A Simple Pooling-Based Design for Real-Time Salient Object Detection

Jiang-Jiang Liu[1*] Qibin Hou[1*] Ming-Ming Cheng[1 †] Jiashi Feng[2] Jianmin Jiang[3]
[1]TKLNDST, College of CS, Nankai University [2]NUS [3]Shenzhen University
{j04.liu, andrewhoux}@gmail.com

Abstract

通过研究如何拓展池化层在卷积神经网络中的作用，我们解决了显著目标检测的问题。基于 *U* 型的网络结构，我们首先在自顶向下的路径上构建了一个全局引导模块（*GGM*），旨在为不同级别的特征级别提供潜在的显著物体的位置。我们进一步设计了特征聚合模块（*FAM*）从而使粗糙的语义信息与源于自顶向下通路的精细特征很好的融合在一起。通过在自顶向下的路径中融合操作后增加多个 *FAM*，从 *GGM* 得到的粗糙的特征可以与各个尺度的特征无缝融合。这两个以池化操作为基础的模块允许高级的语义特征可以被逐步细化，产生细节丰富的显著性图。实验结果显示，相比与以往先进的算法，我们提出的方法可以更加精准地定位显著物体，从而极大地提升了性能。我们的方法也很快，当处理 300×400 尺寸的图像时，能以 *30* 帧以上的速度运行。代码可以在*http://mmcheng.net/poolnet/*上找到。

1. 简介

得益于从给定图片检测最有视觉辨识度物体的能力，显著性物体检测在诸如视觉追踪 [8]、内容感知图像编辑 [4]与机器人导航 [5]等计算机视觉任务中起重要作用。传统方法 [11, 25, 14, 31, 2, 12, 39, 3]大多依赖人工选出的特征来分别地或同时捕捉局部细节与全局上下文，但高层次语义信息的缺失限制了这些方法在复杂场景中检测显著性物体的能力。幸运地是，卷积神经网络（CNNs）由于其在多种尺度空间下提取高层次语义信息与低层次细节特征的能力，极大地促进了显著性物体检测模型的发展。

正如许多前期方法指出 [9, 28, 42]，由于 CNNs 类似金字塔的结构特征，浅层通常有更大的空间尺寸并且可以保留丰富的、细节的低层次信息，而深层包含更多高层次语义知识并在定位显著性物体的确切位置时效果更好。基于上述知识，多种用于显著性物体检测的新构架被设计出来。在这些方法 [9, 17, 37, 10]中，U 型结构 [32, 22] ，由于其通过在分类网络中创建自顶向下通道的方式构建出丰富特征图的能力，收获了最多的注意力。

这类方法尽管获得了较好效果，其中仍有较大改进空间。首先，在 U 型结构中，高层次语义信息逐步传输到浅层，因此被深层捕捉到的定位信息与此同时被逐渐稀释。其次，如 [45]中指出的，CNN 的感受野大小与其层深不成比例。现有方法或是通过向 U 型网络中引入注意力机制 [44, 24]，或是以循环的方式 [23, 44, 35]精炼特征图，或是结合不同尺度的特征信息 [9, 28, 42, 10]，或是向显著性图增加诸如 [28]中边界损失项之类的额外约束，解决上述问题。

在这篇论文中，不同于上述提到的方法，我们研究了如何通过在 U 型网络结构中拓展池化技术来解决这个问题。总体来说，我们的模型以特征金字塔网络（FPNs）为基础，由 2 个主要模块组成。特征金字塔网络即全局引导模块（GGM）与特征聚合模块（FAM）[22]。如图. 1所述，GGM 包含金字塔池化模块（PPM）的一个修改版本和一系列全局引导流（GGFs）。不同

