

用于RGB-D显著物体检测的孪生网络方法及其延伸

傅可人, 范登平*, 季葛鹏, 赵启军, 沈建冰, 朱策, *Fellow, IEEE*

Abstract—现有的RGB-D显著物体检测(SOD)模型通常将RGB和深度视为独立的信息, 并分别为其特征提取设计单独的网络。这种方式可能在很大程度上受限于有限的训练数据或过度依赖精心设计的训练过程。受实际中RGB和深度模态在区分显著性物体方面呈现出一定共性的启发, 本文设计了一种新的联合学习和密集协作融合(JL-DCF)架构, 采用被称为孪生结构(Siamese architecture)的共享主干网络从RGB和深度输入学习特征。本文中, 我们提出了两个有效的组件: 联合学习(JL)和密集协作融合(DCF)。JL模块基于孪生网络提供了鲁棒的显著性特征学习, 而后者DCF用于发掘互补性特征。在五种流行的评价指标上的综合实验表明, 我们所设计的框架能够生成鲁棒且具有很好的泛化性能的RGB-D显著性检测器。在七个具有挑战性的数据集上, JL-DCF较目前最好的模型极大地改进了检测性能, 获得约2.0%(最大F指标)的提升。此外, 我们展示了JL-DCF可直接用于其它相关的多模态检测任务, 包括RGB-T(热红外) SOD和视频SOD, 与现有相关前沿方法相比达到可比甚至更好的性能。另外, 我们也将JL-DCF与RGB-D语义分割领域联系起来, 展示了其可以在RGB-D SOD任务上超越几种语义分割模型。以上事实表明所提出的框架可为各种应用提供一个可行的解决方案, 并且可以为跨模态互补性任务提供更多的启示。

Index Terms—孪生网络, RGB-D SOD, 显著性检测, 显著物体检测, RGB-D语义分割。

1 引言

显著物体检测 (Salient object detection, SOD) 旨在检测场景中人类会自然关注的物体 [2], [3], [4]。其有非常多有用的应用, 包括物体分割和识别 [5], [6], [7], [8], [9], [10]、图像/视频压缩 [11]、视频检测/概括 [12], [13]、基于内容的图像编辑 [14], [15], [16], [17], [18]、发现具有信息量的通用物体 [19], [20], [21]、图像检索 [22], [23], [24]。许多显著物体检测模型是在假设输入是单幅RGB/彩色图像 [25], [26], [27], [28], [29], [30], [31] 或序列 [32], [33], [34], [35], [36]的情况下建立。随着Kinect和RealSense等深度相机越来越普及, 从RGB-D (“D”指深度)输入中进行显著物体检测正成为一个吸引人的研究方向。尽管过去有许多工作都试图探索深度信息在显著性分析中的作用, 一些问题仍然存在:

(i) 基于深度学习的RGB-D SOD方法尚未得到充分探索: 与自从2015年以来发表的100多篇关于RGB显著物体检测模

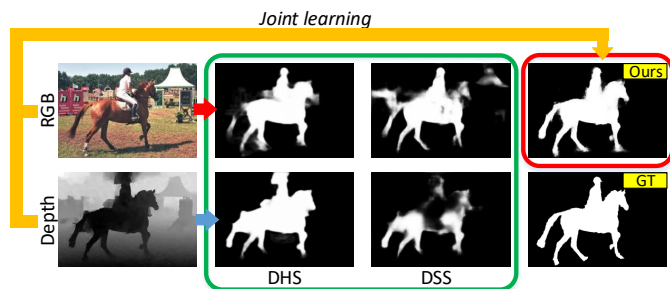


Fig. 1: 应用深度学习显著性模型DHS [37]和DSS [38]的结果。他们的输入为RGB图 (第一行) 或深度图 (第二行)。两个模型均在单一的RGB模态下训练。相比之下, 我们的JL-DCF考虑了两种模态, 能得到更好的结果 (最后一列)。

型的论文相比 [39], [40], [41], [42], [43], 只有很少的基于深度学习的RGB-D SOD工作被提出。第一个在RGB-D显著物体检测中运用卷积神经网络 (CNNs) 的模型 [44]于2017年被提出, 该模型仅采用浅层CNN作为显著图集成模型。从那时至今, 仅有十几个深度学习模型被提出, 如 [45], [46]中所述, 因此RGB-D SOD在性能上仍然有很大提升空间。

(ii) 不够有效的特征提取和融合: 大多数基于学习的模型通过早期融合 [45], [47], [48], [49], [50]或晚期融合 [51], [52]来融合不同模态的特征。尽管这两种简单的策略过去在该领域中取得了令人鼓舞的进展 (如文献 [53]中指出), 但它们在提取具有代表性的多模态特征或有效融合这些特征方面都面临着困难。而其它一些工作则采用中间融合策略 [53],

- 傅可人和赵启军来自四川大学计算机学院, 以及四川大学视觉合成图形图像技术重点实验室。(Email: fkrsuper@scu.edu.cn, qjzhao@scu.edu.cn)
- 范登平来自南开大学计算机学院。(Email: dengpingfan@mail.nankai.edu.cn)
- 季葛鹏来自武汉大学计算机学院。(Email: gepengai.ji@gmail.com)
- 沈建冰来自阿联酋起源人工智能研究院(IIAI)。(Email: shenjianbingcg@gmail.com)
- 朱策来自电子科技大学信息与通信工程学院。(Email: eczhu@uestc.edu.cn)
- 这项工作的初步版本已经在CVPR 2020 [1]上发表。
- 通信作者: 范登平。

[54], [55], 利用单独的CNN进行独立的特征提取和融合, 然而其复杂的网络结构和大量的参数需要依赖精心设计的训练过程和大量的训练数据。不幸的是, 高质量的深度图仍然是稀缺的 [56], 可能导致深度学习模型得到次优解。

研究动机: 为了解决RGB-D SOD, 我们提出一种新的联合学习和密集协作融合 (*JL-DCF*) 结构, 其性能超越现有的基于深度学习的技术。我们的方法采用上述的中间融合策略。然而, 与以往从RGB和深度视角中独立提取特征的方法不同, *JL-DCF* 通过孪生网络 [57] (即权值共享的主干网络), 同时从RGB和深度输入中提取有效的深度层次化特征。其动机是, 尽管深度图和RGB图来自不同的模态, 它们却具有相似的显著性特征/线索, 如强烈的前景-背景对比 [58], [59], [60]、物体轮廓闭合性 [61], [62]和与图像边界的连通性 [63], [64]。这使得跨模态迁移成为可能, 即使对深度学习模型亦是如此。如图1所示, 在单独RGB模态上训练的模型, 如DHS [37], 有时能在深度图上表现良好。然而, 另一个类似的模型如DSS [38], 在没有适当的适配或迁移的情况下在深度图上则可能失效。

据我们所知, 所提出的*JL-DCF* 是第一个在深度学习模型中利用这种可迁移性的方案, 其将一张深度图视为彩色图的特例并用一个共享CNN进行RGB和深度特征的提取。此外, 我们提出了一种密集协作融合策略, 来合理地融合不同模态学习到的特征。总而言之, 本文有三个主要贡献:

- 这项工作首次通过孪生网络结构来利用RGB和深度之间的共性和可转移性的工作。因此我们提出了一个通用的RGB-D SOD 框架, 称为*JL-DCF*。其由两个组件组成: 联合学习和密集协作融合。这两个组件的主要特点是它们的鲁棒性和有效性, 这将有利于未来对计算机视觉中相关多模态任务的建模。特别地, 在七个具有挑战性的数据集上, 我们极大地超越了现有前沿方法, 获得平均约2% (F指标) 的提升。此外, 通过联系RGB和RGB-D SOD 任务来增强我们的模型, 可以获得更多的提升 (参见4.3章节)。代码可在<https://github.com/kerenfu/JLDCF/> 获得。
- 我们对14种现有前沿方法 [44], [45], [47], [51], [52], [53], [55], [56], [61], [65], [66], [67], [68], [69] 进行了全面地评估。此外, 我们进行了全面的消融研究, 包括使用不同的输入模态、学习方式和特征融合策略来说明*JL-DCF* 的有效性。一些有趣的发现也将促进本领域的进一步研究。
- 我们在实验中展示了除了RGB-D SOD任务, *JL-DCF* 可直接用于其它多模态检测任务, 包括RGB-T (热红外) SOD和视频SOD。同样, *JL-DCF* 在这两个任务中和前沿方法相比都取得了相近或是更好的性能, 这也进一步验证了其鲁棒性和泛化性。据我们所知, 这应该是显著性检测领域首次证明一个所提出的框架对如此多样化的任务具有有效性。此外, 我们首次尝

试将*JL-DCF* 和RGB-D语义分割这一密切相关的领域联系起来, 与前沿分割模型进行了比较并得出相关讨论。

本文的其余部分安排如下。第2章节讨论了RGB-D SOD、计算机视觉中的孪生网络和RGB-D语义分割的相关工作。第3章节细节性地介绍所提出的*JL-DCF* 模型。第4章节给出了实验结果、性能评估和比较。最后在第5章节得出结论。

2 相关工作

2.1 RGB-D 显著物体检测

传统模型: RGB-D SOD的开创性工作是由Niu等人 [58]完成的, 他将视差对比度和领域知识引入立体摄影以测量立体显著性。在Niu的工作之后, 一开始被应用于RGB显著物体检测的各种人工特征/假设被扩展到RGB-D情形, 例如中心-周围差异 [65], [67]、对比度 [59], [60], [66]、背景包围性 [61]、中心/边界先验 [60], [63], [64], [70]、紧密性 [66], [70]或各种度量的组合 [47]。所有上述模型都十分依赖于启发式人工特性, 导致在复杂场景中的泛化性能受到局限。

深度学习模型: 通过使用深度学习和CNNs, 这一领域取得了新的进展。Qu等人 [44]首次利用CNN融合不同的底层显著性线索来判断超像素的显著值。Shigematsu等人 [62]首先提取10种基于超像素的手工深度信息特征来捕捉背景包围性、深度对比和直方图距离。这些特征被输入至CNN中, 其输出与RGB特征的输出进行浅层地融合以计算超像素的显著性。

目前, 这一领域的最新趋势是利用全卷积神经网络 (FCNs) [71]。Chen等人 [53]提出一种自底向上/自上而下的架构 [72], 在其自上而下的路径中逐步执行跨模态互补性融合。Han等人 [51]修改并扩展了基于RGB的深度神经网络结构, 使其适用于深度视角, 然后通过一个全连接层融合两个视角的深度表示。文献 [55]提出了一个三分支注意力感知网络, 其通过两个独立的分支从RGB和深度输入中提取层次特征, 然后通过第三分支中的注意力感知模块来逐步组合并选择特征。文献 [68]提出一种具有跨模态相互作用的新型多尺度多路径融合网络。文献 [48]和文献 [49]通过串联RGB和深度信息来形成四通道输入。随后, 该输入被分别输入到单流递归CNN和短连接FCN中。文献 [54]中的模型使用辅助网络来获取深度特征, 并使用它们来增强编码器-解码器结构中的中间特征表示。Zhao等人 [56]提出可生成对比度增强深度图的模型, 该深度图随后被用作流体金字塔集成中的先验图以进行特征增强。Fan等人 [45]构建了一个新的名为SIP (Salient Person)的RGB-D数据集, 并提出深度净化器网络来判断深度图是否应该与RGB图像串联以形成输入信号。Piao等人 [69]提出一种深度引导多尺度循环注意网络, 其中通过深度引导向量对融合的多尺度特征进行重新加权, 随后通过循环注意模块进行处理。Liu等人 [73]在非局部

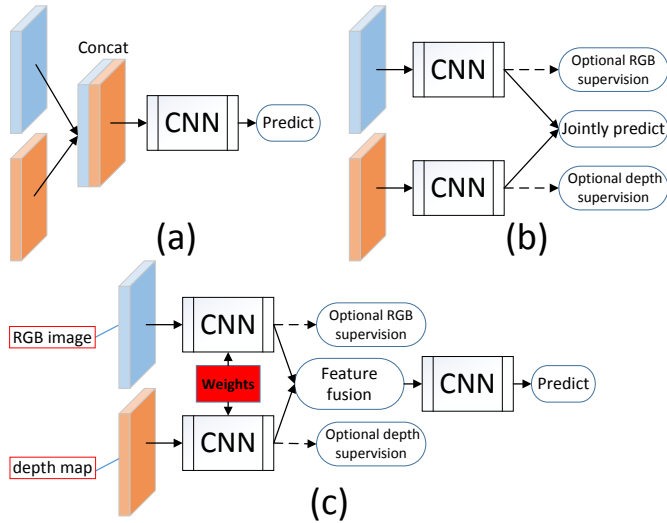


Fig. 2: RGB-D显著性检测的典型方案。(a) 早期融合(b) 晚期融合(c) 中期融合

模型 [74]的启发下, 提出一种选择性的自相互注意力机制用于RGB-D显著性检测。Zhang等人 [75]设计了一种互补的交互模块, 其在RGB和深度数据中进行有区别的选择表示, 之后通过一种新的补偿感知损失来增强学习。Piao等人 [76]提出一种自适应和注意力深度蒸馏器, 通过迁移深度知识来学习一种增强的RGB显著物体检测器。Zhang [46], [77]引入条件变分自编码器对显著性标注中的不确定性进行建模, 从而生成多个潜在的显著性图用于让一个共识模块进行投票选择。

