JL-DCF: RGB-D显著物体检测的联合学习与密集协作融合框架

傅可人[1]　范登平[2,3,*]　季葛鹏[4]　赵启军[1]
[1] 四川大学计算机学院　[2] 南开大学计算机学院
[3] 阿联酋起源人工智能研究院(IIAI)　[4] 武汉大学计算机学院
http://dpfan.net/JLDCF/

Abstract

本文提出一种新颖的用于*RGB-D*显著物体检测的联合学习与密集协作融合框架(***JL-DCF***)。现有的模型通常将*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到有限训练数据或过度依赖精心设计的训练过程的限制。相比之下，本文的*JL-DCF*框架通过孪生网络从*RGB*和深度输入中学习。本文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。在四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泛化性能的*RGB-D*显著性检测器。在六个具有挑战性的数据集上，较目前最好的模型*D3Net*，我们极大地改进了检测性能，获得约*1.9%*（*S*指标）的提升。这表明所提出的框架为实际应用提供了一个可行的解决方案，并且可以为挖掘跨模态互补性的任务提供更多的启示。代码可在*https://github.com/kerenfu/JLDCF/*获得。

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, 9, 78]。其有许多有用的应用，包括物体分割和识别 [27, 32, 39, 51, 70, 79]、图像/视频压缩 [24]、视频检测/概括 [19, 41]、基于内容的图像编辑 [14, 23, 42, 57, 63]、发现具有信息量的通用物体 [71, 72]、图像检索 [8, 22, 37]。许多显著物体检测模型是在假设输入是单幅RGB/彩色图像

通讯作者:范登平(dengpfan@gmail.com)*

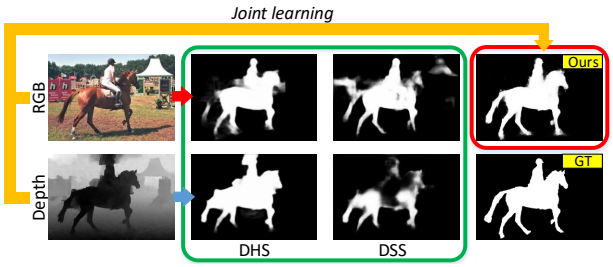

Figure 1: 应用深度学习显著性模型DHS [38]和DSS [29]的结果。他们的输入为RGB图（第一行）或深度图（第二行）。两个模型均在单一的RGB模态下训练。相比之下，我们的*JL-DCF* 考虑了两种模态，能得到更好的结果（最后一列）。

[21,47,66,74–76]或序列 [56,62,67,68]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。尽管许多现有工作都试图探索深度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显著物体检测模型的论文相比 [15, 61, 64, 65, 69]，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 [49]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 [18, 73]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 [18, 31, 40, 55]或晚期融合 [26, 60]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 [4]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 [4, 5, 80]，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 [77]，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（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 [10, 43, 44]、物体轮廓闭合性 [20, 53]和与图像边界的连通性 [36, 59]。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 [38]，有时能在深度图上表现良好。然而，另一个类似的模型如DSS [29]，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。本文有两个主要贡献：

- 提出了一个通用的RGB-D SOD框架，称之为*JL-DCF*，其由两个组件构成：联合学习和密集协作融合。这两部分的主要特点是其鲁棒性和有效性，这将有助于未来对计算机视觉中相关多模态任务的建模。特别地，在六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。

- 我们对14种现有前沿方法 [4–6, 13, 18, 20, 25, 26, 34, 46, 49, 55, 60, 77] 进行了全面地评估，这也是迄今为止该领域最大规模的评测。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。

2. 相关工作

传统方法： RGB-D SOD的开创性工作是由Niu等人[43]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 [25, 34]、对比度 [10, 13, 44]、背景包围性 [20]、中心/边界先验 [10, 12, 36, 59]、紧密性 [12, 13]或各种度量的组合 [55]。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。

深度学习方法： 通过使用深度学习和CNNs，这一领域取得了新的进展。Qu等人 [49]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 [53]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） [52]。Chen等人 [4]提出一种自底向上/自上而下的架构 [48]，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 [26]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 [5]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 [6]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 [40]和文献 [31]通过串联RGB和深度来形成四通道输入。随后，该输入被分别输入到单流递归CNN和短连接FCN中。文献 [80]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 [77]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 [18]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。

总的来说，根据以往文献 [4, 77]的总结，上述方法大可分为三类：（a）早期融合 [18, 31, 40, 55]，（b）晚期融合 [26, 60]和（c）中期融合 [4–6, 80]。中期融合对（a）和（b）进行补足，因为特征提取和后续融合都由

