显著性目标检测中的完整性学习

诸葛鸣晨，范登平，刘念，张鼎文，徐东 *Fellow, IEEE*，邵岭 *Fellow, IEEE*

摘要—尽管当前显著性目标检测已取得重大突破，它们在预测显著区域的“完整性”上仍存在局限性。本文把“完整性”的概念分为微观完整性和宏观完整性两个层面。具体而言，在微观层面上，模型需要找出单个显著目标的所有部分。而在宏观层面上，模型需要发现图片中的所有显著目标。为了达到对显著性目标检测的完整性学习，本文设计了一个新颖的完整性感知网络，该网络探索了三个重要模块。1）与现有模型不同，本文引入了一个多样化特征聚合（DFA）模块，用来生成具有不同感受野（即，拥有不同内核形状和上下文）的特征，进而增强特征的多样性。此多样性是挖掘完整显著目标的基础。2）基于从 DFA 得到的多样化特征，本文又引入了通道完整性增强（ICE）模块。其目的是增强能潜在突出完整显著目标的通道，同时抑制其他的干扰通道。3）在提取了增强的特征后，本文采用了部分-整体验证（PWV）模块来确定目标特征的部分和整体是否有强烈的一致性。这种部分-整体的一致可以进一步提高每个显著目标微观层面的完整性。为了证明 ICON 模型的有效性，本文在七个具有挑战性的基准上进行了全面实验。本文提出的 ICON 在多个指标上都优于基线方法。值得注意的是，文所提出的 ICON 在假阴性率（FNR）方面在六个数据集上都比此前的最佳模型实现了约 10% 的相对提升。代码和结果可见于: https://github.com/mczhuge/ICON.

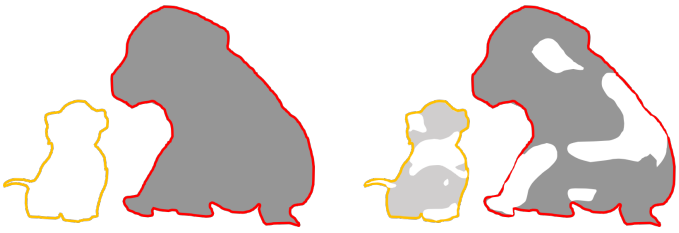
Index Terms—显著性目标检测，胶囊网络，完整性学习。

1 引言

显著性目标检测（SOD）旨在模仿人类的视觉感知系统来捕捉给定图像中最重要的区域。由于 SOD 在计算机视觉领域被广泛应用，它在许多下游任务中起着至关重要的作用，如目标检测 [2]、图像检索 [3]、协同显著检测 [4]、多模态匹配 [5]、VR/AR 应用 [6] 和语义分割 [7]–[9]。

传统的 SOD 方法 [10], [11] 以自下而上的方式预测显著图，并主要基于手工线索，如颜色对比 [12], [13]、边界背景 [14], [15]、或中心先验 [12], [13]。为了提高 SOD 中特征的表示能力，目前的模型采用了卷积神经网络（CNN）或全卷积网络架构，这种强大的特征学习过程可以取代以往的手工特征。这些方法取得了重要的突破，并将 SOD 的性能推到了一个新高度。基于深度学习的 SOD 方法已被相关综述/基准归纳，详情见 [10], [16]–[19]。

当前显著性模型的成功主要依靠多尺度特征聚合、上下文建模、自顶向下建模和边缘指导学习机制。具体而言，使用多尺度特征聚合机制的模型可以增强来自网络不同层和尺度的特征，并融合这些特征生成最终的预测结果。这些方法可以帮助发现不同大小的显著目标，并在粗略语义和精确细节的指导下突出显著区域。例如，Zhang 等人 [21] 提出的网络首先自适应地融合了五个不同尺度的特征，然后用来生成最终预测结果。相似地，Luo 等人 [22] 提出在低和高特征尺度上分别提取全局与局部特征，然后融合它们得到预测结果。



(a) 宏观完整性　　(b) 微观完整性

图 1. 显著性目标检测里的完整性问题（即，(a) 宏观完整性与 (b) 微观完整性）。红色和黄色的线分别代表不同物体的显著性边缘。灰色的区域代表模型预测结果。这张图参考了 [20]。

上下文建模则是另一个在 SOD 里的重要机制。它通过考量周围环境来帮助推断每个局部区域的显著性。目前在 SOD 领域中，通常采用各种注意力模块来探索这种信息。比如，Zhao 等人 [23] 提出一个金字塔特征注意网络，其中引入了通道注意力和空间注意力模块，分别处理高层次和低层次特征，并考虑不同特征通道和空间位置的上下文信息。Liu 等人 [24] 则提出学习像素级的上下文注意力来预测显著目标。用注意力模块优化的深度模型能推断每个像素与其全局/局部上下文的位置相关重要性，从而实现上下文信息的选择性聚合。

对于自顶向下建模，一些 SOD 方法采用精心设计的解码器，在高层语义线索的指导下逐步推断出显著区域。例如，

- 前两位作者对本文贡献相同。
- 范登平，单位为：南开大学计算机学院，天津，中国。(邮箱：*dengpfan@gmail.com*)
- 诸葛鸣晨，刘念，邵岭，单位为：起源人工智能研究院，阿布扎比，阿联酋。(邮件：*mczhuge@gmail.com, liunian228@gmail.com, ling.shao@ieee.org*)
- 张鼎文，单位为：脑与人工智能实验室，西北工业大学，陕西，西安。(邮件：*zhangdingwen2006yyy@gmail.com*)
- 徐东，单位为：电子信息学院，悉尼大学，悉尼，新南威尔士，澳大利亚。(邮件：*Email: dong.xu@sydney.edu.au*)

- 该工作为诸葛鸣晨在起源人工智能研究院实习期间在范登平博士指导下完成的工作。本文通讯作者为：刘念、张鼎文。
- 本文为 *TPAMI2022* 论文 *[1]* 的中文翻译版，由诸葛鸣晨译，范登平、刘念、张鼎文校稿。

(a) 宏观完整性　　(b) 微观完整性

图 2. 完整性是显著性目标检测的一个有效指标。下面是一些来自 SOD 数据集的样本:(a) 中的图像需要模型有宏观层面的完整性判断，而 (b) 组则需要模型有效地提取微观层面的完整性。在每组中：左边为图像，右边为真值。图像顶部则呈现了几个具有不同完整性质量的预测。注意：■ 表示不良预测结果，■ 代表相对好的预测结果。

Wang 等人 [25] 为 SOD 建立了一个迭代和合作推理网络，其中多个自上而下的网络流与自下而上的网络流以迭代推理的方式共同工作。Zhao 等人 [26] 提出了一种门控双分支解码结构，实现了自顶向下流中不同层次特征之间的协同，提高了整个网络的可判别性。在工作 [27] 中，Liu 等人 采用了一个金字塔集合模块来建立全局引导特征，他们引入了该模块来改进对自顶向下特征通路的建模。

为了准确预测显著物体的边界，另一组方法引入了额外的网络分支或目标函数，迫使网络更加关注区分显著目标与周围背景的边缘。例如，Wei 等人 [28] 为 SOD 建立了一个标签解耦框架，该框架明确地将原始的显著图分解为一个主体真值图和一个细节真值图。具体来说，主体真值图专注于显著物体的中心区域，而细节图则专注于物体边界区域。为了提高显著轮廓的预测精度，减少显著物体边缘预测中的局部噪声，Wu 等人 [29] 提出了相互学习策略，分别指导前景轮廓和边缘检测任务。

尽管上述几种机制可以在不同方面提高 SOD 性能，但它们产生的结果仍然不够理想。本文觉得，这可能是由于另一种机制，即对完整性学习的探索不充分所致（见图 1 (a) 与 (b)）。本工作在两个层面上定义了完整性学习机制。在微观层面上，模型应关注单一显著目标内的部分与整体的相关性。在宏观层面上，模型则需要识别给定图像场景中的所有显著目标。图 2展示了一些宏观和微观层面上的完整性质量的例子，其中：顶部的为预测；底部分为两组，左边为宏观完整性相关的图像和真值，右边为宏观完整性相关的图像和真值。很明显，完整性和显著性预测性能之间存在着很强的关联性。

为了追求两种完整性，本文在深度神经网络的设计中引入了三个关键组件。第一个是多样化特征聚合模块（DFA）。与现有模型不同，DFA 更注重特征的可辨别性，它通过聚合来自不同感受野的特征（即，拥有不同的内核形状和上下文），以增加特征的多样性。这样的特征多样性为挖掘完整的显著物体提供了基础，因为它考虑了更丰富的上下文模式来决定每个神经元的激活程度。第二部分被称为通道完整性增强（ICE），其目的是增强有显著物体的特征通道，同时抑制其他分散注意力的特征。由于通过 ICE 增强的通道与真实的显著区域完全匹配的情况很少，本文进一步采用了部分-整体验证（PWV）组件来判断部分特征和整体特征是否有很强的一致性来形成整体目标。这有助于进一步提高微观层面的完整性学习。

值得一提的是，一些现有的工作也试图通过引入额外任务来解决宏观层面的完整性问题，以学习深度显著性目标检测器 [30], [31]。然而，这些方法需要诸如图像内显著目标数量的额外监督信息。相比之下，本文新提出的方法可以在一个统一的、完全不同的学习框架下解决宏观和微观层面的完整性问题，且不需要额外监督信息。

本文的整体框架被称为完整性感知网络（Integrity Cognition Network，ICON），它的细节展示在图 3中。具体来说，ICON 首先利用五个卷积块进行基本特征提取。然后，它将每层的深度特征传递给一个多样性特征聚合模块，以提取不同的基础多样性特征。接下来，从三个相邻的 DFA 中提取到的多样性特征被送到一个通道完整性增强模块。由此，一个完整性指导图被生成，然后用于指导对每个特征通道的注意力加权。最后，从三个特征层中产生的通道完整性增强特征被合并到一起并送入部分-整体验证模块，该模块使用胶囊路由层 [32] 实现。在进一步验证物体部分和整体区域之间的一致性后，显著物体的缺失部分被补足。总之，本文的主要贡献有以下三点：