分类和讨论: 总的来说, 正如前人文献 [53], [56]的总结, 上述方法大可分为三类: (a) 早期融合 [45], [47], [48], [49], (b) 晚期融合 [51], [52]和 (c) 中期融合 [53], [54], [55], [68], [69], [73], [75]。图2 (a)-(c) 展示了这三种融合策略。早期融合 (图2 (a))用简单的串联融合输入。这种方法可能很难捕获RGB和深度视角间的互补和交互作用, 因为两种信息在最靠前的第一阶段即进行混合, 但是其监督信号最终远离混合的输入。其学习过程容易出现局部最优的情况, 只学习到RGB或者深度特征, 因此无法保证融合后效果的提升。此外, 单独从RGB和深度视角进行深监督是不可行的。这使朝正确的方向进行学习变得困难。后期融合 (图2 (b))使用两个并行网络明确地提取RGB和深度特征。这可确保RGB和深度视角都贡献于最终决策。此外, 在这个方案中对单独视图进行监督也非常简单直接。然而, 该方案的缺点是难以挖掘两种视角间复杂的内在关联, 即那些高度非线性的互补规律。中期融合 (图2 (c))是(a)和(b)的补足, 因为特征提取和之后的融合都是由较深的卷积神经网络进行处理。因此, 它可以从两种模态中既学习高层次的概念, 又可以挖掘复杂的集成融合规则。此外, 为RGB和深度数据添加单独的深监督也很简单直接。

本文提出的JL-DCF采用中期融合策略。然而, 与上述方法 [53], [54], [55], [68], [69], [73], [75], [78], [79]采用两个独立的特征提取分支不同的是, 我们提出利用网络结构和网络权重都共享的孪生网络架构 [57], 如图2 (c)的红色部分所示。这将带来两大好处: 1) 通过联合学习可实现跨模态的知识共享; 2) 由于只需要一个共享的网络, 模型参数量极大降低从而更易于学习。

2.2 计算机视觉中的孪生网络

Bromley等人 [80]在上世纪90年代提出“孪生网络”这一概念用于手写签名的验证。他们采用两个具有完全相同参数的卷积神经网络来处理两个手写签名, 同时在学习过程中通过一些距离度量来对获得的特征向量进行约束。孪生网络这一想法后来被扩展到各类计算机视觉任务, 包括人脸验证 [57], [81], 单样本图像识别 [82], 立体匹配 [83], [84], [85], [86], [87], 目标跟踪 [88], [89], [90], [91], [92] 和半监督视频目标分割 [93], [94], [95]。孪生网络的本质和它能够被运用的原因在于它可以从两个相似的输入中, 利用某一距离 (或者相似度) 度量学习通用特征表示, 例如两张人脸图像 [57], 两个图像块 [83], [84], 一对校正后的立体图像 [86], [87], 或者是一张模板图片和一张搜索图片 [88]。在训练后, 孪生网络可以视为用于某一比较函数中的一种特征嵌入。近期工作大都在尝试操控由孪生网络获取的特征, 并以此形成优美的端到端框架 [86], [87], 或是实现更精确的特征学习和匹配 [90], [91], [92]。对孪生网络的全面总结超出了本工作的研究范围, 读者可以参考近期发布的综述工作 [96]来了解更多细节。

与上述所有工作不同的是, 在本文中, 孪生网络旨在开发利用显著性相关的跨模态共性和互补性, 而不是用于匹配或度量距离。换句话说, 为了实现所期望的RGB-D显著性预测, 孪生网络中的RGB和深度线索被融合或合并, 而不是被用于比较。值得注意的是, 孪生网络至今还没有被引入到多模态显著性检测中, 即使在整体显著性检测领域, 也极少有利用孪生网络的工作。

2.3 RGB-D 语义分割

RGB-D语义分割是RGB-D SOD的一个密切相关的研究领域。虽然这两个领域对任务有不同的定义, 但它们都是针对区域分割的。与RGB-D SOD中分割显著物体区域不同的是, RGB-D语义分割是在给定RGB-D输入的前提下, 标记预定义类别的所有像素。作为其中的一项代表性工作, Shelhamer等人 [71]使用FCNs来处理RGB-D语义分割。其通过RGB和深度输入串联作为新的输入来进行前期融合, 以及将RGB和HHA输入的得分平均化来进行后期融合。现有的RGB-D语义分割技术可以分为三类: 1) 将深度信息作为另一种输入源, 将其派生特征与RGB特征相结合 [71], [97], [98], [99], [100], [101], [102], [103], [104], [105]。2) 从RGB-D源中

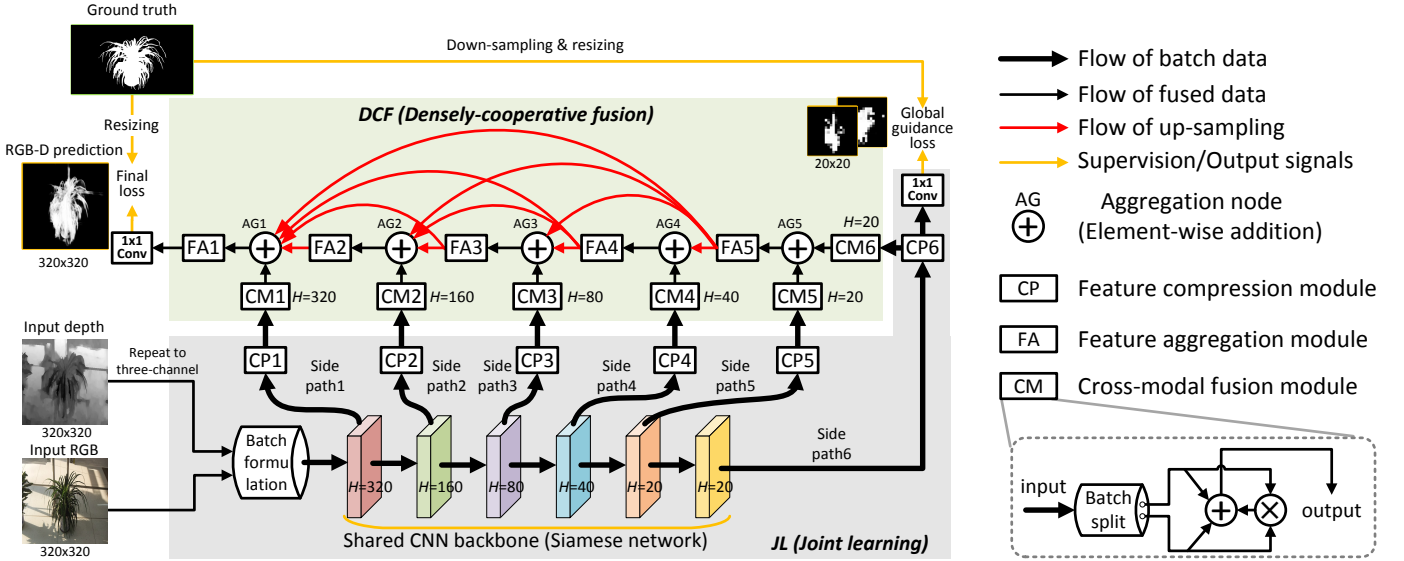


Fig. 3: 本文提出的用于RGB-D SOD的JL-DCF模型框架图。JL（联合学习）组件由灰色区域所示，而DCF（密集协作融合）组件由浅绿色区域所示。CP1~CP6：特征压缩模块。FA1~FA6：特征聚合模块。CM1~CM6：跨模态融合模块。“H”表示在特定阶段输出的特征图的空间大小。具体细节详见第3节。

恢复3D数据，并且使用3D/立体CNN来同时处理外观和几何形状 [106], [107]。3) 利用深度线索作为额外辅助信息来增强RGB模态中的特征提取 [108], [109], [110], [111], [112]，例如深度感知卷积和池化 [109]，2.5D 卷积 [111]和S卷积 [112]。值得注意的是，根据近期文献的总结 [103], [104]，第一类模型同样可以分为早期融合 [71], [97]、中期融合 [99], [100], [103], [104], [105]和晚期融合 [71], [98], [101], [102]三类（如图2所示）。这一事实揭示了RGB-D SOD和语义分割之间的强相关性。我们将在Section 4.7进一步讨论我们提出的方案和RGB-D语义分割之间的关系。

3 本文方法

本文提出的JL-DCF的总体架构如图3所示。它遵循经典的自底向上/自顶向下架构 [72]。为更好地说明，图3绘出了一个具有六个层次的主干网络示例，这样的例子在广泛应用的VGG [113]和ResNet [114]中很常见。所提出的架构由JL组件和DCF组件组成。JL组件使用了一个孪生网络对两种模态进行联合学习。其旨在从“模型共享”的角度发现这两个视角间的共性，因为它们的信息可以通过反向传播融合至模型参数中。如图3所示，由主干网络联合学习的层次化特征输入至后续的DCF组件。DCF致力于特征融合，其各层是以密集协作的方式构建。从这个意义上说，RGB和深度模态间的互补性可以从“特征整合”的角度来探索。为实现跨模态特征融合，在DCF组件中，我们精心设计了一种跨模态融合模块（即图3中的CM模块）。有关JL-DCF框架的细节将在以下章节中给出。

3.1 联合学习(JL)

如图3（灰色部分）所示，JL组件的输入是RGB图像及其对应的深度图。我们首先将深度图归一化至区间[0, 255]，然后通过颜色映射将其转换为3通道图。在实际实施中，我们简单地使用了简易灰度映射，即相当于将单通道图复制到3个通道。需要注意的是其它的颜色映射 [115]或变换，如文献 [51]中使用的方法，也可以考虑用来生成三通道表示。接下来，将三通道RGB图与变换后的深度图进行串联，从而形成一个batch，以便后续的CNN主干网能够进行并行处理。需要提到的是，与以往的早期融合方式不同，因为早期融合通常是将RGB和深度输入在第3通道维度进行串联，而我们的框架则是在第4维度进行串联，该维度通常又被称为batch维度。例如，在本文给出的例子中，转换后的 $320 \times 320 \times 3$ 深度图和 $320 \times 320 \times 3$ RGB图将形成 $320 \times 320 \times 3 \times 2$ 的batch，而不是形成 $320 \times 320 \times 6$ 的结果。

而后，从共享CNN主干网提取的层次化特征以类似文献 [38]中的侧输出方式加以利用。由于侧输出特征具有不同的分辨率和通道数（通常越深，通道越多），我们首先使用一组CP模块（图3中的CP1~CP6，实际上由卷积层和ReLU非线性实现）来将侧输出特征压缩至一个相同且较小的通道数，该通道数表示为 k 。这样做有两个原因：（1）使用大量的特征通道进行后续解码在内存和计算上开销太大；（2）统一的特征通道数易于后续进行各种逐元素操作。需要注意的是CP模块的输出仍然为batches，在图3中用粗黑色箭头表示。

粗略的物体定位能够为接下来的自顶向下的精化提供基础 [72]。此外，对粗定位进行联合学习能引导共享CNN学会同时从RGB和深度视角中提取独立的层次化特征。为了

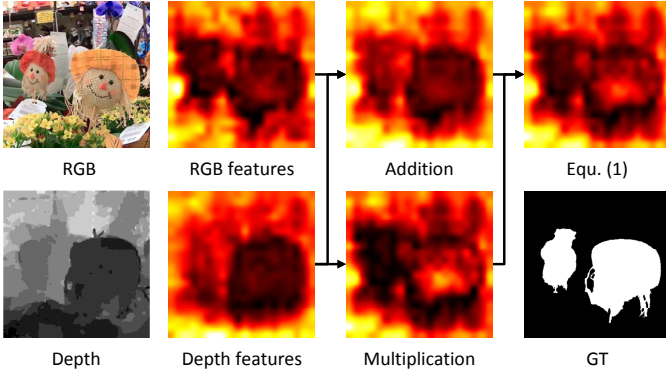


Fig. 4: CM6中间特征的可视化，即批次分离后的RGB和深度特征的可视化。通常加法和乘法收集不同的跨模态线索，使得两个玩偶的特征在运用式 (1) 处理后都得到增强。

使CNN主干能够从RGB和depth视角中同时粗略定位目标物体，我们对JL组件中的最后一个层次使用了深监督。作为实现，如图3所示，我们在CP6模块后添加 $(1 \times 1, 1)$ 卷积层以实现粗预测。深度和RGB对应的输出由下采样的真值图进行监督。这一阶段产生的损失我们称之为全局引导损失 \mathcal{L}_g 。

3.2 密集协作融合(DCF)

如图3（浅绿色部分）所示，从CP模块输出的batch特征包含深度信息和RGB信息。它们被输入至DCF组件，而DCF可被视为执行多尺度跨模态融合的解码器。首先，我们设计一种CM（跨模态融合）模块来分离和融合batch特征（图3中右下角所示）。该模块首先对batch数据进行分离，然后进行“加和乘”特征融合，称之为协同融合。在数学上，一个batch特征用 $\{X_{rgb}, X_d\}$ 表示，其中 X_{rgb} 和 X_d 分别表示RGB特征和depth特征，分别各有 k 个通道。CM模块进行融合的操作如下：

$$CM(\{X_{rgb}, X_d\}) = X_{rgb} \oplus X_d \oplus (X_{rgb} \otimes X_d), \quad (1)$$

其中“ \oplus ”和“ \otimes ”表示逐元素的相加和相乘。从CM模块输出的融合特征仍然有 k 个通道。公式 (1) 强加了由“ \oplus ”和“ \otimes ”表示的显式特征融合，其中“ \oplus ”利用了特征互补性，而“ \otimes ”更强调特征共性。如图4所示，这两种性质直观地收集了不同的线索，并通常在跨模态融合中显得较为重要。另一方面，公式 (1) 可以看作一种结合了“ $A + A \otimes B$ ”和“ $B + B \otimes A$ ”的交互残差注意力机制 [116]，这里 A 和 B 是两种类型的特征（亦即 X_{rgb}, X_d ），它们分别以残差的注意方式和另一类型特征融合。