*表示相同贡献。
†M.M. Cheng (cmm@nankai.edu.cn) 是文章的通讯作者。

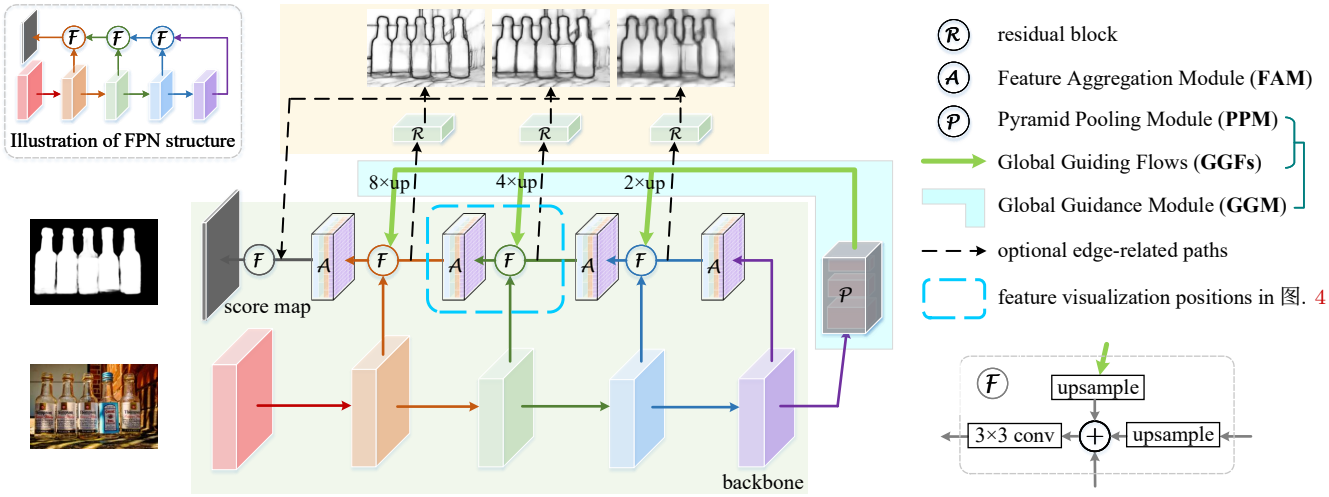


图 1. 我们提出方法整体的流程。为了清楚起见，我们在左上角放置了一个标准 U 型 FPN 结构 [22]。最上边的边缘检测部分是可选的。

于 [36]中直接将 PPM 插入 U 型网络，GGM 是独立的模块。具体而言，PPM 被放置于主干网的顶部，以获取全局引导信息（显著性物体所在位置）。通过引入 GGFs，被 PPM 收集到的高层次语义信息可以被传递到金字塔所有层次的特征图中，纠正了 U 型网络自顶向下信号逐渐被稀释的问题。考虑到从 GGFs 中粗糙特征图与金字塔不同尺度特征图的融合问题，我们进一步提出了特征聚合模块（FAM），其将融合后的特征图作为输入。这个模块首先将融合的特征图转换到多个特征空间，以捕捉不同尺度下局部上下文信息；之后，将融合后的信息进行组合，以更好地对融合后的输入特征图的组成部分进行加权。

由于上述模块均基于池化技术，我们称我们的方法为 PoolNet。据我们所知，这是第一篇旨在研究如何设计多样的基于池化模块以协助提升显著性物体检测性能的文章。作为我们工作的拓展，我们还为我们的架构配置了一个边缘检测分支，通过联合训练我们的模型与边缘检测来进一步锐化显著物体的细节。为了评估我们提出方法的性能，我们汇报了多个流行的显著性物体检测基准的结果。在不附加其他技巧的情况下，PoolNet 在很大程度上超越了所有现有先进方法。此外，我们进行了一系列消融实验以让读者更好的理解在网络中各个组件对性能的影响并且展示了如何与边缘检测联合训练如何有益于增强预测结果的细节。

我们的网络能够在 NVIDIA Titan Xp Gpu 单卡上以 30 帧的速度处理 300×400 大小的图片。当不引入边缘分支时，在 5000 张图像的训练集上训练需要的时间少于 6 小时，这比大多数以往的方法 [24, 41, 28, 42, 43, 9]都快。这主要因为池化技术的有效利用。

2. 相关工作

近来，得益于 CNNs 强大的特征提取能力，大多数基于手工提取特征的传统显著性检测方法 [3, 12, 20, 31]被逐步超越。Li 等人使用从 CNN 提取的多尺度特征来计算每个超像素的显著性值。Wang 等人 [18]提出了一种多内容深度学习框架，其通过运用 2 个独立的 CNNs 来提取局部和全局信息。Zhao 等人 [46]提供了一种多上下文深度学习框架，其可以用 2 个独立的 CNN 分别提取局部和全局的信息。Lee 等人 [6]将诸如色彩直方图与 Gabor 响应的低层次启发性特征与 CNNs 提取的高层次特征结合。所有这些方法将一批图像块作为 CNNs 的输入并因此十分耗时。另外，他们忽略了整个输入图像中至关重要的空间信息。

为了克服上述问题，在全卷积网络 [27]的启发下，对像素级显著性图的预测越来越受研究者关注。Wang 等人 [35]使用低层次线索生成显著性优先图并进一步利用其以指导显著性的循环预测。Liu 等人 [23]提出了一种两阶段网络，其首先产生粗略的显著性图，然后整合局部上下文信息来反复地层次化地改善显著性图。Hou 等人 [9]向多尺度侧输出引入短连接以捕捉细节。Luo 等人 [28]和 Zhang 等人 [42]都发展了 U 型网络结构并利用多层级的上下文信息来精准检测显著性物体。

Zhang 等人 [44]和 Liu 等人 [24]将注意力机制与 U 型网络结合以引导特征整合过程。Wang [37]等人提出了一种网络来循环地定位显著性物体并且用局部上下文信息改进结果。Zhang 等人 [41]使用了一种双向结构来传递 CNN 提取出的不同层级的信息以更好的预测显著性图。Xiao 等人 [38]采用了一个网络先订制分散区域，然后另一个网络进行显著性检测的方法。