- 本文研究了 SOD 中的完整性问题，该问题至关重要但对于该问题的探索是相对空白的。
- 本文介绍了实现完整性 SOD 的三个核心组件，即多样化

特征聚合、通道完整性增强和部分-整体验证。

- 本文设计了一个新网络，即 ICON，它包含了三个模块，并在七个具有挑战性的数据集上证明了其有效性。除了卓越的性能外，本文方法达到了实时预测速度（60fps）。

本文的其余部分组织如下。在第 §2节中，我们讨论相关的工作。随后，在后续章节中我们详细描述所设计的模型（见第 §3节）。实验结果，包括性能评估和比较，在第 §4节中展示。最后，我们在第 §5节中进行总结。

2 相关工作

过去几十年里，众多 SOD 方法被提出，并在各种基准数据集上取得了令人鼓舞的性能。这些现有的 SOD 模型可大致分为基于尺度、基于边界和基于完整性的方法。

2.1 基于尺度的 SOD 方法

尺度变化是 SOD 的主要挑战之一。许多工作都试图从不同角度来处理这个问题。受边缘检测 HED 模型 [33] 的启发，DSS 引入了由深入浅的侧端输出，这些输出具有丰富的语义特征。这一设计使模型浅层能够从背景中区分出真正的显著目标，同时保留高分辨率。此外，Zhang 等人 [21] 设计了一个多层次的特征聚合框架，并采用分层特征作为最终显著性预测的线索。同时，RADF [34] 整合了多层次的特征，并在每一层中用递归方式对其进行细化。这就有效地抑制了低层的非显著性噪声，增加了高层特征的显著性细节。此外，Zhao 等人 [35] 提出使用 F-measure 损失函数，它可以生成精确的对比图来帮助分割多尺度物体。为了有效地提取多尺度特征，Pang 等人 [36] 在解码器单元中嵌入了自交互模块来学习综合信息。最近推出的 GateNet [37] 采用了 Fold-ASPP 来收集多尺度的显著性线索。最后，Liu 等人 [38] 利用了一个集中的信息交互策略来同时处理多尺度特征。

2.2 基于边界学习的 SOD 方法

边界学习对改善 SOD 预测担任着另一个重要角色。早期的工作通过生物启发方法进行边界学习 [13], [39], [40]。然而，这些模型结果会模糊，并通常会丢失完整的显著区域。最近的基于 CNN 的方法，在图像块级别（而不是像素级别）进行操作，由于步长和池化操作，也存在边缘模糊的问题。为了解决这个问题，一些工作（比如, [41]）使用预处理技术（比如, 超像素 [42]）来保留物体的边界，而其他工作，如 DSS [43]、DCL [44] 和 PiCANet [45] 则采用后处理技术（比如, 条件随机场 [46]）来增强边缘细节。但这些方法有个主要缺点，就是推理速度慢。

为了学习内在的边缘信息，PoolNet [27] 采用了一个辅助模块进行边缘检测。此外，许多其他工作通过引入边界感知的损失函数来提高边缘质量。例如，最近的工作 [23], [47]–[50] 明确地使用边界损失来指导边界细节学习。考虑到交叉熵损失更倾向于将像素样本（比如，0 或 1）预测为非整数值，BASNet [51] 引入了一个新的预测-精细化网络和混合损失。为了处理边界模糊的固有缺陷，HRSOD [52] 率先提出了一个高分辨率 SOD 数据集，探索了高分辨率数据如何提高显著物体边缘的性能。F3Net [53] 证明了在损失函数中给边界像素分配较大的权重是处理边界问题的一个简单方法。此外，最近的工作，如 SCRN [54]、LDF [28] 和 VST [55] 建立双流架构，同时对显著物体和边界建模。

2.3 基于完整性的 SOD 方法

完整性学习是 SOD 中一个未被充分探索的研究课题。尚未有大量工作基于此展开，在少许现有模型中，DCL [44] 在像素和图像块两个层面上处理对比信息，以同时整合全局和局部结构线索。CPD [56] 利用有效的解码器来总结辨别特征，并在整体注意力模块的帮助下对整体显著目标进行分割。TSPOANet [57] 对 SOD 中的物体组成进行建模，并在胶囊网络的帮助下提高显著物体预测的整体性和统一性。GCPANet [58] 充分利用全局上下文来捕捉多个显著物体或区域之间的关系，缓解了特征的稀释现象。Wu 等人 [59] 使用一个具有两个特征主干和门控单元的双流网络来融合互补信息。最近，Transformer 已经成为计算机视觉领域的一个研究热点。Mao 等人 [60] 提出了一种基于 Transformer 的架构来解决长序列学习问题，这也可以被认为是一种有利于帮助完整性学习的尝试。

3 模型框架

3.1 ICON 的概述

如图 3所示，本文的方法是基于一个编码器-解码器的架构。编码器使用 ResNet-50 作为骨干来提取多层次的特征。同时，解码器整合这些多层次的特征，生成具有多层监督的显著性预测结果。为了简单起见，从这里开始，把骨干生成的特征表示为一个集合 $\mathcal{F}_{bkb} = \{\mathbf{F}^{(0)}_{bkb}, \mathbf{F}^{(1)}_{bkb}, \mathbf{F}^{(2)}_{bkb}, \mathbf{F}^{(3)}_{bkb}, \mathbf{F}^{(4)}_{bkb}\}$。由于 $\mathbf{F}^{(0)}_{bkb}$ 的空间尺寸较大，为了提高计算效率，在解码器中不使用它。

接下来，模型通过多样化特征聚合（DFA）模块来增强骨干特征，该模块由各种卷积块组成。此后，模型进一步使用完整性通道增强（ICE）模块来加强完整性相关的特征通道并粗略地突出整体性的显著部分。最后，利用部分-整体验证（PWV）模块来验证物体部分和整个突出区域之间的一致性，以进一步完善显著性图。

3.2 多样化特征聚合模块

最近的工作 [61]–[63] 已经证明，增加卷积核的感受野可以帮助网络学习捕捉不同大小物体的特征。在本文工作中，引入了不同形状的卷积核来处理不同物体的形状多样性。具体来说，本文采用新颖的多样化特征聚合（DFA）模块来增强提取的多级特征的多样性，使用三种具有不同内核大小和形状的

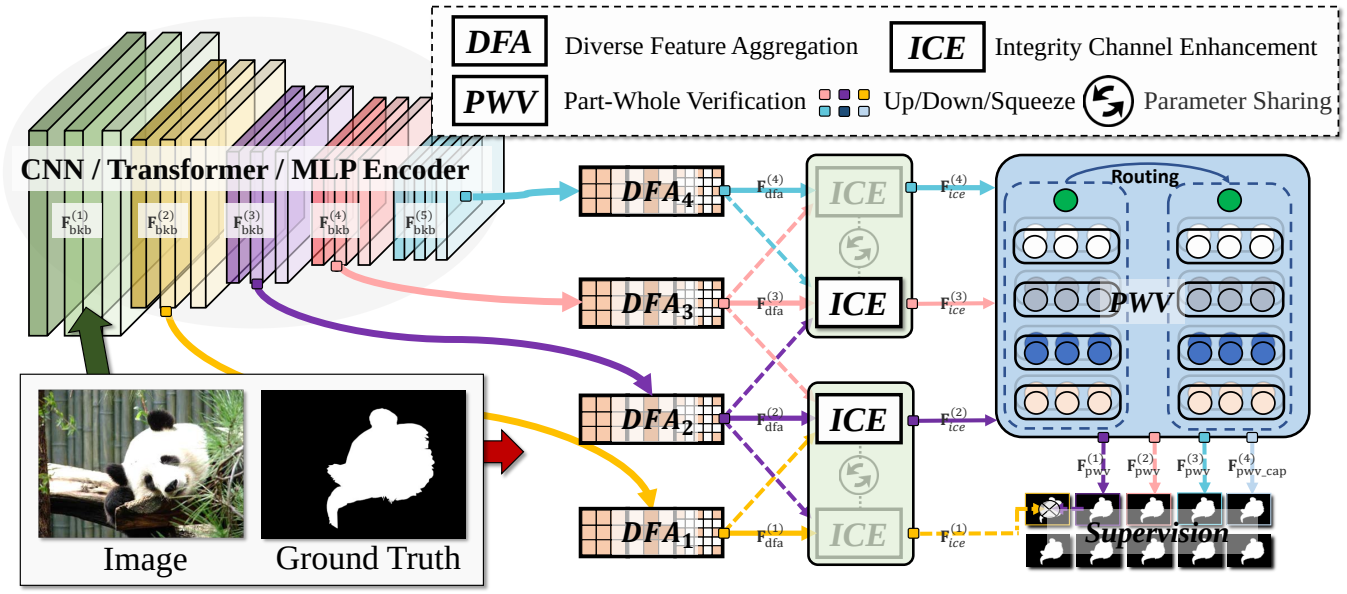


图 3. ICON 的整体架构。特征提取：$\mathbf{F}_{bkb}^{(1)}$ - $\mathbf{F}_{bkb}^{(5)}$ 表示 ResNet-50 的不同输出层。模块 1：DFA 聚合了具有不同感受野的特征。模块 2：ICE 旨在增强能潜在显著目标完整性的特征通道。模块 3：PWV 判断部分和整体特征是否有强烈的一致性。

卷积块，如图 4-(A)。在技术上，利用非对称卷积 [64]、空洞卷积 [65] 和原始卷积的实际组合来捕捉各种空间特征。整个过程总结如下：

$$\mathbf{F}_{dfa}^{(i)} = \text{Concat}\left[\mathcal{X}_{asy}(\mathbf{F}_{bkb}^{(i)}), \mathcal{X}_{atr}^{2}(\mathbf{F}_{bkb}^{(i)}), \mathcal{X}_{ori}(\mathbf{F}_{bkb}^{(i)})\right], \tag{1}$$

其中，$\mathbf{F}_{dfa}^{(i)}$ 为上述过程产生的特征，$\mathcal{X}_{*}$ 为不同类型的模块（即，非对称卷积、空洞卷积和原始卷积），Concat[·] 为连接操作。

