像组成。使用 S_α 指标，对比结果如图 9（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。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。

4.4.2 室内与室外场景的比较

本文评估了不同 RGB-D 显著性检测模型在室内和室外场景中的结果。DES 数据集 [10] 和 NLPR 数据集 [61] 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 9（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 JL-DCF、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。

4.4.3 光照条件

本文在 SIP 数据集 [21] 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 9（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了挑战，这证明了低光照条件对显著性目标检测性能有负面的影响。

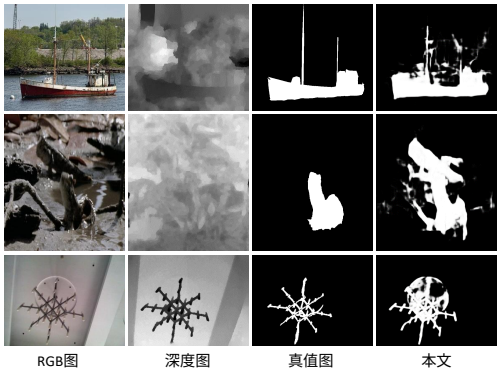

Fig. 11 本文模型的失败案例。

4.4.4 目标的尺寸

为了描述显著性目标的尺寸，本文计算显著性目标区域的尺寸与整个图像的比例 r，并定义三种目标尺寸：小目标，即 $r<0.1$ 时；大目标，即 $r>0.4$ 时；其余情况则视为中等目标。为了评估不同的方法如何处理尺寸变化，本文构建了一个混合数据集，这个数据集包含来自 STERE [57]、NLPR [61]、SSD [107]、DES [10] 和 SIP [21]，总计 2,444 张图像。图 10 展

Tab. 7 使用评价指标 S_α [8] 和 $\mathcal{M}$ [62]，在基准数据集上对伪装检测模型进行评估的评估结果。“↑ / ↓” 分别表示越大或越小越好。

模型	CHAMELEON		CAMO		COD10K	
	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$
FPN [47]	0.794	0.075	0.684	0.131	0.697	0.075
MaskRCNN [29]	0.643	0.099	0.574	0.151	0.613	0.080
PSPNet [92]	0.773	0.085	0.663	0.139	0.678	0.080
PiCANet [49]	0.769	0.085	0.609	0.156	0.649	0.090
BASNet [65]	0.687	0.118	0.618	0.159	0.634	0.105
PFANet [96]	0.679	0.144	0.659	0.172	0.636	0.128
CPD [83]	0.853	0.052	0.726	0.115	0.747	0.059
EGNet [93]	0.848	0.050	0.732	0.104	0.737	0.056
SINet [19]	0.869	0.044	0.751	0.100	0.771	0.051
DANet [98]	0.874	0.043	0.752	0.100	0.765	0.051
HDFNet [59]	0.875	0.032	0.778	0.085	0.779	0.045
SPNet（本文）	**0.895**	**0.027**	**0.795**	**0.082**	**0.797**	**0.042**

示了这种基于属性的评价在显著性目标的尺寸方面的结果。相较于检测显著性大目标，所有方法在检测显著性小目标时都有更好的表现。本文的模型以及 JL-DCF、DANet 和 PGAR 获得了令人满意的结果。

4.5 失败的案例与讨论

本文提出的 SPNet 在大多数情况下显示出良好的 RGB-D 显著性检测性能。然而，在一些具有挑战性的场景中无法检测到显著目标，比如背景复杂的场景和深度图质量较差的情况。本文模型的一些失败案例如图 11所示。在第一行中，深度图的质量很差，导致本文的模型只能粗略地定位船只，而没有识别出精细的细节。这表明，增强或过滤深度图能够改善显著性检测。在第二行中，被标注的显著性目标与场景中的其它物体的外观很相似，因而准确识别显著性目标很有难度。在第三行中，目标有精细的细节，但是本文的模型只定位了主要区域，没有定位精细的细节。本文的模型在应对这种具有精细结构的场景时仍有很大的改进空间。