我们的方法与以往方法大不相同。我们并未探索新的网络结构，而是研究如何将简单的池化技术应用到 CNNs 中，以同时提升性能且加快运行速度。

3. PoolNet

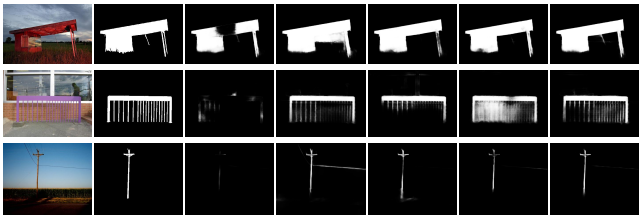
文献 [23, 9, 36, 37] 指出，高级语义特征有助于发现显著物体的具体位置。与此同时，低级和中级特征对于提升网络深层提取出特征同样很重要。基于上述知识，在这一节中，我们提出 2 种互补模型，它们可以精确地捕捉显著性物体的确切位置并同时细化细节。

3.1. 整体流程

我们基于特征金字塔网络（feature pyramid networks，FPNs） [22]建立我们的模型。FPNs 是一种经典的 U 型构架，采用自底向上与自顶向下的方式设计，如图. 1左上角所示。由于源自于分类网络的结合多层次特征的强大能力，这种结构在许多视觉任务（包括显著性物体检测） [7, 33]中被广泛采用。如图. 1所示，我们引入了全局引导模块（GGM），其构造在自底向上通道的顶端。通过将 GGM 提取出的高层次信息聚合到每个特征层级的特征图中，我们的目标是明确地注意到显著对象所在的不同特征层级。在源自 GGM 的引导信息与不同层级的特征融合后，我们进一步引入特征聚合模块以在不同尺度的特征图可以无缝地融合。在下文中，我们描述了上述 2 种模块的结构并详细解释了它们的功能。

3.2. 全局引导模块

FPNs 提供了一个经典的体系结构来组合源自分类主干的多层次特征。然而，由于自顶向下的通道构建于自底向上的主干中，这类 U 型架构的其中一个问题是高层次信息会在它们向低层传递中被逐步稀释。如 [47, 45]中所示，特别是在深层中，CNNs 的实际感受野比理论上要小很多，所以整个网络的感受野并不足



(a) (b) (c) (d) (e) (f) (g)

图 2. 我们提出的 GGM 和 FAM 的不同组合对显著性物体检测的可视化对比。(a) 原图像；(b) 实际结果；(c) FPN 的基础结果；(d) FPN+FAMs 的结果；(e) FPN+PPM 的结果；(f) FPN+GGM 的结果；(g) FPN+GGM+FAMs 的结果。

以捕捉输入图像的全局信息。如图. 2c 所示，最直接的影响是只有显著物体的一部分能被探测到。鉴于自顶向下路径中精细特征图中高层次语义信息的缺乏的问题，我们引入了一个全局引导模块，该模块包含了金字塔池化模块（PPM） [45, 36]和一系列全局引导流（GGFs）来明确地使每一层特征图都知道显著性物体的位置。

具体来说，在 GGM 中的 PPM 由 4 个用于捕捉输入图像语义信息的副分支组成。第一和最后的副分支分别是一个恒等映射层和一个全局平均池化层。对于中间的 2 个副分支，我们采用了自适应平均池化层[1]以保证他们的输出特征图分别具有 3×3 与 5×5 的空间尺寸。给定 PPM，我们现在需要做的是如何保证 PPM 产生的引导信息可以被在自顶向下的通道中合理地融入不同层级的特征图。

不同于以往工作 [36]简单地将 PPM 看做 U 型网络的一部分，我们的 GGM 与 U 型网络独立。通过引入一系列全局引导流（恒等映射），高层次语义信息可以被轻松地传递到不同层级的特征图中（见图. 1中绿色箭头）。以这种方式，我们在每个自顶向下的通道中明显地增加了全局引导信息权重以保证定位信息不会再构建 FPNs 时被稀释。

为了更好地阐释 GMM 的有效性，我们展示了一些可视化对比。如图. 2c 中所述，我们展示了一些 VGGNet 版本的 FPNs 生成的显著性图[2]。易见，对于一些复杂场景，仅有 FPN 骨干的情况下很难定位显著性物体。

[1] https://pytorch.org/docs/stable/nn.html#adaptiveavgpool2d

[2] 类似与 [22]，我们使用 conv2, conv3, conv4, conv5 的输入，并标记为 $\{C_2, C_3, C_4, C_5\}$，来建立 VGGNet [33]上的特征金字塔。通道数分别为 $\{128, 256, 512, 512\}$。

