基于RGB-D的显著物体检测反思：模型，数据集，以及大规模评测

范登平, 林铮, 张钊, 朱梦龙, 程明明

***Abstract*—近年来，RGB-D已被广泛地应用于显著物体检测（Salient Object Detection，SOD）。然而，在真实的人类活动场景中使用RGB-D进行显著物体检测的研究却相对很少。为了填补这一领域的空白，本文做出以下贡献：（1）我们精心构建了一个全新的显著人物（salient person，SIP）数据集，它包含近1,000张高分辨率图像，这些图像涵盖了不同视角、姿态、遮挡、照明和背景下的各种真实场景；（2）我们对现代方法进行了迄今为止最全面的大规模基准测评。此工作不仅填补了该领域的空白，同时可以作为未来研究的基准。我们系统地总结了32个先进模型，并在7个数据集约97,000张图像上对其中18个模型进行了深入评测；（3）本文提出了一个简单通用的框架，称为深度图过滤器的深度网络（Deep Depth-Depurator Network，D^{3}Net）。它由深度图过滤单元（depth depurator unit，DDU）和三分支数据流的特征学习模块（three-stream feature learning module，FLM）组成，分别实现低质量深度图的过滤和跨模态特征学习。这些组件紧密联系在一起，并精心设计以供联合训练。D^{3}Net网络的性能在五个评价指标上均超越以往其他模型，成为推进该领域研究的强有力模型。实验表明D^{3}Net可用于从真实场景中有效提取显著人物、可在单个GPU上以65帧/秒的速度实现有效的背景替换。所有的显著图像、新的SIP数据集、D^{3}Net模型和评估工具等详见https://github.com/DengPingFan/D3NetBenchmark。**

关键词—基准，RGB-D，显著性，显著物体检测（SOD），显著人物数据集（SIP）.

I. 前言

如何设计手机使其可拍摄出高质量的照片，已成为手机制造商之间一个最重要的竞争点。显著物体检测（Salient Object Detection, SOD）方法 [2]–[19] 已被集成到手机中，并被广泛用于通过自动添加大光圈和其他增强效果来生成完美人像。尽管现有的SOD方法 [20]–[36] 已取得显著成功，但大多数方法仅使用RGB图像而忽略了重要的深度信息，而这些深度信息已在当代智能手机（如苹果X系列，华为Mate10和三星盖乐世S10等）中广泛使用。

于2019年7月16日接收初稿；2020年3月9日修订；这项研究于2020年5月16日录用。（通讯作者：程明明。）

范登平就读于中国天津南开大学计算机科学学院。目前就职于阿拉伯联合酋长国阿布扎比的人工智能研究所（IIAI）。

林铮，张钊和程明明就职于中国天津南开大学计算机科学学院。（电子邮箱：cmm@nankai.edu.cn）。

朱梦龙目前就职于美国加利福尼亚州山景城的谷歌人工智能研究院。

本文彩色附图可从http://ieeexplore.ieee.org在线获得。

本文为TNNLS2020 [1]的中文翻译版。

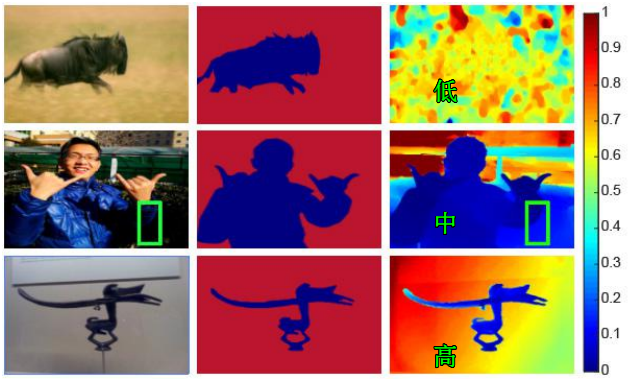


Figure 1. 从左到右：输入图像，真值图和相应的深度图。深度图的质量从低（第一行），中（第二行）到高（最后一行）依次排列。如第二行所示，在输入图像中的边界框区域很难识别人物手臂的边界。但是，在其深度图中却清晰可见。因此，高质量的深度图有助于基于RGB-D的SOD任务。这三个示例分别来自NJU2K [38]，本文的SIP和NLPR [40]数据集。

因此，如何充分使用RGB-D信息来进行显著物体检测已成为当前的研究焦点 [37]–[52]。

现有智能手机摄像头的主要功能之一是通过粗糙的、包围框级或实例级的分割来识别可视场景中的人物。为此，诸如RGBD的显著性检测技术之类的智能解决方案已引起广泛关注。

然而，大多数现有的基于RGB-D的SOD方法都是基于Kinect体感设备 [53]、光场相机 [54] 或光流估计 [55]而获取的RGB和Depth数据，这些设备的特性不同于实际的智能手机摄像头。由于人物是智能手机拍照的关键物体，因此，一个以人为中心、具有真实场景的RGB-D数据集对手机制造商来说更有用。尽管 [38]、 [40]试图通过添加其他物体的方式来扩大场景，但尚不存在以人为中心的、可用于显著物体检测的RGB-D数据集。

此外，尽管深度图可以为识别显著物体提供重要的补充信息，但低质量的深度图往往会导致误检 [56]。虽然现有的基于RGB-D的SOD模型通过不同的策略融合RGB和深度特征 [52]，但仍然没有模型明确提出自动丢弃低质量的深度图（见图 1）。我们认为，此类模型对推动这个领域的

发展潜力巨大。

除了已经提到的当前RGB-D数据集和模型的局限性之外，大多数RGB-D研究还受到以下几个常见的限制：

充分性： 在最近的论文 [40], [57]中，仅对有限的数据集（1∼4）进行了基准测试（见Table III）。在如此少量的数据集上进行实验无法准确验证模型的通用性。

完整性： F-measure [58]、平均绝对误差（MAE）和精度与召回率曲线（PR Curve）是目前使用最为广泛的三个评估指标。但是，正如 [59]和 [60]中所述，这些指标本质上是像素级的评价指标。因此，很难从定量评估 [61]中得出全面而可靠的结论。

公平性： 一些研究如 [50], [52], [62]，虽然都使用的是F-measure，但并未明确描述具体的统计值（例如，均值还是最大值），这样容易导致不公平的比较和不一致的性能。同时，对于F-measure不同的阈值化策略（例如：255个变化的阈值 [52], [62], [63]、自适的显著性阈值 [40], [42]和自适应阈值 [44]）会得出不同的性能。因此，使用同一套评价指标对各个RGB-D显著物体检测模型进行广泛评估，从而提供一个公平的比较是至关重要的。

A. 贡献

为了解决上述问题，本文做出以下三个贡献。

1) 我们构建了一个全新的显著人物数据集（**Salient Person**，***SIP***）（见 Fig. 2, Fig. 3）。该数据集由929张具有精准标注的高分辨率图像组成，每张图像均包含多个显著人物。值得一提的是，深度图是由实际的智能手机采集。我们认为这样的数据集非常有价值，并且有助于将RGB-D模型应用于移动设备。此外，数据集经过精心设计，涵盖各种场景和各种极具挑战的情况（例如遮挡和外观变化），并且精心标注像素级的真值图（GT）。本文 SIP 数据集的另一个显著特征是，提供了双目相机采集的 RGB 和灰度图像，这有助于其他多个研究方向，例如立体匹配，深度估计和以人为中心的检测等。

2) 通过本文的SIP和六个现有的RGB-D数据集 [38]–[40], [64]–[66]，我们对32个经典的RGB-D的SOD模型进行全面总结，并给出了对18个最先进（SOTA）算法的大规模（约97,000张图像）的全面评估 [38]–[48], [50], [56], [67]–[69]，使该工作成为一个全面的RGB-D基准。为了进一步推动该领域的发展，我们还提供了一个存有测试集的在线评估平台。

3) 本文提出了一个简单通用的模型，称为深度图过滤器的深度网络（Deep Depth-Depurator Network，**D^{3}Net**），该模型创新地引入深度图过滤单元（Depth Depurator Unit，DDU）从而自动丢弃低质量的深度图。由于采用了门连接机制，本文的D^{3}Net可以更准确地预测出显著物体。大量实验表明，本文的D^{3}Net网络在许多有挑战性的数据集上的性能明显优于之前的工作。这样的通用框架设计有助于RGB和深度图的跨模式特征学习。

总之，本文贡献了一个完整的评测系统。如，用于全面评估RGB-D模型的评测工具和对基于RGB-D建模任务的深入分析并为该领域的研究指明了方向。

B. 章节安排

在§ II中，我们首先回顾RGB-D SOD任务的数据集以及最具代表性的模型。然后，在 § III中详细介绍本文的显著人物数据集SIP。在 § IV中，描述了本文提出的，能够显式滤除低质量的深度图的RGB-D SOD模型，我们称之为D^{3}Net。

在§ V中，我们对本文算法进行了定量和定性的实验分析。具体来说，在§ V-A中，我们提供了有关实验设置的更多细节，包括基准测试模型、数据集和运行时间。在§ V-B中，详细描述了5个评价指标（E测评法（E-measure）[60]、S测评法（S-measure）[59]、平均绝对误差（MAE）、精度与召回率曲线（PR Curve）和 F测评法（F-measure）[58]）。在§ V-C中，我们提供了不同数据集的平均统计数据，并在Table II中进行总结。在7个数据集*STERE* [64]、*LFSD* [66]、*DES* [39]、*NLPR* [40]、*NJU2K* [38]、*SSD* [65]和SIP（本文的）上，通过对18个基于RGB-D的最先进的SOD模型的实验结果比较，清晰地展示了本文D^{3}Net模型的鲁棒性和效率。此外，在§ V-D中，我们对传统模型和深度模型进行了性能比较。我们对实验结果也进行了更深入的探讨。在§ V-E中，我们提供了结果的可视化，并展示了在各种挑战性的场景下生成的显著图像。在§ VI中，我们讨论了有关人类活动的一些潜在应用，并在变换背景的应用中提供了一个有趣且现实的D^{3}Net使用场景。为了更好地理解DDU在本文的D^{3}Net网络中的贡献，在§ VII中，我们介绍了DDU的上限和下限。总而言之，大量的实验结果清晰地表明，我们的D^{3}Net模型在五个不同的指标上超过了任何其他竞争对手。在§ VII-B中，我们讨论了工作的局限性。最后，§ VIII总结了全文。