4.6 应用于 RGB-D 伪装目标检测

SPNet 最初是为 RGB-D 显著性检测任务设计的，它可以很容易地扩展到其它相关的 RGB-D 任务，例如，基于 RGB-D 的伪装目标检测任务（COD）。伪装目标检测的目的是识别“无缝”嵌入其背景环境中的物体。目标物体和背景 [19, 45, 73] 之间具有很高的内在相似性，因此这是一项非常具有挑战性的任务。最近的研究 [88] 表明，深度图可以提供有效的空间信息来改善伪装检测的结果。因此，可以将本文的 SPNet 扩展到 RGB-D 伪装目标检测任务中。

本文在三个公开的基准数据集上进行了这个用于伪装物体检测的实验。(i) CHAMELEON 数据集 [19]，包括 76 张伪装图像；(ii) CAMO 数据集 [39] 包括 8 个类别的 1,250 张图像（其中 1,000 张用于训练，250 张用于测试）；(iii) COD10K 数据集 [19]，包括 5 个超级类别和 69 个子类别的 5,066 张伪装图像（其中，3,040 用于训练，2,026 用于测试）。按照与 [20] 相同的设置，将训练集和测试集分开，然后在训练集上训练本文的模型。

将本文的方法与现有的其它伪装目标检测模型进行比较，其它模型包括 FPN [47]、MaskRCNN [29]、PSPNet [92]、PiCANet [49]、BASNet [65]、PFANet [96]、CPD [83]、EGNet [93] 和 SINet [20]（结果来自 [20]）。由于 RGB-D 伪装目标检测的工作很少，本文还在实验中对比了两个最近的 RGB-D 显著性检测方法，DANet [98] 和 HDFNet [59]。本文使用了 RGB 图和深度图像重新