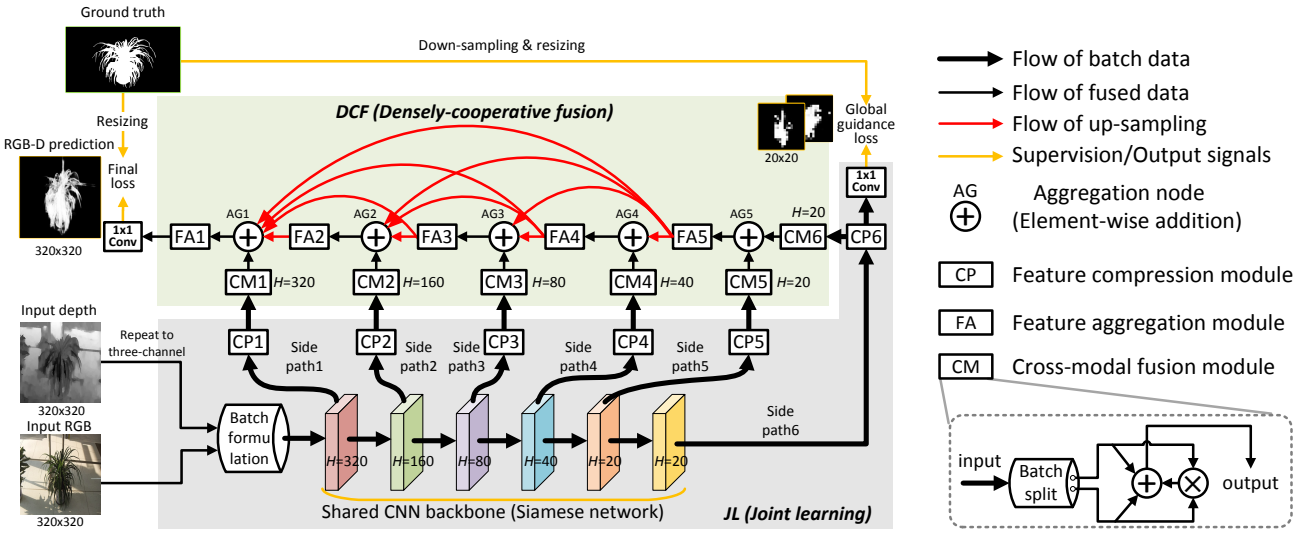

Figure 2: 本文提出的用于RGB-D SOD的*JL-DCF* 模型框架图。JL（联合学习）组件由灰色区域所示，而DCF（密集协作融合）组件由浅绿色区域所示。CP1∼CP6：特征压缩模块。FA1∼FA6：特征聚合模块。CM1∼CM6：跨模态融合模块。“*H*”表示在特定阶段输出的特征图的空间大小。具体细节详见第3节。

相对较深的CNNs进行处理。因此，可从两种模态中学习高层语义，同时挖掘更加复杂的集成规则。此外，对RGB和depth进行独立的深监督也很直接。本文提出的*JL-DCF* 框架属于中期融合策略。

然而，与上述方法 [4–6,80]不同，其两个特征提取分支是独立的，我们提出采用孪生网络架构 [11]，其网络结构和网络权值是共享的。这将带来两大好处：1）通过联合学习可实现跨模态知识共享；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。

3. 本文方法

本文提出的*JL-DCF* 的总体架构如图2所示。它遵循经典的自底向上/自顶向下架构 [48]。为更好地说明，图2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 [54]和ResNet [28]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“基于模型”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2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“基于特征”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2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2（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，我们使用了简单的灰度映射，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 [1]或变换，如文献 [26]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 [29]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2中的CP1∼CP6）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2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 [48]。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2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2（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2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和depth部分，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。与利用特征互补性的逐元素加法“$\oplus$”相比，逐元素乘法“$\otimes$”更强调特征共性。而这两种性质在跨模态融合中通常而言较为重要。

有人也许会说这样的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，其中仅是RGB特征主导了最终预测。而值得注意的是，如4.4节将所示，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。