值得注意的是，本文用 $\mathcal{X}_{atr}^{r}$ 表示使用不同扩张率 r 的空洞卷积操作，比如，$\mathcal{X}_{atr}^{2}$ 是扩张率为 2 的空洞卷积，并使用有十字形状卷积核的非对称卷积（$\mathcal{X}_{asy}$）[64]。$\mathcal{X}_{asy}$ 包含三个层次，一个是正常的 3×3 元的卷积核 $\mathbf{K}_{3\times3}$，一个是水平的 1×3 的卷积核 $\mathbf{K}_{1\times3}$，还有一个是垂直的 3×1 的卷积核 $\mathbf{K}_{3\times1}$，它们共享在同一个滑动窗口。它可以被描述为：

$$\mathcal{X}_{asy}(\mathbf{I}) = (\mathbf{I} \star \mathbf{K}_{3\times3}) \oplus (\mathbf{I} \star \mathbf{K}_{1\times3}) \oplus (\mathbf{I} \star \mathbf{K}_{3\times1}), \tag{2}$$

其中 $\star$ 是二维卷积算子，$\oplus$ 是元素相加，$\mathbf{I}$ 表示输入特征。

这样一来，DFA 模块可以通过融合从十字卷积核、空洞卷积核和普通卷积核中学习到的知识来丰富特征空间。因此，DFA 可以覆盖多个上下文中的不同显著区域，增强完整性。本文将 DFA 处理的特征标记为 $\mathcal{F}_{dfa} = \{\mathbf{F}_{dfa}^{(1)}, \mathbf{F}_{dfa}^{(2)}, \mathbf{F}_{dfa}^{(3)}, \mathbf{F}_{dfa}^{(4)}\}$。

3.3 通道完整性增强模块

最近的一些研究 [66]–[69] 通过使用空间或通道注意机制取得了可喜的视觉分类结果。尽管这些方法是由不同动机驱动的，但它们本质上都是为了建立不同特征之间的对应关系，以突出最重要的物体部分。然而，如何挖掘隐藏在不同通道特征中的完整性信息，仍然没得到充分研究。为了解决这个问题，本文提出了一个简单的 ICE 模块，以进一步挖掘不同通道内的关系，并增强能潜在突出完整目标的通道。

这里，本文考虑每三个相邻特征的多尺度信息。首先，对下一个和上一个特征层进行重新缩放，并使用上采样和下采样操作，将其调整为 $H \times W$ 的空间分辨率。然后，通过连接三个输入特征来生成融合图 $\mathbf{F}_{fuse}^{(i)}$。

$$\mathbf{F}_{fuse}^{(i)} = \text{Concat}\left[\mathbf{F}_{dfa}^{(i-1)}, \mathbf{F}_{dfa}^{(i)}, \mathbf{F}_{dfa}^{(i+1)}\right]. \tag{3}$$

之后，我们通过对 $\mathbf{F}_{fuse}^{(i)}$ 应用 l_2 正则化，提取完整性特征 $\mathbf{I}_{emb}^{(i)}$。接下来，为了进一步整合完整性信息，我们使用参数高效的 Bottleneck 设计来学习 $\mathbf{I}_{emb}$。由于信道变换会略微增加优化的难度，在两个卷积层内（ReLU 之前）加入了层归一化，以缓解优化难度。该设计思路与 [68] 类似。

$$\mathbf{F}_{ice}^{(i)} = \mathbf{F}_{fuse}^{(i)} \otimes \mathcal{X}_{ori}(\text{ReLU}(\text{LN}(\mathcal{X}_{ori}(\mathbf{I}_{emb}^{(i)})))), \tag{4}$$

其中，$\otimes$ 是元素间的乘法运算，LN 表示层归一化操作。

通过使用本文的 ICE 模块，具有更好完整性的通道可以被有效地增强。从图 5中可以看出，在将特征送入 ICE 后，前景区域与背景被明显区分开来，ICE 产生的特征倾向于突出

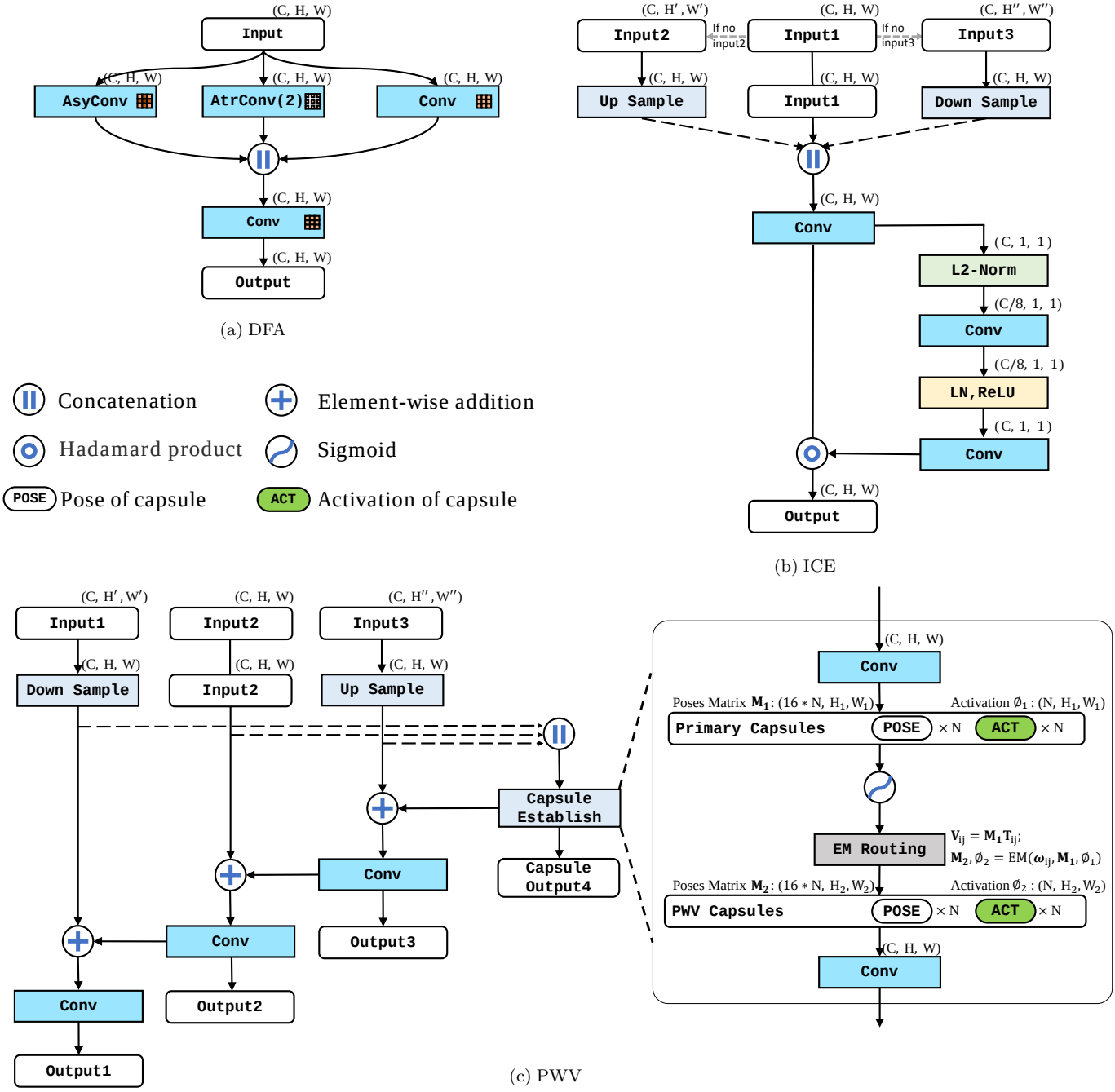


图 4. 设计模块的细节。(a) 多样化特征聚合 (DFA) 模块结合了不同的卷积核，以增强框架的表示能力。(b) 完整性通道增强 (ICE) 组件，挖掘通道中的完整性信息。(c) 部分-整体验证 (PWV) 组件被设计用来对部分和整体目标之间的关系进行建模。AsyConv = 非对称卷积块 [64], AtrConv = 空洞卷积块 [65], Conv = 卷积块。所有这些块包括卷积、BatchNorm 和 ReLU 激活函数。EM Routing= 期望最大的路由机制 [32]。C、H、W 分别表示特征张量的通道数、高度和宽度。

微观和宏观层面上的完整显著目标。在本文的实现中，ICE 在第一层和最后一层没有足够的多层次特征输入，模型用当前层的特征来填补。此外，模型使用两个共享参数的 ICE 模块来帮助本文的 ICON 模型整合多层次特征线索。

3.4 部分整体验证模块

PWV 模块旨在通过评估显著物体部分和整个区域之间的一致性来增强所学习到的完整性特征。为了实现这一目标，本文采用了一个胶囊网络 [32], [70]，该网络在建模部分-整体关系方面已被证明是有效的。受先前工作 SegCaps [71] 的成功启发，本文将胶囊网络嵌入到模型中。

在 PWV 中，一个关键的问题是如何实现从低级胶囊分配给高级胶囊的投票。高层胶囊需要通过聚合相关低层胶囊的目标部分来形成整个目标的表示，此处模型使用 EM 路由 [32]，以类似聚类的方式对低层和高层胶囊之间的关系进

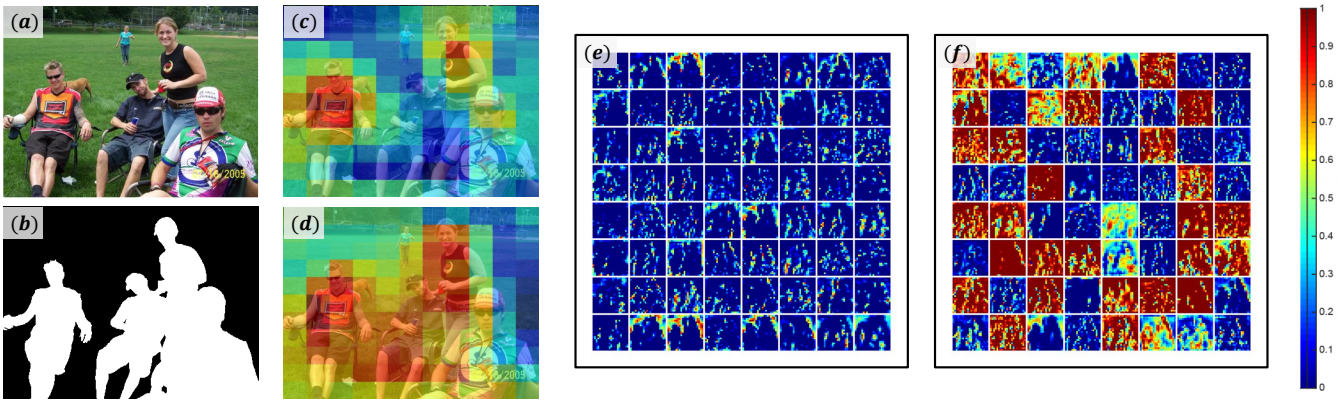


图 5. ICE 热力图视觉比较。(a) 输入图像。(b) 真值图。通过 ICE 模块之前对的特征热图 (c)，和之后的特征热图 (d)。经过 ICE 模块 (e) 之前和 (f) 之后的通道可视化。本文的 ICE 模块有助于网络更多地关注显著区域的完整性并将前景与背景区分开。(d) 中呈现的特征热图清楚地表明了 ICE 在宏观和微观层面上捕捉完整性的能力。通道的可视化是通过挤压所有通道生成，本文使用 Matplotlib 的"jet" 伪彩色进行着色。放大观看更佳。