II. 相关工作

A. *RGB-D* 数据集

在过去几年中，SOD领域已经出现多个RGB-D数据集。这些数据集的统计数据详见Table III。具体来说，*STERE* [64]数据集是该领域中的第一个立体图像集。*GIT* [37]，*LFSD* [66] 和*DES* [66]是三个小规模数据集。*GIT* 和*LFSD*是根据特定目的设计的，比如，基于显著性的通用物体分割和光场上的显著性检测。*DES* 包含微软Kinect [53]传感设备捕获的135张室内图像。尽管这些数据集在很大程度上促进了该领域的发展，但数据集的小规模和低分辨率也严重制约了该领域的发展。为了克服这些障碍，Peng等人创建了*NLPR* [40]，这是一个大规模RGB-D数据集，图像分辨率为640×480。后来，Ju等

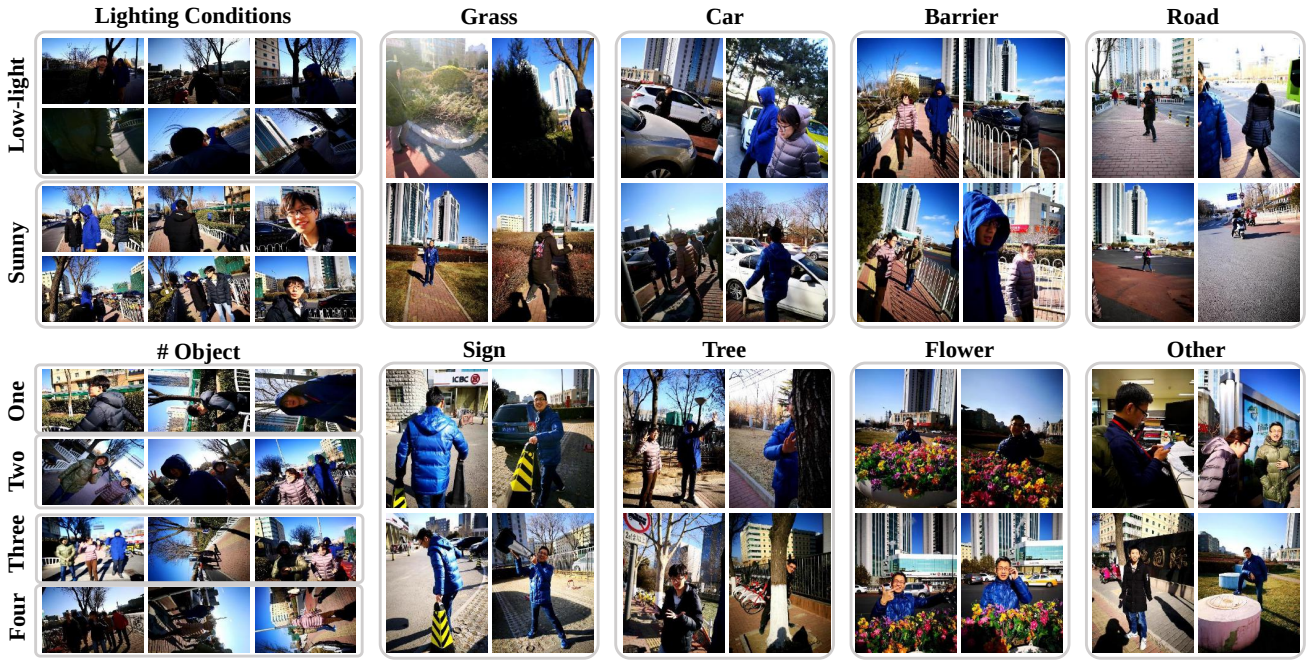


Figure 2. *SIP*中的代表性子集。*SIP*中的图像根据背景对象（例如草，汽车，障碍物，道路，标志，树木，花朵等）、不同的光照条件（即弱光和晴朗且物体边界清晰）、以及物体的数量（即1、2、≥3）分为八个子集。

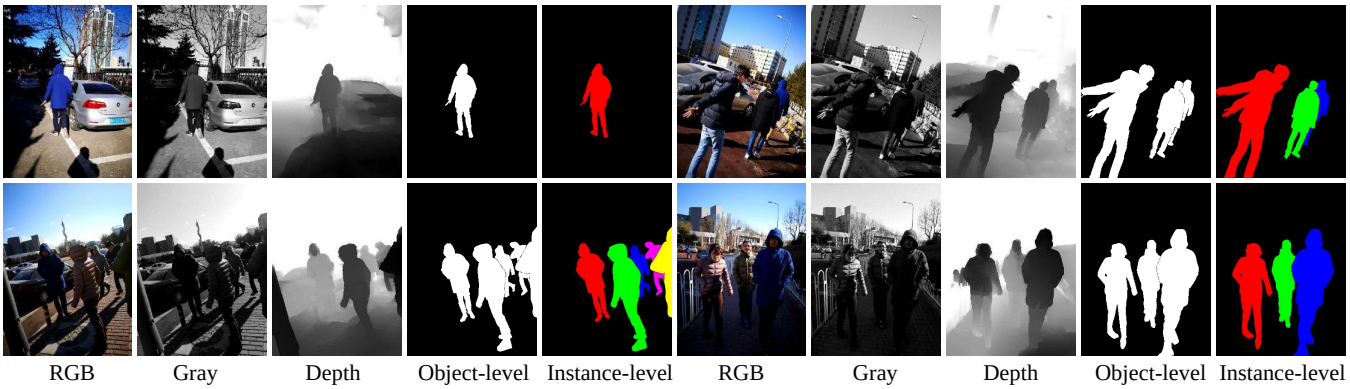


Figure 3. *SIP*数据集中具有不同数量的显著物体，物体尺寸，物体位置，场景复杂度和照明条件的图像，深度图和标注（即对象级和实例级的）的示例。请注意，“ RGB”和“灰色”图像是由两个距离很近的单眼相机分别捕获的。因此，“灰色”图像与通过彩色（RGB）图像灰度化而获得的灰度图像略有不同。我们的*SIP*数据集可提供一个新的研究方向，例如根据“ RGB”和“灰色”图像进行深度估计以及实例级的RGB-D显著物体检测任务。

人构建的*NJU2K* [38]数据集，已成为最受欢迎的RGB-D数据集之一。最近的*SSD* [65]数据集部分程度弥补了*NLPR*和*NJU2K*分辨率的限制。但是，它仅包含80张图像。尽管现有的RGB-D数据集取得了进步，但它们仍然受到以下限制：无法提供在实际智能手机上采集的深度图，从而不能反映真实的环境条件（例如，照明或到物体的距离）。

与以前的数据集相比，本文*SIP*数据集具有三个不同：

- 数据集包括929幅各种来自室外场景的图像，他们包含了各种极具挑战性的情况 [83]（如，黑暗的背景，遮挡，外观变化和超出视野）。
- 具有双摄像头的智能手机可以采集RGB，灰度图像和估计的深度图。由于SOD在移动电话上主要应用于人物，因此我们也将重点放在这一点上，从而首次强调了现实世界场景中的显著人物。
- 提出了对数据集质量的详细定量分析（例如，中心偏向和物体尺寸分布），而在以前的基于RGB-D的研究中并未对此进行详细研究。

B. RGB-D 模型

传统的模型很大程度上依赖手工提取的特征（如：对比度 [39], [40], [73], [75]，形状 [37]等）。通过编码经典原则（例如，空间偏向 [39]，中心暗通道 [47]，3D [77]和背景 [41], [48]），利用高斯差分 [38]，区域分类 [63]，SVM [46], [73]，图知识 [56]，元胞自动机 [43]和马尔可夫随机场 [41], [75]等方式建模的模型表明，特定的手工特征可实现良好的性能。一些研究还探索了各种组合策略的使用，

Table I
31个基于RGB-D的经典算法和本文模型（D^{3}Net）之间的比较。**Train/Val Set. (#)**表示训练集或测试集。*NLR* = *NLPR* [40]。*NJU* = *NJU2K* [38]。*MK* = *MSRA10K* [70]。*O* = *MK* + *DUTS* [71]。**Basic:** 4Priors 代表四种先验，即，区域、背景、深度和表面方向等先验信息。IPT: 初始化参数的传递。LGBS先验：局部对比、全局对比、背景和空间先验。RFR [72]：随机森林回归。MCFM：多约束特征匹配。CLP：交叉标签传播。在**Type:**中，T表示传统方法，D表示深度学习。**SP.**表示超像素，即是否是使用超像素的方法。**E测评法:** Table II中是七个数据集的得分范围。评估工具:https://github.com/DengPingFan/E-measure.