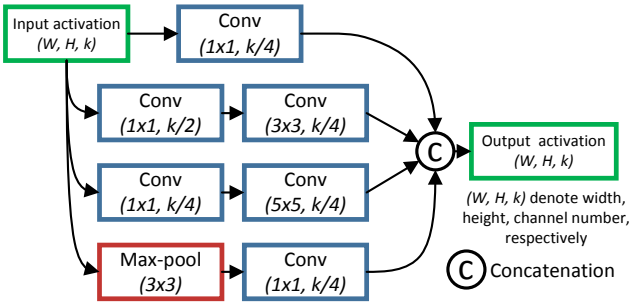

Figure 3: 用于图2中FA模块的Inception结构。所有的卷积层和最大池层的步长为1，因此保持空间特征大小不变。与最初的Inception模块 [58]不同，我们对它进行了调整，使其具有相同的输入/输出通道数k。

如图2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 [30]。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 [50]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的特征聚合模块。为此我们使用了图3所示的Inception模块 [58]，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模块是比较灵活的。未来还可考虑使用其它模块来提高性能。

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^c_{rgb}和S^c_d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：

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

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$S \in \{S^c_{rgb}, S^c_d, S^f\}$.

Table 1: 定量指标: 前沿方法以及本文提出的*JL-DCF* 在六个RGB-D数据集上的S指标(S_α) [16], 最大F指标($F_\beta^{\max}$) [3], 最大E指标($E_\phi^{\max}$) [17]，MAE(M) [45]。最好的结果用**粗体**突出表示。

	Metric	ACSD [34]	LBE [20]	DCMC [13]	MDSF [55]	SE [25]	DF [49]	AFNet [60]	CTMF [26]	MMCI [6]	PCF [4]	TANet [5]	CPFP [77]	DMRA [46]	D3Net [18]	*JL-DCF* Ours
NJU2K [34]	$S_\alpha \uparrow$	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**
	$F_\beta^{\max} \uparrow$	0.711	0.748	0.715	0.775	0.748	0.804	0.775	0.845	0.852	0.872	0.874	0.877	0.886	0.889	**0.903**
	$E_\phi^{\max} \uparrow$	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**
	$M \downarrow$	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**
NLPR [44]	$S_\alpha \uparrow$	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**
	$F_\beta^{\max} \uparrow$	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**
	$E_\phi^{\max} \uparrow$	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**
	$M \downarrow$	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**
STERE [43]	$S_\alpha \uparrow$	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**
	$F_\beta^{\max} \uparrow$	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**
	$E_\phi^{\max} \uparrow$	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**
	$M \downarrow$	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**
RGBD135 [10]	$S_\alpha \uparrow$	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**
	$F_\beta^{\max} \uparrow$	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**
	$E_\phi^{\max} \uparrow$	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**
	$M \downarrow$	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**
LFSD [35]	$S_\alpha \uparrow$	0.727	0.729	0.746	0.694	0.692	0.783	0.730	0.788	0.779	0.786	0.794	0.820	0.839	0.824	**0.854**
	$F_\beta^{\max} \uparrow$	0.763	0.722	0.813	0.779	0.786	0.813	0.740	0.787	0.767	0.775	0.792	0.821	0.852	0.815	**0.862**
	$E_\phi^{\max} \uparrow$	0.829	0.797	0.849	0.819	0.832	0.857	0.807	0.857	0.831	0.827	0.840	0.864	0.893	0.856	**0.893**
	$M \downarrow$	0.195	0.214	0.162	0.197	0.174	0.146	0.141	0.127	0.139	0.119	0.118	0.095	0.083	0.106	**0.078**
SIP [18]	$S_\alpha \uparrow$	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**
	$F_\beta^{\max} \uparrow$	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**
	$E_\phi^{\max} \uparrow$	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**
	$M \downarrow$	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**

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 [34]（2000个样本），NLPR [44]（1000个样本），STERE [43]（1000个样本），RGBD135 [10]（135个样本），LESD [35]（100个样本）和SIP [18]（929个样本）。参照文献 [77]，我们从NLPR和NJU2K中分别选择了相同的700个和1500个样本来训练我们的算法。其余样本用于测试。为公平比较，我们将在该训练集上训练的模型应用于其它数据集上。评测时，我们采用了四个广泛采用的评测指标，即S指标(S_α) [16,77]、最大F指标($F_\beta^{\max}$) [3,29]、最大E指标($E_\phi^{\max}$) [17,18]和MAE (M) [3,45]。此处略去了这些指标的详细定义，对于这些指标的定义读者可参考相关论文。需要注意的是，由于E指标最初被文献 [17] 提出是用于评估二值图，为了将其扩展用于比较非二值显著性图和二值真值图，我们采用类似于$F_\beta^{\max}$的策略。具体而言，首先采用区间[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，然后报告所有前景图中最大的E指标。

