PoolNet：基于池化设计的显著性目标检测

程自强
petecheng@zju.edu.cn
学号: 21821136
浙江大学计算机学院

2019 年 5 月 28 日

摘要

这项发表于 CVPR 2019 [1] 的工作提出了 PoolNet，通过研究如何扩展卷积在卷积神经网络中的作用来解决显着目标检测问题。基于 U 形结构，该模型首先在自下而上的路径上构建全局引导模块（GGM），旨在为不同特征层提供潜在显著对象的位置信息。然后进一步设计了一个特征聚合模块（FAM），使粗粒度语义信息与自上而下的路径中的细粒度特征很好地融合在一起。通过在自上而下路径中的融合操作之后添加 FAM，来自 GGM 的粗粒度特征可以与各种尺度的特征无缝地合并。这两个基于池的模块允许逐步细化高级语义特征，从而产生细节丰富的显著性映射。实验结果表明，提出的 PoolNet 可以更精确地定位具有锐化细节的显著对象，因此与先前的现有技术相比明显改善了目标检测的性能。

关键词— 目标检测，池化，全局引导模块 (GGM)，特征聚合，显著性检测

1 背景和相关工作

1.1 背景

得益于从给定图像中检测最具特色的对象的能力，显著性目标检测在许多计算机视觉任务中起着重要作用，例如视觉跟踪 [2]，内容感知图像编辑 [3] 等。传统方法 [4, 5, 6, 7, 8, 9] 主要依靠手工特征来分别或同时捕捉局部细节和全局背景，但缺乏高级语义信息这一缺点限制了这些模型在复杂场景中检测整体显著目标的能力。

幸运的是，卷积神经网络（CNN）极大地促进了显著对象检测模型的发展，因为它们能够在多尺度空间中提取高级语义信息和低级细节特征。正如许多以前的方法 [10, 11, 12] 所指出的，由于 CNN 的金字塔状结构特征，较浅的阶段通常具有较大的空间大小并保持丰富、详细的低级信息，而较深的阶段包含更多的高级别语义知识，更好地定位显著目标的准确位置。基于上述知识，[10, 13, 14] 设计了各种用于显著目标检测的新架构。在这些方法中，基于 U 形的结构 [15, 16] 受到最多的关注，因为它们能够通过在分类网络上构建自上而下的路径来构建丰富的特征图。

尽管这一类方法取得了良好的性能，但仍有很大的改进空间。首先，在 U 形结构中，高级语义信息逐渐传输到较浅层，因此可以逐渐稀释由较深层捕获的位置信息。其次，正如 [17] 所指出的，CNN 的感知视野大小与其层深度不成比例。现有方法通过将注意力机制 [18, 19] 引入 U 形结构，以循环方式细化特征图 [20, 21, 18]，结合多尺度特征信息来解决上述问题 [10, 11, 12]，或者在显著映射中添加额外约束，如 [11] 中的边界损失项。

而和以上提及的方法不同的是，PoolNet 通过探究池化技术在基于 U 形结构的架构中能够起到的不同作用来解决显著性目标检测问题。这应该是目前第一篇旨在研究如何设计各种基于池的模块以帮助提高显著目标检测性能的论文。作为这项工作的延伸，作者还为该模型设计了边缘检测分支，通过联合训练模型和边缘检测，进一步锐化显著目标的细节。为了评估本文提出的方法的性能，作者报告了多个流行的显著目标检测基准的结果。可以看到，PoolNet 在很大程度上超越了所有以前最先进的方法。此外，作者进行了一系列消融实验，更好地展示了 PoolNet 架构中每个组件对性能的影响，并阐述了如何通过边缘检测进行联合训练有助于增强预测结果的细节。

1.2 相关工作

最近，受益于 CNN 强大的特征提取能力，大多数基于手工特征的传统显著目标性检测方法 [5, 7, 9, 22] 已逐渐被超越。[23] 使用从 CNN 提取的多尺度特征来计算每个超像素的显著性值。[24] 采用了两个 CNN，旨在将局部超像素估计和全局建议搜索结合起来，以产生显著性映射。[25] 提出了一个多语境深度学习框架，通过使用两个独立的 CNN 提取本地和全球背景信息。[26] 结合了低级启发式特征，如颜色直方图和 Gabor 响应，以及从 CNN 中提取的高级特征。所有这些方法都将图像补丁作为 CNN 的输入，因此非常耗时。而且，它们忽略了整个输入图像的基本空间信息。