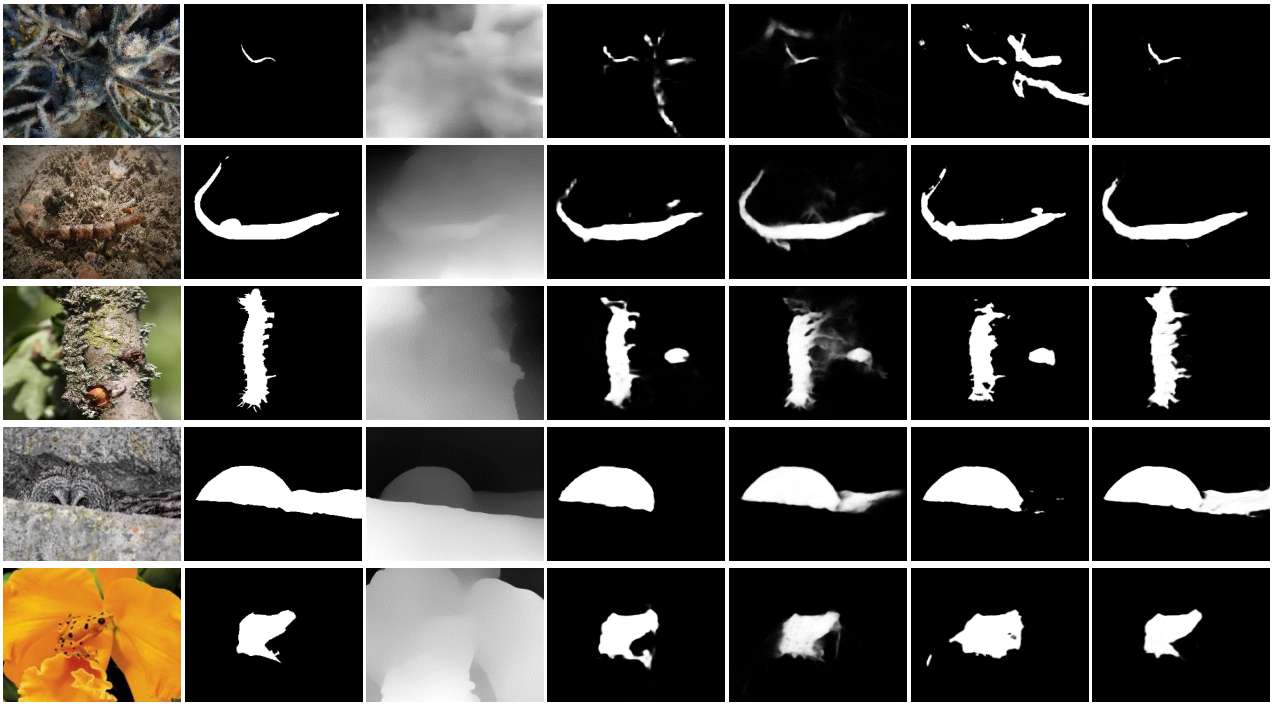

Fig. 12 本文的 SPNet 和三种前沿的伪装目标检测方法的结果对比，这三种方法是：SINet [20]，DANet [98] 和 HDFNet [59]

训练了本文的模型和这两个 RGB-D 显著性检测模型。

表 7 显示了在三个公开数据集上的定量分析结果。本文的模型比其它伪装检测方法更好。本文的模型和两种 RGB-D 伪装检测方法使用了深度线索，而且比没有使用深度线索的其它方法效果好，这表明深度线索可以提供空间信息来改善伪装检测结果。图 12 显示了不同伪装检测方法的定性分析结果。与其它伪装检测模型相比，本文的 SPNet 模型能够更加准确的检测伪装目标的边界，从而得到更好的结果。

5 结论

本文提出了一个新颖的基于属性特征保留网络的 RGB-D 显著性目标检测框架。现有大多数模型主要侧重于学习共享特征，与这些模型不同，本文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。为了学习两种模态的共享特征，本文引入了一个交叉增强融合模块（CIM）来融合跨模态的特征，每个 CIM 的输出被传播到下一层，用来探索丰富的跨层信息。本文进一步采用多模态特征聚合模块（MFA）来整合学习到的特定模态特征，从而增强多模态信息的互补性。在基准数据集上的大量结果表明，与其它最前沿的 RGB-D 显著性目标检测方法相比，SPNet 模型是有效的。此外，本文验证了 SPNet 关键组成部分的有效性，并且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。最后，本文将 SPNet 扩展到最近提出的 RGB-D 伪装目标检测任务中，并且验证了本文方法的有效性。

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[2] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136, 2009.

[3] H. Chen and Y. Li. Progressively complementarity-

aware fusion network for RGB-D salient object detection. In *CVPR*, pages 3051–3060, 2018.

[4] H. Chen and Y. Li. Three-stream attention-aware network for RGB-D salient object detection. *IEEE TIP*, 28(6):2825–2835, 2019.

[5] H. Chen, Y. Li, Y. Deng, and G. Lin. CNN-based RGB-D salient object detection: Learn, select, and fuse. *IJCV*, pages 1–21, 2021.

[6] H. Chen, Y. Li, and D. 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.

[7] H. Chen, Y.-F. Li, and D. Su. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In *IROS*, pages 6821–6826, 2018.

[8] M.-M. Chen and D.-P. Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129:2622–2638, 2021.

[9] S. Chen and Y. Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In *ECCV*, 2020.

[10] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao. Depth enhanced saliency detection method. In *ICIMCS*, pages 23–27, 2014.

[11] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE TCYB*, 2019.

[12] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *SPL*, 23(6):819–823, 2016.

[13] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690, 2018.

[14] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013.

[15] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *IEEE TMM*, 17(11):2049–2058, 2015.

[16] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang. Depth-aware saliency detection using convolutional neural networks. *JVCIR*, 61:1–9, 2019.

[17] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *IROS*, pages 681–687, 2015.

[18] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018.

[19] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE TPAMI*, 2021.

[20] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020.

[21] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE TNNLS*, 32(5):2075–2089, 2021.

[22] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao. Bbs-net: RGB-D salient object detection with a bifurcated backbone strategy network. In *ECCV*, pages 275–292, 2020.