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 [54] 和ResNet-101 [28]构造了两个版本的*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具有13个卷积层的VGG-16，侧*path1*∼*path6*依次连接到*conv1_2*、*conv2_2*、*conv3_3*、*conv4_3*、*conv5_3*和*pool5*。受文献 [29]启发，我们在侧*path1*∼*path6* 中额外添加了两个卷积层。为了增强侧 *path6*粗糙特征图的分辨率且保持感受野不变，我们使*pool5* 的步长为1，而对其侧路上两个额外的卷积层使用膨胀卷积 [7] ，膨胀率设置为2。总体上，如图2所示，我们最终修改后的VGG-16主干网络生成的最粗糙的特征的空间尺寸为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为20×20。由于ResNet的第一个卷积层的步长已经为2，所以其最浅层的特征图大小为160×160。为了获得完整大小（320×320）的特征图而不依赖于简单的上采样，我

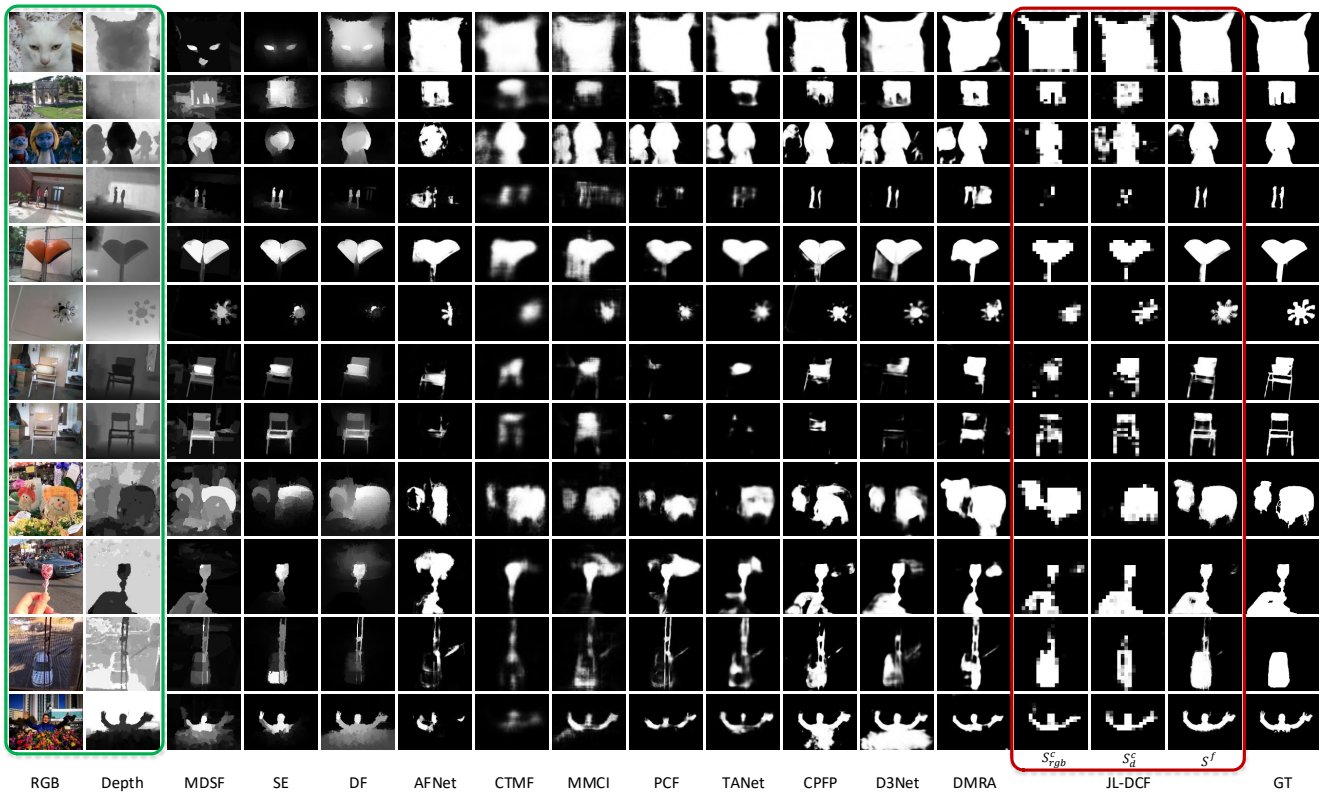

Figure 4: *JL-DCF* 与前沿的RGB-D显著性物体检测模型的可视化比较。从RGB和深度中联合学习到的粗预测图（S^c_{rgb}和S^c_d）也与*JL-DCF* 的最终结果（S^f）一同进行展示。