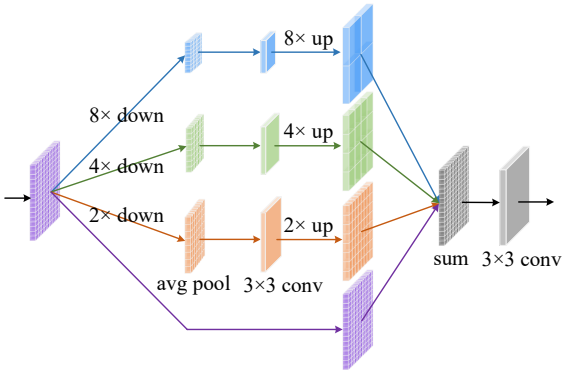


图 3. 特征聚合模块（FAM）的细节示意图。其包含 4 个副分支，每个都工作在独立的尺度空间中。在上采样后，所有副分支均被组合并输入卷积层。

也有一些结果只检测到显著性物体的一部分。然而，当并入 GMM 后，显著性图的质量显著提高。如图. 2f 所示，显著性物体可以被准确地检测到，这说明了 GMM 的重要性。

3.3. 特征聚合模块

GGM 的使用允许将全局引导信息在金字塔的不同层级运输到特征图中。然而，如何使源自 GGM 的粗略特征图在金字塔的不同尺度无缝融入特征图是一个值得考虑的问题。以 VGGNet 版本的 FPNs 为例，在金字塔中对应的特征图为 $C = C_2, C_3, C_4, C_5$ 相对输入的下采样比率分别为 $2, 4, 8, 16$。FPNs 初始的自顶向下的通道中，较粗糙分辨率的特征图被 2 倍上采样。因此，在融合操作后增加一个卷积核为 3×3 的卷积层能够有效地减少上采样的锯齿效应。然而，GGFs 需要更大的上采样率（例如：8 倍）。这对有效并高效地弥合 GGFs 与不同尺度的特征图间的差距是至关重要的。

为了这个目的，我们提出了一系列特征聚合模块，如图. 3所示，其中每个都包含 4 个副分支。在前向传播时，输入特征图通过将其输入不同下采样率的平均池化层最先被转换为不同尺度空间。从不同副分支的上采样特征图被融合到一起，接着是一个 3×3 卷积层。

总的来说，FAM 有 2 大优势。首先，其有助于模型减小上采样的锯齿效应，尤其在上采样率较大时（例如：8 倍）。并且，其允许每个空间位置查看不同尺度空间下的局部上下文，进一步增大整个网络空间的感受野。据我们所知，这是首个揭示 FAM 有助于减少上采样的锯齿效应的工作。为了证实我们提出的 FAMs 的有效性，我们在图. 4中具象化了接近 FAMs 的特征图。

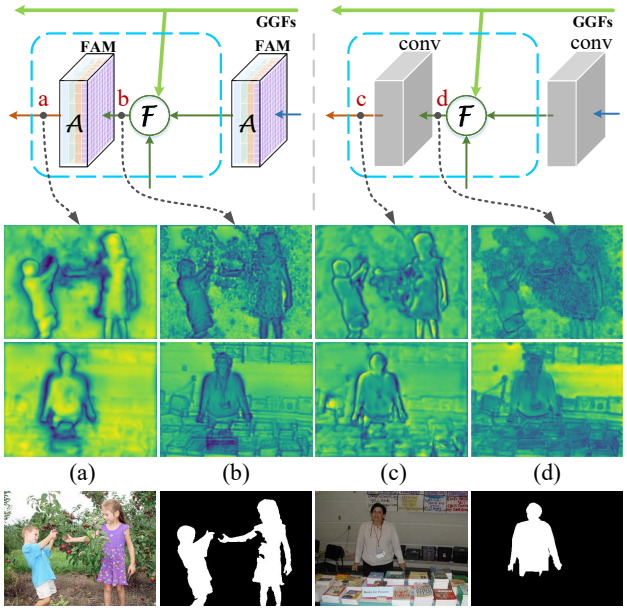


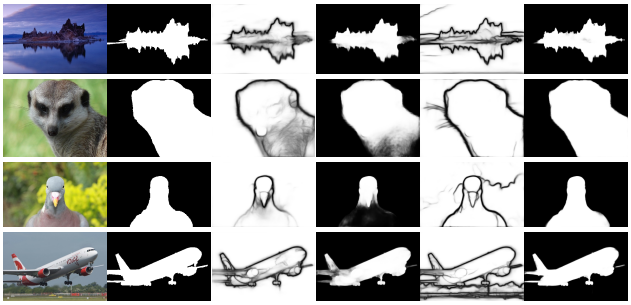
图 4. FAMs 周围特征图的可视化。左边展示的特征图源于包含 FAMs 的模型，右边展示的特征图源于用 2 个卷积层代替 FAMs 的模型。最后一行为原图与对应的实际标注。(a-d) 为不同位置特征图的可视化。可见，当使用 FAMs 时，相较于经过 2 个卷积层的特征图（c 列），FAMs 后的特征图更精准地捕捉了显著性物体的位置和细节信息（a 列）。