序号	模型	年份	出版物	训练集/验证机(#)	测试集(#)	基本要素	类别	超像素	E评测法↑ [60]
1	LS [37]	2013	BMVC	Without training dataset	One	Markov Random Field	T	✓	Not Available
2	RC [73]	2013	BMVC	Without training dataset	One	Region Contrast, SVM [74]	T		Not available
3	LHM [40]	2014	ECCV	Without training dataset	One	Multi-Context Contrast	T	✓	0.653∼0.771
4	DESM [39]	2014	ICIMCS	Without training dataset	One	Color/Depth Contrast, Spatial Bias Prior	T		0.770∼0.868
5	ACSD [38]	2014	ICIP	Without training dataset	One	Difference of Gaussian	T	✓	0.780∼0.850
6	SRDS [75]	2014	DSP	Without training dataset	One	Weighted Color Contrast	T		Not available
7	GP [41]	2015	CVPRW	Without training dataset	Two	Markov Random Field, 4Priors	T	✓	0.670∼0.824
8	PRC [63]	2016	Access	Without training dataset	Two	Region Classification, RFR	T		Not available
9	LBE [42]	2016	CVPR	Without training dataset	Two	Angular Density Component	T	✓	0.736∼0.890
10	DCMC [56]	2016	SPL	Without training dataset	Two	Depth Confidence, Compactness, Graph	T	✓	0.743∼0.856
11	SE [43]	2016	ICME	Without training dataset	Two	Cellular Automata	T	✓	0.771∼0.856
12	MCLP [67]	2017	Cybernetic	Without training dataset	Two	Addition, Deletion and Iteration Scheme	T	✓	Not available
13	TPF [65]	2017	ICCVW	Without training dataset	Four	Cellular Automata, Optical Flow	T	✓	Not available
14	CDCP [47]	2017	ICCVW	Without training dataset	Two	Center-dark Channel Prior	T	✓	0.700∼0.820
15	DF [45]	2017	TIP	*NLR* (0.75K) + *NJU* (1.0K)	Three	Laplacian Propagation, LGBS Priors	D	✓	0.759∼0.880
16	BED [76]	2017	ICCVW	*NLR* (0.80K) + *NJU* (1.6K) + MK (9K)	Two	Background Enclosure Distribution	D	✓	Not available
17	MDSF [46]	2017	TIP	*NLR* (0.50K) + *NJU* (0.5K)	Two	SVM [74], RFR, Ultrametric Contour Map	T		0.779∼0.885
18	MFF [77]	2017	SPL	Without training dataset	One	Minimum Barrier Distance, 3D prior	T		Not available
19	Review [57]	2018	TCSVT	Without training dataset	Two	Without model introduced	T		Not available
20	HSCS [68]	2018	TMM	Without training dataset	Two	Hierarchical Sparsity, Energy Function	T	✓	Not available
21	ICS [69]	2018	TIP	Without training dataset	One	MCFM, CLP	T	✓	Not available
22	CDB [48]	2018	NC	Without training dataset	One	Background Prior	T	✓	0.698∼0.830
23	SCDL [78]	2018	DSP	*NLR* (0.75K) + *NJU* (1.0K)	Two	Silhouette Feature, Spatial Coherence Loss	D		Not available
24	PCF [50]	2018	CVPR	*NLR* (0.70K) + *NJU* (1.5K)	Three	Complementarity-Aware Fusion module [50]	D		0.827∼0.925
25	CTMF [44]	2018	Cybernetic	*NLR* (0.65K) + *NJU* (1.4K)	Four	HHA [79], IPT, Hidden Structure Transfer	D		0.829∼0.932
26	ACCF [80]	2018	IROS	*NLR* (0.65K) + *NJU* (1.4K)	Three	Attention-Aware	D		Not available
27	PDNet [49]	2019	ICME	*NLR* (0.50K) + *NJU* (1.5K) + *O* (21K)	Five	Depth-Enhanced Net [49]	D		Not available
28	AFNet [62]	2019	Access	*NLR* (0.70K) + *NJU* (1.5K)	Three	Switch map, Edge-Aware loss	D		0.807∼0.887
29	MMCI [81]	2019	PR	*NLR* (0.70K) + *NJU* (1.5K)	Three	HHA [79], Dilated Convolutional	D		0.839∼0.928
30	TANet [82]	2019	TIP	*NLR* (0.70K) + *NJU* (1.5K)	Three	Attention-Aware Multi-Modal Fusion	D		0.847∼0.941
31	CPFP [52]	2019	CVPR	*NLR* (0.70K) + *NJU* (1.5K)	Five	Contrast Prior, Fluid Pyramid	D		0.852∼0.932
32	**D^{3}Net (Ours)**	2020		*NLR* (0.70K) + *NJU* (1.5K)	Seven	Depth Depurator Unit	D		0.862∼0.953

例如，角密度 [42]，随机森林回归器 [46], [63]和最小障碍距离 [77]来集成RGB和深度特征的方法。Table I中显示了更多详细信息。

为了克服手工特征的表达局限性，近期的一些工作 [44], [45], [49], [50], [52], [62], [76], [78], [80]–[82] 已经提出使用卷积神经网络（CNN）去预测RGB-D图像中的显著物体。BED [76] 和DF [45] 是将深度学习应用在RGB-D显著物体检测任务中的两个开拓性工作。最近，Huang 等人 [78] 开发了一种更有效的具有修正损失函数的端到端模型。为了解决训练数据的不足，Zhu等人 [49]提出了一个鲁棒的先验模型，带有针对SOD的引导深度增强模块。另外，Chen等人 [50]为此领域开发了一系列新颖的方法，例如隐藏结构转移 [44]，互补融合模块 [50]，注意感知组 [80], [82]和膨胀卷积 [81]。

然而，这些工作都致力于通过各种策略来提取深度特征或信息。我们认为，并非深度图中的所有信息都能提供SOD信息，低质量的深度图通常会引入明显的噪声（Fig. 1中的第一行）。因此，我们改为设计一个简单的通用框架D^{3}Net，该框架配有深度图过滤单元，以在学习互补特征时明确排除低质量的深度图。

III. 本文数据集

A. 数据集概述

SIP数据集是第一个以人类活动为主的显著人物检测数据集。我们的数据集包含在八个不同背景场景下的929张RGB-D图像，他们在两种不同的对象边界条件下，扮演了多重角色。每幅图像中每个人都穿着不同的衣服。参照 [83]，我们精心地挑选图像使其涵盖各种极具挑战的情况（例如，外观变化，遮挡和复杂的形状）。在Fig. 2 和Fig. 3中展示了这些例子。从网站http://dpfan.net/SIPDataset/可以下载整个数据集。

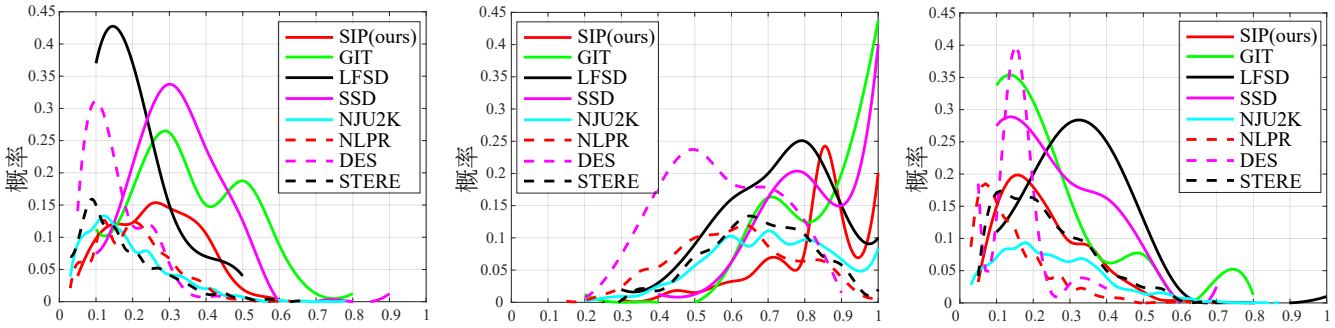
B. 传感器和数据采集

图像采集： 我们使用华为Mate 10来采集图像。Mate 10的后置摄像头采用了徕卡SUMMILUX-H镜头，光圈为f/1.6，并结合了12MP RGB和20MP单色（灰度）传感器。深度图是由Mate 10自动估计的。我们请了9个人，他们身着不同颜色的衣服，在真实的日常场景中表演特定的动作。给出了如何执行动作以覆盖不同具有挑战性的情况（例如遮挡和视野外）的说明，但没有提供关于风格、角度或速度的说明，以便记录真实的数据。

Table II
在*SIP* 和6个经典数据集 [38]–[40], [64]–[66] 上对18个先进的RGB-D方法的基准测试结果。↑ & ↓ 分别表示越大越好和越小越好。"-T" 表示对应数据集的测试集。对于传统模型，统计数据基于总体数据集而不是测试集。"RANK" 表示每个模型在特定评价指标上的排名。"ALL RANK" 表示在特点数据集上(排名均值)。性能最突出的以加粗显示。

*	Model	2014-2017											2018-2019						
		LHM [40]	CDB [48]	DESM [39]	GP [41]	CDCP [47]	ACSD [38]	LBE [42]	DCMC [56]	MDSF [46]	SE [43]	DF [45]†	AFNet [62]†	CTMF [44]†	MMCI [81]†	PCF [50]†	TANet [82]†	CPFP [52]†	**D^{3}Net Ours**†
	Time (s)	2.130	-	7.790	12.98	>60.0	0.718	3.110	1.200	>60.0	1.570	10.36	0.030	0.630	0.050	0.060	0.070	0.170	**0.015**
	Code	M	-	M	M&C	M&C	C	M&C	M	C	M&C	M&C	Tf	Caffe	Caffe	Caffe	Caffe	Caffe	Pytorch
NJU-T [38]	$S_\alpha \uparrow$	.514	.624	.665	.527	.669	.699	.695	.686	.748	.664	.763	.772	.849	.858	.877	.878	.879	**.900**
	$F_\beta \uparrow$	.632	.648	.717	.647	.621	.711	.748	.715	.775	.748	.804	.775	.845	.852	.872	.874	.877	**.900**
	$E_\xi \uparrow$	.724	.742	.791	.703	.741	.803	.803	.799	.838	.813	.864	.853	.913	.915	.924	.925	.926	**.950**
	$M \downarrow$	.205	.203	.283	.211	.180	.202	.153	.172	.157	.169	.141	.100	.085	.079	.059	.060	.053	**.041**
	Rank	17	16	14	17	15	12	10	13	9	11	7	7	6	5	4	3	2	1
STERE [64]	$S_\alpha \uparrow$	.562	.615	.642	.588	.713	.692	.660	.731	.728	.708	.757	.825	.848	.873	.875	.871	.879	**.899**
	$F_\beta \uparrow$	.683	.717	.700	.671	.664	.669	.633	.740	.719	.755	.757	.823	.831	.863	.860	.861	.874	**.891**
	$E_\xi \uparrow$	.771	.823	.811	.743	.786	.806	.787	.819	.809	.846	.847	.887	.912	.927	.925	.923	.925	**.938**
	$M \downarrow$	.172	.166	.295	.182	.149	.200	.250	.148	.176	.143	.141	.075	.086	.068	.064	.060	.051	**.046**
	Rank	16	12	14	18	13	15	17	10	11	9	8	7	6	3	4	5	2	1
DES [39]	$S_\alpha \uparrow$	.578	.645	.622	.636	.709	.728	.703	.707	.741	.741	.752	.770	.863	.848	.842	.858	.872	**.898**
	$F_\beta \uparrow$	.511	.723	.765	.597	.631	.756	.788	.666	.746	.741	.766	.728	.844	.822	.804	.827	.846	**.885**
	$E_\xi \uparrow$	.653	.830	.868	.670	.811	.850	.890	.773	.851	.856	.870	.881	.932	.928	.893	.910	.923	**.946**
	$M \downarrow$	.114	.100	.299	.168	.115	.169	.208	.111	.122	.090	.093	.068	.055	.065	.049	.046	.038	**.031**
	Rank	18	13	14	17	16	12	10	15	11	9	7	8	3	5	6	4	2	1
NLR-T [40]	$S_\alpha \uparrow$	.630	.629	.572	.654	.727	.673	.762	.724	.805	.756	.802	.799	.860	.856	.874	.886	.888	**.912**
	$F_\beta \uparrow$	.622	.618	.640	.611	.645	.607	.745	.648	.793	.713	.778	.771	.825	.815	.841	.863	.867	**.897**
	$E_\xi \uparrow$	.766	.791	.805	.723	.820	.780	.855	.793	.885	.847	.880	.879	.929	.913	.925	.941	.932	**.953**
	$M \downarrow$	.108	.114	.312	.146	.112	.179	.081	.117	.095	.091	.085	.058	.056	.059	.044	.041	.036	**.030**
	Rank	14	15	16	18	12	17	10	13	7	11	8	8	5	6	4	3	2	1
SSD [65]	$S_\alpha \uparrow$	.566	.562	.602	.615	.603	.675	.621	.704	.673	.675	.747	.714	.776	.813	.841	.839	.807	**.857**
	$F_\beta \uparrow$	.568	.592	.680	.740	.535	.682	.619	.711	.703	.710	.735	.687	.729	.781	.807	.810	.766	**.834**
	$E_\xi \uparrow$	.717	.698	.769	.782	.700	.785	.736	.786	.779	.800	.828	.807	.865	.882	.894	.897	.852	**.910**
	$M \downarrow$	.195	.196	.308	.180	.214	.203	.278	.169	.192	.165	.142	.118	.099	.082	.062	.063	.082	**.058**
	Rank	16	17	15	11	17	13	14	9	12	9	7	8	6	4	2	2	5	1
LFSD [66]	$S_\alpha \uparrow$	.553	.515	.716	.635	.712	.727	.729	.753	.694	.692	.783	.738	.788	.787	.786	.801	**.828**	.825
	$F_\beta \uparrow$	.708	.677	.762	.783	.702	.763	.722	.817	.779	.786	.813	.744	.787	.771	.775	.796	**.826**	.810
	$E_\xi \uparrow$	.763	.766	.811	.824	.780	.829	.797	.856	.819	.832	.857	.815	.857	.839	.827	.847	**.872**	.862
	$M \downarrow$	.218	.225	.253	.190	.172	.195	.214	.155	.197	.174	.145	.133	.127	.132	.119	.111	**.088**	.095
	Rank	17	18	16	12	15	11	14	6	13	9	5	10	4	7	8	3	1	2
SIP (Ours)	$S_\alpha \uparrow$	.511	.557	.616	.588	.595	.732	.727	.683	.717	.628	.653	.720	.716	.833	.842	.835	.850	**.860**
	$F_\beta \uparrow$	.574	.620	.669	.687	.505	.763	.751	.618	.698	.661	.657	.712	.694	.818	.838	.830	.851	**.861**
	$E_\xi \uparrow$	.716	.737	.770	.768	.721	.838	.853	.743	.798	.771	.759	.819	.829	.897	.901	.895	.903	**.909**
	$M \downarrow$	.184	.192	.298	.173	.224	.172	.200	.186	.167	.164	.185	.118	.139	.086	.071	.075	.064	**.063**
	Rank	17	16	14	12	18	6	9	14	10	11	13	7	8	5	3	4	2	1
	All Rank	18	17	15	14	16	13	12	11	10	9	7	8	6	5	4	3	2	1