们借用了VGG-16中的*conv1_1*和*conv1_2*层进行特征提取。将侧*path1*∼*path6*分别连接至*conv1_2*和ResNet-101的*conv1*、*res2c*、*res3b3*、*res4b22*、*res5c*。同样，我们也将*res5a*块的步长从2改为1，但随后使用了膨胀率为2的膨胀卷积。

解码器配置： 图2中的所有CP模块都为3×3且通道数$k = 64$的卷积，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2所示，为了对齐解码器中的特征尺寸，FA模块的输出被进行各种倍数的上采样。一个极端的情况是，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 [33]上实现了*JL-DCF* 。训练期间，主干网络 [28, 54]采用DSS [29]的预训练参数进行初始化，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反射进行增强，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。

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。在比较者中，DF [49]、AFNet [60]、CTMF [26]、MMCI [6]、PCF [4]、TANet [5]、CPFP [77]、D3Net [18]、DMRA [46]为最近的基于深度学习的方法，而ACSD [34]、LBE [20]、DCMC [13]、MDSF [55]、SE [25]为使用了各种人工特征/假设的传统方法。定量结果如表1所示。*JL-DCF* 相对于现有的和近期提出的技术（如CPFP [77], D3Net [18] 和DMRA [46]），在所有四个指标上的性能都有显著提高。这验证了*JL-DCF* 的有效性和泛化性能。可视化示例如图4所示。*JL-DCF* 在利用深度信息进行跨模态互补方面似乎更加有效，在RGB-D模式下能够更好地检测出目标物体。此外图4中还给出了深监督的粗糙预测结果。可以看出，它们为后续的跨模态精化提供了基本的物体定位支撑，