通过比较左边（w/ FAMs）与右边（w/o FAMs），FAM 后的特征图（a 列）可以比没有 FAMs 的特征图（c 列）更好地捕捉显著性物体。除了可视化中间特征图，我们也在图. 2中展示了一些不同设定下模型生成的显著性图。通过对比 f 列（w/o FAMs）中与 g 列（w/ FAMs）中的结果，显然多次引入 FAM 使我们的网络能更好地锐化显著性物体的细节。通过观察图. 2的第 2 行，这一现象尤其明显。所有上述讨论证实了 FAMs 在更好地融合不同尺度的特征图上的重大影响。在我们的实验部分，我们将给出更多数值结果。

4. 边缘检测联合训练

第3节中描述的构架已经超越了全部以往 SOTA 的单模型在许多热门显著性检测基准测试上的结果。尽管如此，通过观察我们模型产生的显著性图，我们发现许多由不清晰物体边界导致的不准确预测（不完全或过度预测）。

首先，我们通过在第3节中展示的构架上增加额外的预测分支以推测显著性物体的边界来解决这个问题。



(a) (b) (c) (d) (e) (f)

图 5. 与边缘检测联合训练的可视化结果。(a) 原图；(b) 客观事实；(c-d) 使用显著性物体边界作为边缘分支客观事实得到的边缘图和显著性图；(e-f) 通过与边缘数据集联合训练得到的边缘图与显著性图。通过对比 d 列和 f 列的结果，我们可以容易地观察到，使用高质量的边缘数据集进行联合训练可以显著改善检测到的显著性物体的细节。

细节的结构可以在图. 1的顶端找到。我们在自顶向下的通道中，在 FAMs 后增加了 3 个残差模块用于信息转换。这些残差模块与 [7]中的设计类似，从精细层次到粗略层次，通道数依次为 $128, 256, 512$。如 [26]中所做的，每个卷积块紧跟着一个用于特征压缩的 16 通道 3×3 的卷积层和一个用于边缘检测的单通道 1×1 卷积层。我们还将这 3 个 16 通道 3×3 卷积层串联起来，送入 3 个连续 48 通道的 3×3 卷积层，将捕捉到的边缘信息传输到显著性目标检测分支中进行细节增强。

类似于 [17]，在训练阶段，我们使用显著性物体的边缘作为客观事实以进行联合训练。但是，这个过程没有带来任何性能增益。一些结果仍缺少物体边缘的细节信息。例如，如图. 5c 列展示的，对于前景与背景对比度较低的场景，显著性图结果和边缘图依然模糊不清。其原因或许是从显著性物体提取出的客观事实边缘图仍缺乏显著性物体的大部分细节信息。它们只告诉我们哪里是显著性物体最外围的边界，尤其是对于显著性物体重叠的情况。

考虑到上述讨论，我们尝试使用如 [26]中相同的边缘检测数据集 [1, 29] 对边缘检测任务进行联合训练。在训练期间，源于显著性检测数据集的图片与边缘检测数据集的图片交替输入。从图. 5中可见，与边缘检测任务进行联合训练极大地提升了探测到显著性物体的细节。我们将在实验部分进一步提供更多定量分析。

5. 实验结果

在这节中，我们先描述了实验设置，包括实现细节、所用数据集与评估标准。我们之后进行了一系列消融实验以证明我们提出的方法的每个组件对性能的影响。最后，我们报告了我们方法的性能并与以往 SOTA 方法进行对比。

5.1. 实验设置

实现细节. 我们提出的框架基于 Pytorch[3]实现。所有实验使用 Adam [13]优化器（权重衰减 5e-4，初始学习率 5e-5 每 15 个周期除以 10）。我们的网络一共训练了 24 个周期。网络的主干参数（例如：VGG-16 [33]与 ResNet-50 [7]）使用在 ImageNet 数据集 [16] 上预训练的对应模型进行初始化，剩余参数进行随机初始化。如 [17]中所做，若无特意解释，我们的消融实验默认基于 VGG-16 主干并使用 MSRA-B [25]与 HKU-IS [18]联合数据集。我们仅使用简单的随机水平翻转进行数据增强。如 [9]中所作，在训练与测试中输入图像的尺寸保持不变。

数据集和损失函数. 为了衡量我们提出框架的性能，我们在常用的 6 个数据集上进行了实验，包括 ECSSD [39], PASCAL-S [21], DUT-OMRON [40], HKU-IS [18], SOD [30]和 DUTS [34]。为了方便，如果没有明显的冲突，我们有时使用数据集的首字母作为缩写。我们使用标准二元交叉熵损失进行显著性检测并使用平衡二元交叉熵函数进行边缘检测。

评价准则. 我们使用 3 个广泛使用的标准来评估我们方法与其他方法性能：精度-召回率曲线（PRC），F 得分（F-measure score）与平均绝对值误差（MAE）。F 得分，写作 F_β，是一种整体性能度量，是由精度与召回率计算加权调和平均值得出的：