行建模。PWV 的输入是三个不同的 ICE 特征（$\mathcal{F}_{ice}$）。具体来说，首先将每个层次的 ICE 特征统一到相同的分辨率，即，22×22，以减少计算成本。

接下来，模型建立主要胶囊层。具体来说，我们用八个姿态向量来建立一个姿态矩阵 $\mathbf{M}$，和一个激活 $\phi \in [0, 1]$ 来代表每个胶囊。姿态矩阵包含实例化参数，以反映物体部分或整个物体的属性，而激活表示物体的存在概率。来自主要胶囊层的胶囊通过一个路由协议机制将信息传递给随后的 PWV 胶囊层的胶囊。具体来说，来自上级的胶囊为后一级的胶囊生成投票，票数 ω_{ij} 是通过学习的变换矩阵 $\mathbf{T}_{ij}$ 和前一级的姿态矩阵 $\mathbf{M}_i$ 之间的矩阵乘法运算得到的，其中 i 和 j 分别是前一级和后一级胶囊的索引。一旦获得这些票数，它们就被用于 EM 路由算法 [32]，以获得具有姿态矩阵 $\mathbf{C}_j$ 和激活 ϕ_j 的后一级胶囊 $\mathbf{M}_j$。之后，模型得到部分-整体验证过的特征。随后，我们通过点加和上采样操作，自底而上地融合这些相邻层次的部分整体验证特征，从而鼓励多尺度特征之间的交互。经过 PWV 模块，模型生成：$\mathcal{F}_{pwv} = \{\mathbf{F}^{(1)}_{pwv}, \mathbf{F}^{(2)}_{pwv}, \mathbf{F}^{(3)}_{pwv}, \mathbf{F}^{(4)}_{pwv\_cap}\}$.

3.5 监督策略

在这项工作中，除了 BCE 损失外，我们还使用了 IoU 损失 [51], [72]。具体来说，所提出模型的整体损失被表述为 $\mathcal{L}_{CPR}(P, G)$。其中 P 是生成的显著性预测图，G 是真值图。$\mathcal{L}_{CPR}$ 包含了 BCE 损失和 IoU 损失，即，$\mathcal{L}_{CPR} = \mathcal{L}_{BCE} + \mathcal{L}_{IoU}$。具体来说，$\mathcal{L}_{BCE}$ 的表述如下:

$$\begin{aligned}\mathcal{L}_{BCE} = -\sum_{x=1}^{H}\sum_{y=1}^{W}[G(x,y)\log(P(x,y)) \\ + (1-G(x,y))\log(1-P(x,y))],\end{aligned} \tag{5}$$

其中 W 和 H 分别是图像的宽度和高度。同时，L_{IoU} 被定义为:

$$\mathcal{L}_{IoU} = 1 - \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} P(x,y)G(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W}[P(x,y) + G(x,y) - P(x,y)G(x,y)]}, \tag{6}$$

其中 $G(x, y)$ 和 $P(x, y)$ 分别是位置 (x, y) 的真值标签和预测标签。训练过程中，本文采用了该领域广泛使用的多级监督策略 [27], [36], [53], [58]。除了使用 $\mathcal{F}_{pwv}$ 中的四个特征外，本文通过点乘将 $\mathbf{F}^{(1)}_{pwv}$ 和 $\mathbf{F}^{(1)}_{ice}$ 融合在一起，作为被监督的额外特征。在推理期，这个特征也被用来生成最终预测图。在训练和推理阶段，为了将预测图和真值图相匹配，它们的特征通道将被缩减为一维，空间大小将被恢复为与输入图像相同。

4 实验

4.1 数据集

本文在 **DUTS-TR** [73] 数据集上训练模型，该数据集常被用作 SOD 训练集，包含 10,553 张图像。本文在七个流行基准上评估模型。**ECSSD** [74]、**HKU-IS** [75]、**OMRON** [14]、**PASCAL-S** [76]、**DUTS-TE** [73]、**SOD** [77] 和基于属性的 **SOC** [16]，它们都有像素级标签的注释。具体来说，ECSSD 是由 1,000 张有丰富语义的图像组成。HKU-IS 包括 4,447 张图像，包含多个前景物体。OMRON 由 5,168 张至少有一个物体的图像组成。这些物体通常在结构上很复杂。PASCAL-S 是由最初用于语义分割的数据集建立的，它由 850 张具有挑战性的图像组成。DUTS 是一个相对较大的数据集，有两个子集。DUTS-TR 中的 10,553 张图像被用于训练，DUTS-TE 中的 5,019 张图像被用于测试。SOD 包括 300 张极具挑战性的图像。SOC 包含复杂的场景，比其他六个 SOD 数据集中的场景更具挑战性。

4.2 实施细节

本文所有实验在开源的 Pytorch1.5.0 平台上运行。一台配备英特尔酷睿 i7-9700K CPU (4.9GHz Turbo boost)、16GB 3000

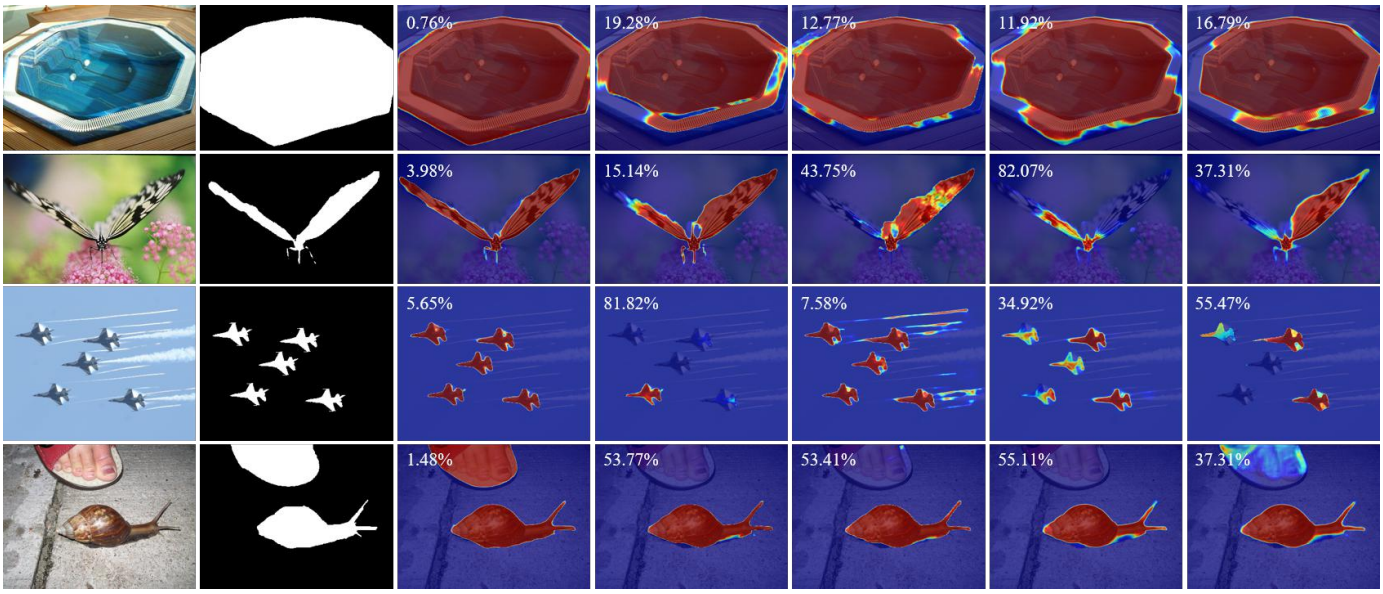


图 6. 假阴性率（FNR）视觉比较。可见，FNR 在很大程度上反映了完整性的情况。

MHz 内存和 RTX 2080Ti GPU 卡（11GB 内存）的八核 PC 被用于训练和测试。在网络训练过程中，每幅图像首先被调整为 352×352（对于 VGG [78]/ResNet [79]/PVT [80], [81]）或 384×384（对于 Swin [82]/CycleMLP [83]）。诸如归一化、裁剪和翻转的数据增强方法被使用。一些编码器参数从 VGG-16, ResNet-50、PVT, Swin-B 和 CycleMLP-B4 中初始化。本文用 0 或 1 初始化 PWV 的一些层，而其他卷积层则按照 [84] 进行初始化。本文使用 SGD 优化器 [85] 来训练本文的网络，其超参数设置为：初始学习率 lr = 0.05，momen = 0.9，eps = 1e-8，weight_decay = 5e-4。学习率调整采用 Warm-up 和线性衰减策略。批量大小设置为 32 (ResNet)，10 (PVTv2/CycleMLP) 或 8 (VGG/Swin)，最大 epochs 数设置为 60 (基于 ResNet 的训练需要 ~2.5 小时)。此外，本文还使用了 apex[1]和 fp16 来加速训练过程。同时使用梯度剪裁来防止梯度爆炸。基于 ResNet 的架构的推理过程对于 352×352 的图像，包括 IO 时间只需花费 0.0164s。

4.3 评估指标

本文用五个指标来评估所有模型：

(1) MAE (M) 它评估预测图 (P) 和真值图 (G) 之间的平均像素差异。计算时将 P 和 G 归一化为 $[0,1]$，因此 MAE 得分可以计算为 $M = \frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|P(x,y)-G(x,y)|$。