[23] D. Feng, N. Barnes, S. You, and C. McCarthy. Local background enclosure for RGB-D salient object detection. In *CVPR*, pages 2343–2350, 2016.

[24] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu. Siamese network for RGB-D salient object detection and beyond. *IEEE TPAMI*, 2021.

[25] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 2020.

[26] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *JMLR*, 12:2211–2268, 2011.

[27] J. Guo, T. Ren, and J. Bei. Salient object detection for RGB-D image via saliency evolution. In *ICME*, pages 1–6, 2016.

[28] J. Han, H. Chen, N. Liu, C. Yan, and X. Li. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE TCYB*, 48(11):3171–3183, 2017.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[30] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017.

[31] J. Hu, J. Lu, and Y.-P. Tan. Sharable and individual multi-view metric learning. *IEEE TPAMI*, 40(9):2281–2288, 2017.

[32] Z. Huang, H.-X. Chen, T. Zhou, Y.-Z. Yang, and B.-Y. Liu. Multi-level cross-modal interaction network for rgb-d salient object detection. *Neurocomputing*, 452:200–211, 2021.

[33] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, et al. Calibrated RGB-D salient object detection. In *CVPR*, pages 9471–9481, 2021.

[34] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu. Accurate RGB-D salient object detection via collaborative learning. In *ECCV*, 2020.

[35] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo. cmsalgan: RGB-D salient object detection with cross-view generative adversarial networks. *IEEE TMM*, 2020.

[36] Z. Jiang and L. S. Davis. Submodular salient region detection. In *CVPR*, pages 2043–2050, 2013.

[37] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119, 2014.

[38] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, pages 101–115, 2012.

[39] T. Le, T. Nguyen, Z. Nie, M. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019.

[40] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE TCYB*, 2020.

[41] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy. RGB-D salient object detection with cross-modality modulation and selection. In *ECCV*, 2020.

[42] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling. Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3528–3542, 2021.

[43] G. Li, Z. Liu, and H. Ling. Icnet: Information conversion network for RGB-D based salient object detection. *IEEE TIP*, 29:4873–4884, 2020.

[44] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling. Cross-modal weighting network for RGB-D salient object detection. In *ECCV*, 2020.

[45] L. Li, B. Dong, E. Rigall, T. Zhou, J. Dong, and G. Chen. Marine animal segmentation. *IEEE TCSVT*, 2021.

[46] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. *Neurocomputing*, 275:2227–2238, 2018.

[47] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[48] D. Liu, Y. Hu, K. Zhang, and Z. Chen. Two-stream refinement network for RGB-D saliency detection. In *ICIP*, pages 3925–3929, 2019.

[49] N. Liu, J. Han, and M. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[50] N. Liu, J. Han, and M.-H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018.

[51] N. Liu, N. Zhang, and J. Han. Learning selective self-mutual attention for RGB-D saliency detection. In *CVPR*, 2020.

[52] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han. Visual saliency transformer. In *ICCV*, 2021.

[53] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao. Salient object detection for RGB-D image by single stream recurrent convolution neural network. *Neurocomputing*, 363:46–57, 2019.

[54] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020.

[55] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu. Cascade graph neural networks for RGB-D salient object detection. In *ECCV*, 2020.

[56] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[57] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012.

[58] Y. Pang, L. Zhang, X. Zhao, and H. Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In *ECCV*, 2020.

[59] Y. Pang, L. Zhang, X. Zhao, and H. Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In *ECCV*, 2020.

[60] Y. Pang, X. Zhao, L. Zhang, and H. Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.

[61] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. RGBD salient object detection: a benchmark and algorithms. In *ECCV*, pages 92–109, 2014.

[62] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012.

[63] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019.

[64] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *CVPR*, 2020.

[65] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

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

[67] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *CVPR*, pages 1454–1461, 2009.

[68] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang. Exploiting global priors for RGB-D saliency detection. In *CVPRW*, pages 25–32, 2015.

[69] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[70] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[71] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, pages 218–234, 2016.