有人也许会说上述的CM模块可被通道串联所取代，它将会产生 $2k$ 通道串联的特征。然而，我们发现这样的选择倾向于导致学习过程陷入局部最优，即偏向于只去学习RGB信息。其原因似乎是通道串联实际上涉及到的是特征选择，而非明确的特征融合。这导致了学习结果的降质，其中RGB特征很容易主导最终的预测。而值得注意的是，如4.4节将所示，

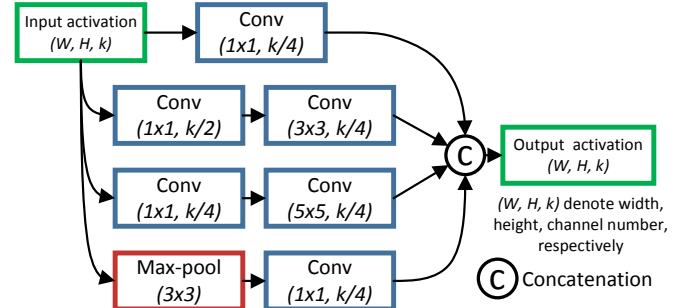


Fig. 5: 用于图3中FA模块的Inception结构。所有的卷积层和最大池层的步长为1，因此保持空间特征大小不变。与最初的Inception模块 [117]不同，我们对它进行了调整，使其具有相同的输入/输出通道数 k 。

在本文框架中仅使用RGB输入也可获得较好的性能。第4.4节将给出我们的CM模块和通道串联间的实验比较。

如图3所示，来自CM1~CM6的融合特征被送入密集连接增强后的解码器 [118]。使用密集连接可促进不同尺度上的深度特征和RGB特征的融合。因此，与传统UNet类似的解码器 [119]不同，一个聚合模块FA会接收比自身更深的所有层次（的输出）作为输入。特别地，FA表示一个执行非线性聚合和变换的特征聚合模块。为此我们使用了图5所示的Inception模块 [117]，其使用大小为 $1 \times 1, 3 \times 3, 5 \times 5$ 的滤波器以及最大池化执行多尺度卷积操作。值得一提，在我们框架中FA模块是非常灵活的。未来还可考虑使用其它模块（例如 [74], [120], [121], [122]）来提高性能。

最后，输出最细化特征的FA模块表示为FA1，其输出被送入 $(1 \times 1, 1)$ 的卷积层产生最终的激励信号，而后得到最终显著图。在训练期间，该显著图由调整大小后的真值图（GT）进行监督。我们将这一阶段产生的损失表示为 \mathcal{L}_f 。

3.3 损失函数

我们的框架的总体损失函数由全局引导损失 \mathcal{L}_g 和最终损失 \mathcal{L}_f 组成。假设 G 表示来自GT的监督， S_{rgb}^c 和 S_d^c 表示CP6之后得到batch中包含的粗预测图，而 S^f 是模块FA1之后的最终预测图。总体损失函数定义为：

$$\mathcal{L}_{total} = \mathcal{L}_f(S^f, G) + \lambda \sum_{x \in \{rgb, d\}} \mathcal{L}_g(S_x^c, G), \quad (2)$$

其中， λ 平衡了全局引导损失的权重，同时 \mathcal{L}_g 和 \mathcal{L}_f 采用被广泛使用的交叉熵损失函数：

$$\mathcal{L}(S, G) = - \sum_i [G_i \log(S_i) + (1 - G_i) \log(1 - S_i)], \quad (3)$$

其中 i 表示像素索引，而 $S \in \{S_{rgb}^c, S_d^c, S^f\}$ 。

3.4 联系RGB和RGB-D SOD

因为在 $JL-DCF$ 中, RGB和深度模态在提取特征时共享一个主干卷积神经网络, 因此很容易将 $JL-DCF$ 适配到单模态的情况(例如RGB或深度), 其只需要将所有的批次相关的操作(如图3中的批次构建和CM模块)替换成恒等映射, 同时保持包括密集解码器和深监督在内的其它设置不变。通过这种方式, 我们可以从RGB或深度输入中获得全分辨率的显著性估计结果。受此启发, 我们可以在模型训练时从数据角度联系RGB和RGB-D SOD任务。这样做的动机是为了使用更多的RGB数据来增强 $JL-DCF$ 中JL组件的泛化性, 因为JL组件被RGB和深度模态所共享。新结合进的基于RGB的知识可以帮助改进用于RGB和深度模态的孪生网络。

为此, 我们提出以多任务方式进一步扩展JL组件, 将RGB和RGB-D SOD视为两个同时的任务。如图6所示, JL组件被RGB和RGB-D SOD共享, 并且由两个任务的数据源(即训练数据)共同优化。需要提到的是, 当前可以用于训练的RGB SOD数据集比RGB-D SOD数据集大得多, 致使泛化性能得到潜在的提升。在实际实施中, 我们从JL组件为RGB SOD任务获得一张粗略显著图, 因此整体的损失函数, 即该情况下的 \mathcal{L}_{total}^* , 可以表示为两个任务的损失函数之和:

$$\mathcal{L}_{total}^* = \mathcal{L}_f(S^f, G) + \lambda \sum_{x \in \{rgb, d, rgb^*\}} \mathcal{L}_g(S_x^c, G), \quad (4)$$

其中 $S_{rgb^*}^c$ 表示RGB SOD任务对应的粗略显著图, 而其余的符号含义和公式(2)相同。更具体地, RGB SOD任务对应的RGB图像与RGB-D数据在批次维度进行串联从而形成单个批次, 然后将该批次送到CNN主干。RGB SOD任务对应的粗预测通过批次分离获得, 然后由相应的真值图进行监督。和RGB-D任务中对 S_{rgb}^c 的监督一样, 我们对RGB SOD任务同样采用标准交叉熵损失。最后, 值得一提的是, 我们的上述方法旨在采用额外的RGB SOD数据来增强RGB-D SOD。而相比之下, 最近的工作[123]旨在使用额外的RGB-D SOD数据来增强RGB SOD。

4 实验

4.1 数据集和评测指标

我们在六个公开的RGB-D评测数据集上进行了实验: NJU2K [65] (2000个样本), NLPR [59] (1000个样本), STERE [58] (1000个样本), RGBD135 [60] (135个样本), LFS [128] (100个样本)和SIP [45] (929个样本), 以及最近提出的数据集DUT-RGBD (仅测试子集, 100个样本)。参照文献[56], 我们从NLPR和NJU2K中分别选择了相同的700个和1500个样本, 合起来总共2200个样本来训练我们的算法。其余样本用于测试。此外, 当使用RGB-D和RGB数据源联合训练 $JL-DCF$ 时, 我们使用DUTS [129]的训练集来作为RGB数据集。DUTS是目前最大的、具有明确训练和测

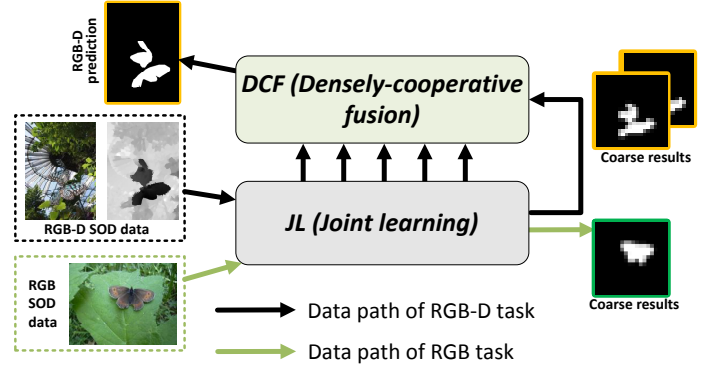


Fig. 6: 通过 $JL-DCF$ 联系RGB和RGB-D SOD任务, 其中JL和DCF组件的细节在图3中已给出。在训练中, $JL-DCF$ 的网络以一种在线的方式同时针对这两种任务进行训练/优化。

试评估协议的显著性检测数据集, 并被广泛用于训练近期的RGB SOD模型 [28], [29], [40]。为公平比较, 我们将在以上训练集上训练的模型应用于其它数据集上。

评测时, 我们采用了五个广泛采用的评测指标, 即PR曲线 [2], [125], [130]、S指标(S_α) [124]、最大F指标(F_β^{\max}) [38], [125]、最大E指标(E_ϕ^{\max}) [126]和MAE (M) [125], [127]。给出显著性图 S_{map} 和真值图 G , 这些指标的定义如下所示:

- 1) 精确率-召回率(PR) [2], [125], [130] 被定义为:

$$\text{Precision}(T) = \frac{|M(T) \cap G|}{|M(T)|}, \quad \text{Recall}(T) = \frac{|M(T) \cap G|}{|G|}, \quad (5)$$

其中 $M(T)$ 是由阈值 T 直接对显著图 S_{map} 阈值化得到的二进制掩码, $|\cdot|$ 是图中掩码的总面积。通过变化 T 可以得到精确率-召回率曲线。

- 2) S指标(S_α) [124]用来衡量显著图的空间结构相似性:

$$S_\alpha = \alpha * S_o + (1 - \alpha) * S_r, \quad (6)$$

其中 α 是对象感知结构相似性 S_o 和区域感知结构相似性 S_r 之间的平衡参数。我们参照[124]将 α 设为0.5。

- 3) F指标(F_β) [38], [125] 被定义为:

$$F_\beta = \frac{(1 + \beta^2) \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \quad (7)$$

其中 β 是用于平衡精确率和召回率的权值。因为通常情况下精确率的权重大于召回率, 所以一般将 β^2 设为0.3。为了得到单值指标, 通常利用阈值将显著图二值化为前景掩码图。本文中, 我们采用通过利用所有阈值(即[0, 255]区间)从PR曲线中计算得到的最大F指标, 即 F_β^{\max} 。

- 4) E指标(E_ϕ) [126] 是用来比较两个二进制图的强化指标。这个指标首先根据全局均值对齐两个二进制图,

TABLE 1: 定量指标: 前沿方法以及本文提出的 $JL-DCF$ 和 $JL-DCF^*$ (RGB和RGB-D数据集联合训练) 在六个RGB-D数据集上的S指标(S_α) [124], 最大F指标(F_β^{\max}) [125], 最大E指标(E_ϕ^{\max}) [126], MAE(M) [127]。最好的和次好的结果分别用**粗体**和**斜体**突出表示^a。

Metric	Traditional					Deep Learning											
	ACSD [65]	LBE [61]	DCMC [66]	MDSF [47]	SE [67]	DF [44]	AFNet [52]	CTMF [51]	MMCI [68]	PCF [53]	TANet [55]	CPFP [56]	DMRA [69]	D3Net [45]	$JL-DCF$ Ours	$JL-DCF^*$ Ours	
NJL2K [65]	$S_\alpha \uparrow$	0.699	0.695	0.686	0.748	0.664	0.763	0.772	0.849	0.858	0.877	0.878	0.879	0.886	0.895	0.903	0.911
	$F_\beta^{\max} \uparrow$	0.711	0.748	0.715	0.775	0.748	0.804	0.775	0.845	0.852	0.874	0.877	0.886	0.889	0.903	0.913	
	$E_\phi^{\max} \uparrow$	0.803	0.803	0.799	0.838	0.813	0.864	0.853	0.913	0.915	0.924	0.925	0.926	0.927	0.932	0.944	0.948
	$M \downarrow$	0.202	0.153	0.172	0.157	0.169	0.141	0.100	0.085	0.079	0.059	0.060	0.053	0.051	0.051	0.043	0.040
NLPR [59]	$S_\alpha \uparrow$	0.673	0.762	0.724	0.805	0.756	0.802	0.799	0.860	0.856	0.874	0.886	0.888	0.899	0.906	0.925	0.926
	$F_\beta^{\max} \uparrow$	0.607	0.745	0.648	0.793	0.713	0.778	0.771	0.825	0.815	0.841	0.863	0.867	0.879	0.885	0.916	0.917
	$E_\phi^{\max} \uparrow$	0.780	0.855	0.793	0.885	0.847	0.880	0.879	0.929	0.913	0.925	0.941	0.932	0.947	0.946	0.962	0.964
	$M \downarrow$	0.179	0.081	0.117	0.095	0.091	0.085	0.058	0.056	0.059	0.044	0.041	0.036	0.031	0.034	0.022	0.023
STERE [58]	$S_\alpha \uparrow$	0.692	0.660	0.731	0.728	0.708	0.757	0.825	0.848	0.873	0.875	0.871	0.879	0.886	0.891	0.905	0.911
	$F_\beta^{\max} \uparrow$	0.669	0.633	0.740	0.719	0.755	0.757	0.823	0.831	0.863	0.860	0.861	0.874	0.886	0.881	0.901	0.907
	$E_\phi^{\max} \uparrow$	0.806	0.787	0.819	0.809	0.846	0.847	0.887	0.912	0.927	0.925	0.923	0.925	0.938	0.930	0.946	0.949
	$M \downarrow$	0.200	0.250	0.148	0.176	0.143	0.141	0.075	0.086	0.068	0.064	0.060	0.051	0.047	0.054	0.042	0.039
RGBD135 [60]	$S_\alpha \uparrow$	0.728	0.703	0.707	0.741	0.741	0.752	0.770	0.863	0.848	0.842	0.858	0.872	0.900	0.904	0.929	0.936
	$F_\beta^{\max} \uparrow$	0.756	0.788	0.666	0.746	0.741	0.766	0.728	0.844	0.822	0.804	0.827	0.846	0.888	0.885	0.919	0.929
	$E_\phi^{\max} \uparrow$	0.850	0.890	0.773	0.851	0.856	0.870	0.881	0.932	0.928	0.893	0.910	0.923	0.943	0.946	0.968	0.975
	$M \downarrow$	0.169	0.208	0.111	0.122	0.090	0.093	0.068	0.055	0.065	0.049	0.046	0.038	0.030	0.030	0.022	0.021
LFSO [128]	$S_\alpha \uparrow$	0.734	0.736	0.753	0.700	0.698	0.791	0.738	0.796	0.787	0.794	0.801	0.828	0.847	0.832	0.862	0.863
	$F_\beta^{\max} \uparrow$	0.767	0.726	0.817	0.783	0.791	0.817	0.744	0.792	0.771	0.779	0.796	0.826	0.857	0.819	0.866	0.862
	$E_\phi^{\max} \uparrow$	0.837	0.804	0.856	0.826	0.840	0.865	0.815	0.865	0.839	0.835	0.847	0.872	0.901	0.864	0.901	0.900
	$M \downarrow$	0.188	0.208	0.155	0.190	0.167	0.138	0.134	0.119	0.132	0.112	0.111	0.088	0.075	0.099	0.071	0.071
SIP [45]	$S_\alpha \uparrow$	0.732	0.727	0.683	0.717	0.628	0.653	0.720	0.716	0.833	0.842	0.835	0.850	0.806	0.864	0.879	0.892
	$F_\beta^{\max} \uparrow$	0.763	0.751	0.618	0.698	0.661	0.657	0.712	0.694	0.818	0.838	0.830	0.851	0.821	0.862	0.885	0.900
	$E_\phi^{\max} \uparrow$	0.838	0.853	0.743	0.798	0.771	0.759	0.819	0.829	0.897	0.901	0.895	0.903	0.875	0.910	0.923	0.949
	$M \downarrow$	0.172	0.200	0.186	0.167	0.164	0.185	0.118	0.139	0.086	0.071	0.075	0.064	0.085	0.063	0.051	0.046