为了克服上述问题，更多的研究注意力被用于预测像素显著图，这是由完全卷积网络 [27] 引起的。[21] 使用低级别线索生成显著性先验地图，并进一步利用它来指导反复预测显著性。[20] 提出了一个两阶段网络，它首先生成粗略显著图，然后整合本地上下文信息，以便循环和分层地改进它们。[10] 将短连接引入多尺度侧输出以捕获精细细节。[11] 和 [12] 两者都推进了 U 形结构，并利用多层次的背景信息来准确检测显著物体。[18, 28] 将注意力机制与 U 形模型相结合，以指导特征整合过程。[14] 提出了一个网络，以反复定位显著对象，然后用本地环境信息改进它们。[29] 使用双向结构在 CNN 提取的多级特征之间传递消息，以更好地预测显著性映射。[30] 采用一个网络首先确定分散注意力的区域，然后使用另一个网络进行显著性检测。

2 模型和算法

PoolNet 主要通过两个模块，即全局引导模块 (Global Guidance Model, GGM) 和特征金字塔网络 (Feature Pyramid Networks, FPNs)，来实现对显著对象的特征挖掘，其整体模型架构如图-1所示：

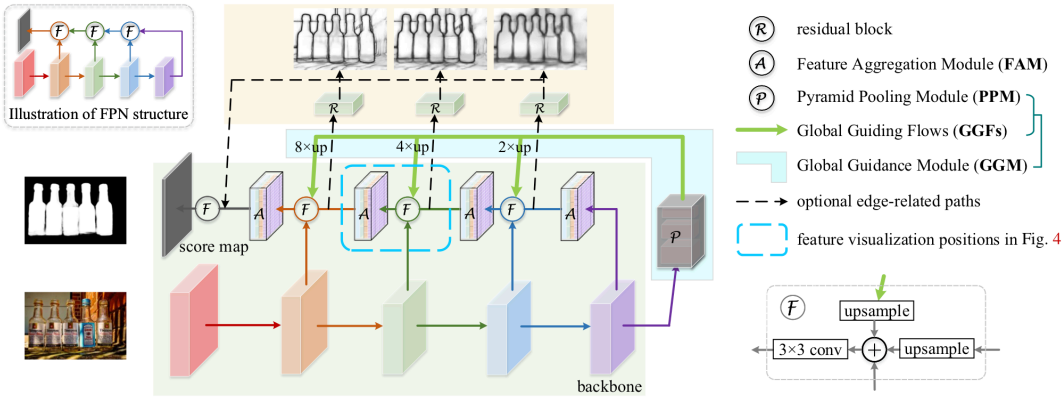

图 1: PoolNet 模型架构示意图

2.1 全局引导模块 GGM

以往的工作指出，高层语义特征对挖掘显著对象的详细位置是很有帮助的，但是中低层的语义特征也可以提供必要的细节。因为在自上而下的过程中，高层语义信息被稀释，而且实际上的感知视野也是小于理论感知视野的，所以对于全局信息的捕捉十分缺乏，导致显著物体被背景吞噬。

因此作者提出了 GGM 模块，GGM 其实是 PPM（Pyramid Pooling module，金字塔池化模块）的改进，并且加上了一系列的 GGFs（Global Guiding Flows，全局引导流）。这样做的好处是，在特征图上的每层都能关注到显著物体；另外与 PPM 不同的是，GGM 是一个独立的模块，而 PPM 是 U 型架构里在基础网络（backbone）中参与引导全局信息的过程。