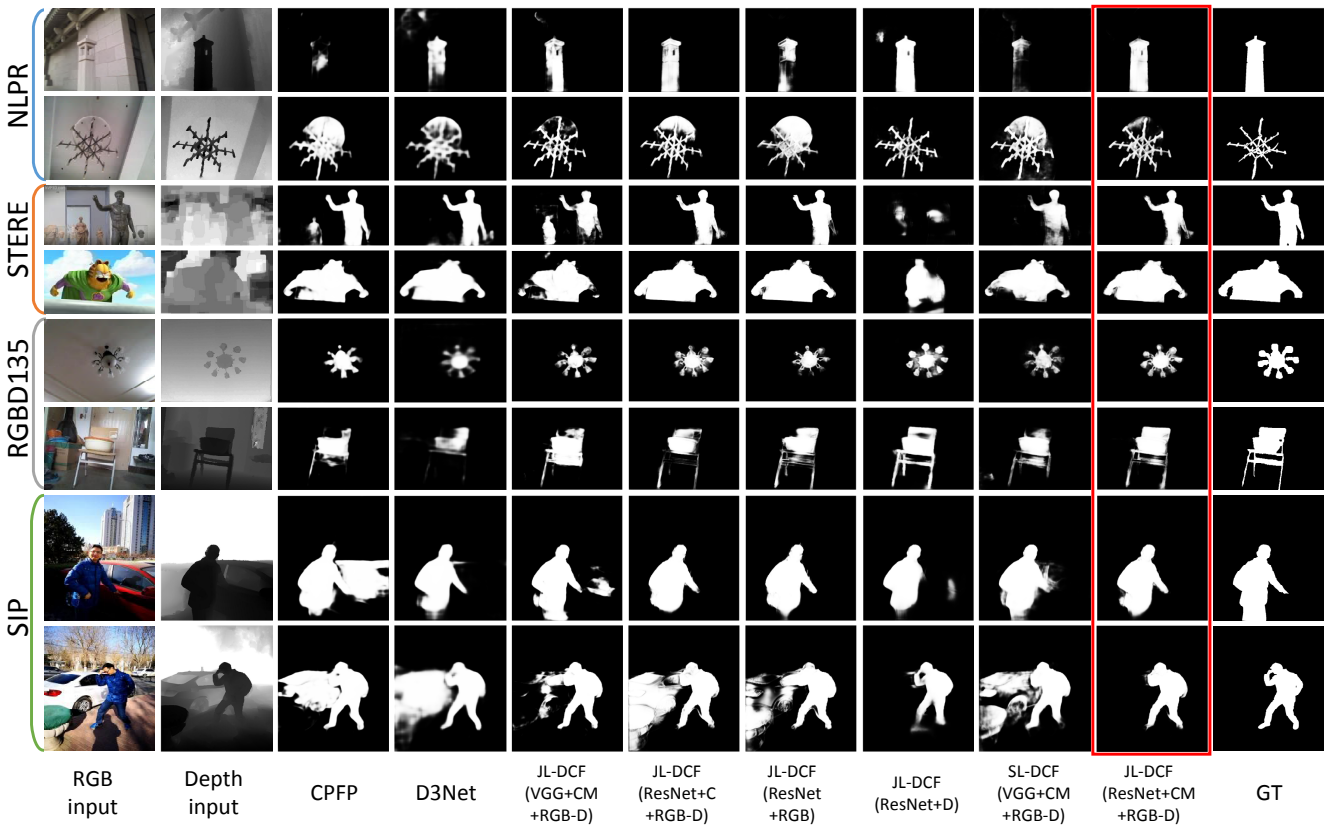

Figure 5: 消融实验的可视化示例，示例图片来自NLPR、STERE、RGB135和SIP数据集。总体上，*JL-DCF* 的完整实现（即ResNet+CM+RGB-D，红色方框高亮）获得了最接近真值的结果。

而我们的密集协作融合架构学习了一种自适应且“图像依赖”的方式来将这种支撑信息与层次化多视角特征进行融合。这证明了融合过程并没有退化至两种视角之一（RGB或深度），从而使融合后性能得到提升。

4.4. 消融实验

通过从*JL-DCF* 的完整实现中移除或替换组件来进行消融。以*JL-DCF* 的ResNet版本为参考，将各种消融结果与其进行对比。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”表示同时输入RGB和深度。

首先，为了比较不同的主干网络，训练了“*JL-DCF* (VGG+CM+RGB-D)”模型，其将ResNet主干网络替换为VGG并保持其它设置不变。为验证CM模块的有效性，我们训练了“*JL-DCF* (ResNet+C+RGB-D)”模型，将CM模块替换为通道串联操作。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)”模型，其中将图2中所有batch处理相关操作（如CM模块）替换为恒等映射，而密集解码器、深监督等设置保持不变。这个验证是十分重要的，它表明我们的网络的确学习到了RGB和深度的互补信息。最后，为了说明联合学习的好处，我们训练了“SL-DCF (VGG+CM+RGB-D)”模型，其对RGB和深度使用两个独立的主干网。“SL”代表“分别学习”，对标所提出的“联合学习”。为此我们采用相对更小的VGG-16网络，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2展示了各种评测指标的定量比较，同时列出CPFP [77]和D3Net [18]两种前沿方法作为参照。图5展示视觉上的消融对比结果。可观察得五个方面：

ResNet-101 *vs.* **VGG-16:** 从表2中“A”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，该结论也与前人工作一致。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）和D3Net（ResNet主干）。

CM模块的有效性: 比较列“A”和“C”可以看出，将CM模块改为通道串联操作会一定程度降低性能。

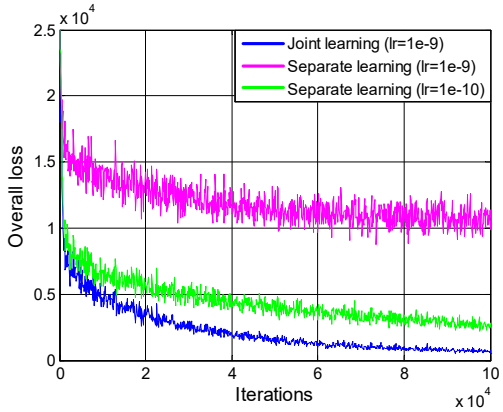

Figure 6: 联合(*JL-DCF*)与分别学习(SL-DCF)的学习曲线。

可能的原因是整个网络倾向于将学习偏向于RGB信息，而忽略了深度信息，因为这样也能够在大多数的数据集上获得较好的结果(列“D”)。尽管串联是一种流行的融合特征的方式，如果学习过程缺乏正确的引导，则可能会陷入次优。相比之下，CM模块对RGB和depth模态实施的是“显式的融合操作”。

RGB与Depth结合的有效性： 在大多数数据集上可见，结合RGB和depth来提高性能这一做法得到了验证（比较列“A”与列“D”和“E”）。唯一例外的是在STERE [43]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。可视化示例见图5中的第3、第4行。我们发现很多来自STERE的深度图过于粗糙，物体边界非常不准确，与真正的物体边界不对齐。融合这种不可靠的深度信息可能反而会降低性能。从表2“E”列（STERE数据集）中可以看到定量证明，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（与RGB相比时，在$S_\alpha/F_\beta^{\max}$ 上低了16%/20%）。