^a我们还用了不同的骨干网络 (ResNet-101, ResNet-50, 和VGG-16) 实现了 $JL-DCF$ 的Pytorch版本。与这张表中的模型相比, 他们都达到了前沿的性能。由于深度学习平台之间的差异, 通常会发现Pytorch版本比Caffe实现的性能要更好。相关结果和模型都可以在我们的项目网站上找到。

然后计算它们的局部像素相关性, 最后得到与二进制制图大小相同的增强对齐矩阵 ϕ 。 E_ϕ 则被定义为:

$$E_\phi = \frac{1}{W \cdot H} \sum_{x=1}^W \sum_{y=1}^H \phi(x, y), \quad (8)$$

其中 $\phi(x, y)$ 表示像素位置 (x, y) 处的矩阵项。 W 和 H 代表 S_{map} 的宽和高。 E_ϕ 的取值范围为区间[0,1]。为了将其扩展成可以比较非二值显著图和二值真值图的指标, 我们采用和 F_β^{\max} 相似的策略。具体地讲, 首先利用区间[0, 255]中所有可能的阈值将显著性图进行二值化得到一系列前景图, 然后选取所有前景图中最大的 E_ϕ , 即 E_ϕ^{\max} 。

5) 平均绝对误差(MAE) [125], [127]被定义为:

$$M = \frac{1}{W \cdot H} \sum_{x=1}^W \sum_{y=1}^H |S_{map}(x, y) - G(x, y)|, \quad (9)$$

其中 $S_{map}(x, y)$ 和 $G(x, y)$ 表示像素点 (x, y) 处显著性图和真值图的数值。 W 和 H 代表 S_{map} 的宽和高。

综上, 对于以上五种指标, 更高的精确率-召回率曲线、 S_α 、 F_β^{\max} 、 E_ϕ^{\max} 和更低的 M 表示具有更好的性能。

4.2 实施细节

本文提出的 $JL-DCF$ 框架与采用何种主干网络无关。在本工作中, 我们分别基于VGG-16 [113] 和ResNet-101 [114]构造

TABLE 2: 插入侧 $path1 \sim path6$ 的两个其它卷积层的细节 (对于VGG-16和ResNet-101配置)。下面括号中的参数从左到右依次为: 卷积核尺寸、通道数、步长、膨胀率和填充。

No. \ Layers	1 st Conv. layer	2 nd Conv. layer
Side path1	(3, 128, 1, 1, 1)	(3, 128, 1, 1, 1)
Side path2	(3, 128, 1, 1, 1)	(3, 128, 1, 1, 1)
Side path3	(5, 256, 1, 1, 2)	(5, 256, 1, 1, 2)
Side path4	(5, 256, 1, 1, 2)	(5, 256, 1, 1, 2)
Side path5	(5, 512, 1, 1, 2)	(5, 512, 1, 1, 2)
Side path6	(7, 512, 1, 2, 6)	(7, 512, 1, 2, 6)

了两个版本的 $JL-DCF$ 。我们将网络的输入尺度固定为 $320 \times 320 \times 3$, 并采用简单的灰度映射将深度图转换为三通道图。

VGG-16配置: 对于已去除全连接层且同时具有13个卷积层的VGG-16, 侧 $path1 \sim path6$ 依次连接到 $conv1_2$ 、 $conv2_2$ 、 $conv3_3$ 、 $conv4_3$ 、 $conv5_3$ 和 $pool5$ 。为了提高侧 $path6$ 粗糙特征图的分辨率且保持感受野不变, 我们将 $pool5$ 的步长设置为1, 而对于其对应的两个侧卷积层使用膨胀率为2的膨胀卷积 [131]。表2给出了额外侧卷积层的详细信息。总的来说, 如图3所示, 我们修改后的VGG-16主干网络生成的最粗糙的特征图的空间尺寸为 20×20 。

ResNet-101配置: 与上面的VGG-16类似, 我们修改后的ResNet-101 主干网络生成的最粗糙的特征图的大小也是 20×20 。由于ResNet的第一个卷积层的步长已经是2, 所以其最浅层的特征图大小为 160×160 。为了在不使用简单

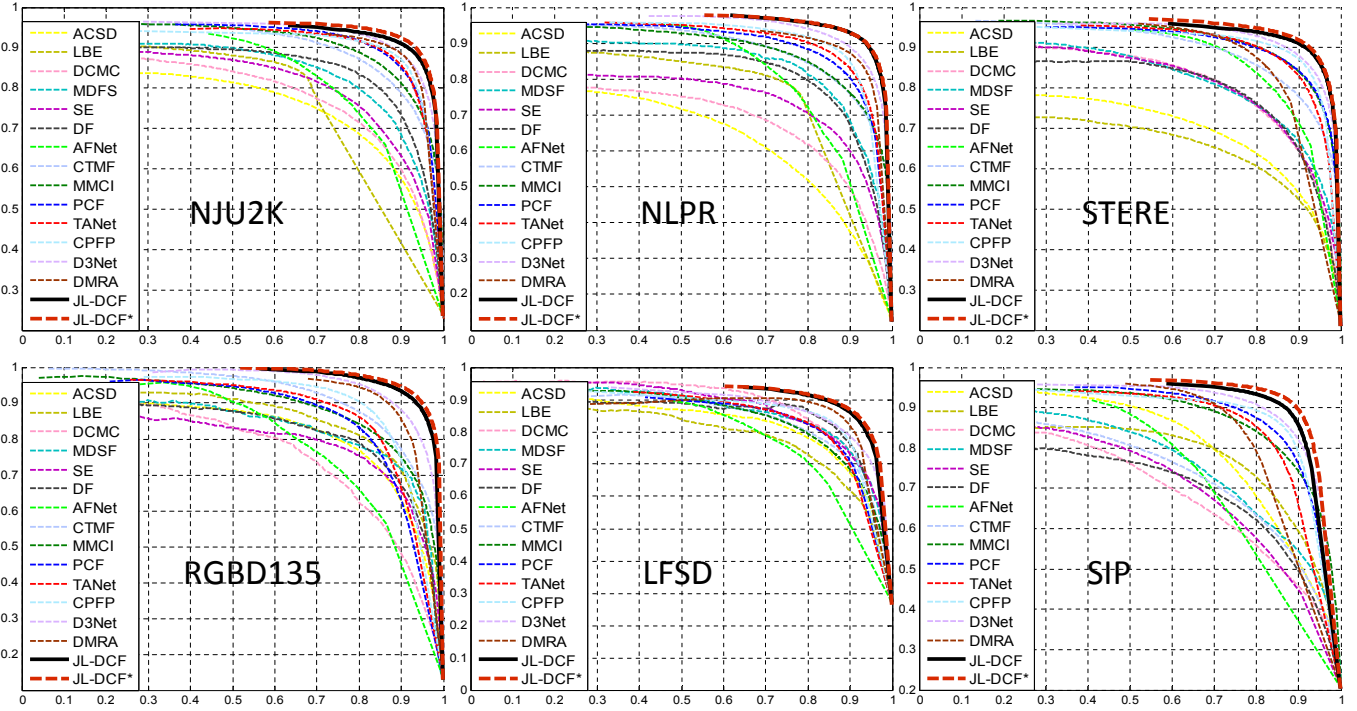


Fig. 7: 6个数据集上前沿方法、*JL-DCF* 和 *JL-DCF** 的精确率-召回率曲线

上采样的情况下获得完整大小（即 320×320 ）的特征图，我们借用了VGG-16中的 $conv1_1$ 和 $conv1_2$ 层进行特征提取。将侧 $path1 \sim path6$ 分别连接至 $conv1_2$ 和ResNet的 $conv1$ 、 $res2c$ 、 $res3b3$ 、 $res4b22$ 、 $res5c$ 。同样，我们也将 $res5a$ 块的步长从2改为1，但在后续使用了膨胀率为2的膨胀卷积。

解码器配置: 图3中的所有CP模块都是大小为 3×3 且通道数 $k = 64$ 的卷积层，所有FA模块都为前述的Inception模块。上采样操作通过简单的双线性插值实现。如图3所示，为了对应解码器中特征的尺寸，FA模块的输出会被进行不同倍数的上采样。一种极端的情况是FA5的输出被上采样2、4、8和16倍。FA1的最终输出大小为 320×320 ，与最开始的输入大小一致。

训练设置: 我们在Caffe [132]上实现了*JL-DCF*。训练期间，主干网络 [113], [114]采用预训练模型进行初始化，其他层采用随机初始化。我们通过端到端联合学习对整个网络进行了微调。训练的数据使用镜面翻转进行增强，进而产生两倍的数据量。动量参数设为0.99，学习率设为 $lr = 10^{-9}$ ，权重衰减为0.0005。式(2)中的权重 λ 设为256 ($=16^2$)来平衡高低分辨率间的损失。学习采用随机梯度下降算法并在NVIDIA 1080Ti GPU上加速。在ResNet-101/VGG-16配置下，历经40个周期的训练时间约为20/18小时。在ResNet-101上进行合并RGB数据的多任务训练，训练相同周期需要额外多7个小时。

4.3 与前沿方法对比

我们将*JL-DCF* (ResNet配置)与14种前沿方法进行比较。在比较的对象中，DF [44]、AFNet [52]、CTMF [51]、

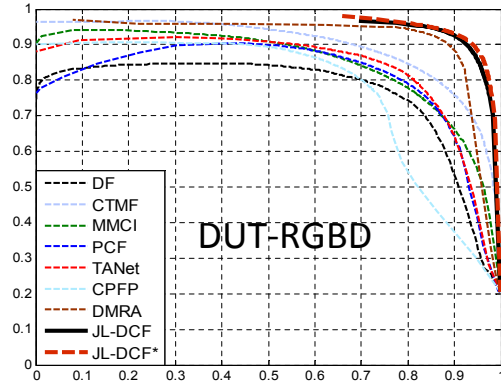


Fig. 9: 前沿方法和*JL-DCF* 在DUT-RGBD数据集上的精确率-召回率曲线 [69]。

MMCI [68]、PCF [53]、TANet [55]、CPFP [56]、D3Net [45]、DMRA [69]是最新的基于深度学习的方法，而ACSD [65]、LBE [61]、DCMC [66]、MDSF [47]、SE [67]是使用了各种人工特征/假设的传统方法。具体来说，“*JL-DCF*”指的是仅使用了RGB-D数据训练的模型，而“*JL-DCF**”指的是用RGB-D和RGB数据联合训练的模型。在6个广泛使用的数据集上的定量结果如表1所示¹。根据所有的4个指标可以看出，*JL-DCF* 相比于现有的以及最新提出的技术（例如CPFP [56]，D3Net [45]和DMRA [69]）都有显著的提升。这验证了*JL-DCF* 的一致性的有效性和泛化性。除此之外，如表1所示，

1. 我们之前的会议版本 [1]对于LFSD数据集上的评分存在一点小错误，因为我们后来发现由于格式转换损坏了真值图“29.png”。这个错误导致所有模型的性能略有下降，但没有改变它们的相对排名。我们已经修正了这个真值图以及所有的评分。