数据标注: 在采集了5,269张图像及其深度图之后，首先手动选择了大约2,500张图像，每个图像都包含一个或多个显著人物。与许多著名的SOD数据集 [20], [58], [70], [71], [85]–[91] 一致，六名观察者根据指示，以他们的第一直觉画出最引人注目的人物周围的包围盒(bboxes)。我们采用 [40]中描述的投票机制来丢弃投票一致性较低的图像，并选择了前1,000张最满意的图像。然后另外五个标注者根据包围盒标注显著物体的准确轮廓。我们丢弃了一



(a) 物体中心到图像中心的距离 (b) 边缘到图像中心的距离 (c) 归一化的物体大小

Figure 4. （a）归一化的物体中心距图像中心的距离分布。（b）归一化的物体边缘（物体中最远点）到图像中心的距离分布。（c）归一化的物体大小分布。

Table III
现有RGB-D数据集的对比：年份（YEAR）、出版物（PUB.）、数据集大小（DS.）、图像中的物体数量（#OBJ.）、场景类型（TYPES.）、深度传感器（SENSOR.）和深度图质量（**DQ.**，例如高质量的深度图受到较少的随机噪声。详见FIG. 1的最后一行）、标注质量（**AQ.**，见FIG. 12），是否提供由单目相机采集的灰度级图像（**GI.**），中心偏置B（**CB.**, 详见FIG. 4 (A)-(B)）和分辨率（以像素为单位）。H & W分别表示图像的高度和宽度。

No.	Dataset	Year	Pub.	DS.	#Obj.	Types.	Sensor.	DQ.	AQ.	GI.	CB.	Resolution (H×W)
1	*STERE* [64]	2012	CVPR	1K	∼one	internet	Stereo camera+sift flow [55]		High	No	High	[251∼1200]×[222∼900]
2	*GIT* [37]	2013	BMVC	0.08K	multiple	home environment	Microsoft Kinect [53]		High	No	Low	640 × 480
3	*LFSD* [66]	2014	CVPR	0.1K	one	60 indoor/40 outdoor	Lytro Illum camera [54]		High	No	High	360 × 360
4	*DES* [39]	2014	ICIMCS	0.135K	one	135 indoor	Microsoft Kinect [53]	High		No	High	640 × 480
5	*NLPR* [40]	2014	ECCV	1K	multiple	indoor/outdoor	Microsoft Kinect [53]	High		No	High	640 × 480, 480 × 640
6	*NJU2K* [38]	2014	ICIP	1.985K	∼one	3D movie/internet/photo	FujiW3 camera+optical flow [84]		High	No	High	[231∼1213]×[274∼828]
7	*SSD* [65]	2017	ICCVW	0.08K	multiple	three stereo movies	Sun's optical flow [84]			No	Low	960 × 1080
8	***SIP (Ours)***	**2020**	TNNLS	**0.929K**	**multiple**	**person in the wild**	**Huawei Mate10**	**High**	**High**	**Yes**	**Low**	**992×744**

Table IV
*SIP*数据集中关于相机/物体运动和显著物体实例数量的统计。

SIP (Ours)	Background Objects								Object Boundary		# Object		
	car	flower	grass	road	tree	signs	barrier	other	dark	clear	1	2	≥3
#Img	107	9	154	140	97	25	366	32	162	767	591	159	179

些低质量标注的图像，最终得到929张高质量标注的真值图像。

C. 数据集统计

中心偏向： 中心偏向被视为显著性检测数据集的最重要偏向之一 [92]。发生这种情况是因为受试者倾向于注视屏幕中心 [93]。如 [83]所述，简单地重叠数据集中的所有图并不能很好地描述中心偏向的程度。

参考 [83]，我们在Fig. 4（a）和（b）中给出了两个距离R_o和R_m的统计数据，其中R_o和R_m分别表示物体中心和物体边缘（最远）的点离图像中心的距离。我们的SIP和现有的 [37]–[40], [64]–[66] 数据集的中心偏向如Fig. 4（a）和（b）所示。除了我们的SIP和两个小规模数据集（*GIT*和*SSD*）之外，大多数数据集都具有很高的中心偏向，即物体的中心靠近图像中心。

物体大小： 我们将物体大小定义为图像中显著物体像素与像素总数之比。SIP中归一化后的物体尺寸分布（见Fig. 4（c））为0.48%∼66.85%（平均值为：20.43%）。

背景物体: 如Table IV所示，SIP包含各种背景物体（例如汽车，树木和草地）。在这种数据集上测试的模型可以更好地处理现实场景，因此更加实用。

物体边界条件: 在Table IV中，我们在SIP数据集中显示了不同的物体边界条件（例如，黑暗和清晰）。如Fig. 3所示，黑暗场景会经常出现在日常场景中。在弱光条件下获得的深度图不可避免地会为检测显著物体带来更多的挑战。

显著物体的数量： 从Table III中，我们注意到现有数据集的显著物体的数量不足（例如，它们通常只有一个）。然而，先前的研究 [94]表明，人类可以准确地枚举至少五个物体而无需计数。因此，我们的SIP设计为每张图像最多包含五个显著物。Table IV中的(# Object)显示了每张图像中带标签的物体的统计信息。

IV. 本文模型

根据Fig. 1描述的动机，我们非常需要跨模态特征提取和深度图过滤单元。因此，我们提出了一个简单

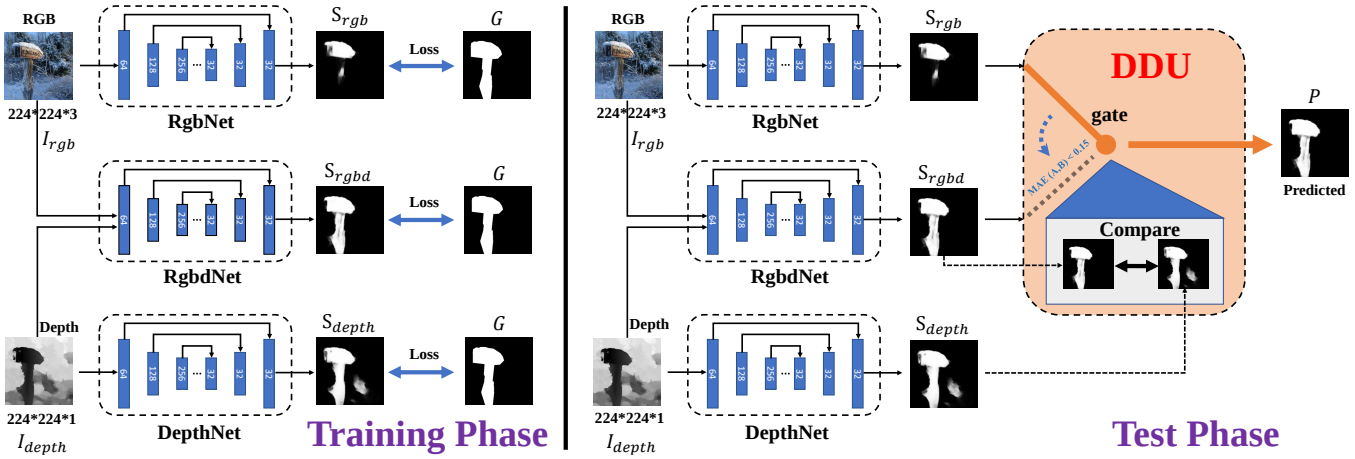


Figure 5. 本文的D^{3}Net网络图。在训练阶段（左），使用三个并行子网络（例如RgbNet，RgbdNet和DepthNet）处理输入的RGB和深度图像。三个子网络都是基于相同的改进结构后的特征金字塔网络（FPN）（详见§ IV-A）。我们引入这些子网以获得三个显著性图（即S_{rgb}, S_{rgbd}和S_{depth}），这些显著图同时考虑了输入的粗糙和精细的细节。在测试阶段（右），本工作中首次使用新型DDU（详见§ IV-B）显式地丢弃（如S_{rgbd}）显著图或保留（如S_{rgbd}）引入了深度信息的显著图。在训练/测试阶段，这些组件形成一个紧密的结构，并经过精心设计（例如，DDU中的门连接），以自动从RGB图像和深度图像中联合推断出显著物体。