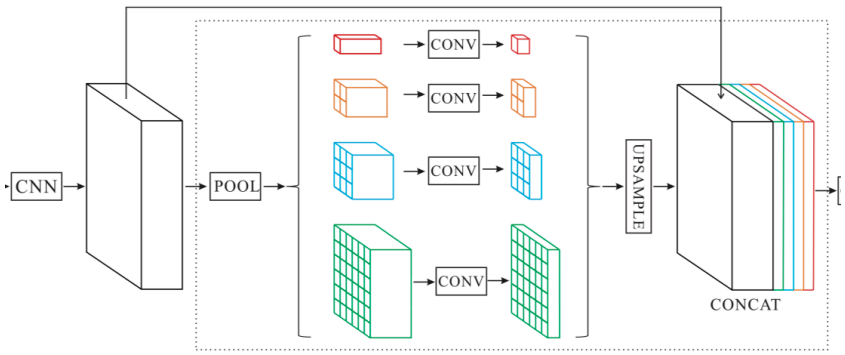

图 2: PPM 架构示意图

图-2展示了金字塔池化模块 [17] 的结构：该 PPM 模块融合了四种不同金字塔尺度的特征，第一行红色是最粗糙的特征–全局池化生成单个 bin 输出，后面三行是不同尺度的池化特征。为了保证全局特征的权重，如果金字塔共有 N 个级别，则在每个级别后使用 1×1 的卷积将对于级别通道降为原本的 $\frac{1}{N}$；再通过双线性插值获得未池化前的大小，最终将这些特征拼接到一起。

GGM 则是在 PPM 结构上的改进。PPM 是对每个特征图都进行了金字塔池化，所以作者说是嵌入在 U 型结构中的，但是这里 PoolNet 中加入了全局引导流 GGFs，即图-1中绿色的箭头，这一步骤引入了

对每级特征的不同程度的上采样映射（identity mapping），可以看作是个独立的模块。简单地说，PoolNet 想要 FPNs 在自上而下的路径上不被稀释语义特征，所以在每次横向连接的时候都加入高层的语义信息，这样做也是一个十分直接主观的想法。

2.2 特征整合模块 (Feature Aggregate Model, FAM)

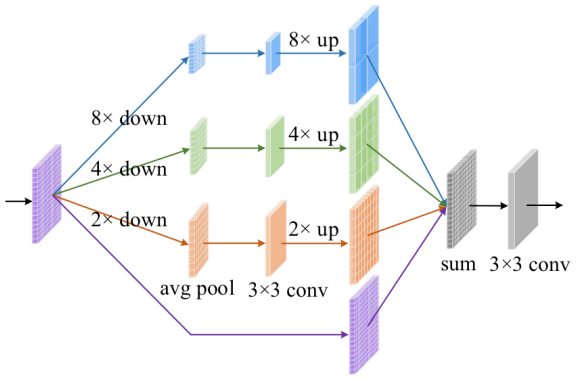

图 3: 特征整合模块 FAM 示意图

特征整合模块这里也充分利用了池化技术。如图-3所示。首先把 GGM 得到的高层语义与该级特征分别上采样之后横向连接得到 FAM 的输入，然后采取的操作是先把拼接得到的特征用 2,4,8 的三种下采样得到下采样的特征图（分别对应图-3中红、绿、蓝三个分支），然后利用平均池化再上采样回原来尺寸，最后将蓝绿红紫（紫色分支是 FAM 的原始输入）四个分支像素相加得到整合后的特征图。

具体来说，FAM 有以下两个优点：

- 帮助模型降低上采样（upsample）导致的混叠效应（aliasing）。混叠效应其实相当于引入杂质，GGFs 从基础网络最后得到的特征图经过金字塔池化之后需要最高是 8 倍上采样才能与前面的特征图融合，而这样高倍数的采样确实容易引入杂质。
- 从不同的多角度的尺度上纵观显著物体的空间位置，放大整个网络的感知视野。

因此 PoolNet 提出 FAM 进行特征整合，先把特征用不同倍数的下采样、池化之后，再用不同倍数的上采样，最后叠加在一起。因为单个高倍数上采样容易导致失真，所以补救措施就是高倍数上采样之后，再下采样，再池化上采样，最后平均下来以达到降低采样引入杂质的影响，并且多角度捕捉显著目标的空间位置。

3 实验结果

3.1 实验设定

为了评估提出的框架 PoolNet 的性能，作者对 6 个常用数据集进行了实验，包括 ECSSD [31]，PASCAL-S [32]，DUT-OMRON [33]，HKU-IS [23]，SOD [34] 和 DUTS [35]。作者使用标准二元交叉熵损失进行显著物体检测，并使用平衡二元交叉熵损失进行边缘检测。。

为了评估实验结果，作者使用了准确率-召回率曲线 (PR curves)，F 得分 (式-(3.1.1)) 和平均绝对误差 (MAE, 式-(3.1.1))