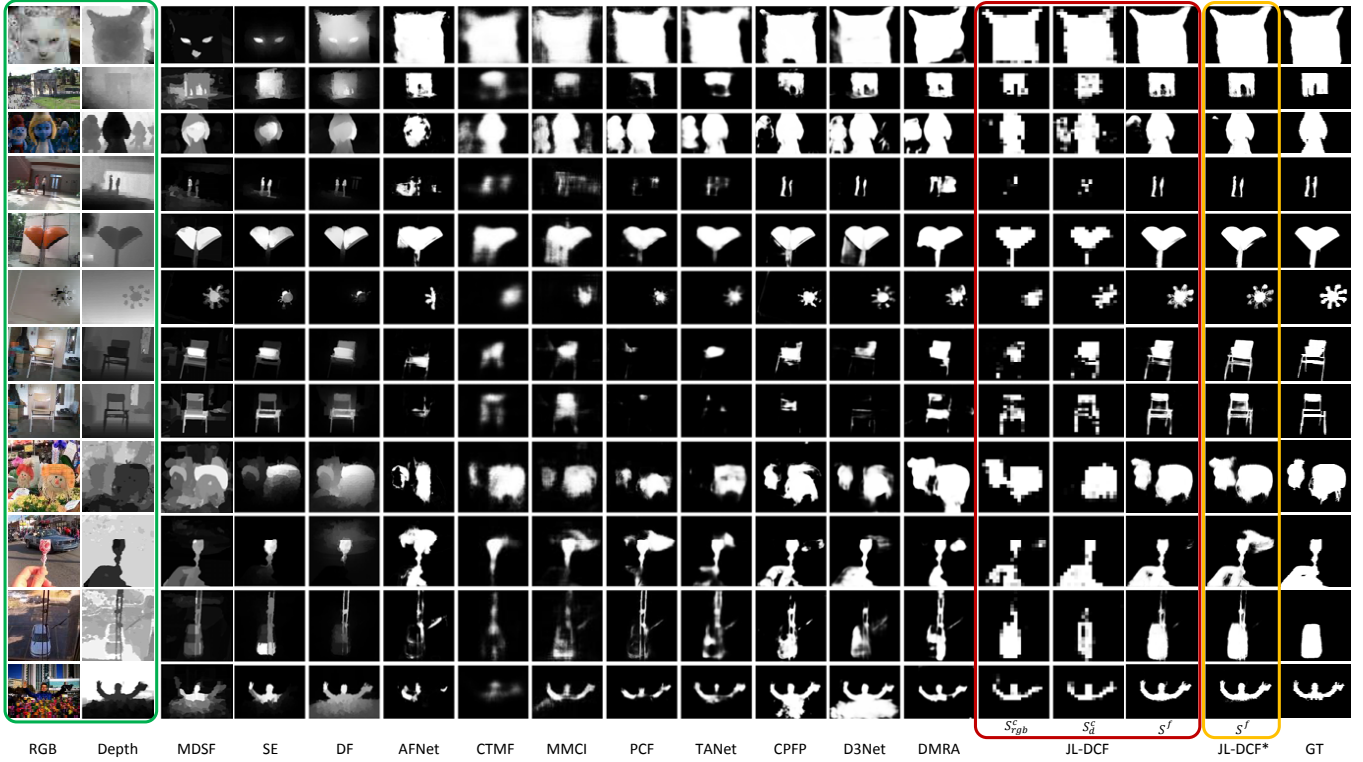


Fig. 8: $JL-DCF$ (仅使用RGB-D数据训练)和 $JL-DCF^*$ (用RGB-D数据和RGB数据共同训练)与前沿RGB-D显著性模型的比较。由RGB和深度模态联合学习的粗预测图(S_{rgb}^c and S_d^c)也和 $JL-DCF$ 的最终预测图(S^f)一起展示。

$JL-DCF^*$ 在大多数数据集上都比 $JL-DCF$ 性能更好, 这表明将知识从RGB任务迁移到RGB-D任务确实有利于后者, 并且可以带来实质性的提升。例如, 在全部的6个数据集上, $JL-DCF^*$ 的 S_α 平均提高了0.6%。图7给出精确率-召回率曲线的对比。与现有的模型相比, $JL-DCF$ 和 $JL-DCF^*$ 取得了最好的效果。

图8展示了可视化示例。 $JL-DCF$ 和 $JL-DCF^*$ 在运用深度信息进行跨模态补充方面似乎更加有效, 使其更适合在RGB-D模式下检测目标物体。此外, 图8列出了 $JL-DCF$ 的深监督粗预测图。由图可以看出, 粗预测结果为后续的跨模态细化提供了基本的物体定位支持。同时我们的密集协作融合框架学习了一种自适应的且“图像依赖”的方法, 通过此方法可以将这种物体定位支持与层次化多模态特征进行融合。这证明了融合过程对于这两种模态(RGB/深度)都没有产生退化, 从而提高了融合后的性能。

表3和图9进一步显示了最新的DUT-RGBD数据集[69]的比较结果。我们的 $JL-DCF$ 再次显示出相比于所有前沿方法的优越性能。需要注意的是, 该数据集上的实验结果清楚地验证了 $JL-DCF$ 的良好泛化性。因为虽然它并没有在含有800对RGB和深度图像的DUT-RGBD训练集上进行额外的训练, 但仍然可以以明显的提升超过训练数据已经包含了DUT-RGBD训练集的DMRA方法。

4.4 消融实验

我们通过从 $JL-DCF$ 的完整实现中移除或替换组件来进

TABLE 3: DUT-RGBD测试集(400张图片)[69]上的定量评估。被用于比较的模型是其结果在该数据集上公开可用的模型, 包括DF[44], CTMF[51], MMCI[68], PCF[53], TANet[55], CPFP[56], DMRA[69], $JL-DCF$ (Ours) 和 $JL-DCF^*$ (Ours*)

Metric	[44]	[51]	[68]	[53]	[55]	[56]	[69]	Ours	Ours*
$S_\alpha \uparrow$	0.730	0.831	0.791	0.801	0.808	0.749	0.889	0.905	0.913
$F_\beta^{\max} \uparrow$	0.734	0.823	0.767	0.771	0.790	0.718	0.898	0.911	0.916
$E_\phi^{\max} \uparrow$	0.819	0.899	0.859	0.856	0.861	0.811	0.933	0.943	0.949
$M \downarrow$	0.145	0.097	0.113	0.100	0.093	0.099	0.048	0.042	0.039

行消融研究。以 $JL-DCF$ 的ResNet版本(仅使用RGB-D数据进行训练)为参考, 将各种消融/变化结果与该版本进行对比。首先将该参考模型称为“ $JL-DCF$ (ResNet+CM+RGB-D)”, 其中“CM”表示使用了CM模块, 而“RGB-D”表示用RGB和深度图像作为输入。

首先, 为了比较不同的主干网络, 我们通过将ResNet主干网络替换为VGG并保持其它设置不变, 训练了一个“ $JL-DCF$ (VGG+CM+RGB-D)”版本。为验证协同融合模块的有效性, 我们通过将CM模块替换为通道串联操作, 训练了另一个版本“ $JL-DCF$ (ResNet+C+RGB-D)”。为说明RGB和深度信息结合的有效性, 我们分别训练了“ $JL-DCF$ (ResNet+RGB)”和“ $JL-DCF$ (ResNet+D)”两个模型, 其中将图3中所有batch相关操作(如CM模块)替换为恒等映射, 而密集解码器、深监督等其它设置保持不变。注意这个验证

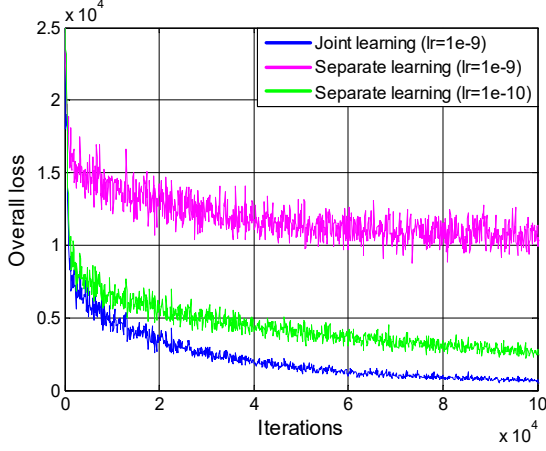


Fig. 11: 联合(JL-DCF)与分别学习(SL-DCF)的学习曲线。

是十分重要的，它表明我们的网络通过融合RGB和深度，学习到了互补信息。最后，为了说明联合学习的好处，我们训练了分别对RGB和深度使用独立主干网的“SL-DCF (VGG+CM+RGB-D)”模型。“SL”代表“分别学习”，对标所提出的“联合学习”。在该测试中，我们采用相对更小的VGG-16网络，是因为使用两个独立的主干网将使得模型的大小增加几乎一倍。

表4展示了各种评测指标的定量比较。同时列出CPFP [56]和D3Net [45]两种前沿方法作为参照。图10展示了可视化的消融对比结果。通过以上实验，可得以下五个方面的观察现象：

I. ResNet-101 vs. VGG-16: 从表4的“A”列和“B”列的比较可知，ResNet主干网络相较于VGG-16主干网络的优势是明显的，该结论也与人工作结论一致。值得注意的是，我们的VGG版本仍然优于最好的前沿方法CPFP（VGG-16主干网络）和D3Net（ResNet主干网络）。

II. CM模块的有效性: 比较“A”列和“C”列可以看出，将CM模块改为通道串联操作会降低一定程度的性能。这可能是因为整个网络倾向于只学习RGB信息，而忽略了深度信息。因为这样也能够在大数目的数据集上获得较好的结果（“D”列）。尽管串联是一种常用的特征融合的方式，但如果缺乏正确的引导，则可能会陷入次优。相比之下，CM模块对RGB和depth模态实施的则是“显式的融合操作”。

III. RGB和Depth结合的有效性: 通过在多数数据集上的持续改进清楚地验证了结合RGB和depth来提升性能的有效性（比较“A”列、“D”列和“E”列）。唯一例外的是在STERE [58]上，原因是这个数据集的深度图的质量比其他数据集要差得多。可视化示例参见图10中的第3、第4行。我们发现STERE中的很多深度图过于粗糙，而且物体边界非常不准确，甚至与真正的物体边界对不齐。融合这种不可靠的深度信息可能反而会降低性能。从表4的“E”列（STERE数据集）可以看到定量证明，即单独使用深度信息获得的性能比在其它数据集上差得多（和RGB相比， S_α/F_β^{\max} 指标下降了16%/20%）。

TABLE 4: 第4.4节所述的消融实验的定量评估结果。不同列对应不同的配置，“A”：JL-DCF (ResNet+CM+RGB-D)，“B”：JL-DCF (VGG+CM+RGB-D)，“C”：JL-DCF (ResNet+C+RGB-D)，“D”：JL-DCF (ResNet+RGB)，“E”：JL-DCF (ResNet+D)，“F”：SL-DCF (VGG+CM+RGB-D)。

	Metric	CPFP	D3Net	A	B	C	D	E	F
NJU2K [65]	$S_\alpha \uparrow$	0.878	0.895	0.903	0.897	0.900	0.895	0.865	0.886
	$F_\beta^{\max} \uparrow$	0.877	0.889	0.903	0.899	0.898	0.892	0.863	0.883
	$E_\phi^{\max} \uparrow$	0.926	0.932	0.944	0.939	0.937	0.937	0.916	0.929
	$M \downarrow$	0.053	0.051	0.043	0.044	0.045	0.046	0.063	0.053
NLPR [59]	$S_\alpha \uparrow$	0.888	0.906	0.925	0.920	0.924	0.922	0.873	0.901
	$F_\beta^{\max} \uparrow$	0.868	0.885	0.916	0.907	0.914	0.909	0.843	0.881
	$E_\phi^{\max} \uparrow$	0.932	0.946	0.962	0.959	0.961	0.957	0.930	0.946
	$M \downarrow$	0.036	0.034	0.022	0.026	0.023	0.025	0.041	0.033
STERE [58]	$S_\alpha \uparrow$	0.879	0.891	0.905	0.894	0.906	0.909	0.744	0.886
	$F_\beta^{\max} \uparrow$	0.874	0.881	0.901	0.889	0.899	0.901	0.708	0.876
	$E_\phi^{\max} \uparrow$	0.925	0.930	0.946	0.938	0.945	0.946	0.834	0.931
	$M \downarrow$	0.051	0.054	0.042	0.046	0.041	0.038	0.110	0.053
RGBD135 [60]	$S_\alpha \uparrow$	0.872	0.904	0.929	0.913	0.916	0.903	0.918	0.893
	$F_\beta^{\max} \uparrow$	0.846	0.885	0.919	0.905	0.906	0.894	0.906	0.876
	$E_\phi^{\max} \uparrow$	0.923	0.946	0.968	0.955	0.957	0.947	0.967	0.950
	$M \downarrow$	0.038	0.030	0.022	0.026	0.025	0.027	0.027	0.033
LFSD [128]	$S_\alpha \uparrow$	0.820	0.832	0.862	0.841	0.860	0.853	0.760	0.834
	$F_\beta^{\max} \uparrow$	0.821	0.819	0.866	0.844	0.858	0.850	0.768	0.832
	$E_\phi^{\max} \uparrow$	0.864	0.864	0.901	0.885	0.901	0.897	0.824	0.872
	$M \downarrow$	0.095	0.099	0.071	0.084	0.071	0.076	0.119	0.093
SIP [45]	$S_\alpha \uparrow$	0.850	0.864	0.879	0.866	0.870	0.855	0.872	0.865
	$F_\beta^{\max} \uparrow$	0.851	0.862	0.885	0.873	0.873	0.857	0.877	0.863
	$E_\phi^{\max} \uparrow$	0.903	0.910	0.923	0.916	0.916	0.908	0.920	0.913
	$M \downarrow$	0.064	0.063	0.051	0.056	0.055	0.061	0.056	0.061

IV. 仅使用RGB vs. 仅使用Depth: 表4中“D”列和“E”列的对比表明在大多数情况下使用RGB数据进行检测比使用深度数据效果更好，这说明RGB视角通常能提供更多的信息。然而在SIP [45]和RGBD135 [60]上，使用深度信息反而比RGB效果更好，如图10中所示。这说明来自这两个数据集的深度图具有相对较好的质量。

V. JL组件的有效性: 现有模型通常采用独立学习的方式分别从RGB和深度数据中提取特征。相比之下，我们的JL-DCF采用联合学习策略同时从RGB和深度图中获取特征。通过比较这两种学习策略，我们发现独立学习（两个独立的主干网络）会增加训练难度。图11展示了两种策略的一个典型的学习曲线。在独立学习的设置中，当初始学习率设为 $lr = 10^{-9}$ 时，网络会陷入局部解且损失较大，而采用联合学习设置（共享主干网络）则可以很好地收敛。此外，对于独立学习，如果将学习率设为 $lr = 10^{-10}$ ，学习过程虽然可以脱离局部振荡，但与联合学习策略相比收敛得更慢。从表4的“B”列和“F”列可以看出，经历40个周期后得到的收敛模型的性能和JL-DCF相比更差，即 S_α/F_β^{\max} 指标整体下降1.1%/1.76%。我们将JL-DCF具有更好的性能归功于RGB和深度数据的联合学习。