的通用D^{3}Net模型（Fig. 5），该模型包含两个组件，例如，三流特征学习模块（FLM）（请参阅§ IV-A）和DDU（请参阅§ IV-B）。FLM用于从不同的模态中提取特征，而DDU则充当显式滤除低质量深度图的开关。如果DDU决定过滤掉该深度图，则数据流将与RgbNet一起传递。这些组件经过精心设计，形成了紧密的结构，在各种具有挑战的数据集上实现强大的性能和较高的泛化性。

A. 特征学习模块

现有的大多数模型 [95]–[97]都表现出在多种应用中物体检测性能的提高。这些模型通常都使用类似的FPN结构 [98]。基于这种动机，我们决定在D^{3}Net 基线中引入诸如FPN之类的组件，以金字塔的方式有效地提取特征。由于DDU选择仅在测试阶段使用，因此整个D^{3}Net 模型分为训练阶段和测试阶段。

如 Fig. 5所示，设计的FLM出现在训练和测试阶段。FLM由三个子网组成，即 *RgbNet, RgbdNet*, 和*DepthNet*。请注意，三个子网工作具有相同的结构，但是输入的特征通道数目不同。具体而言，每个子网都以224×224的分辨率接收缩放后的图像$I \in \{I_{rgb}, I_{rgbd}, I_{depth}\}$。FLM的目标是获得相应的预测图$S \in \{S_{rgb}, S_{rgbd}, S_{depth}\}$。

如 [98]中所述，我们还使用了自下而上，自上而下的路径以及横向连接来提取特征。然后，将在多个级别按比例组织输出。FPN独立于骨干网，因此，为简单起见，我们采用VGG-16 [99]架构作为基本卷积网络来提取空间特征。未来我们还可以进一步探索采用更强大的 [100]特征提取器。一些研究，例如 [101]，表明更深的层保留了更多的语义信息以定位对象。基于此观察，我们在五层VGG-16结构的基础上引入了一个包含两个3×3卷积核的层，以实现此目标。

如Fig. 6所示，我们建立了自上而下的特征。对于不同的图层（例如，较粗的图层），我们首先使用最近邻操作

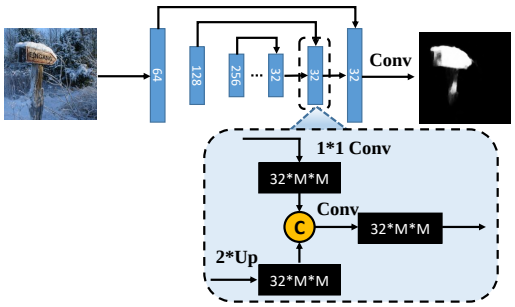


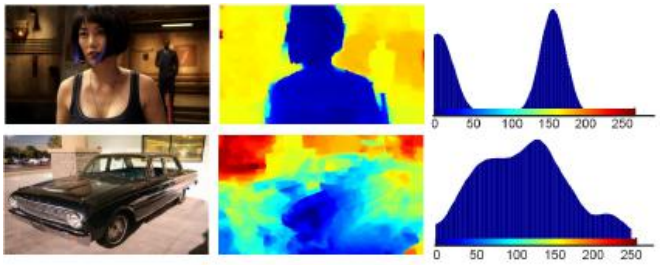
Figure 6. 引入特征金字塔网络（FPN）以提取上下文感知信息。与 [98]不同，我们在VGG-16的基础上进一步添加了第六层，并且信息合并的策略是拼接而不是相加。更多详细信息，请参见§ IV-A。

进行2倍上采样。然后，将上采样的特征与更精细的特征图连接起来以获得丰富的特征。在与粗略图连接之前，较细图经过1×1转换操作以减少通道。例如，令$\mathbf{I_{rgbd}} \in \mathbb{R}^{W \times H \times 4}$ 表示RgbdNet输入的4-D特征张量。然后，我们在不同的层上定义一组锚，以便获得具有$C_i \times W_i \times H_i$的一组金字塔特征张量，即$\{64 \times 224 \times 224, 128 \times 112 \times 112, 256 \times 56 \times 56, 512 \times 28 \times 28, 512 \times 14 \times 14, 32 \times 7 \times 7, 32 \times 14 \times 14, 32 \times 28 \times 28, 32 \times 56 \times 56, 32 \times 112 \times 112, 32 \times 224 \times 224\}$ 分别对应$\{F_i, \text{i} \in [1,11]\}$特征。注意，$\{F_1, F_2, F_3, F_4, F_5\}$ 又对应于VGG-16的五个卷积层（即$\{C_1, C_2, C_3, C_4, C_5\}$）。

B. 深度图过滤单元(*DDU*)

在测试阶段，我们进一步采用新的门连接策略以获得最佳预测图。对比有用信息的线索，低质量的深度图将会给预测结果带来更多的噪声。闸门连接的目的是将深度图分类为合理的和低质量的图，并且在整个框架中不使用较差的深度图。

如图Fig. 7（b）所示，高质量深度图中的单个显著物体通常会带有明显的闭合边界特征，并在其直方图分布中显



(a) RGB (b) 深度图 (c) 直方图

Figure 7. 高质量（第一行）和低质量（第二行）深度图的平滑直方图。

示出清晰双峰效应。现有数据集 [38]–[40], [64]–[66]中的深度图统计数据也支持以下观察："高质量的深度图通常包含清晰的物体，而低质量的深度图通常含有许多混乱的元素（如Fig. 7中的第二行）。"为了过滤低质量的深度图，我们提出DDU。

更具体地说，在测试阶段，首先将RGB和深度图调整为固定大小（例如，与训练阶段224×224相同），以降低计算复杂度。如Fig. 5（右）所示，DDU通过栅极连接实现。用三个预测图$\mathbf{S} \in \{S_{rgb}, S_{rgbd}, S_{depth}\}$表示输入图像，然后DDU的目标是确定哪个预测图$\mathbf{P} \in [0,1]^{W \times H}$）是最优

$$\mathbf{P} = F_{ddu}(\{S_{rgb}, S_{rgbd}, S_{depth}\}). \quad (1)$$

直觉上，有两种方案可以实现此目标，即后处理和预处理。我们提出了一种简单但通用的DDU后处理方案。它运行在测试阶段而不是训练阶段考虑DDU，它利用比较单元F_{cu}来评估分别从DepthNet和RgbdNet生成的S_{depth}和S_{rgbd}之间的相似性。

$$F_{cu} = \begin{cases} 1, & \delta(S_{rgbd}, S_{depth}) \le t \\ 0, & otherwise, \end{cases} \quad (2)$$

$\delta(\cdot)$表示距离函数，t表示固定阈值。注意，比较单元F_{cu}充当决定应该使用哪个子网（RgbNet或RgbdNet）的索引。

比较单元的关键是DDU。本文利用比较单元F_{cu}作为门连接来确定最终/最佳预测结果$\mathbf{P}$。因此，我们的F_{ddu}模块可以表示为：

$$\mathbf{P} = F_{cu} \cdot S_{rgbd} + \bar{F}_{cu} \cdot S_{rgb}, \quad (3)$$

其中，$\bar{F}_{cu} = 1 - F_{cu}$。$F_{cu}$可以看作是固定权重。一个更优雅的方式(自适应权重)将是我们未来工作的一部分。

C. 实施细节

1) DDU: DDU是我们D^{3}Net 网络的关键组成部分。在这项工作中，我们展示了一个简单但功能强大的距离函数，该函数由(Eq. 2)公式化。我们利用平均绝对误差（MAE）指标（与(Eq. 5)相同）来评估两张结果图之间的距离。其基本思想是，如果高质量深度包含清晰的物体，DepthNet将轻松检测到S_{depth}的这些物体（请参见Fig. 7中的第一行）。I_{depth}的深度图的质量越高，S_{rgbd}和S_{depth}之间的相似度就越高。换句话说，来自RgbdNet的预测图S_{rgbd}考虑了I_{depth}的特征。如果深度图的质量较低，则RgbdNet的预测图将与DepthNet生成的图不同。我们在(Eq. 2)中测试了一组固定阈值t的值，例如0.01、0.02、0.05、0.10、0.15和0.20，但发现$t = 0.15$达到了最佳性能。

2) 损失函数: 我们采用广泛使用的交叉熵损失函数L来训练模型:

$$L(\mathbf{S}, \mathbf{G}) = -\frac{1}{N}\sum_{i=1}^{N}\Big(g_i \log(s_i) + (1-g_i)\log(1-s_i)\Big), \quad (4)$$

其中$\mathbf{S} \in [0,1]^{224\times224}$和$\mathbf{G} \in \{0,1\}^{224\times224}$分别表示估计的显著性图(即, S_{rgb}, S_{rgbd}, 或S_{depth})和真值图. $g_i \in \mathbf{G}$, $s_i \in \mathbf{S}$, N表示像素总数。

3) 训练设置: 为了公平地进行比较，我们遵循 [52] 中所述的相同训练设置。从*NJU2K* [38]数据集中选择1485对图像，从*NLPR* [40]数据集中选择了700对图像，作为训练数据(请参考网站上的*Trainlist.txt*)。本文的D^{3}Net采用Pytorch工具，使用Python实现。采用Adam作为优化器，初始学习率为1e-4，批处理大小设置为8。在带有12 GB内存的GTX TITAN X GPU上，总共训练迭代30次。

4) 数据扩充: 由于现有数据集规模有限，我们通过水平翻转图像来增强训练样本，以克服过拟合的风险。

V. 基准评测结果

在本文中，我们对大约97k张图像（5398图像×18个模型）进行了基准测试，使其成为迄今为止最大、最全面的基于RGB-D的SOD基准测试。

A. 实验设置

1) 模型: 我们以18个SOTA模型（请参见Table II）为基准，包括10个传统模型和8个基于CNN的模型。

2) 数据集: 我们在7个数据集上进行实验（详见Table II）。使用*NJU2K* [38] 和*NLPR* [40]的测试集以及整个*STERE* [64]、*DES* [39]、*SSD* [65]、*LFSD* [66] 和*SIP*数据集进行测试。

3) 运行时间: 在Table II中，我们对现有方法进行了总结。在同一平台上测试了时间：Intel Xeon(R) E5-2676v3 2.4GHz×24 and GTX TITAN X。由于 [44], [48], [50], [67]–[69], [80]–[82] 尚未公开其代码，它们运行时间来自原文或由作者提供。D^{3}Net网络未使用后处理（例如CRF），因此，对于224×224的图像，计算仅需花费约0.015 s。