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{3.1.1}$$

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)| \tag{3.1.2}$$

在式-(3.1.1) 中，S 为显著性映射 (saliency map)，G 为真实值 (ground truth)。

	Training		ECSSD [41]		PASCAL-S [21]		DUT-O [42]		HKU-IS [18]		SOD [30]		DUTS-TE [35]	
Model	#Images	Dataset	MaxF ↑	MAE ↓	MaxF ↑	MAE ↓	MaxF ↑	MAE ↓	MaxF ↑	MAE ↓	MaxF ↑	MAE ↓	MaxF ↑	MAE ↓
VGG-16 backbone														
DCL [19]	2,500	MB	0.896	0.080	0.805	0.115	0.733	0.094	0.893	0.063	0.831	0.131	0.786	0.081
RFCN [36]	10,000	MK	0.898	0.097	0.827	0.118	0.747	0.094	0.895	0.079	0.805	0.161	0.786	0.090
DHS [23]	9,500	MK+DTO	0.905	0.062	0.825	0.092	-	-	0.892	0.052	0.823	0.128	0.815	0.065
MSR [17]	5,000	MB + H	0.903	0.059	0.839	0.083	0.790	0.073	0.907	0.043	0.841	0.111	0.824	0.062
DSS [9]	2,500	MB	0.906	0.064	0.821	0.101	0.760	0.074	0.900	0.050	0.834	0.125	0.813	0.065
NLDF [28]	3,000	MB	0.903	0.065	0.822	0.098	0.753	0.079	0.902	0.048	0.837	0.123	0.816	0.065
UCF [45]	10,000	MK	0.908	0.080	0.820	0.127	0.735	0.131	0.888	0.073	0.798	0.164	0.771	0.116
Amulet [44]	10,000	MK	0.911	0.062	0.826	0.092	0.737	0.083	0.889	0.052	0.799	0.146	0.773	0.075
GearNet[10]	5,000	MB + H	0.923	0.055	-	-	0.790	0.068	0.934	0.034	0.853	0.117	-	-
PAGR [46]	10,553	DTS	0.924	0.064	0.847	0.089	0.771	0.071	0.919	0.047	-	-	0.854	0.055
PiCANet [24]	10,553	DTS	0.930	0.049	0.858	0.078	0.815	0.067	0.921	0.042	0.863	**0.102**	0.855	0.053
PoolNet (Ours)	2,500	MB	0.918	0.057	0.828	0.098	0.783	0.065	0.908	0.044	0.846	0.124	0.819	0.062
PoolNet (Ours)	5,000	MB + H	0.930	0.053	0.838	0.093	0.806	0.063	**0.936**	**0.032**	0.861	0.118	0.855	0.053
PoolNet (Ours)	10,553	DTS	0.936	0.047	0.857	0.078	0.817	0.058	0.928	0.035	0.859	0.115	0.876	0.043
PoolNet† (Ours)	10,553	DTS	**0.937**	**0.044**	**0.865**	**0.072**	**0.821**	**0.056**	0.931	0.033	**0.866**	0.105	**0.880**	**0.041**
ResNet-50 backbone														
SRM [37]	10,553	DTS	0.916	0.056	0.838	0.084	0.769	0.069	0.906	0.046	0.840	0.126	0.826	0.058
DGRL [38]	10,553	DTS	0.921	0.043	0.844	0.072	0.774	0.062	0.910	0.036	0.843	0.103	0.828	0.049
PiCANet [24]	10,553	DTS	0.932	0.048	0.864	0.075	0.820	0.064	0.920	0.044	0.861	0.103	0.863	0.050
PoolNet (Ours)	10,553	DTS	0.940	0.042	0.863	0.075	0.830	0.055	0.934	0.032	0.867	**0.100**	0.886	0.040
PoolNet† (Ours)	10,553	DTS	**0.945**	**0.038**	**0.880**	**0.065**	**0.833**	**0.053**	**0.935**	**0.030**	**0.882**	0.102	**0.892**	**0.036**

MB: MSRA-B [25], MK: MSRA10K [3], DTO: DUT-OMRON [42], H: HKU-IS [18], DTS: DUTS-TR [35].