进一步消融分析: 除了上述的五个方面，JL-DCF中还有其它灵活的部分值得讨论，例如FA模块和密集连接。因此，我们也做了以下额外的配置“G”~“J”，其

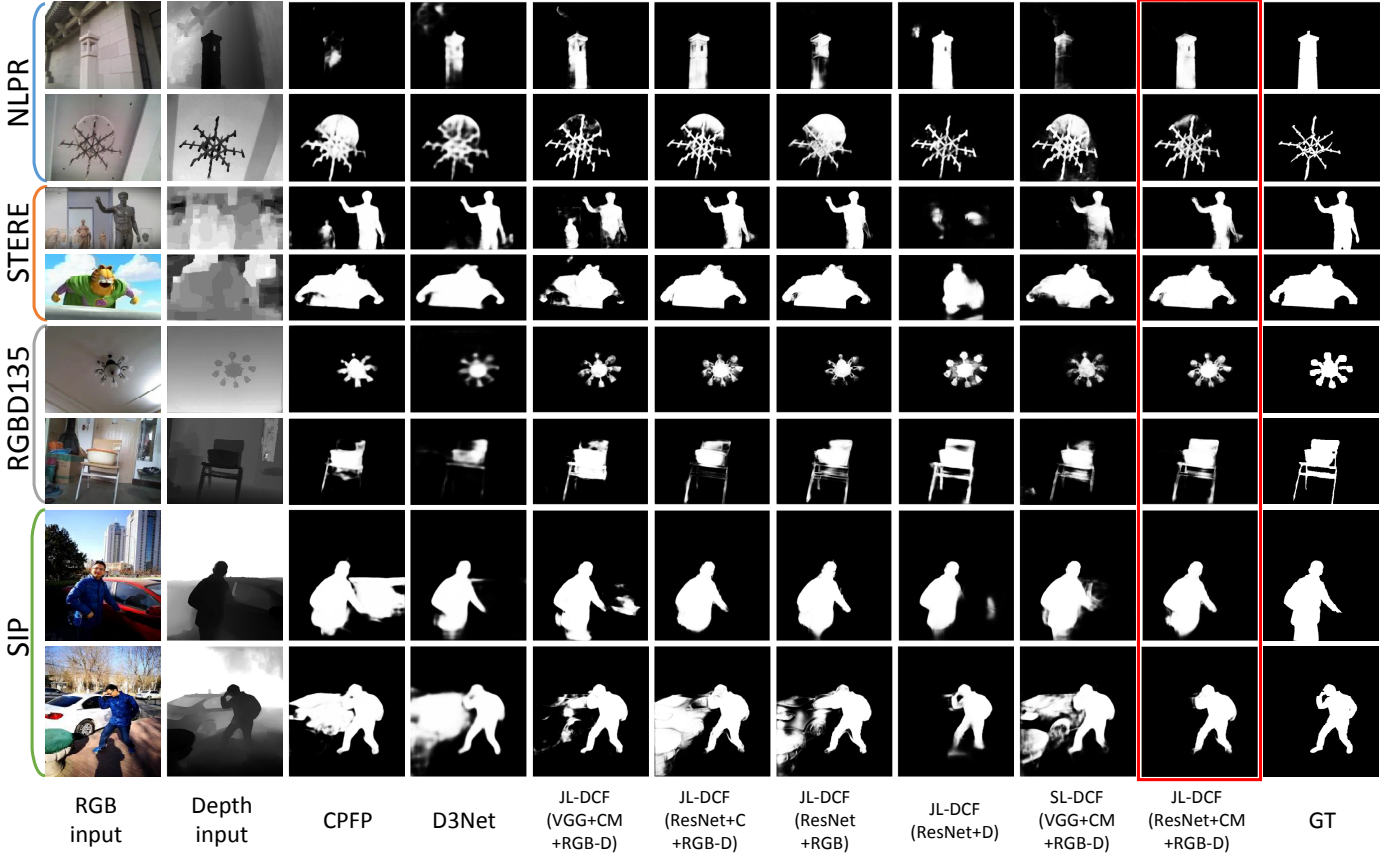


Fig. 10: 消融实验的可视化示例，示例图片来自NLPR、STERE、RGBD135和SIP数据集。总体上，*JL-DCF* 的完整实现（即ResNet+CM+RGB-D，红色方框高亮）获得了最接近真值的结果。

中“G”：从“A²”中移除所有的FA模块来得到一种退化的解码器，它采用的是线性相加所有尺度的跳层输出；“H”：从“A”中移除所有的密集连接；“I”：从“A”中移除所有的密集连接，但是只留下FA5到FA1的跳跃连接作为一种残差方式；“J”：用一种更强的DenseNet-161 [118]主干网络来替换ResNet-101主干网络，这样可以显示*JL-DCF*的潜在提升是否可以通过采用其它先进的主干网络获得。对于“J”，合并DenseNet到*JL-DCF*中则通过连接侧 $path1 \sim path6$ 到VGG-16的 $conv1_2$ （类似于4.2章中ResNet的配置）、DenseNet的 $conv0$ 、 $denseblock1 \sim denseblock4$ 来实现。

上述实验的结果展示在表5中。简而言之，我们发现添加FA模块（即“A”）作为非线性聚合会提升网络性能，而移除所有的FA模块（即“G”）会导致 F_{β}^{\max} 平均下降~1.38%。关于所采用的密集连接，可以看到“A”在大多数数据集上都比“H”的性能有提升（除了在RGBD135上获得相近的结果），这表明密集连接可以在一定程度上增强网络的鲁棒性。另一个有趣的发现是残差连接“I”在NJU2K、STERE和RGBD135上效果还不错，是可比的。这是因为虽然残差连接是从密集连接中简化的，但它减轻了深度位置信

2. 它表示前面表4中提到的“A”，且下同。

TABLE 5: 进一步消融分析，其中“G”~“J”的细节见4.4章中。这里 F_{β} 指的是 F_{β}^{\max} ，因为节省空间省略了上标。

	NJU2K		NLPR		STERE		RGBD135		LFSD		SIP	
	$S_{\alpha} \uparrow$	$F_{\beta} \uparrow$	$S_{\alpha} \uparrow$	$F_{\beta} \uparrow$	$S_{\alpha} \uparrow$	$F_{\beta} \uparrow$	$S_{\alpha} \uparrow$	$F_{\beta} \uparrow$	$S_{\alpha} \uparrow$	$F_{\beta} \uparrow$	$S_{\alpha} \uparrow$	$F_{\beta} \uparrow$
A	0.903	0.903	0.925	0.916	0.905	0.901	0.929	0.919	0.862	0.866	0.879	0.885
G	0.893	0.893	0.911	0.894	0.893	0.884	0.924	0.912	0.855	0.852	0.870	0.872
H	0.902	0.902	0.922	0.911	0.904	0.898	0.930	0.923	0.854	0.857	0.874	0.879
I	0.904	0.906	0.924	0.913	0.905	0.901	0.929	0.921	0.859	0.861	0.876	0.881
J	0.917	0.917	0.934	0.924	0.909	0.905	0.934	0.926	0.863	0.868	0.894	0.903

息的逐渐稀释并提供了额外的高层指导，正如 [133]中所观察到的那样。关于“J”，通过将主干网进一步从ResNet-101切换到DenseNet-161，我们看到了巨大的提升。这表明更强大的主干网络是可以在我们的*JL-DCF*框架中发挥作用的。

4.5 计算效率

我们在一台配置为Intel I7-8700K CPU (3.7GHz), 16G RAM, NVIDIA 1080Ti GPU的台式机上评估*JL-DCF*的计算时间。*JL-DCF*在Caffe [132]上实现。我们用LFSD的100个样本（大小调整为 320×320 ）开展测试，并且利用Caffe的Matlab接口来测试我们模型的推算时间。表6给出了平均GPU推算时间。

TABLE 6: *JL-DCF* 的平均GPU推算时间

Backbones\Components	Overall	JL	DCF
VGG-16	0.089	0.065	0.024
ResNet-101	0.111	0.087	0.024

可以发现, *JL-DCF* 中包含的孪生网络的JL (联合学习) 组件消耗了绝大多数时间, 而DCF (密集协作融合) 组件仅消耗了0.024s。注意后者实际上意味着我们引入的CM、FA模块以及密集连接只会产生很小的计算负载, 因为整个DCF组件整体上是高效的。比如说, 额外的密集连接只会增加0.008秒。此外, ResNet-101由于网络参数更多, 其计算速度比VGG-16慢了0.022s。这表明在*JL-DCF* 中, 主干网络在时间成本上占主导地位, 所以一种优化的方式是采用轻量级主干。但同时也要考虑这种优化方式对检测精度的影响。

4.6 应用于其它多模态融合任务

尽管我们提出的*JL-DCF* 的设计和评估是基于RGB-D SOD任务, 但由于其利用了跨模态共性和互补性的通用设计, 它可以应用于其它密切相关的多模态SOD任务, 例如RGB-T (“T”指热红外) SOD [134], [135], [136], [137]和视频SOD (VSOD) [13], [34], [138], [139], [140], [141]。直观上看, 显著物体会在热红外图像 (图12上半部分) 和光流图像 (图12下半部分) 中呈现出相似的显著特征, 正如它们通常在RGB图像中呈现的那样。因此对于SOD, 热力/光流图和RGB图像之间存在一定的共性, 许多传统的基于手工特征的模型 [142], [143], [144] 也都指明了这一点。解释这个概念的示例如图12所示。为了将*JL-DCF* 应用于RGB-T SOD和VSOD, 我们仅需要将*JL-DCF* 的训练数据由成对的RGB和深度数据改为成对的RGB和热力/光流数据, 而不需要对框架进行其它任何修改。此外, 因为热力图和光流图通常都会被转换为三通道RGB格式, 因此将孪生网络用于RGB vs. 热力/光流时是简单直接的。

RGB-T (热红外) SOD. 由于迄今为止与RGB-T SOD相关的工作很少 [134], [135], [136], [137], 因此缺乏被广泛认可的评估协议和基准。根据最近的一项工作 [134], 我们在VT821 [136]上测试了*JL-DCF*。VT821是一个具有821个对齐RGB和热图像样本的RGB-T SOD数据集。同时我们将得到的结果和 [134]的作者给出的结果进行了对比。我们采用包含了1000个样本的VT1000 [135]作为训练集。[134]中提出的方法被称为MIED。同时, 文献 [134]也为我们提供了在VT1000上重新训练DMRA [69]的结果图。

表7展示了VT821上的定量评估结果。这里我们报告了四个版本的*JL-DCF*: “*JL-DCF*”, “*JL-DCF(T)*”, “*JL-DCF**”和“*JL-DCF*(T)*”。“*JL-DCF*”和“*JL-DCF**”与表1中测试的模型相同, 它们是基于RGB-D SOD任务训练的。“*JL-DCF(T)*”和“*JL-DCF*(T)*”指的是在RGB-T数据 (即VT1000) 上重新训练

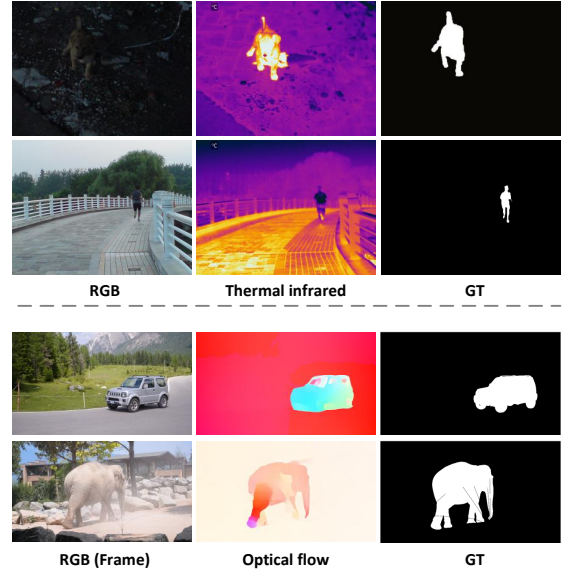


Fig. 12: 热红外图像 (上两行) /光流图像 (下两行) 和RGB SOD的共性和互补性的图示。它们和RGB视角的互补性表现在, 显著物体有时候更容易在这两种视角中进行区分。同时, 显著物体在两种视角中可能会很“突出”, 就像其有时在RGB视角表现的那样, 这表明它们和RGB模态间具有一定的共性。

TABLE 7: *JL-DCF* 和现有的RGB-T SOD 模型在VT821 [136]数据集上的对比。

Metric	MIED	DMRA	<i>JL-DCF</i>	<i>JL-DCF(T)</i>	<i>JL-DCF*</i>	<i>JL-DCF*(T)</i>
$S_{\alpha} \uparrow$	0.866	0.844	0.873	0.876	0.885	0.892
$M \downarrow$	0.053	0.049	0.037	0.037	0.031	0.033

的*JL-DCF* 模型 (训练了40个周期, 初始化方式与RGB-D任务一致), 其中后者是指用RGB-T和RGB数据联合训练的模型, 类似于前面提到的RGB-D情形。从表7可以看出, 首先, 我们的四个模型在两个指标上始终优于MIED和DMRA。令人惊讶的是, 即使是经过RGB-D数据训练的模型 (例如*JL-DCF* 和*JL-DCF**) 也可以很好地泛化到这个RGB-T SOD任务中, 这进一步验证了我们框架的鲁棒性和泛化性。同样, 使用更多的RGB数据可以提高检测精度, 而使用RGB-T数据重新训练*JL-DCF* 可以让其更好地适用于特定的任务。毫无疑问, “*JL-DCF*(T)*”具有最好的性能, 其在 S_{α} 指标上超过MIED 2.6%。