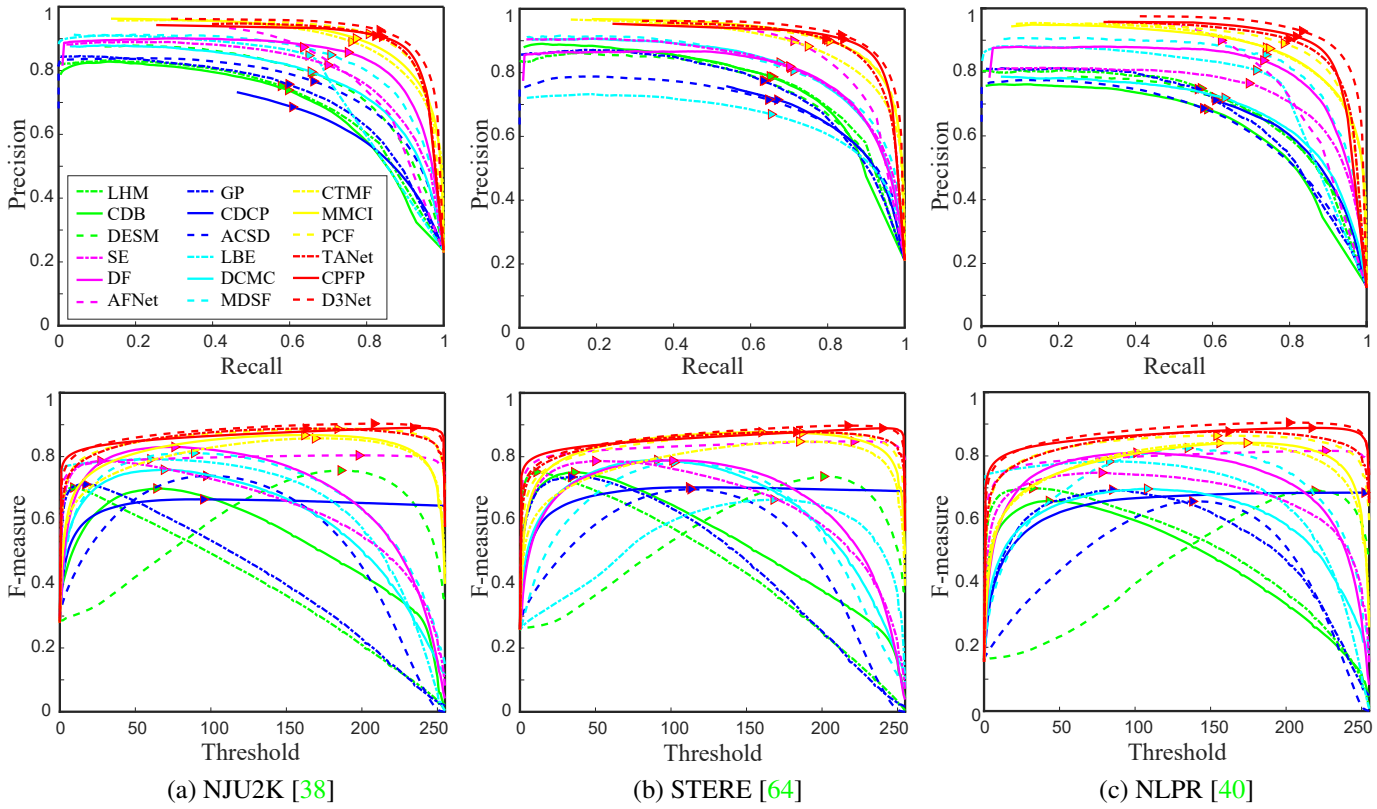


Figure 8. PR曲线（上）和F指标（下），来自NJU2K、STERE、和NLPR数据集上多个固定阈值的18个模型的测试结果。

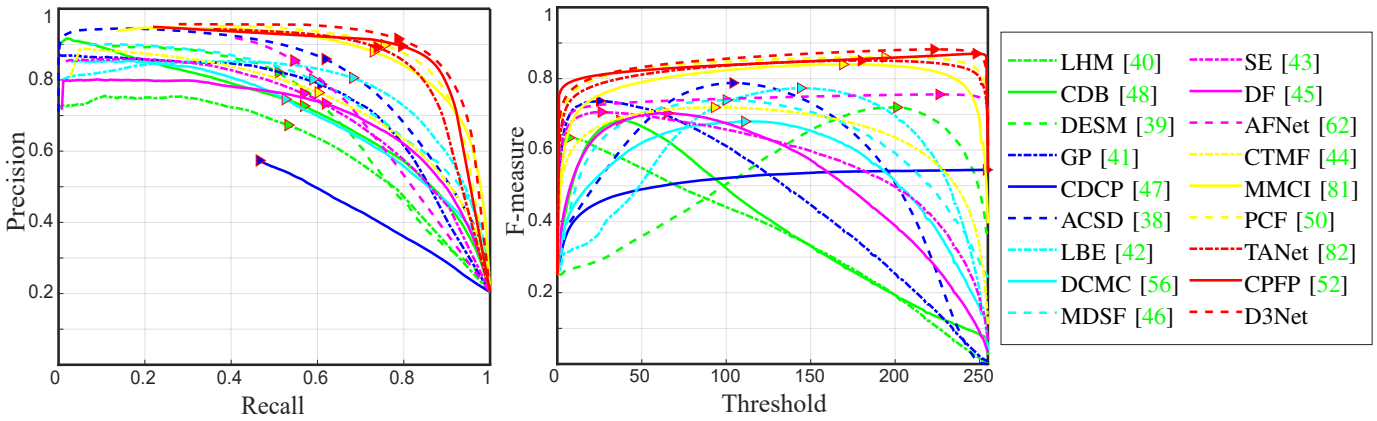


Figure 9. PR曲线（左）和F指标（右），来自SIP数据集上不同阈值的结果。

B. 评估指标

1) MAE指标: 我们参考Perazzi等人的工作 [102]并评估所有像素的实际显著性图Sal和二进制真值图G之间的平均绝对误差:

$$\mathrm{MAE} = \frac{1}{N}|Sal - G|, \tag{5}$$

其中，N是像素总数。MAE估计显著图和真值图之间的近似度，并将其归一化为$[0,1]$。MAE提供了对预测图和真值图之间一致性的直接估算。但是，对于MAE度量，自然会为小物体分配较小的误差，而为大对象分配较大的误差。度量标准也无法确定错误发生的位置 [103]。

2) PR曲线: 同Borji等人 [6]一样，我们也采用PR曲线。我们从0到255变化的固定阈值划分显著性图S。对每个阈值，都会计算一对召回率和精度，然后将其组合以形成描述模型在不同情况下的性能的精确召回曲线。PR曲线的总体评价结果示于Fig. 8（顶部）和Fig. 9（左侧）。

3) F指标F_β: F指标本质上是基于区域的相似性度量。继Cheng 和Zhang等人的工作 [6], [104]之后。我们还提供了使用各种固定（0-255）阈值的最大F指标。Fig. 8 （底部）和Fig. 9 （右侧）显示了每个数据集在不同阈值下总体F指标的评估结果。

4) S指标S_α: MAE和F指标都忽略了重要的结构信息。然而，行为视觉研究表明，人类视觉系统对场景中的结构高度敏感 [59]。因此，我们增加了结构度量（S指标 [59]）。S指标结合了区域感知（S_r）和物体感知(S_o)得

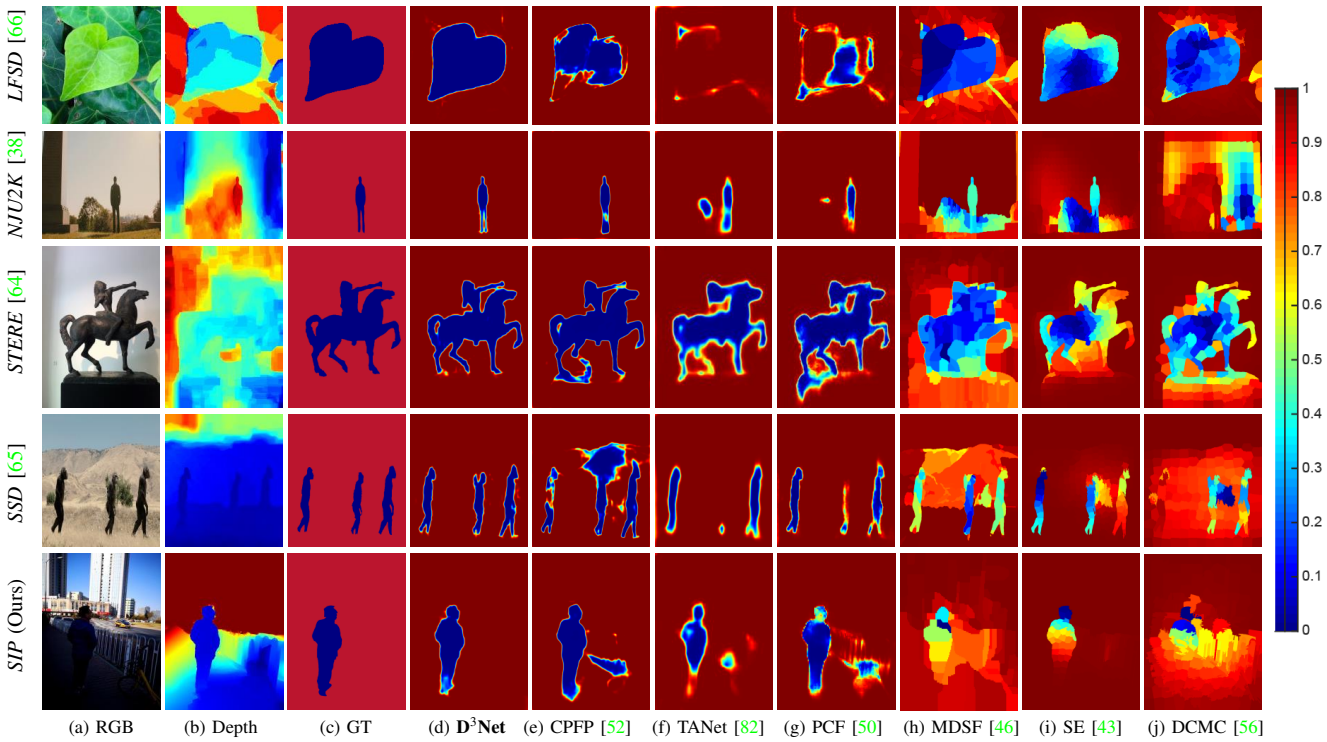


Figure 10. 在五个数据集上的视觉对比：排名前三的基于CNN的模型（CPFP [52]，TANet [82]，和PCF [50]）和三个经典的非深度方法（MDSF [46], SE [46] and DCMC [56]）。进一步的结果可查看http://dpfan.net/D3NetBenchmark。