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

如之前工作，这里 β^2 设置为 0.3 来加权使精度权重大于召回率。MAE 显示显著性图 S 与客观事实 G 有多么相近：

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)| \quad (2)$$

[3] https://pytorch.org

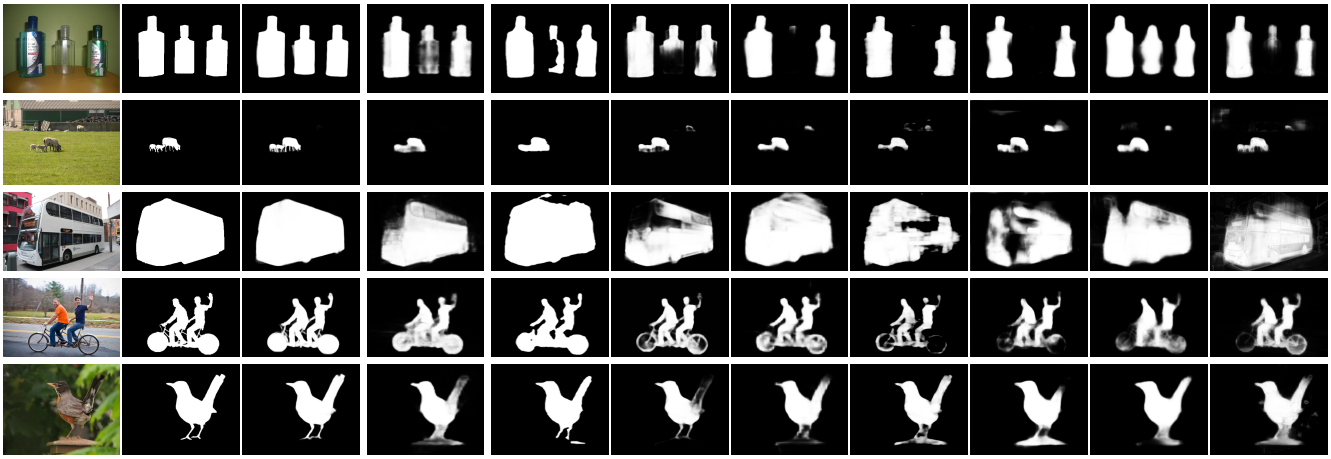


图 6. 与以往先进方法的定性比较。显然，相较于其他方法，我们的方法不仅能定位完整的显著性物体而且能改进检测到显著性物体的细节。这使我们的结果显著性图与客观事实标注十分相近。

No.	GGM + FAMs			DUT-O [40]		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**

表 1. 在 2 个流行的数据集上对所提出框架的消融分析。所有实验均基于 VGG-16 主干并在 MSRA-B [25]与 HKU-IS [18]联合数据集上训练。默认情况下，我们的基准是 VGG-16 版本的 FPN [22]。可见，在我们构架中的每个组件均起到重要作用并有助于性能提升。每列最好的结果被标红。

这里 W 和 H 分别表示 S 的长和宽。

5.2. 消融实验

在这一小节，我们首先调查我们提出的 GGM 与 FAMs 的有效性。然后，我们进行了更多对 GGM 于 FAMs 配置的实验。最后，我们展示了边缘检测联合训练对系统性能的影响。

GGM 与 FAMs 的有效性. 了证明我们提出的 GGM 与 FAMs 的有效性，我们以 VGG-16 为主干进行了以 FPN 为基准的消融实验。除了 GGM 与 FAMs 的不同组合，其他设置均相同。表 1展示了 2 个具有挑战性的数据集 DUT-O 与 SOD 上的性能。对应的可视化结果在图. 2中可见。

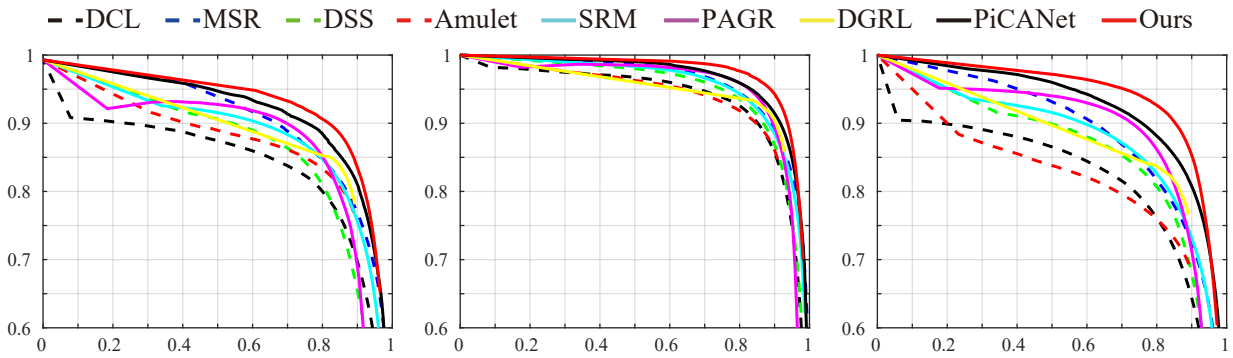
- **仅 GGM.** 相较于 FPN 基准，GGM 的加入(表 1第 4 行）在两个数据集上均带来 F 得分与 MAE 两方面性能的提升。GGM 产生的全局引导信息允许网络更多地关注显著性物体的整体，极大地提升了结果显著性图的质量。因此，显著性物体的细节更加细化，但这也许会错误地估计为具有有限感受野模型的背景。（例如：图. 2最后一行）