图 4: 在 6 个常用数据集上的显著性目标检测实验结果。

目标检测实验结果如图-4所示，可以看到 PoolNet 在所有数据集上几乎都达到了最好的性能，而图-5从 PR 曲线的角度同样印证了这一点。

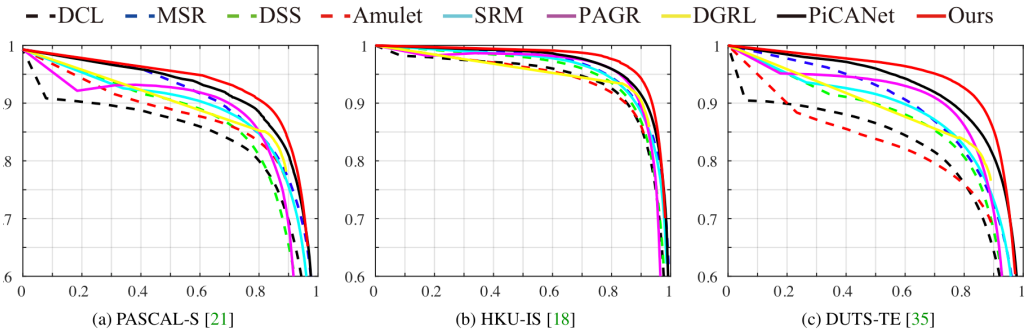

图 5: 在三个数据集上的 PR 曲线。

此外，图-6还对模型的各个组成模块进行消融实验，验证了各模块对目标检测效果的影响。可以看到，组合了 GGM 和 FAM 的模型在检测表现上是最好的。最后，作者还给出了显著目标检测的可视化结果图-7。

No.	GGM + FAMs			DUT-O [42]		SOD [30]	
	PPM	GGFs	FAMs	MaxF ↑	MAE ↓	MaxF ↑	MAE ↓
1				0.770	0.076	0.838	0.124
2	✓			0.783	0.071	0.847	0.125
3		✓		0.772	0.076	0.843	0.121
4	✓	✓		0.790	0.069	0.855	0.120
5			✓	0.798	0.065	0.852	0.118
6	✓	✓	✓	**0.806**	**0.063**	**0.861**	**0.117**

(a) 不同模块的消融实验。

Settings	PASCAL-S [21]		DUT-O [42]		SOD [30]	
	MaxF	MAE	MaxF	MAE	MaxF	MAE
Baseline (B)	0.838	0.093	0.806	0.063	0.861	0.117
B + SalEdge	0.835	0.096	0.805	0.063	0.863	0.120
B + StdEdge	**0.849**	**0.077**	**0.808**	**0.059**	**0.872**	**0.105**

(b) 平衡特征方法的影响。

图 6: 模型不同部分的对比实验（一）。

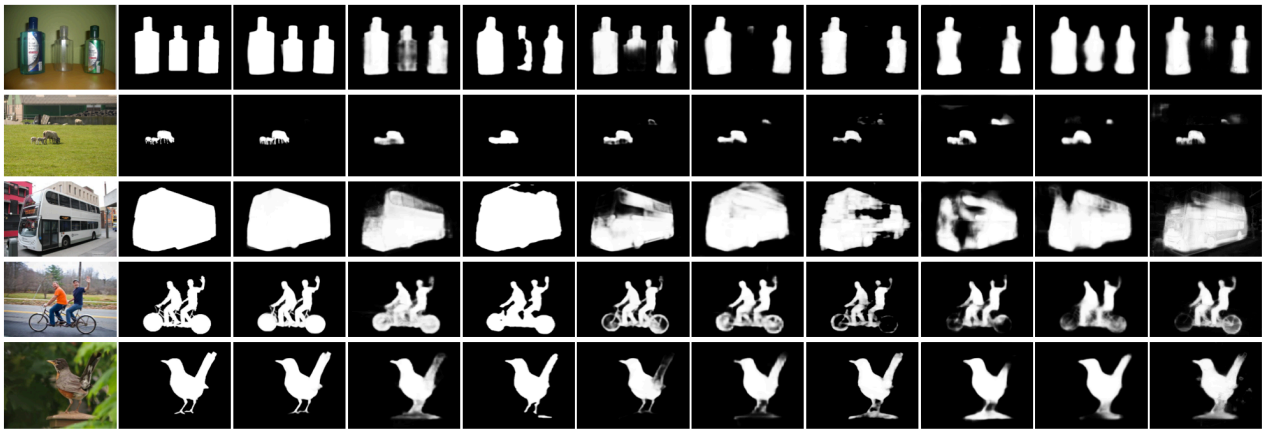

图 7: 显著目标检测可视化

4 结论

本文提出了两种基于池化技术的模块 GGM（全局引导模块）和 FAM（特征整合模块），改进 FPN 在显著性检测的应用，而且这两个模块也能应用在其他金字塔模型中，具有普遍性，并且能够锐化显著物体细节，检测速度能够达到 30FPS。总体上说是一个优秀的工作。但是 FAM 的整合过程带来的损失可能代价太大，也许能够寻找更为优雅的解决办法。本文主要参考 [1] 和 PaperWeekly 博客[1]。