到最终的结构相似性指标:

$$S_\alpha = \alpha * S_o + (1-\alpha) * S_r, \tag{6}$$

其中$\alpha \in [0,1]$ 是平衡参数，设置为：0.5。

*5) E*指标E_ξ*:* E指标是二值图评估领域中最近提出的增强的对齐度量指标 [60]。该方法基于认知视觉研究，将局部像素值与图像级平均值结合在一个表达式中，共同获取图像级统计信息和局部像素匹配信息。在这里，我们引入最大E指标以提供更全面的评估。

C. 评价指标统计

对于给定的评价指标$\zeta \in \{S_\alpha, F_\beta, E_\xi, M\}$，我们考虑不同的统计量。$I_j^i$ 表示来自特定数据集D_i的图像。因此，$D_i = \{I_1^i, I_2^i, \ldots, I_j^i\}$，$\overline{\zeta}(I_j^i)$是图像$I_j^i$的评估得分。平均值是定义为$M_\zeta(D_i) = \frac{1}{|D_i|}\sum \overline{\zeta}(I_j^i)$的平均数据集统计量。其中$|D_i|$ 是D_i数据集上的图像总数。Table II总结了不同数据集的平均统计量。

D. 性能比较与分析

1) 传统模型的性能: 基于Table II中列出的总体性能，我们观察到"*SE [43], MDSF [46],* 和*DCMC [56]* 是排名前三的传统算法。" SE和DCMC都采用超像素技术，从RGB图像中显式提取区域对比度特征。相比之下，MDSF将将SOD定义为像素化的二值标记问题，并采用SVM解决。

2) 深度模型的性能: 我们的D^{3}Net，CPFP [52]和TANet [82]是所有领先方法中排名前三的深度模型，展现了深度学习在此任务上的强大特征表示能力。

3) 传统模型vs.深度模型: 从Table II中，我们观察到大多数深度模型的性能均优于传统算法。有趣的是，MDSF [65]优于*NLPR*数据集上的两个深度模型（即DF [45] 和AFNet [62]）。

E. 与*SOTA*模型的比较

我们在表II中将D^{3}Net 与17个SOTA模型进行了比较。总的来说，我们的模型在六个数据集上的表现优于现有公开的最好结果（CPFP [52]-CVPR'19），幅度为1.0% ∼ 5.8%。值得注意的是，在本文提出的*SIP*数据集上获得了1.4%的显著性能提升。

我们还将报告在各种具有挑战性的场景中生成的显著性图像，以显示D^{3}Net的视觉优势。Fig. 10中展示出了一些代表性示例，诸如当深度图中的显著物体的结构被部分（例如，第一，第四和第五行）或显著地（即，第二至第三行）损坏时。具体地，在第三和第五行中，显著物体的深度与背景场景局部连接。另外，第四行包含多个隔离的显著物体。对于这些具有挑战性的情况，大多数现有的顶级模型由于使用了低质量的深度图或融合方法不充分而不太可能找到显著物体。尽管CPFP [52]，TANet [82]，和PCF [50]可以比其他模型生成更准确的显著性图，但显

Table V
在本文的*SIP* 和STERE 数据集上的S测评法↑。↑表示分数越高，模型的性能越好，反之亦然。详见§ VII。

Aspects	Model	*SIP (Ours)*	*STERE* [64]	DES [39]	LFSD [66]	SSD [65]	NJU2K [38]	NLPR [40]
w/o DDU	RgbNet	0.831	0.893	0.881	0.810	0.839	0.888	0.911
	RgbdNet	0.862	0.898	0.896	0.836	0.857	0.898	0.910
	DepthNet	0.862	0.713	0.911	0.724	0.811	0.857	0.864
DDU	Lower Bound	0.822	0.881	0.870	0.788	0.817	0.875	0.897
	D^{3}Net (Ours)	*0.860*	*0.899*	*0.898*	*0.825*	*0.857*	*0.900*	*0.912*
	Upper Bound	0.872	0.910	0.907	0.858	0.879	0.912	0.924

著物体经常会引入明显的不同背景（第三到第五行），或者由于缺乏跨模式学习能力，显著物体的精确细节会丢失（第一行）。相比之下，我们的D^{3}Net可以消除低质量的深度图的影响，并从RGB和深度图像中自适应选择互补信息，以推断出真正的显著物体并突出其细节。

(a) 原图 (b) 模板 (c) 显著物体 (d) 结果

Figure 11. 书籍封面变换的示例。详见§ VI。

VI. 应用

A. 人类活动

如今，手机通常具有深度感应摄像头。使用RGB-D SOD，用户可以更好地实现以下功能：目标提取，散景效果，移动用户识别等。许多监视探头还具有深度传感器，并且RGB-D SOD有助于发现可疑对象。例如，自动驾驶汽车中有一个LiDAR探头，旨在获取深度信息。因此，RGB-D SOD有助于检测基本物体，例如这些车辆中的行人和车牌。大多数工业机器人中也都有深度传感器，因此RGBD-SOD可以帮助他们更好地感知环境并采取某些措施。

B. 背景变换应用

背景变更技术对于艺术设计师来说，利用日益增加的可用图像数据库变得至关重要。传统设计师利用photoshop设计产品。这是一项非常耗时的任务，需要大量的技术知识。绝大多数潜在用户无法掌握艺术设计中的高技能技术。因此，需要一个易于使用的应用程序。

为了克服上述缺点，SOD技术可能是一种潜在的解决方案。以前的类似作品，例如视觉文本应用程序的自动生成 [105], [106]，促使我们为书籍封面版面创建了一个背景变化的应用程序。我们提供了一个原型演示，如Fig. 11所示。首先，用户可以上传图像作为候选设计图像（Fig. 11（a）中的输入图像）。然后，考虑基于内容的图像特征，例如基于RGB-D的显著性图，以便自动生成显著对象。最后，该系统允许我们从专业设计的书籍封面版式库中进行选择（Fig. 11（b）中的模板）。通过结合高级模板约束和低级图像特征，我们获得了背景变化的书的封面（Fig. 11（d）中的结果）。

由于设计一个完整的软件系统不是本文的重点，因此未来的研究人员可按照 [105]设置具有指定主题的视觉背景图像 [106]。在阶段二，根据D^{3}Net模型的预测结果，调整输入图像的大小以匹配目标样式的大小并保留显著区域。

VII. 讨论

基于我们全面的基准测试结果，本文总结了一些最重要问题的结论，这些问题可能会使学术界重新思考用于显著性物体检测任务的RGB-D图像。

A. 消融研究.

现在，我们对本文的基准D^{3}Net模型进行详细分析。为了验证深度图过滤单元（DDU）的有效性，我们得出了两种消融研究：w/o DDU和DDU，指的是不使用DDU和包含DDU的D^{3}Net。对于不带DDU的网络，我们将在D^{3}Net的测试阶段进一步测试三个子网的性能。在Table V中，我们观察到RgbdNet在*SIP*，*STERE*，*DES*，*LFSD*，*SSD*，*NJU2K*数据集上的性能优于RgbNet。这表明跨模态（RGB和深度）对于RGB-D图像的表达学习特别有希望。但是，在大多数情况下，DepthNet的性能低于DepthNet和RgbNet。它表明，仅基于单个模态，模型很难在图像中建立几何结构。

从Table V中，我们还观察到DDU的使用在*STERE*，*DES*，*NJU2K*，和*NLPR*数据集上在一定程度的性能（与RgbdNet相比）提升。我们认为这归因于DDU能够丢弃低质量的深度图并选择一条最佳路径（RgbNet或RgbdNet）。但是，对于SSD数据集，DDU的性能可与单流网络（即RgbdNet）相媲美。值得一提的

Figure 12. 与以前目标级标注的数据集进行比较，它们使用多边形方式进行标注（*NLPR* [40]中的脚垫），粗糙的边界（如, *DES* [39]中的椅子）, 以及部分缺失的对象（如*SSD* [65]中的人）。相反，本文的对象/实例级的 *SIP*数据集的标记具有平滑，精细的边界。而且我们还考虑了遮挡问题（例如，阻挡区域）。

是，D^{3}Net优于任何用于SOD的现有方法，而没有任何通常用于提高得分的后处理技术（例如CRF）。为了了解D^{3}Net的上下界，我们还选择了D^{3}Net的最佳路径（RgbdNet或RgbNet）。例如，对于特定的RGB（I_{rgb}）和深度图（I_{depth}），可以分别评估两个预测图，即S_{rgb}和S_{rgbd}。因此，对于每个输入，我们都知道现有网络中的最佳输出。我们汇总所有最佳和最差结果，并实现D^{3}Net的上限和下限。从Table V中列出的现有结果来看，与上限相关的Table V平均性能差距仍为∼1.6%。

B. 局限性

首先，值得指出的是，与大多数RGB SOD数据集相比，SIP数据集中的图像数量相对较少。建立此数据集的背后目标是探索基于智能手机的应用程序的潜在方向。从基准测试结果和第六节中描述的演示应用程序可以看出，在真实的人类活动场景上进行显著物体检测是一个有前途的方向。我们计划在更具挑战性的、各种前景人物的情况下继续增长数据集。

其次，我们简单的通用框架D^{3}Net由三个子网组成，这可能会增加轻型设备上的内存。在实际环境中，可以考虑采用多种策略来避免这种情况，例如用MobileNet V2 [107]取代骨干网络，数据降维 [108]或使用最近发布的ESPNet V2 [109]模型。

第三，我们给出DDU的上下限。最佳上限是通过将输入RgbdNet或RgbNet来获得的，从而使预测图最佳。如Table V 所示，我们的DDU模块未在当前训练子集上达到最佳上限。因此，仍有机会设计更好的DDU来进一步提高性能。

VIII. 结论

我们通过以下方法对基于RGB-D的SOD进行系统研究：（1）引入新的面向人类的*SIP*数据集，以反映现实的户外移动使用场景；（2）设计一个新颖的D^{3}Net。（3）进行了迄今为止最大规模的基准测试（∼97K）。与现有数据集相比，*SIP*涵盖了真实环境中人类的若干挑战（例如，背景多样性和遮挡）。此外，提出的基准取得了可喜的结果。它是最快的方法之一，使其成为RGB-D SOD的实用解决方案。全面的基准测试结果包括32个总结的SOTA和18个评估的传统/深层楷模。我们希望该基准测试不仅将加速该领域的发展，而且还将加速其他领域的发展（例如，立体估计/匹配 [110]，多个显著人物检测，显著实例检测 [20]，敏感对象检测 [111]和图像分割 [112]）。请注意，在我们的D^{3}Net基线中使用的方法很简单，并且更复杂的组件（例如 [113]中的PDC）或训练策略 [114] 有望提高性能。未来，我们计划将最近提出的技术（例如，加权三元组损失 [115]，分层深度特征 [116]和视觉问题驱动的显著性 [117]）纳入我们的D^{3}Net网络，以进一步提高性能。在提交此论文之后，已经发布了许多有趣的模型，例如UCNet [118]、JL-DCF [119]、GFNet [120]、DMRA [121]、ERNet [122] 和 BiANet [123]。有关更多详细信息，请参考我们的在线排行榜（http://dpfan.net/d3netbenchmark/）。该网站将不断更新。我们预见到这项研究将推动SOD走向具有多个显著人员并通过移动设备（例如智能手机或平板电脑）进行复杂交互的实际应用场景。

REFERENCES

[1] "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE TNNLS*, 2020.