- **仅 FAMs.** 如图. 1所示，简单地向 FPN 中嵌入 FAMs（表 1第 5 行）在两个数据集上同样能提升 F 得分和 MAE 得分。

  这或许因为与基准网络相比在 FAMs 中的池化操作同样增大了整个网络的感受野，而且 FPN 基准仍需要融合不同层级的特征图，这表明 FAMs 在解决上采样的锯齿效应时的有效性。

- **GGM & FAMs.** 通过向基准网络中引入 GGM 和 FAMs，与上述 2 种情况相比，F 得分和 MAE 得分均进一步有所提升。(表 1的最后一行) 这种现象表明 GGM 与 FAMs 是 2 个互补的模块。如图. 2所示，它们的使用使我们的方法在准确发现显著性物体与改善细节方面拥有更强大的能力。图. 6展示了更多的定性结果。

GMM 的设置. 为了能更好的了解 GGM 的组成，我们进行了 2 项消融实验，分别对应了表 1的第二行与第三行。我们交替地移除 GGFs 与 PPM 其中之一并保持另一个不变。可见，相较于两者都加入时的结果（第



(a) PASCAL-S [21] (b) HKU-IS [18] (c) DUTS-TE [34]

图 7. 在 3 个流行显著性物体数据集上的精度（纵轴）召回率（横轴）曲线。

	PASCAL-S [21]		DUT-O [40]		SOD [30]	
Settings	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**

表 2. 使用不同边缘信息时我们方法的消融实验。这里的基准网络为加入 GGM+FAM 的 VGG-16 版本 FPN。我们还使用了 MSRA-B [25]与 HKU-IS [18]的组合作为训练集。“SalEdge”指显著性物体边界，“StdEdge”指边缘检测的标准数据集，如 [26, 15]中，其包括了 BSDS500 [1]与 PASCAL VOC Context [29]。

4 行），2 种操作均引起性能下降。这些数值结果表明 PPM 与 GGFs 在 GGM 中均起重要作用。任一个的缺失都会损害我们方法的性能。

联合训练的影响. 为了进一步提升我们方法产生显著性图的质量，我们尝试以联合训练的方式结合边缘检测与显著性物体检测。在表 2中，我们列出了考虑到 2 种显著性物体的边界信息时的结果。可见，采用显著性物体边界作为监督对结果没有改善，而使用标准边界能够极大地提升在全部三个数据集上的性能，尤其是 MAE 得分。这表明引入细节的边缘信息有助于显著性物体检测。

5.3. 与 SOTA 算法的对比

在这节中，我们将我们的 Poolnet 与 13 种先进方法进行了对比，包括 DCL [19], RFCN [35], DHS [23], MSR [17], DSS [9], NLDF [28], UCF [43], Amulet [42], GearNet[10], PAGR [44], PiCANet [24], SRM [36], and DGRL [37]。为了公平起见，这些方法的显著性图通过原作者放出的代码生成或直接由他们提供。另外，所有结果均直接从单模型测试，不依赖任何后处理工具并且所有预测显著性图都采用统一评测代码进行评估。

定量对比. 定量结果在表 3列出。我们采用 VGG-16 [33]与 ResNet-50 [7]作为主干并展示了 2 者的结果。另外，我们在不同的数据集下进行实验以消除潜在的性能波动。从表 3，我们可以观察到，在相同的主干与训练数据集下 PoolNet 超越了近乎所有以往先进的方法在所有数据集上的结果。不同方法的平均速度（FPS）对比（在相同环境下进行测试）在表 4中展示。显然，我们的方法可以实时运行并比其他方法速度快。

PR 曲线. 除了其他数值结果，我们还在图. 7中展示了 3 个数据集下的 PR 曲线。可见，我们方法的 PR 曲线（红色）相较于其他以往方法尤为突出。当召回率接近 1 时，我们的精确度远高于其他方法。这个现象显示了在我们显著性图中假阳性很少。

可视化对比. 了进一步阐述我们方法的优越性，我们在图. 6中展示了一些定性结果。自顶向下，分别为透明物体、小型物体、大型物体、复杂纹理与前景背景对比度低的图片。易见，我们的方法不仅标记出正确的显著性物体而且在几乎所有场景都保留了清晰的边界。

6. 结论

在这篇论文中，我们通过设计 2 种简单的基于池化的模块（全局引导模块 GGM 与特征聚合模块 FAM），探索了池化在显著性物体检测上的潜力。通过将这些模块插入 FPN 结构中，我们展示了我们提出的 PoolNet 可以在 6 个常用的显著性物体基准测试上超越所有以

Model	Training		ECSSD [39]		PASCAL-S [21]		DUT-O [40]		HKU-IS [18]		SOD [30]		DUTS-TE [34]	
	#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 [35]	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 [43]	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 [42]	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 [44]	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 [36]	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 [37]	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 [40], H: HKU-IS [18], DTS: DUTS-TR [34].