(2) Weighted F-measure (F_β^ω) [86] 该指标提供了一个直观的概括，定义为：$F_\beta^\omega = \frac{(1+\beta^2)\text{Precision}^\omega\cdot\text{Recall}^\omega}{\beta^2\cdot\text{Precision}^\omega+\text{Recall}^\omega}$。作为一个广泛采用的指标 [16], [21], [44], [48], [52], [57], [87]–[90]，F_β^ω 可以解决存在于 MAE 和 F-measure [91] 中导致评价不准确的插值、依赖性和同等重要性问题。和 [10] 中所建议的一样，本文将 β^2 设置为 0.3，以强调精度而非召回率。遵循特定的位置和邻域信息，通过给不同的错误分配不同的权重 (ω)，F_β^ω 将 F-measure 扩展到非二元评价。

(3) S-measure (S_m) [92] 它侧重于评估结构相似性，这更接近于人类的视觉感知。计算公式为 $S_m = ms_o+(1-m)s_r$，其中，s_o 和 s_r 分别表示目标感知和区域感知的结构相似度。遵循 [92]，m 被设置为 0.5。

(4) E-measure (E_ξ) [93] 该指标同时考虑局部像素值与图像级别的平均值，可以计算为：$E_\xi = \frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\theta(\xi)$，其中，$\xi$ 是对齐矩阵，$\theta(\xi)$ 表示增强对齐矩阵。本文采用平均 E-measure (E_ξ^m) 作为此处的评价指标。

(5) FNR 为假阴性率。本文用它进一步评估完整性，它可以检测预测结果是否与显著像素成完整一体。FNR 的计算方法是：

$$FN(x,y) = \begin{cases} 1, G(x,y)=1 \ \&\ P(x,y)=0, \\ 0, \ others. \end{cases} \tag{7}$$

$$FNR = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}FN(x,y)}{\sum_{x=1}^{W}\sum_{y=1}^{H}G(x,y)}\times 100\%, \tag{8}$$

其中 FN 是像素级指示器，决定了一个像素是否是假阴性。本文在图 6中展示了几个 FNR 的例子。它清楚而准确地反映了预测图的完整性，在宏观和微观层面上都很敏感。

4.4 性能比较

本文将所提出的模型与 14 个最近性能优越的模型进行比较，包括：Condinst [95]、PointRend [96]、PiCANet [45]、RAS [97]、AFNet [98]、BASNet [51]、CPD [56]、EGNet [50]、SCRN [54]、F3Net [53]、MINet [36]、ITSD [94]、GateNet [37] 和 VST [55]。

1. https://github.com/NVIDIA/apex

表 1

在六个数据集上量化结果。最佳性能以粗体显示。符号 “↑”/“↓’ 意味着分数越高/越低越好。’-V/VGG-Based’:VGG16 [78], ’-R/ResNet-Based’: ResNet50 [79], ’-P’: PVTv2-1K [81], ’-S’: Swin-B-22k [82], ’-M’: CycleMLP-B4 [83]。

Summary			ECSSD [74]				PASCAL-S [76]				DUTS [73]				HKU-IS [75]				OMRON [14]				SOD [77]			
Method	MACs	Params	S_m ↑	E_ξ^m ↑	F_β^w ↑	M↓	S_m ↑	E_ξ^m ↑	F_β^w ↑	M↓	S_m ↑	E_ξ^m ↑	F_β^w ↑	M↓	S_m ↑	E_ξ^m ↑	F_β^w ↑	M↓	S_m ↑	E_ξ^m ↑	F_β^w ↑	M↓	S_m ↑	E_ξ^m ↑	F_β^w ↑	M↓
VGG16-Based Methods																										
RAS	-	-	.893	.914	.857	.056	.799	.835	.731	.101	.839	.871	.740	.059	.887	.920	.843	.045	.814	.843	.695	.062	.767	.791	.718	.123
CPD	59.46	29.23	.910	.938	.895	.040	.845	.882	.796	.072	.867	.902	.800	.043	.904	.940	.879	.033	.818	.845	.715	.057	.771	.787	.718	113
EGNet	149.89	108.07	**.919**	.936	.892	.041	.848	.877	.788	.077	.878	.898	.797	.044	.910	.938	.875	.035	**.836**	.853	.728	**.057**	.788	.803	.736	.110
ITSD	14.61	17.08	.914	.937	.897	.040	.856	.891	.811	.068	.877	.905	.814	.042	.906	.938	.881	.035	.829	.853	.734	.063	.797	.826	.764	.098
MINet	94.11	47.56	**.919**	.943	**.905**	**.036**	.854	.893	.808	**.064**	.875	.907	.813	**.039**	.912	.944	.889	**.031**	.822	.846	.718	**.057**	-	-	-	-
GateNet	108.34	100.02	.917	.932	.886	.041	.857	.886	.797	.068	.870	.893	.786	.045	.910	.934	.872	.036	.821	.840	.703	.061	-	-	-	-
Ours-V	64.90	19.17	**.919**	**.946**	**.905**	**.036**	**.861**	**.902**	**.820**	**.064**	**.878**	**.915**	**.822**	.043	**.915**	**.950**	**.895**	.032	.833	**.865**	**.743**	.065	**.814**	**.848**	**.784**	**.089**
ResNet50-Based Methods																										
CondInst	-	-	.721	.717	.603	.115	.813	.852	.757	.084	.760	.764	.631	.070	.748	.754	.648	.093	.646	.629	.433	.114	.670	.657	.535	.148
PointRend	-	-	.753	.766	.667	.111	.810	.850	.763	.099	.774	.794	.667	.081	.784	.805	.715	.091	.651	.647	.453	.125	.693	.793	.589	.141
PiCANet	54.05	47.22	.917	.925	.867	.046	.854	.870	.772	.076	.869	.878	.754	.051	.904	.916	.840	.043	.832	.836	.695	.065	.793	.799	.722	.103
AFNet	21.66	35.95	.913	.935	.886	.042	.849	.883	.797	.070	.867	.893	.785	.046	.905	.935	.869	.036	.826	.846	.717	.057	-	-	-	-
BASNet	127.36	87.06	.916	.943	.904	.037	.838	.879	.793	.076	.866	.895	.803	.040	.909	.943	.889	.032	.836	.865	.751	.056	.772	.801	.728	.112
CPD	17.77	47.85	.918	.942	.898	.037	.848	.882	.794	.071	.869	.898	.795	.043	.905	.938	.875	.034	.825	.847	.719	.056	.771	.782	.713	.110
EGNet	157.21	111.69	.925	.943	.903	.037	.852	.881	.795	.074	.887	.907	.815	.039	.918	.944	.887	.031	.841	.857	.738	**.053**	.807	.822	.767	.097
SCRN	15.09	25.23	.927	.939	.900	.037	**.869**	.892	.807	.063	.885	.900	.803	.040	.916	.935	.876	.034	.837	.848	.720	.056	-	-	-	-
F3Net	16.43	25.54	.924	.948	.912	.033	.861	.898	.816	**.061**	**.888**	.920	.835	**.035**	.917	.952	.900	**.028**	.838	.864	.747	**.053**	.806	.834	.775	.091
ITSD	15.96	26.47	.925	.947	.910	.034	.859	.894	.812	.066	.885	.913	.823	.041	.917	.947	.894	.031	.840	.865	.750	.061	.809	.836	.777	.093
MINET	87.11	126.38	.925	.950	.911	.033	.856	.896	.809	.064	.884	.917	.825	.037	.919	.952	.897	.029	.833	.860	.738	.056	-	-	-	-
GateNet	162.13	128.63	.920	.936	.894	.040	.858	.886	.797	.067	.885	.906	.809	.040	.915	.937	.880	.033	.838	.855	.729	.055	-	-	-	-
Ours-R	20.91	33.09	**.929**	**.954**	**.918**	**.032**	.861	**.899**	**.818**	.064	**.888**	**.924**	**.836**	.037	**.920**	**.953**	**.902**	.029	**.844**	**.876**	**.761**	.057	**.824**	**.854**	**.794**	**.084**
Transformer-Based Methods																										
VST	23.16	44.63	.932	.951	.910	.033	.872	.902	.816	.061	.896	.919	.828	.037	.928	.952	.897	.029	.850	.871	.755	.058	.820	.846	.778	.086
Ours-P	34.70	65.68	.940	.964	.933	.024	.882	.921	.847	.051	**.917**	.950	.882	**.022**	**.935**	.967	**.925**	**.022**	.865	.896	.793	.047	**.832**	**.864**	**.813**	**.078**
Ours-S	52.59	94.30	**.941**	**.966**	**.936**	**.023**	**.885**	**.924**	**.854**	**.048**	**.917**	**.954**	**.886**	.025	**.935**	**.968**	**.925**	**.022**	**.869**	**.900**	**.804**	**.043**	.825	.856	.802	.083
MLP-Based Methods																										
Ours-M	26.13	54.92	.940	.964	.934	.025	.873	.912	.838	.056	.909	.942	.874	.029	.935	.966	.926	.022	.855	.886	.783	.051	.821	.853	.803	.081