[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, no. 2, pp. 117–150, 2019.

[3] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3127–3135.

[4] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based rgbd image co-segmentation with mutex constraint," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4428–4436.

[5] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE T. Image Process.*, 2019.

[6] 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.

[7] 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.*, 2019.

[8] 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.

[9] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Multi-source weak supervision for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[10] R. Wu, M. Feng, W. Guan, and D. Wang, "A mutual learning method for salient object detection with intertwined multi-supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[11] L. Zhang, i. JZhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[12] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[13] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *AAAI Conference on Artificial Intelligence*, 2018.

[14] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE T. Pattern Anal. Mach. Intell.*, 2018.

[15] Y. Xu, X. Hong, F. Porikli, X. Liu, J. Chen, and G. Zhao, "Saliency integration: An arbitrator model," *IEEE T. Multimedia*, vol. 21, no. 1, pp. 98–113, 2019.

[16] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3089–3098.

[17] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3Net: recurrent residual refinement network for saliency detection," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 684–690.

[18] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, 2019.

[19] D. Tao, J. Cheng, M. Song, and X. Lin, "Manifold ranking-based matrix factorization for saliency detection," *IEEE T. Neur. Net. Lear.*, vol. 27, no. 6, pp. 1122–1134, 2015.

[20] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 247–256.

[21] M. A. Islam, M. Kalash, M. Rochan, N. Bruce, and Y. Wang, "Salient object detection using a context-aware refinement network," in *Brit. Mach. Vis. Conf.*, 2017.

[22] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6609–6617.

[23] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep Image Saliency Computing via Progressive Representation Learning." *IEEE T. Neur. Net. Lear.*, vol. 27, no. 6, pp. 1135–1149, 2016.

[24] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 660–668.

[25] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1265–1274.

[26] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 234–250.

[27] Y. Zhuge, Y. Zeng, and H. Lu, "Deep embedding features for salient object detection," in *AAAI Conference on Artificial Intelligence*, 2019.

[28] J. Su, J. Li, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," *arXiv preprint arXiv:1812.10066*, 2018.

[29] P. Jiang, Z. Pan, N. Vasconcelos, B. Cheng, and J. Peng, "Super diffusion for salient object detection," *arXiv preprint arXiv:1811.09038*, 2018.

[30] Z. Li, C. Lang, Y. Chen, J. Liew, and J. Feng, "Deep reasoning with multi-scale context for salient object detection," *arXiv preprint arXiv:1901.08362*, 2019.

[31] S. Jia and N. D. Bruce, "Richer and deeper supervision network for salient object detection," *arXiv preprint arXiv:1901.02425*, 2019.

[32] X. Huang and Y.-J. Zhang, "300-fps salient object detection via minimum directional contrast," *IEEE T. Image Process.*, vol. 26, no. 9, pp. 4243–4254, 2017.

[33] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 355–370.

[34] M. Kummerer, T. S. A. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Eur. Conf. Comput. Vis.* Springer, 2018.

[35] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns," in *Int. Conf. Comput. Vis.*, 2017.

[36] M. Amirul Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7142–7150.

[37] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *Brit. Mach. Vis. Conf.*, 2013.

[38] 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.

[39] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Int. Conf. Internet Multi. Comput. Serv.*, 2014, p. 23.

[40] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 92–109.

[41] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting Global Priors for RGB-D Saliency Detection," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2015, pp. 25–32.

[42] 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.

[43] 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.

[44] 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.*, 2018.

[45] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE T. Image Process.*, vol. 26, no. 5, pp. 2274–2285, 2017.

[46] 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.

[47] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Int. Conf. Comput. Vis. Worksh.*, 2017.

[48] 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.

[49] 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.

[50] 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.

[51] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE T. Vis. Comput. Gr.*, vol. 23, no. 8, pp. 2014–2027, 2017.

[52] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[53] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE T. Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[54] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan *et al.*, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report (CSTR)*, vol. 2, no. 11, pp. 1–11, 2005.

[55] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, 2011.

[56] 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 Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.

[57] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE T. Circuit Syst. Video Technol.*, 2018.

[58] 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.

[59] 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.

[60] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," in *International Joint Conferences on Artificial Intelligence*, 2018, pp. 698–704.

[61] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 248–255.

[62] N. Wang and X. Gong, "Adaptive Fusion for RGB-D Salient Object Detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[63] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu, "Improving rgbd saliency detection using progressive region classification and saliency fusion," *IEEE Access*, vol. 4, pp. 8987–8994, 2016.

[64] 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.

[65] C. Zhu and G. Li, "A Three-pathway Psychobiological Framework of Salient Object Detection Using Stereoscopic Technology," in *Int. Conf. Comput. Vis. Worksh.*, 2017, pp. 3008–3014.

[66] 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.

[67] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for rgbd images," *IEEE T. Cybern.*, no. 99, pp. 1–14, 2017.

[68] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "Hscs: Hierarchical sparsity based co-saliency detection for rgbd images," *IEEE T. Multimedia*, 2018.

[69] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE T. Image Process.*, vol. 27, no. 2, pp. 568–579, 2018.

[70] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. 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.

[71] 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.

[72] P. Sauer, T. F. Cootes, and C. J. Taylor, "Accurate regression procedures for active appearance models." in *Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[73] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *Brit. Mach. Vis. Conf.*, 2013.

[74] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM T. Intel. Syst. Tec.*, vol. 2, no. 3, p. 27, 2011.

[75] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *IEEE Conf. Dig. Sig. Process.*, 2014, pp. 454–458.

[76] 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.

[77] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 663–667, 2017.

[78] P. Huang, C.-H. Shen, and H.-F. Hsiao, "Rgbd salient object detection using spatially coherent deep learning framework," in *IEEE Conf. Dig. Sig. Process.*, 2018, pp. 1–5.

[79] 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.* Springer, 2014, pp. 345–360.

[80] H. Chen, Y.-F. Li, and D. Su, "Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection," in *IEEE Int. Conf. Intell. Rob. Syst.*, 2018, pp. 6821–6826.

[81] 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 Recogn.*, vol. 86, pp. 376–385, 2019.

[82] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE T. Image Process.*, 2019.

[83] D.-P. Fan, J.-J. Liu, S.-H. Gao, Q. Hou, A. Borji, and M.-M. Cheng, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 1597–1604.

[84] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2432–2439.

[85] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[86] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang, "Joint Salient Object Detection and Existence Prediction," *Front. Comput. Sci.*, 2017.

[87] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5455–5463.

[88] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[89] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Int. Conf. Comput. Vis.*, vol. 2, 2001, pp. 416–423.

[90] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[91] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1155–1162.

[92] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 280–287.

[93] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[94] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann, "The discrimination of visual number," *The American Journal of Psychology*, vol. 62, no. 4, pp. 498–525, 1949.

[95] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6399–6408.

[96] X. Chen, R. Girshick, K. He, and P. Dollár, "Tensormask: A foundation for dense object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2061–2069.

[97] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, "Upsnet: A unified panoptic segmentation network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8818–8826.

[98] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2117–2125.

[99] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

[100] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE T. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[101] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3085–3094.

[102] 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.

[103] D. Tsai, M. Flagg, and J. Rehg, "Motion coherent tracking with multi-label mrf optimization, algorithms," in *Brit. Mach. Vis. Conf.*, 2010.

[104] 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.

[105] X. Yang, T. Mei, Y.-Q. Xu, Y. Rui, and S. Li, "Automatic generation of visual-textual presentation layout," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 2, p. 33, 2016.

[106] A. Jahanian, J. Liu, Q. Lin, D. Tretter, E. O'Brien-Strain, S. C. Lee, N. Lyons, and J. Allebach, "Recommendation system for automatic design of magazine covers," in *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 2013, pp. 95–106.

[107] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4510–4520.

[108] J. Zhang, J. Yu, and D. Tao, "Local deep-feature alignment for unsupervised dimension reduction," *IEEE T. Image Process.*, vol. 27, no. 5, pp. 2420–2432, 2018.

[109] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[110] G.-Y. Nie, M.-M. Cheng, Y. Liu, Z. Liang, D.-P. Fan, Y. Liu, and Y. Wang, "Multi-level context ultra-aggregation for stereo matching," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[111] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE T. Info. Foren. Secur.*, vol. 12, no. 5, pp. 1005–1016, 2016.

[112] J. Shen, X. Dong, J. Peng, X. Jin, L. Shao, and F. Porikli, "Submodular function optimization for motion clustering and image segmentation," *IEEE T. Neur. Net. Lear.*, 2019.

[113] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[114] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct scans," *IEEE T. Med. Imag.*, 2020.

[115] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition," *IEEE T. Neur. Net. Lear.*, 2019.

[116] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE T. Pattern Anal. Mach. Intell.*, 2019.

[117] S. He, C. Han, G. Han, and J. Qin, "Exploring duality in visual question-driven top-down saliency," *IEEE T. Neur. Net. Lear.*, 2019.

[118] 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.

[119] K. F. 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.

[120] Z. Liu, W. Zhang, and P. Zhao, "A Cross-modal Adaptive Gated Fusion Generative Adversarial Network for RGB-D Salient Object Detection," *Neurocomputing*, 2020.

[121] 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.

[122] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and Replace: An Asymmetrical Two-Stream Architecture for Versatile Light Field Saliency Detection," in *AAAI Conference on Artificial Intelligence*, 2020.

[123] Z. Zhang, Z. Lin, J. Xu, W. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *arXiv preprint arXiv:2004.14582*, 2020.