视频SOD. *JL-DCF* 也可以用于VSOD。我们首先使用FlowNet 2.0 [149] (一种用于光流估计的前沿深度学习模型) 计算RGB帧的前向光流图。一个计算出的光流图最初由两个通道表示并指示运动位移。为了将其放入*JL-DCF* 的JL组件中, 我们使用常见的流场颜色编码技术 [149]将其转换为三通道颜色图。我们使用了官方的训练集DAVIS (30个片段) [145] 和FBMS (29个片段) [146]来训练我们的模型, 这最终得到2373个样本, 每个样本包含成对的RGB和光流图。此外, 我们发现在这个任务中, 和RGB数据联合训练是

TABLE 8: *JL-DCF* 和现有VSOD模型在5个广泛使用的基准数据集上的对比。

Model	DAVIS-T [145]		FBMS-T [146]		ViSal [147]		VOS [148]		DAVSOD [13]	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
DLVS [138]	0.794	0.061	0.794	0.091	0.881	0.048	0.760	0.099	0.657	0.129
FGRN [139]	0.838	0.043	0.809	0.088	0.861	0.045	0.715	0.097	0.693	0.098
MBNM [140]	0.887	0.031	0.857	0.047	0.898	0.020	0.742	0.099	0.637	0.159
PDBM [34]	0.882	0.028	0.851	0.064	0.907	0.032	0.818	0.078	0.698	0.116
SSAV [13]	0.893	0.028	0.879	0.040	0.943	0.020	0.819	0.073	0.724	0.092
PCSA [141]	0.902	0.022	0.866	0.041	0.946	0.017	0.827	0.065	0.741	0.086
<i>JL-DCF*</i>	0.903	0.022	0.884	0.044	0.940	0.017	0.825	0.063	0.756	0.091

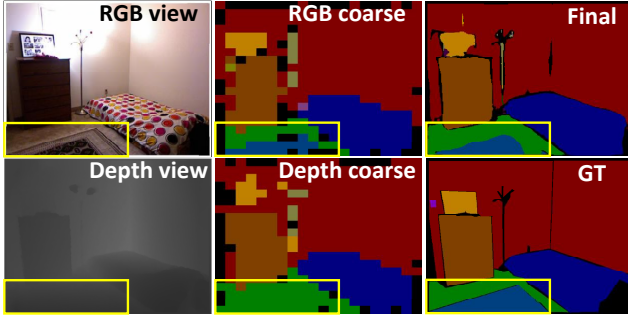


Fig. 13: 一个将*JL-DCF* 应用于RGB-D语义分割的实例。尽管类“地板垫”（黄色框中所示）在深度图中几乎难以区分，但它的主要部分仍然在最终预测中被正确识别出。

模型泛化的关键³，因为训练样本的场景多样性是非常有限的⁴。根据 [13]，实验评估基于五个广泛使用的基准数据集：FBMS-T [146] (30个片段)，DAVIS-T [145] (20个片段)，ViSal [147] (17个片段)，MCL [150] (9个片段)，UVSD [151] (18个片段)，VOS [148] ([13]中选择的40个片段)，DAVSOD [13] (含有35个片段的简单子数据集)。从表8可以看出，虽然*JL-DCF* 不是专门为VSOD设计的（没有任何长期时间上的考虑 [13], [34], [141]），但通过从RGB和运动图像中学习，它可以达到和前沿方法可比的性能，甚至在10个分数指标中有6个是最高的。这再一次应证了*JL-DCF* 可能成为解决多模态特征学习和融合问题的通用统一框架，因为它是首次同时利用跨模态共性和互补性的工作。图14展示了该任务上几项可视化对比。

4.7 联系RGB-D语义分割

据我们所知，目前尚没有模型采用孪生网络来进行RGB-D语义分割。相比之下，它们大多采用双流中间融合的方式 [99], [100], [103], [104], [105]。而*JL-DCF* 能通过简单地替换预测头来适应这个任务 [152]，即将粗预测/最终预测之前的两个 $(1 \times 1, 1)$ 卷积层替换为 $(1 \times 1, C)$ 卷积，其中 C 表示语义分割的种类数。然后，网络训练采用对抗真值图的逐像素多分类交叉熵损失。按照NYUDv2上的标准40类训练/测试协议 [98], [103], [105]，我们直接运用*JL-DCF* (Fig. 13)并且不采

3. 注意现有的绝大多数基于深度学习的VSOD工作在训练过程中都采用了RGB SOD数据，例如 [13], [34], [141]。

4. 大多数样本是在相似背景下有相同物体的连续的帧。

取任何其他改变，最后得到35.0% mIOU。而根据文献 [71]，这样的结果在该任务上表明是可行的。

我们注意到使用孪生网络进行RGB-D语义分割的一个潜在挑战是，RGB和深度模态在这个任务上存在巨大的共性鸿沟，这是因为RGB-D语义分割是识别特定类别的区域，而其中RGB和深度存在较大的差距（即微弱的共性）。这与RGB-D SOD形成了明显的对比，因为在RGB-D SOD任务的两种模态中，类别无关的显著物体会一致性地“弹出”（即变得显眼），正如图1和图12所示。这实际上引出了一个问题：除了RGB-D SOD和本文中的应用之外，在其它何种任务中孪生网络也是适用的。我们相信这将是一个有趣的问题，并值得未来深入研究。

为了更好地理解*JL-DCF* 和现有的语义分割模型相比效果如何，以及联系这两个领域，我们仔细地将几种开源的前沿分割模型（包括PSPNet [153]，RDFNet [103]，DANet [154]，SA-Gate [105]和SGNet [112]）应用于RGB-D SOD任务。我们将它们的多类分类头替换为相应的显著性预测头（如上所述），并在RGB-D SOD数据集上进行评估。注意RDFNet [103]，SA-Gate [105]和SGNet [112]是三种RGB-D语义分割模型，它们可以直接应用和迁移。而PSPNet [153]和DANet [154]是两种有代表性的RGB语义分割模型，我们通过文献 [71]中所采用的后期融合策略来对它们进行适配。同时我们的实验为了公平比较，一些RGB-D语义分割模型（例如RDFNet和SA-Gate）中原先使用的HHA图 [98], [103], [105]被替换成三通道深度图作为输入。表9展示了比较性结果，其中所有的模型都基于ResNet-101，并被用与*JL-DCF* 相同的训练数据进行重新训练。我们可以看出*JL-DCF* 在五个代表性的数据集上总体上优于这些语义分割模型。有趣的是，我们还发现了一些前沿模型获得还不错的效果，尤其是最新的SA-Gate，其甚至比表1中的那些为RGB-D SOD设计的模型表现得更好。这一事实从实验角度揭示了这两个领域之间的内在联系和可迁移性，我们相信这也是未来在研究上比较有趣的一个问题。此外，它们间的差异也可能是存在的，正如SGNet的不太好的表现所表明的那样。SGNet在这个任务上的性能下降可能是因为它依赖深度信息引导来对RGB特征进行过滤。然而在RGB-D SOD任务中，深度信息可能会变得不太可靠。我们观察到这些模型存在的另一个问题是它们因为输出步长大，导致预测结果粗糙，这会产生不准确的边界细节。

5 结论

本文提出了一种RGB-D SOD框架，称为*JL-DCF*，该框架基于联合学习和密集协作融合。实验结果表明，为RGB和深度视角同时学习一个孪生网络来进行显著性目标物体定位和检测是可行的，并可得到较为准确的预测。此外，所采用的密集协作融合策略对实现跨模态互补是有效的。*JL-DCF* 在七个评测数据集上展示出了较前沿方法更好的性

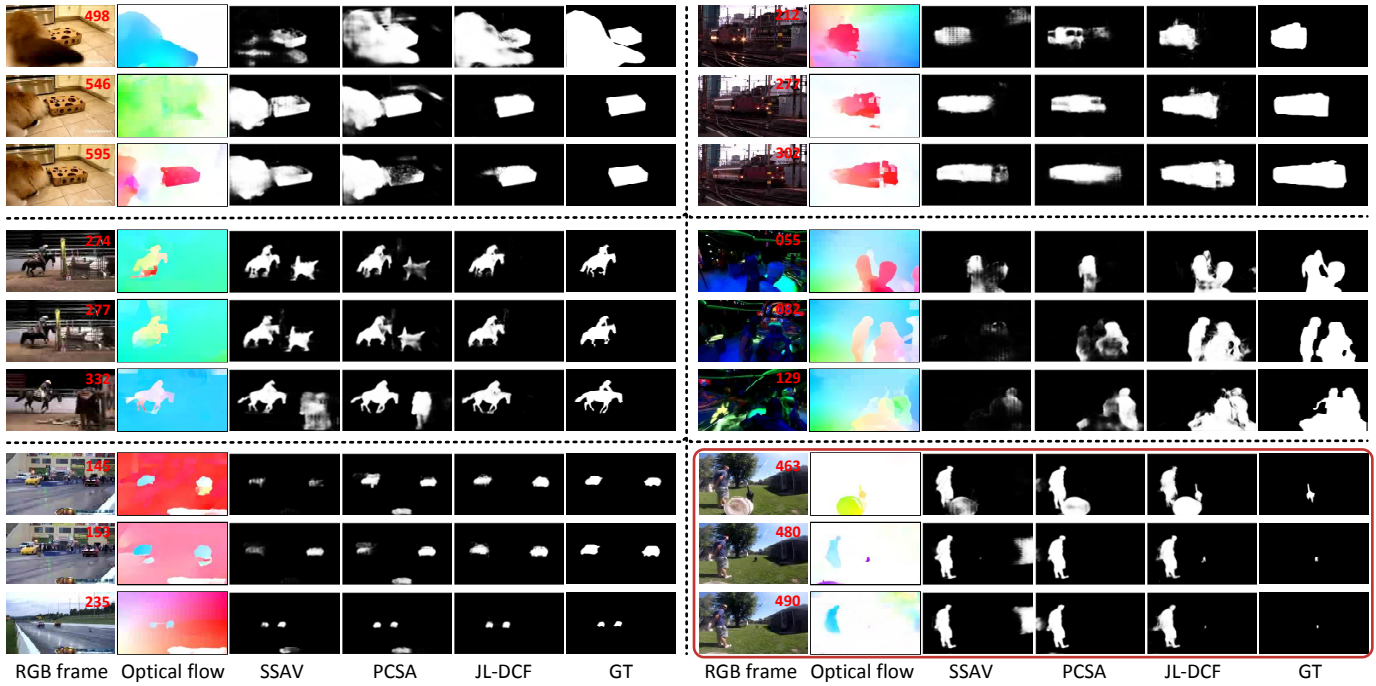


Fig. 14: *JL-DCF* 与两个最新的前沿VSOD模型SSAV-CVPR19 [13] 和PCSA-AAAI20 [141]的可视化对比。右下方的一组图展现了一个所有模型都检测到其它物体的失败案例。但是只有*JL-DCF* 对真正的目标，即小狗目标有响应。

TABLE 9: 将*JL-DCF* 与现有的迁移到RGB-D显著检测任务的语义分割模型进行比较。符号“†”表示通过后融合策略 [71]适应此任务的RGB语义分割模型。

Model	NJU2K [65]		NLPR [59]		STERE [58]		RGBD135 [60]		SIP [45]	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
PSPNet† [153]	0.901	0.045	0.918	0.028	0.899	0.046	0.909	0.026	0.856	0.066
RDFNet [103]	0.891	0.050	0.910	0.031	0.897	0.047	0.919	0.027	0.875	0.055
DANet† [154]	0.900	0.044	0.912	0.027	0.889	0.048	0.896	0.027	0.870	0.056
SA-Gate [105]	0.898	0.051	0.923	0.028	0.896	0.054	0.941	0.022	0.874	0.059
SGNet [112]	0.873	0.060	0.888	0.039	0.883	0.055	0.899	0.034	0.832	0.075
<i>JL-DCF</i>	0.903	0.043	0.925	0.022	0.905	0.042	0.929	0.022	0.879	0.051

能，并得到了全面的消融实验的验证。我们框架的泛化性和鲁棒性也在两个密切相关的任务，即RGB-Thermal (RGB-T) SOD和VSOD中得到了验证，并且与前沿语义分割模型进行了比较。*JL-DCF* 的先进性能表明它可以成为用于多模态特征学习和融合任务的统一框架，同时我们希望这项工作能够在未来成为促进很多跨模态研究任务的催化剂。

REFERENCES

- [1] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, “JL-DCF: joint learning and densely-cooperative fusion framework for rgb-d salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3052–3062.
- [2] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, “Global contrast based salient region detection,” *IEEE T. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.
- [3] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, “Salient object detection: A survey,” *Comput. Vis. Media*, pp. 1–34, 2019.
- [4] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, “EGNet: Edge guidance network for salient object detection,” in *Int. Conf. Comput. Vis.*, 2019, pp. 8779–8788.
- [5] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, “Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut,” *IEEE T. Multimedia*, vol. 14, no. 4, pp. 1275–1289, 2012.
- [6] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, “Salient object segmentation via effective integration of saliency and objectness,” *IEEE T. Multimedia*, vol. 19, no. 8, pp. 1742–1756, 2017.
- [7] T. Zhou, H. Fu, C. Gong, J. Shen, L. Shao, and F. Porikli, “Mutual consistency induced transfer subspace learning for human motion segmentation,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 10277–10286.
- [8] K. R. Jerripothula, J. Cai, and J. Yuan, “Image co-segmentation via saliency co-fusion,” *IEEE T. Multimedia*, vol. 18, no. 9, pp. 1896–1909, 2016.
- [9] U. Rutishauser, D. Walther, C. Koch, and P. Perona, “Is bottom-up attention useful for object recognition,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, pp. II–II.
- [10] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, “Unsupervised extraction of visual attention objects in color images,” *IEEE T. Circuit Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, 2006.
- [11] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” *IEEE T. Image Process.*, vol. 19, no. 1, pp. 185–198, 2010.
- [12] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE T. Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.
- [13] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, “Shifting more attention to video salient object detection,” in *IEEE Conf. Comput.*