[1] https://www.jiqizhixin.com/articles/2019-05-27-12

参考文献

[1] Jiang-Jiang Liu et al. “A Simple Pooling-Based Design for Real-Time Salient Object Detection”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[2] Seunghoon Hong et al. “Online tracking by learning discriminative saliency map with convolutional neural network”. In: *International conference on machine learning*. 2015, pp. 597–606.

[3] Ming-Ming Cheng et al. “Repfinder: finding approximately repeated scene elements for image editing”. In: *ACM Transactions on Graphics (TOG)*. Vol. 29. 4. ACM. 2010, p. 83.

[4] Ali Borji and Laurent Itti. “Exploiting local and global patch rarities for saliency detection”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 478–485.

[5] Ming-Ming Cheng et al. “Global contrast based salient region detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2014), pp. 569–582.

[6] Laurent Itti, Christof Koch, and Ernst Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 11 (1998), pp. 1254–1259.

[7] Huaizu Jiang et al. “Salient object detection: A discriminative regional feature integration approach”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 2083–2090.

[8] Dominik A Klein and Simone Frintrop. “Center-surround divergence of feature statistics for salient object detection”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2214–2219.

[9] Federico Perazzi et al. “Saliency filters: Contrast based filtering for salient region detection”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 733–740.

[10] Qibin Hou et al. “Deeply supervised salient object detection with short connections”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3203–3212.

[11] Zhiming Luo et al. “Non-local deep features for salient object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6609–6617.

[12] Pingping Zhang et al. “Amulet: Aggregating multi-level convolutional features for salient object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 202–211.

[13] Guanbin Li et al. “Instance-level salient object segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2386–2395.

[14] Tiantian Wang et al. “Detect globally, refine locally: A novel approach to saliency detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3127–3135.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[16] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2117–2125.

[17] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2881–2890.

[18] Xiaoning Zhang et al. “Progressive attention guided recurrent network for salient object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 714–722.

[19] Tie Liu et al. “Learning to detect a salient object”. In: *IEEE Transactions on Pattern analysis and machine intelligence* 33.2 (2010), pp. 353–367.

[20] Nian Liu and Junwei Han. “Dhsnet: Deep hierarchical saliency network for salient object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 678–686.

[21] Linzhao Wang et al. “Saliency detection with recurrent fully convolutional networks”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 825–841.

[22] Xiaohui Li et al. “Saliency detection via dense and sparse reconstruction”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2976–2983.

[23] Guanbin Li and Yizhou Yu. “Visual saliency based on multiscale deep features”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5455–5463.

[24] Lijun Wang et al. “Deep networks for saliency detection via local estimation and global search”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3183–3192.

[25] Rui Zhao et al. “Saliency detection by multi-context deep learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1265–1274.

[26] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. “Deep saliency with encoded low level distance map and high level features”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 660–668.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

[28] Nian Liu, Junwei Han, and Ming-Hsuan Yang. “PiCANet: Learning pixel-wise contextual attention for saliency detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3089–3098.

[29] Lu Zhang et al. “A bi-directional message passing model for salient object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1741–1750.

[30] Huaxin Xiao et al. “Deep salient object detection with dense connections and distraction diagnosis”. In: *IEEE Transactions on Multimedia* 20.12 (2018), pp. 3239–3251.

[31] Qiong Yan et al. “Hierarchical saliency detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 1155–1162.

[32] Yin Li et al. “The secrets of salient object segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 280–287.

[33] Chuan Yang et al. “Saliency detection via graph-based manifold ranking”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3166–3173.

[34] Vida Movahedi and James H Elder. “Design and perceptual validation of performance measures for salient object segmentation”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE. 2010, pp. 49–56.

[35] Lijun Wang et al. “Learning to detect salient objects with image-level supervision”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 136–145.