仅使用RGB *vs.* **仅使用Depth：** 比较表2中“D”和“E”列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更优，说明RGB视角通常能提供更多的信息。然而在SIP [18]和RGBD135 [10]上，使用深度信息反而比RGB效果更好，如图5中所示。这意味着这两个数据集的深度图质量相对较好。

JL组件的有效性: 现有模型通常采用分别学习的方式分别从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和深度图获取特征。比较这两种学习策略，发现分别学习（两个独立的主干网络）会增加训练难度。图6展示了典型的学习曲线。在分别学习的设置中，当初始学习率

Table 2: 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 [34]	$S_\alpha \uparrow$	.878	.895	**.903**	.897	.900	.895	.865	.886
	$F_\beta^{\max} \uparrow$	.877	.889	**.903**	.899	.898	.892	.863	.883
	$E_\phi^{\max} \uparrow$	.926	.932	**.944**	.939	.937	.937	.916	.929
	$M \downarrow$	.053	.051	**.043**	.044	.045	.046	.063	.053
NLPR [44]	$S_\alpha \uparrow$	.888	.906	**.925**	.920	.924	.922	.873	.901
	$F_\beta^{\max} \uparrow$	.868	.885	**.916**	.907	.914	.909	.843	.881
	$E_\phi^{\max} \uparrow$	.932	.946	**.962**	.959	.961	.957	.930	.946
	$M \downarrow$	.036	.034	**.022**	.026	.023	.025	.041	.033
STERE [43]	$S_\alpha \uparrow$	.879	.891	.905	.894	.906	**.909**	.744	.886
	$F_\beta^{\max} \uparrow$	.874	.881	**.901**	.889	.899	.901	.708	.876
	$E_\phi^{\max} \uparrow$	.925	.930	**.946**	.938	.945	.946	.834	.931
	$M \downarrow$	.051	.054	.042	.046	.041	**.038**	.110	.053
RGBD135 [10]	$S_\alpha \uparrow$	.872	.904	**.929**	.913	.916	.903	.918	.893
	$F_\beta^{\max} \uparrow$	.846	.885	**.919**	.905	.906	.894	.906	.876
	$E_\phi^{\max} \uparrow$	.923	.946	**.968**	.955	.957	.947	.967	.950
	$M \downarrow$	.038	.030	**.022**	.026	.025	.027	.027	.033
LFSD [35]	$S_\alpha \uparrow$	.820	.832	**.854**	.833	.852	.845	.752	.826
	$F_\beta^{\max} \uparrow$	.821	.819	**.862**	.840	.854	.846	.764	.828
	$E_\phi^{\max} \uparrow$	.864	.864	**.893**	.877	.893	.889	.816	.864
	$M \downarrow$	.095	.099	**.078**	.091	.078	.083	.126	.101
SIP [18]	$S_\alpha \uparrow$	.850	.864	**.879**	.866	.870	.855	.872	.865
	$F_\beta^{\max} \uparrow$	.851	.862	**.885**	.873	.873	.857	.877	.863
	$E_\phi^{\max} \uparrow$	.903	.910	**.923**	.916	.916	.908	.920	.913
	$M \downarrow$	.064	.063	**.051**	.056	.055	.061	.056	.061

为$lr = 10^{-9}$时，网络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$lr = 10^{-10}$，学习过程虽然可脱离局部振荡，但与联合学习策略相比收敛更慢。从表2的“B”和“F”列可以看出，40个周期收敛后模型的性能较*JL-DCF* 更差，即$S_\alpha/F_\beta^{\max}$指标整体下降1.1%/1.76%。我们将*JL-DCF* 更好的性能归功于RGB和深度数据的联合学习。

5. 结论

本文提出了一种RGB-D SOD框架，称之为*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和深度视角学习一个共享的网络来进行准确的显著性目标物体定位和检测是可行的。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6个评测数据集上展示出了较前沿方法更好的性能，并得到了较全面的消融实验的支撑。该框架相对灵活且通用，其内部模块可以被相应的模块替换以进一步提升检测结果。

References

[1] Jamil Al Azzeh, Hussein Alhatamleh, Ziad A Alqadi, and Mohammad Khalil Abuzalata. Creating a color map to be used to convert a gray image to color image. *International Journal of Computer Applications*, 153(2):31–34, 2016.

[2] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *CVM*, pages 1–34, 2019.