- Vis. Pattern Recog.*, 2019, pp. 8554–8564.
- [14] W. Wang and J. Shen, “Deep cropping via attention box prediction and aesthetics assessment,” in *Int. Conf. Comput. Vis.*, 2017, pp. 2186–2194.
- [15] F. Stentiford, “Attention based auto image cropping,” in *International Conference on Computer Vision Systems*, 2007.
- [16] L. Marchesotti, C. Cifarelli, and G. Csurka, “A framework for visual saliency detection with applications to image thumbnailing,” in *Int. Conf. Comput. Vis.*, 2009, pp. 2232–2239.
- [17] Y. Ding, J. Xiao, and J. Yu, “Importance filtering for image retargeting,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 89–96.
- [18] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2376–2383.
- [19] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, “Detection of co-salient objects by looking deep and wide,” *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.
- [20] D. Zhang, D. Meng, and J. Han, “Co-saliency detection via a self-paced multiple-instance learning framework,” *IEEE T. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, 2017.
- [21] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu, “Adaptive graph convolutional network with attention graph clustering for co-saliency detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9050–9059.
- [22] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, “Sketch2photo: Internet image montage,” *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 1–10, 2006.
- [23] Y. Gao, M. Shi, D. Tao, and C. Xu, “Database saliency for fast image retrieval,” *IEEE T. Multimedia*, vol. 17, no. 3, pp. 359–369, 2015.
- [24] G. Liu and D. Fan, “A model of visual attention for natural image retrieval,” in *IEEE International Conference on Information Science and Cloud Computing Companion*, 2013, pp. 728–733.
- [25] W. Wang, J. Shen, L. Shao, and F. Porikli, “Correspondence driven saliency transfer,” *IEEE T. Image Process.*, vol. 25, no. 11, pp. 5025–5034, 2016.
- [26] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, “Amulet: Aggregating multi-level convolutional features for salient object detection,” in *Int. Conf. Comput. Vis.*, 2017, pp. 202–211.
- [27] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, “Learning uncertain convolutional features for accurate saliency detection,” in *Int. Conf. Comput. Vis.*, 2017, pp. 212–221.
- [28] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, “Progressive attention guided recurrent network for salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 714–722.
- [29] M. Feng, H. Lu, and E. Ding, “Attentive feedback network for boundary-aware salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1623–1632.
- [30] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, “Deep light-field-driven saliency detection from a single view,” in *Int. Jt. Conf. Artif. Intell.*, 2019, pp. 904–911.
- [31] K. Fu, Q. Zhao, I. Y.-H. Gu, and J. Yang, “Deepside: A general deep framework for salient object detection,” *Neurocomputing*, vol. 356, pp. 69–82, 2019.
- [32] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. Hoi, and H. Ling, “Learning unsupervised video object segmentation through visual attention,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3064–3074.
- [33] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, “Zero-shot video object segmentation via attentive graph neural networks,” in *Int. Conf. Comput. Vis.*, 2019, pp. 9236–9245.
- [34] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, “Pyramid dilated deeper convlstm for video salient object detection,” in *Eur. Conf. Comput. Vis.*, 2018, pp. 715–731.
- [35] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, “Revisiting video saliency prediction in the deep learning era,” *IEEE T. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, 2021.
- [36] Y. Fang, G. Ding, W. Wen, F. Yuan, Y. Yang, Z.-J. Fang, and W. Lin, “Salient object detection by spatiotemporal and semantic features in real-time video processing systems,” *IEEE Transactions on Industrial Electronics*, vol. 67, no. 11, pp. 9893–9903, 2020.
- [37] N. Liu and J. Han, “Dhsnet: Deep hierarchical saliency network for salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 678–686.
- [38] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, “Deeply supervised salient object detection with short connections,” *IEEE T. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, 2019.
- [39] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, “Salient objects in clutter: Bringing salient object detection to the foreground,” in *Eur. Conf. Comput. Vis.*, 2018, pp. 196–212.
- [40] 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 T. Pattern Anal. Mach. Intell.*, 2021.
- [41] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, “An iterative and cooperative top-down and bottom-up inference network for salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 5968–5977.
- [42] W. Wang, J. Shen, X. Dong, and A. Borji, “Salient object detection driven by fixation prediction,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1711–1720.
- [43] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, “Salient object detection with pyramid attention and salient edges,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1448–1457.
- [44] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, “Rgb-d salient object detection via deep fusion,” *IEEE T. Image Process.*, vol. 26, no. 5, pp. 2274–2285, 2017.
- [45] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M.-M. Cheng, “Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks,” *IEEE T. Neural Netw. Learn. Syst.*, 2021.
- [46] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Sadat Saleh, T. Zhang, and N. Barnes, “UC-Net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8582–8591.
- [47] 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 T. Image Process.*, vol. 26, no. 9, pp. 4204–4216, 2017.
- [48] 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*, vol. 363, pp. 46–57, 2019.
- [49] P. Huang, C.-H. Shen, and H.-F. Hsiao, “Rgb-d salient object detection using spatially coherent deep learning framework,” in *International Conference on Digital Signal Processing*, 2018, pp. 1–5.
- [50] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, “Bilateral attention network for rgb-d salient object detection,” *IEEE T. Image Process.*, vol. 30, pp. 1949–1961, 2021.

- [51] 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 T. Cybern.*, vol. 48, no. 11, pp. 3171–3183, 2017.
- [52] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.
- [53] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3051–3060.
- [54] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: prior-model guided depth-enhanced network for salient object detection," in *Int. Conf. Multimedia and Expo*, 2019, pp. 199–204.
- [55] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE T. Image Process.*, vol. 28, no. 6, pp. 2825–2835, 2019.
- [56] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgb-d salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3927–3936.
- [57] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 539–546.
- [58] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 454–461.
- [59] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgb-d salient object detection: A benchmark and algorithms," in *Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.
- [60] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Int'l Conference on Internet Multimedia Computing and Service*, 2014, pp. 23–27.
- [61] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for rgb-d salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2343–2350.
- [62] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features," in *Int. Conf. Comput. Vis. Worksh.*, 2017, pp. 2749–2757.
- [63] A. Wang and M. Wang, "Rgb-d salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663–667, 2017.
- [64] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing, "Stereoscopic saliency model using contrast and depth-guided-background prior," *Neurocomputing*, vol. 275, pp. 2227–2238, 2018.
- [65] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *IEEE Int. Conf. Image Process.*, 2014, pp. 1115–1119.
- [66] 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," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, 2016.
- [67] J. Guo, T. Ren, and J. Bei, "Salient object detection for rgb-d image via saliency evolution," in *Int. Conf. Multimedia and Expo*, 2016, pp. 1–6.
- [68] 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 Recognit.*, vol. 86, pp. 376–385, 2019.
- [69] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 7254–7263.
- [70] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from rgb to rgb-d saliency: A depth-guided transformation model," *IEEE T. Cybern.*, vol. 50, no. 8, pp. 3627–3639, 2020.
- [71] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.
- [72] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Eur. Conf. Comput. Vis.*, 2016, pp. 75–91.
- [73] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 756–13 765.
- [74] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7794–7803.
- [75] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3472–3481.
- [76] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9060–9069.
- [77] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *IEEE T. Pattern Anal. Mach. Intell.*, 2021.
- [78] 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 *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 275–292.
- [79] 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," *arXiv preprint arXiv:2007.02713v2*, 2020.
- [80] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Adv. Neural Inform. Process. Syst.*, 1994, pp. 737–744.
- [81] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1701–1708.
- [82] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Int. Conf. on Mach. Learn. Worksh.*, vol. 2, 2015.
- [83] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4353–4361.
- [84] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1592–1599.
- [85] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 5695–5703.
- [86] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Int. Conf. Comput. Vis.*, 2017, pp. 66–75.
- [87] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Eur. Conf. Comput. Vis.*, 2018, pp. 573–590.
- [88] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

- [89] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 104–119, 2021.
- [90] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "Siamcar: Siamese fully convolutional classification and regression for visual tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6269–6277.
- [91] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "Siam r-cnn: Visual tracking by re-detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6578–6588.
- [92] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 6668–6677.
- [93] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang, "Fast and accurate online video object segmentation via tracking parts," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7415–7424.
- [94] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. Joo Kim, "Fast video object segmentation by reference-guided mask propagation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7376–7385.
- [95] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1328–1338.
- [96] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [97] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *Int. Conf. Learn. Represent.*, 2013.
- [98] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Eur. Conf. Comput. Vis.*, 2014, pp. 345–360.
- [99] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asia. Conf. Comput. Vis.*, 2016, pp. 213–228.
- [100] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in *Eur. Conf. Comput. Vis.*, 2016, pp. 664–679.
- [101] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in *Eur. Conf. Comput. Vis.*, 2016, pp. 541–557.
- [102] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3029–3037.
- [103] S.-J. Park, K.-S. Hong, and S. Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *Int. Conf. Comput. Vis.*, 2017, pp. 4980–4989.
- [104] L. Deng, M. Yang, T. Li, Y. He, and C. Wang, "Rfbnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation," *arXiv preprint arXiv:1907.00135*, 2019.
- [105] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *Eur. Conf. Comput. Vis.*, 2020, pp. 561–577.
- [106] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgb-d semantic segmentation," in *Int. Conf. Comput. Vis.*, 2017, pp. 5199–5208.
- [107] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1746–1754.
- [108] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of rgb-d images," in *Int. Conf. Comput. Vis.*, 2017, pp. 1311–1319.
- [109] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Eur. Conf. Comput. Vis.*, 2018, pp. 135–150.
- [110] Y. Chen, T. Mensink, and E. Gavves, "3d neighborhood convolution: Learning depth-aware features for rgb-d and rgb semantic segmentation," in *Int. Conf. 3D Vis. (3DV)*, 2019, pp. 173–182.
- [111] Y. Xing, J. Wang, and G. Zeng, "Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing," in *Eur. Conf. Comput. Vis.*, 2020, pp. 555–571.
- [112] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time rgb-d semantic segmentation," *IEEE T. Image Process.*, vol. 30, pp. 2313–2324, 2021.
- [113] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.
- [114] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
- [115] J. Al Azzeh, H. Alhatamleh, Z. A. Alqadi, and M. K. Abuzalata, "Creating a color map to be used to convert a gray image to color image," *International Journal of Computer Applications*, vol. 153, no. 2, pp. 31–34, 2016.
- [116] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3156–3164.
- [117] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1–9.
- [118] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4700–4708.
- [119] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015, pp. 234–241.
- [120] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7132–7141.
- [121] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: convolutional block attention module," in *Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [122] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2Net: a new multi-scale backbone architecture," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, 2021.
- [123] J. Zhao, Y. Zhao, J. Li, and X. Chen, "Is depth really necessary for salient object detection?" in *ACM Int. Conf. Multimedia*, 2020, pp. 1745–1754.
- [124] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.
- [125] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE T. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

- [126] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function (in chinese)," *SCIENTIA SINICA Informationis*, 2021.
- [127] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 733–740.
- [128] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2806–2813.
- [129] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 136–145.
- [130] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1597–1604.
- [131] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.
- [132] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
- [133] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3917–3926.
- [134] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive encoder-decoder network for rgbt salient object detection," *arXiv preprint arXiv:2005.02315*, 2020.
- [135] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE T. Multimedia*, vol. 22, no. 1, pp. 160–173, 2019.
- [136] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "Rgbt salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE T. Circuit Syst. Video Technol.*, vol. 30, no. 12, pp. 4421–4433, 2019.
- [137] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE T. Image Process.*, vol. 29, pp. 3321–3335, 2019.
- [138] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE T. Image Process.*, vol. 27, no. 1, pp. 38–49, 2017.
- [139] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3243–3252.
- [140] S. Li, B. Seybold, A. Vorobyov, X. Lei, and C.-C. Jay Kuo, "Unsupervised video object segmentation with motion-based bilateral networks," in *Eur. Conf. Comput. Vis.*, 2018, pp. 207–223.
- [141] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *AAAI Conf. Art. Intell.*, vol. 34, no. 7, 2020, pp. 10 869–10 876.
- [142] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE T. Image Process.*, vol. 28, no. 10, pp. 4819–4831, 2019.
- [143] M. Xu, B. Liu, P. Fu, J. Li, and Y. H. Hu, "Video saliency detection via graph clustering with motion energy and spatiotemporal objectness," *IEEE T. Multimedia*, vol. 21, no. 11, pp. 2790–2805, 2019.
- [144] M. Xu, B. Liu, P. Fu, J. Li, Y. H. Hu, and S. Feng, "Video salient object detection via robust seeds extraction and multi-graphs manifold propagation," *IEEE T. Circuit Syst. Video Technol.*, vol. 30, no. 7, pp. 2191–2206, 2020.
- [145] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 724–732.
- [146] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1187–1200, 2013.
- [147] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE T. Image Process.*, vol. 24, no. 11, pp. 4185–4196, 2015.
- [148] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE T. Image Process.*, vol. 27, no. 1, pp. 349–364, 2017.
- [149] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2462–2470.
- [150] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE T. Image Process.*, vol. 24, no. 8, pp. 2552–2564, 2015.
- [151] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE T. Circuit Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, 2016.
- [152] X. Li, L. Zhao, L. Wei, M. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE T. Image Process.*, vol. 25, no. 8, pp. 3919–3930, 2016.
- [153] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2881–2890.
- [154] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3146–3154.