表 3. 在 6 个广泛使用的数据集下定量显著性分割结果。不同主干的最好结果分别用**蓝色**和**红色**高亮。†: 边缘检测联合训练。可见，我们的方法以 F 得分与 MAE 衡量在几乎所有数据集上得到了最好的结果。

	Ours	PiCANet [24]	DGRL [37]	SRM [36]	Amulet [42]
Size	400 × 300	224 × 224	384 × 384	353 × 353	256 × 256
FPS	32	7	8	14	16
	UCF [43]	NLDF [28]	DSS [9]	MSR [17]	DHS [23]
Size	224 × 224	400 × 300	400 × 300	400 × 300	224 × 224
FPS	23	12	12	2	23

表 4. 我们的方法（ResNet-50，w/ edge）与以往 SOTA 方法的平均速度（FPS）对比。

往最先进的方法。另外，我们还揭示了以端到端的训练方式与标准边缘检测任务联合训练可以极大地增强探测到显著性物体的细节。我们的模块与网络结构无关，因此可以灵活的应用于任何基于金字塔的模型。这些发展也为提高显著性图的质量提供了有希望的方法。

致谢. 本研究受 NSFC（61620106008，61572264），国家青年人才支持计划，天津市自然科学基金（17JCJQJC43700，18ZXZNGX00110）与中央高校基本科研业务费专项基金（南开大学，NO. 63191501）资助。

参考文献

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2011. 5, 7

[2] Ali Borji and Laurent Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012. 1

[3] Ming Cheng, Niloy J Mitra, Xumin Huang, Philip HS Torr, and Song Hu. Global contrast based salient region detection. *IEEE TPAMI*, 2015. 1, 2, 8

[4] Ming-Ming Cheng, Fang-Lue Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Repfinder: finding approximately repeated scene elements for image editing. *ACM TOG*, 29(4):83, 2010. 1

[5] Celine Craye, David Filliat, and Jean-François Goudou. Environment exploration for object-based

visual saliency learning. In *ICRA*, pages 2303–2309, 2016. 1

[6] Lee Gayoung, Tai Yu-Wing, and Kim Junmo. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5, 7

[8] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *ICML*, pages 597–606, 2015. 1

[9] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019. 1, 2, 3, 5, 6, 7, 8

[10] Qibin Hou, Jiang-Jiang Liu, Ming-Ming Cheng, Ali Borji, and Philip HS Torr. Three birds one stone: A unified framework for salient object segmentation, edge detection and skeleton extraction. *arXiv preprint arXiv:1803.09860*, 2018. 1, 7, 8

[11] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 1

[12] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, pages 2083–2090, 2013. 1, 2

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[14] Dominik A Klein and Simone Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011. 1

[15] Iasonas Kokkinos. Pushing the boundaries of boundary detection using deep learning. *arXiv preprint arXiv:1511.07386*, 2015. 7

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[17] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017. 1, 5, 6, 7, 8

[18] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015. 2, 5, 6, 7, 8

[19] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016. 6, 7, 8

[20] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, pages 2976–2983, 2013. 2

[21] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 5, 7, 8

[22] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1, 2, 3, 6

[23] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016. 1, 2, 3, 7, 8

[24] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018. 1, 2, 3, 6, 7, 8

[25] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011. 1, 5, 6, 7, 8

[26] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *CVPR*, 2017. 5, 7

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[28] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 1, 2, 7, 8

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 5, 7

[30] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR*, pages 49–56, 2010. 5, 6, 7, 8

[31] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast

based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 1, 2

[32] 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*, pages 234–241, 2015. 1

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3, 5, 7

[34] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 5, 7, 8

[35] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016. 1, 2, 7, 8

[36] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, pages 4019–4028, 2017. 2, 3, 6, 7, 8

[37] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, pages 3127–3135, 2018. 1, 3, 6, 7, 8

[38] Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. Deep salient object detection with dense connections and distraction diagnosis. *IEEE Transactions on Multimedia*, 2018. 3

[39] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013. 1, 5, 8

[40] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 5, 6, 7, 8

[41] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018. 2, 3

[42] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 1, 2, 6, 7, 8

[43] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 2, 7, 8

[44] 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. 1, 3, 6, 7, 8

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 3

[46] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015. 2

[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015. 3