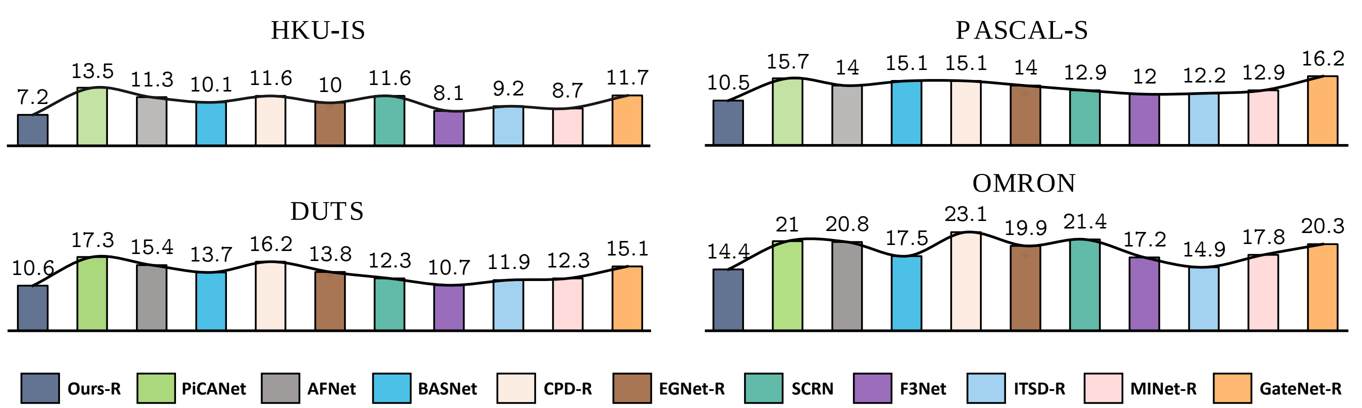


图 7. FNR 统计，包含 11 个模型在不同数据集上的数据。

4.4.1 量化评估

表 1报告了在六个传统基准数据集上的定量结果，与 14 种最先进的算法在 S-measure、E-measure、加权 F-measure 和 MAE 方面进行了比较。本文的模型明显优于其他模型。此外，本文还在图 7中展示了本文和基线方法的 FNR 结果。可以看出，本文方法在所有数据集上都取得了最低的 FNR 分数（越低代表效果越好）。视觉比较（见图 6）也证明了它在捕捉整体目标方面的有效性。事实上，ICON 在所有数据集上且所有的评估指标方面都比现有的方法表现得好。这证明了它在处理具有挑战性的输入方面有更强的能力。此外，本文在图 8中展示了 PR 曲线 [12] 和 F-measure 曲线 [91]。可以看到，红色实线（代表所提出的 ICON 模型）明显高于其他曲线，这进一步证明了本文模型的有效性和完整性学习能力。

4.4.2 视觉比较

图 9 展示了本文方法和基线方法之间的视觉比较。可以看出，ICON 在各种具有挑战性的情况下产生了更准确的显著图，比如: 小物体（见第一行）、大物体（见第二行）、精细结构（见第三行）、低对比度（见第四行）以及多个物体（见第五行）。此外，本文模型架可以完整地、无噪音地检测显著目标。上述结果证明了所提方法的准确性和稳健性。

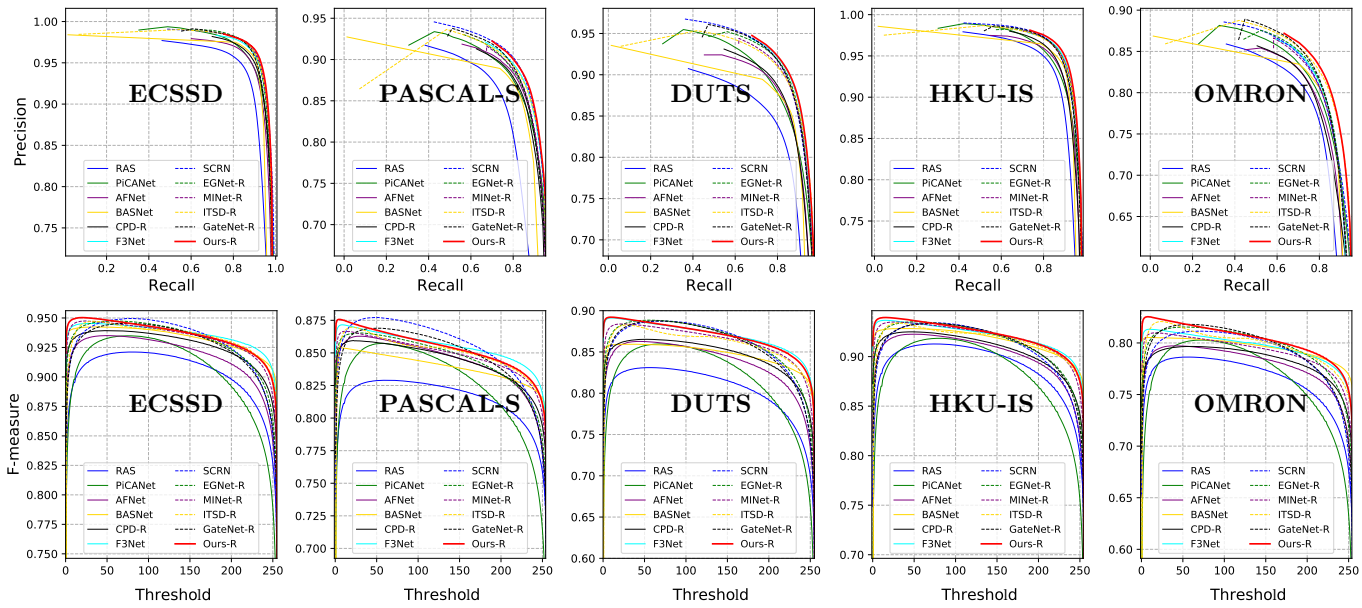


图 8. 在五个流行的 SOD 数据集上，所提出的方法和其他 SOTA 算法的 PR 曲线和 F-measure 曲线。

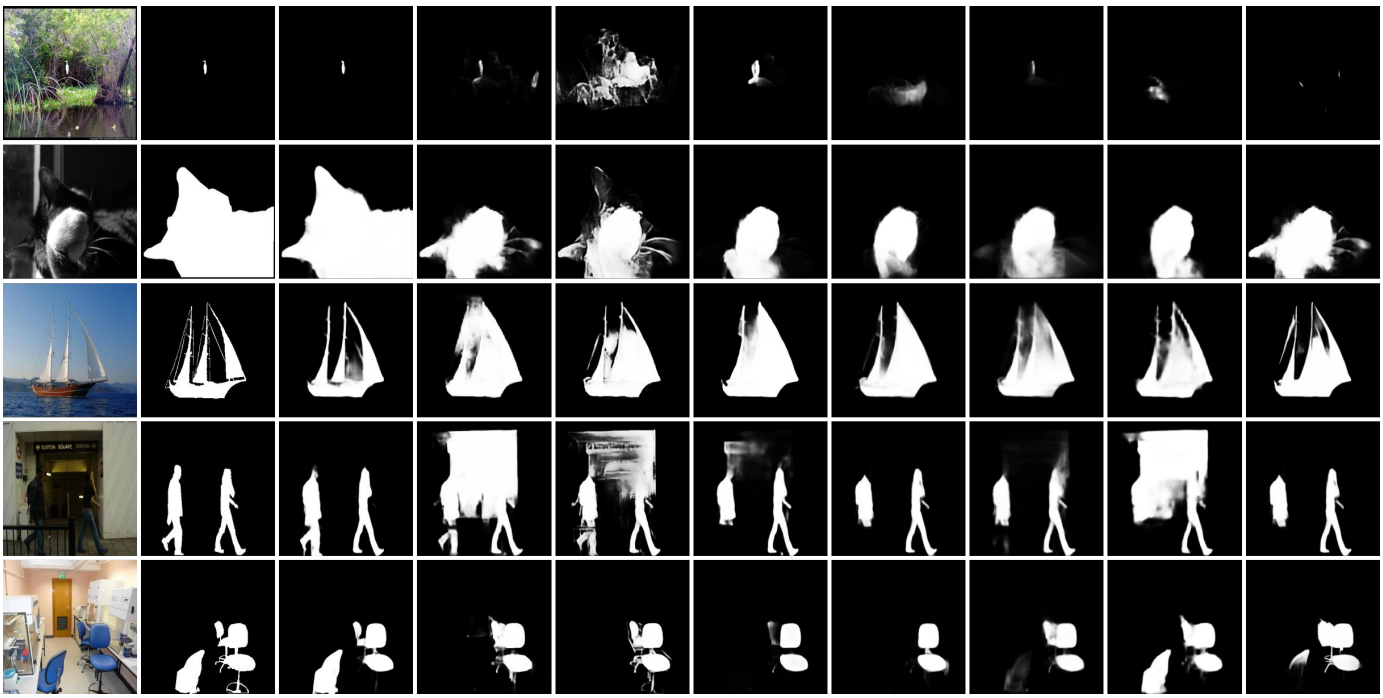


图 9. 本文模型与 7 个模型的定性比较。与其他模型不同，本文方法不仅能准确定位显著物，还能在各种场景中产生更清晰的边缘并减少背景干扰。

4.4.3 基于属性的分析

除了最常用的显著性目标检测数据集，本文还在另一个具有挑战性的 SOC 数据集 [16], [104] 上测试了本文的模型。与之前的六个 SOD 数据集相比，这个数据集包含了许多更复杂的场景。此外，SOC 数据集根据九个不同的属性对图像进行分类，包括：AC（外观变化）、BO（大物体）、CL（杂乱）、HO（异质物体）、MB（运动模糊）、OC（遮挡）、OV（超出视野）、SC（形状复杂）和 SO（小物体）。

在表 2中，展示了本文模型和 16 个最先进的模型之间在面向不同属性时的性能比较，包括：Amulet [21]、DSS [43]、NLDF [47]、C2SNet [48]、SRM [99]、R3Net [100]、BMPM [101]、DGRL [102]、PiCANet-R（PiCA-R） [24]、RANet [103]、AFNet [98]、CPD [56]、PoolNet [27]、EGNet [50]、BANet [49] 和 SCRN [54]。从表 2中可以看出，与现有的方法相比，本文的模型取得了明显的性能改进。

表 2
本文方法与其他 SOTA 模型在 SOC 测试集上的比较。对于 ↑ 和 ↓，较高和较低的分数分别表示更好的结果。† 表示结果由在 SOC 训练集上训练的模型生成。