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

[4] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *CVPR*, pages 3051–3060, 2018.

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

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

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017.

[8] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM TOG*, 28(5):1–10, 2006.

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

[10] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Int'l Conference on Internet Multimedia Computing and Service*. ACM, 2014.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005.

[12] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from rgb to rgbd saliency: A depth-guided transformation model. *IEEE T-CYB*, 2019.

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

[14] Yuanyuan Ding, Jing Xiao, and Jingyi Yu. Importance filtering for image retargeting. In *CVPR*, pages 89–96, 2011.

[15] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 196–212, 2018.

[16] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *ICCV*, pages 4548–4557, 2017.

[17] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018.

[18] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *IEEE TNNLS*, 2020.

[19] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019.

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

[21] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019.

[22] Yuan Gao, Miaojing Shi, Dacheng Tao, and Chao Xu. Database saliency for fast image retrieval. *IEEE TMM*, 17(3):359–369, 2015.

[23] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010.

[24] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1):185–198, 2010.

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

[26] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE TCYB*, 48(11):3171–3183, 2017.

[27] Junwei Han, King Ngi Ngan, Mingjing Li, and Hong-Jiang Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE TCSVT*, 16(1):141–145, 2006.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[29] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019.

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[31] Posheng Huang, Chin-Han Shen, and Hsu-Feng Hsiao. Rgbd salient object detection using spatially coherent deep learning framework. In *International Conference on Digital Signal Processing*, pages 1–5, 2018.

[32] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Image co-segmentation via saliency co-fusion. *IEEE TMM*, 18(9):1896–1909, 2016.

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

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

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

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

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

[38] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016.

[39] Zhi Liu, Ran Shi, Liquan Shen, Yinzhu Xue, King Ngi Ngan, and Zhaoyang Zhang. Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut. *IEEE TMM*, 14(4):1275–1289, 2012.

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

[41] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. *IEEE TMM*, 7(5):907–919, 2005.

[42] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, pages 2232–2239, 2009.

[43] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012.

[44] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *ECCV*, pages 92–109, 2014.

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

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

[47] Yongri Piao, Zhengkun Rong, Miao Zhang, Xiao Li, and Huchuan Lu. Deep light-field-driven saliency detection from a single view. In *IJCAI*, pages 904–911, 2019.

[48] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *ECCV*, pages 75–91, 2016.

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

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

[51] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition. In *CVPR*, pages II–II, 2004.

[52] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 39(4):640–651, 2017.

[53] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features. In *ICCVW*, pages 2749–2757, 2017.

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[55] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE TIP*, 26(9):4204–4216, 2017.

[56] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018.

[57] Fred Stentiford. Attention based auto image cropping. In *ICVS*, 2007.

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[59] Anzhi Wang and Minghui Wang. Rgb-d salient object detection via minimum barrier distance transform and saliency fusion. *IEEE SPL*, 24(5):663–667, 2017.

[60] Ningning Wang and Xiaojin Gong. Adaptive fusion for rgb-d salient object detection. *IEEE Access*, 7:55277–55284, 2019.

[61] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.

[62] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, pages 9236–9245, 2019.

[63] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *ICCV*, pages 2186–2194, 2017.

[64] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, pages 5968–5977, 2019.

[65] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *CVPR*, pages 1711–1720, 2018.

[66] Wenguan Wang, Jianbing Shen, Ling Shao, and Fatih Porikli. Correspondence driven saliency transfer. *IEEE TIP*, 25(11):5025–5034, 2016.

[67] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE TPAMI*, 2019.

[68] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, pages 3064–3074, 2019.

[69] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *CVPR*, pages 1448–1457, 2019.

[70] Linwei Ye, Zhi Liu, Lina Li, Liquan Shen, Cong Bai, and Yang Wang. Salient object segmentation via effective integration of saliency and objectness. *IEEE TMM*, 19(8):1742–1756, 2017.

[71] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *IJCV*, 120(2):215–232, 2016.

[72] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE TPAMI*, 39(5):865–878, 2017.

[73] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. UC-Net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, 2020.

[74] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017.

[75] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221, 2017.

[76] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, pages 714–722, 2018.

[77] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and flu-

id pyramid integration for rgbd salient object detection. In *CVPR*, pages 3927–3936, 2019.

[78] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019.

[79] Tao Zhou, Huazhu Fu, Chen Gong, Jianbing Shen, Ling Shao, and Fatih Porikli. Multi-mutual consistency induced transfer subspace learning for human motion segmentation. In *CVPR*, 2020.

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