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

[73] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu. Context-aware cross-level fusion network for camouflaged object detection. *IJCAI*, 2021.

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

[75] N. Wang and X. Gong. Adaptive fusion for RGB-D salient object detection. *IEEE Access*, 7:55277–55284, 2019.

[76] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE TPAMI*, 2021.

[77] W. Wang, J. Shen, R. Yang, and F. Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2017.

[78] X. Wang, H. Ma, X. Chen, and S. You. Edge preserving and multi-scale contextual neural network for salient object detection. *IEEE TIP*, 27(1):121–134, 2017.

[79] X. Wang, L. Zhu, S. Tang, H. Fu, P. Li, F. Wu, Y. Yang, and Y. Zhuang. Boosting rgb-d saliency detection by leveraging unlabeled rgb images. *IEEE Transactions on Image Processing*, 31:1107–1119, 2022.

[80] J. Wei, S. Wang, and Q. Huang. F3Net: Fusion, feedback and focus for salient object detection. *AAAI*, 2019.

[81] M. White, X. Zhang, D. Schuurmans, and Y.-l. Yu. Convex multi-view subspace learning. In *NeurIPS*, pages 1673–1681, 2012.

[82] Z. Wu, L. Su, and Q. Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.

[83] Z. Wu, L. Su, and Q. Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[84] Y. Zhai, D.-P. Fan, J. Yang, A. Borji, L. Shao, J. Han, and L. Wang. Bifurcated backbone strategy for rgb-d salient object detection. *IEEE TIP*, 2021.

[85] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao. Latent

multi-view subspace clustering. In *CVPR*, pages 4279–4287, 2017.

[86] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes. Uncertainty inspired RGB-D saliency detection. *IEEE TPAMI*, 2021.

[87] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao. RGB-D saliency detection via cascaded mutual information minimization. In *ICCV*, 2021.

[88] J. Zhang, Y. Lv, M. Xiang, A. Li, Y. Dai, and Y. Zhong. Depth confidence-aware camouflaged object detection. *arXiv preprint arXiv:2106.13217*, 2021.

[89] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018.

[90] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu. Select, supplement and focus for RGB-D saliency detection. In *CVPR*, 2020.

[91] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017.

[92] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[93] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng. Egnet: Edge guidance network for salient object detection. In *CVPR*, 2019.

[94] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *CVPR*, pages 3927–3936, 2019.

[95] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *IEEE TPAMI*, 39(2):356–370, 2016.

[96] T. Zhao and X. Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019.

[97] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *ECCV*, 2020.

[98] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *ECCV*, pages 646–662, 2020.

[99] Y. Zhao, J. Zhao, J. Li, and X. Chen. Rgb-d salient object detection with ubiquitous target awareness. *IEEE Transactions on Image Processing*, 30:7717–7731, 2021.

[100] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu. Salient region detection via integrating diffusion-based compactness and local contrast. *IEEE TIP*, 24(11):3308–3320, 2015.

[101] T. Zhou, D.-P. Fan, G. Chen, Y. Zhou, and H. Fu. Specificity-preserving rgb-d saliency detection. In *Computational Visual Media*, 2022.

[102] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao. RGB-D salient object detection: A survey. *CVMJ*, pages 1–33, 2021.

[103] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. *IEEE TMI*, 39(9):2772–2781, 2020.

[104] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao. Specificity-preserving RGB-D saliency detection. In *ICCV*, 2021.

[105] T. Zhou, C. Zhang, X. Peng, H. Bhaskar, and J. Yang. Dual shared-specific multiview subspace clustering. *IEEE TCYB*, 50(8):3517–3530, 2019.

[106] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *ICME*, pages 199–204, 2019.

[107] C. Zhu and G. Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *ICCVW*, pages 3008–3014, 2017.

[108] C. Zhu, G. Li, W. Wang, and R. Wang. An innovative salient object detection using center-dark channel prior. In *ICCVW*, pages 1509–1515, 2017.

[109] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE TPAMI*, 37(4):862–875, 2014.

[110] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, pages 2814–2821, 2014.