Attr	Metrics	Amulet [21]	DSS [43]	NLDF [47]	C2SNet [48]	SRM [99]	R3Net [100]	BMPM [101]	DGRL [102]	PiCA-R [45]	RANet [103]	AFNet [98]	CPD [56]	PoolNet [27]	EGNet [50]	BANet [49]	SCRN [54]	ICON-R (Ours)
AC	S_m ↑	0.752	0.753	0.737	0.755	0.791	0.713	0.780	0.790	0.792	0.708	0.796	0.799	0.795	0.806	0.806	0.809	**0.835**/**0.835**†
	E_ξ^m ↑	0.791	0.788	0.784	0.807	0.824	0.753	0.815	0.853	0.815	0.765	0.852	0.843	0.846	0.854	0.858	0.849	**0.891**/0.889†
	F_β^w ↑	0.620	0.629	0.620	0.647	0.690	0.593	0.680	0.718	0.682	0.603	0.712	0.727	0.713	0.731	0.740	0.724	**0.784**/**0.784**†
	M ↓	0.120	0.113	0.119	0.109	0.096	0.135	0.098	0.081	0.093	0.132	0.084	0.083	0.094	0.085	0.086	0.078	**0.062**/0.064†
BO	S_m ↑	0.574	0.561	0.568	0.654	0.614	0.437	0.604	0.684	**0.729**	0.421	0.658	0.647	0.561	0.528	0.645	0.698	0.714/0.713†
	E_ξ^m ↑	0.551	0.537	0.539	0.661	0.616	0.419	0.620	0.725	0.741	0.404	0.698	0.665	0.554	0.528	0.650	0.706	0.740/**0.743**†
	F_β^w ↑	0.612	0.614	0.622	0.730	0.667	0.456	0.670	0.786	**0.799**	0.453	0.741	0.739	0.610	0.585	0.720	0.778	0.794/0.794†
	M ↓	0.346	0.356	0.354	0.267	0.306	0.445	0.303	0.215	0.200	0.454	0.245	0.257	0.353	0.373	0.271	0.224	0.200/**0.199**†
CL	S_m ↑	0.763	0.722	0.713	0.742	0.759	0.659	0.761	0.770	0.787	0.624	0.768	0.773	0.760	0.757	0.784	0.795	0.789/**0.802**†
	E_ξ^m ↑	0.789	0.763	0.764	0.789	0.793	0.710	0.801	0.824	0.794	0.715	0.802	0.821	0.801	0.790	0.824	0.820	0.829/**0.855**†
	F_β^w ↑	0.663	0.617	0.614	0.655	0.665	0.546	0.678	0.714	0.692	0.542	0.696	0.724	0.681	0.677	0.726	0.717	0.732/**0.754**†
	M ↓	0.141	0.153	0.159	0.144	0.134	0.182	0.123	0.119	0.123	0.188	0.119	0.114	0.134	0.139	0.117	0.113	0.113/**0.102**†
HO	S_m ↑	0.791	0.767	0.755	0.768	0.794	0.740	0.781	0.791	0.809	0.713	0.798	0.803	0.815	0.802	0.819	0.823	0.818/**0.830**†
	E_ξ^m ↑	0.810	0.796	0.798	0.805	0.819	0.782	0.813	0.833	0.819	0.777	0.834	0.845	0.846	0.829	0.850	0.842	0.852/**0.865**†
	F_β^w ↑	0.688	0.660	0.661	0.668	0.696	0.633	0.684	0.722	0.704	0.626	0.722	0.751	0.739	0.720	0.754	0.743	0.752/**0.771**†
	M ↓	0.119	0.124	0.126	0.123	0.115	0.136	0.116	0.104	0.108	0.143	0.103	0.097	0.100	0.106	0.094	0.096	0.092/**0.087**†
MB	S_m ↑	0.712	0.719	0.685	0.720	0.742	0.657	0.762	0.744	0.775	0.696	0.734	0.754	0.751	0.762	0.764	0.792	0.774/**0.821**†
	E_ξ^m ↑	0.739	0.753	0.740	0.778	0.778	0.697	0.812	0.823	0.813	0.761	0.762	0.804	0.779	0.789	0.803	0.817	0.828/**0.866**†
	F_β^w ↑	0.561	0.577	0.551	0.593	0.619	0.489	0.651	0.655	0.637	0.576	0.626	0.679	0.642	0.649	0.672	0.690	0.699/**0.768**†
	M ↓	0.142	0.132	0.138	0.128	0.115	0.160	0.105	0.113	0.099	0.139	0.111	0.106	0.121	0.109	0.104	0.100	0.100/**0.076**†
OC	S_m ↑	0.735	0.718	0.709	0.738	0.749	0.653	0.752	0.747	0.765	0.641	0.771	0.750	0.756	0.754	0.765	0.775	0.771/**0.791**†
	E_ξ^m ↑	0.763	0.760	0.755	0.784	0.780	0.706	0.800	0.808	0.784	0.718	0.820	0.810	0.801	0.798	0.809	0.800	0.817/**0.831**†
	F_β^w ↑	0.607	0.595	0.593	0.622	0.630	0.520	0.644	0.659	0.638	0.527	0.680	0.672	0.659	0.658	0.678	0.673	0.683/**0.710**†
	M ↓	0.143	0.144	0.149	0.130	0.129	0.168	0.119	0.116	0.119	0.169	0.109	0.115	0.119	0.121	0.112	0.111	0.106/**0.100**†
OV	S_m ↑	0.721	0.700	0.688	0.728	0.745	0.624	0.751	0.762	0.781	0.611	0.761	0.748	0.747	0.752	0.779	0.774	0.779/**0.802**†
	E_ξ^m ↑	0.751	0.737	0.736	0.790	0.779	0.663	0.807	0.828	0.810	0.664	0.817	0.803	0.795	0.802	0.835	0.808	0.834/**0.846**†
	F_β^w ↑	0.637	0.622	0.616	0.671	0.682	0.527	0.701	0.733	0.721	0.529	0.723	0.721	0.697	0.707	0.752	0.723	0.749/**0.768**†
	M ↓	0.173	0.180	0.184	0.159	0.150	0.216	0.136	0.125	0.127	0.217	0.129	0.134	0.148	0.146	0.119	0.126	0.120/**0.108**†
SC	S_m ↑	0.768	0.761	0.745	0.756	0.783	0.716	0.799	0.772	0.784	0.724	0.808	0.793	0.807	0.793	0.807	0.809	0.803/**0.824**†
	E_ξ^m ↑	0.794	0.799	0.788	0.806	0.814	0.765	0.841	0.837	0.799	0.792	0.854	0.858	0.856	0.844	0.851	0.843	0.860/**0.882**†
	F_β^w ↑	0.608	0.599	0.593	0.611	0.638	0.550	0.677	0.669	0.627	0.594	0.696	0.708	0.695	0.678	0.706	0.691	0.714/**0.745**†
	M ↓	0.098	0.098	0.101	0.100	0.090	0.114	0.081	0.087	0.093	0.110	0.076	0.080	0.075	0.083	0.078	0.078	0.080/**0.073**†
SO	S_m ↑	0.718	0.713	0.703	0.706	0.737	0.682	0.732	0.736	0.748	0.682	0.746	0.745	0.768	0.749	0.755	0.767	0.763/**0.801**†
	E_ξ^m ↑	0.745	0.756	0.747	0.752	0.769	0.732	0.780	0.802	0.766	0.759	0.792	0.804	0.814	0.784	0.801	0.797	0.816/**0.848**†
	F_β^w ↑	0.523	0.524	0.526	0.531	0.561	0.487	0.567	0.602	0.566	0.518	0.596	0.623	0.626	0.594	0.621	0.614	0.634/**0.689**†
	M ↓	0.119	0.109	0.115	0.116	0.099	0.118	0.096	0.092	0.095	0.113	0.089	0.091	0.087	0.098	0.090	0.082	0.087/**0.073**†

4.5 失败案例

尽管本文的 ICON 优于其他 SOD 算法，并且很少产生完全不正确的预测结果，但仍有一些失败的情况，如图 10所示。具体来说，在第一行，显示的是一个整洁的房间，本文方法仍被枕头、床和墙是否是显著的物体所迷惑。同时，在第二张图片中，三盏灯是显著区域，但本文的方法无法检测到它们。不过，其他 SOTA 方法对这些样本的预测也是失败的。造成这些失败案例可能有几个潜在原因：(1) 强烈的色彩对比影响了模型的判断（见第一行）；(2) 缺乏足够的训练样本（见第二行）；(3) 有争议的注释（见第一行）。

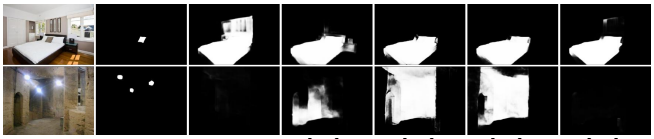


图 10. 失败案例。第一列和第二列是输入图像以及真值，其他是 ICON 和对比模型的预测结果。

4.6 消融实验

4.6.1 模块的有效性

为了证明本文的模型中不同模块的有效性，本文报告了模型的几个简化版本的定量性能。本文从编码器-解码器基线（一个有跳层连接的近似 UNet 网络）开始，逐步用不同模块来扩展它，包括 DFA、ICE 和 PWV。如表 3所示，本文首先测试了基线（ID：1）和在其基础上添加 DFA（ID：2）模块，它表现出明显的性能提升。这是合理的，因为 DFA 有能力通过不同线索搜索目标。然后，加入 ICE 模块（ID：3），它再次显示出大幅提升。最后，正如预期的那样，将所有的组件（ID：4）加入到本文提出的模型中实现了最好的性能。

表 3
所设计模块的消融分析。最佳性能以**粗体**表示（下同）。

ID	Component Settings	OMRON [14]				HKU-IS [75]				DUTS-TE [73]			
		$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
1	Baseline	0.832	0.854	0.731	0.064	0.902	0.930	0.866	0.043	0.861	0.879	0.801	0.049
2	+DFA	0.837	0.857	0.740	0.063	0.913	0.939	0.875	0.035	0.879	0.886	0.818	0.046
3	+DFA+ICE	0.840	0.869	0.753	0.059	0.918	0.951	0.895	0.031	0.887	0.916	0.825	0.038
4	+DFA+ICE+PWV	**0.844**	**0.876**	**0.761**	**0.057**	**0.920**	**0.953**	**0.902**	**0.029**	**0.888**	**0.924**	**0.836**	**0.037**

表 4
不同特征增强方法（FEMs）的消融分析。

ID	FEMs Settings	OMRON [14]				HKU-IS [75]				DUTS-TE [73]			
		$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
2	+DFA	0.837	**0.857**	**0.740**	0.063	**0.913**	0.939	**0.875**	0.035	**0.879**	**0.886**	**0.818**	0.046
5	+Inception [62]	0.839	0.853	**0.740**	0.064	0.909	0.936	0.869	0.037	0.875	**0.886**	0.817	**0.043**
6	+ASPP [65]	**0.840**	0.855	0.738	**0.061**	0.912	**0.943**	0.869	**0.034**	0.877	0.882	**0.818**	**0.043**
7	+PSP [63]	0.835	0.851	0.738	0.063	0.906	0.935	0.870	0.036	0.878	0.884	0.816	0.045
8	+DFA (3xOriConv)	0.833	0.855	0.733	0.062	0.909	0.938	0.868	0.035	0.874	0.875	0.810	0.046
9	+DFA (3xAtrConv [65])	0.830	0.849	0.729	0.066	0.906	0.933	0.869	0.038	0.872	0.880	0.811	0.047
10	+DFA (3xAsyConv [64])	0.837	0.854	0.737	0.064	0.909	0.937	0.873	0.035	**0.879**	0.885	0.817	0.044

表 5
ICE 和相关注意机制的消融分析。

ID	Attention Settings	OMRON [14]				HKU-IS [75]				DUTS-TE [73]			
		$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
3	+DFA+ICE	0.840	**0.869**	**0.753**	0.059	**0.918**	**0.951**	**0.895**	**0.031**	0.887	**0.916**	0.825	**0.038**
11	+DFA+SE [66]	0.839	0.861	0.720	0.061	0.909	0.943	0.877	0.034	**0.888**	0.895	0.831	0.039
12	+DFA+CBAM [67]	**0.842**	0.864	0.739	**0.058**	0.917	0.946	0.891	**0.031**	0.885	0.907	**0.832**	0.039
13	+DFA+GCT [69]	0.838	0.857	0.712	0.062	0.901	0.937	0.874	0.033	0.883	0.905	0.821	0.041

4.6.2 *DFA* 与其他特征增强方法的比较

DFA、ASPP [65]、Inception [62] 和 PSP [63] 是四种特征增强方法（FEM），它们在学习代表性特征方面有一些共通之处。不同的是，DFA 的设计是在不扩大感受野的情况下增强特征子空间，从而产生更多样化的表征。在表 4中，DFA 明显优于或与其他 FEM 相当[2]。然而，DFA 也带来了一些缺点。例如，与其他 FEM 相比，它会导致 MAE 分数更高。本文认为，一个可能的原因是 DFA 不仅带来了特征多样性，也带来了一些噪声。此外，本文的实验（ID：2 vs. ID：8∼10）显示，结合三种不同类型的卷积可以达到最好的分数。同时，只使用 3xAsyConv 比只使用 3xOriConv 或 3xAtrConv 产生更好的结果。

4.7 ICE 与其它注意力方法对比

在 表 5中，本文做了组对照实验（即，ID：3，11∼13）来验证 ICE 机制带来的改进。按照相同设置（ID：3），本文进行实验，将 ICE 与 SE [66]、CBAM [67] 和 GCT [69] 进行比较。可以观察到，CBAM 取得了可相提并论的性能，在这些模块中排名第二。然而，使用 SE 和 GCT 的替代方法将导致性能的明显下降。一个可解释的原因是 ICE 可以加强特征的完整性，并通过本文设计的注意机制突出潜在包含完整性信息的特征通道。

2. DFA 只有 4 个卷积或计算操作块，而其他 FEM 至少有 5 个块。

4.7.1 对不同路由算法的评估

为了评估 EM 路由（ID：4）的性能 [32]，本文还进行了额外实验（见表 7），用动态路由（DR）[70] 和自路由（SR）[105] 取代它。本文观察到，前者（ID：14）也取得了合理的性能，但与使用 EM 路由相比，后者（ID：15）产生的性能更差。一个可能的原因是，SR 没有路由协议机制，使其与本文的 PWV 方案不兼容。

4.7.2 损失函数的评估

为了证明 L_{CPR} 损失的有效性，本文进行损失函数的消融实验。在相同的模型设置下，本文将 L_{CPR} 与 L_{BCE} 进行了比较，在表 6中报告的结果表明，训练过程中使用 L_{CPR} 损失后，模型可以明显提高所有指标下的 SOD 性能。请注意，结合 IoU 和 BCE 损失是一种常见的训练设置，这也被用于最近的许多工作中 [51], [60]。

表 6
损失函数的消融分析。

ID	Loss Settings	OMRON [14]				HKU-IS [75]				DUTS-TE [73]			
		$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
4	ICON+L_{CPR}	**0.844**	**0.876**	**0.761**	**0.057**	**0.920**	**0.953**	**0.902**	**0.029**	0.888	**0.924**	**0.836**	**0.037**
16	ICON+L_{BCE}	0.840	0.866	0.757	0.060	0.918	0.950	0.899	0.031	**0.889**	0.918	0.831	**0.037**

表 7
PWV 中路由机制的消融分析。

ID	Routing Settings	OMRON [14]				HKU-IS [75]				DUTS-TE [73]			
		$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_m\uparrow$	$E_\xi^m\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
4	+DFA+ICE+PWV	**0.844**	**0.876**	**0.761**	**0.057**	0.920	**0.953**	**0.902**	**0.029**	**0.888**	**0.924**	**0.836**	**0.037**
14	+DFA+ICE+PWV (DR) [70]	**0.844**	0.868	0.757	0.058	**0.923**	0.950	**0.902**	0.030	**0.888**	0.919	0.832	0.039
15	+DFA+ICE+PWV (SR) [105]	0.837	0.862	0.745	0.060	0.912	0.843	0.895	0.030	0.881	0.903	0.831	0.042

5 结论

本文提出了一种新型显著性方法，即完整性感知网络（**ICON**），用来检测给定图像场景中的显著物体。该模型的思想是：挖掘整体特征（在微观和宏观层面）有利于显著目标的检测。具体来说，在这项工作中，本文设计了三个新的网络模块：多样化特征聚合模块、完整性通道增强模块和部分-整体验证模块。通过整合这些模块，ICON 能够在不同特征层次上捕获多样化的特征，并突出潜在的包含完整性信息的特征通道，以及进一步验证挖掘出的显著区域内的部分与整体的关联性。本研究工作在七个基准数据集上进行了详实全面的实验，结果证明了每个新提出的模块的有效性，以及该模型的卓越性能。

参考文献

[1] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE TPAMI*, 2022.

[2] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, 2019.

[3] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *Int. Conf. Inf. Sci. Cloud Comput. Companion*, 2013, pp. 728–733.

[4] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[5] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 12 647–12 657.

[6] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.

[7] L. Hoyer, M. Munoz, P. Katiyar, A. Khoreva, and V. Fischer, "Grid saliency for context explanations of semantic segmentation," in *Adv. Neural Inform. Process. Syst.*, 2019, pp. 6462–6473.

[8] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, 2016.

[9] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7223–7233.

[10] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[11] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, 2017.

[12] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2014.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[14] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.

[15] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2814–2821.

[16] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.*, 2018, pp. 186–202.

[17] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, pp. 1–34, 2014.

[18] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[19] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, 2018.

[20] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary iou: Improving object-centric image segmentation evaluation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 15 334–15 342.

[21] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[22] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P. Jodoin, "Non-local deep features for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6593–6601.

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

[24] N. Liu, J. Han, and M.-H. Yang, "Picanet: Pixel-wise contextual attention learning for accurate saliency detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6438–6451, 2020.

[25] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 5968–5977.

[26] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Eur. Conf. Comput. Vis.*, 2020, pp. 35–51.

[27] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3917–3926.

[28] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 025–13 034.

[29] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8150–8159.

[30] 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.

[31] S. He, J. Jiao, X. Zhang, G. Han, and R. W. H. Lau, "Delving into salient object subitizing and detection," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1059–1067.

[32] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *Int. Conf. Learn. Represent.*, 2018.

[33] S. Xie and Z. Tu, "Holistically-nested edge detection," *IJCV*, vol. 125, no. 1-3, pp. 3–18, 2017.

[34] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *AAAI Conf. Art. Intell.*, 2018, pp. 6943–6950.

[35] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng, "Optimizing the f-measure for threshold-free salient object detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8849–8857.

[36] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9413–9422.

[37] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Eur. Conf. Comput. Vis.*, 2020, pp. 35–51.

[38] J.-J. Liu, Z.-A. Liu, P. Peng, and M.-M. Cheng, "Rethinking the u-shape structure for salient object detection," *TIP*, vol. 30, pp. 9030–9042, 2021.

[39] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *ACM Int. Conf. Multimedia*, 2003, pp. 374–381.

[40] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Adv. Neural Inform. Process. Syst.*, 2006.

[41] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2300–2309.

[42] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 330–344, 2015.

[43] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," *PAMI*, vol. 41, no. 4, pp. 815–828, 2019.

[44] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 478–487.

[45] 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.

[46] J. D. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[47] 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. 6593–6601.

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

[49] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.

[50] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8779–8788.

[51] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7479–7489.

[52] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7234–7243.

[53] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *AAAI Conf. Art. Intell.*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[54] Z. Wu, L. Su, and Q. Huang, "Stacked Cross Refinement Network for Edge-Aware Salient Object Detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7264–7273.

[55] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *IEEE Int. Conf. Comput. Vis.*, 2021.

[56] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast

and accurate salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3907–3916.

[57] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1232–1241.

[58] Zuyao Chen and Qianqian Xu and Runmin Cong and Qingming Huang, "Global Context-Aware Progressive Aggregation Network for Salient Object Detection," in *AAAI Conf. Art. Intell.*, vol. 34, no. 07, 2020, pp. 10 599–10 606.

[59] Z. Wu, S. Li, C. Chen, A. Hao, and H. Qin, "A deeper look at image salient object detection: Bi-stream network with a small training dataset," *IEEE Trans. Multimedia*, 2020.

[60] Y. Mao, J. Zhang, Z. Wan, Y. Dai, A. Li, Y. Lv, X. Tian, D.-P. Fan, and N. Barnes, "Generative transformer for accurate and reliable salient object detection," *arXiv preprint arXiv:2104.10127*, 2021.

[61] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2818–2826.

[63] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2881–2890.

[64] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1911–1920.

[65] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[66] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *PAMI*, vol. 42, no. 08, pp. 2011–2023, 2020.

[67] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[68] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *IEEE Int. Conf. Comput. Vis. Worksh.*, 2019, pp. 0–0.

[69] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 11 794–11 803.

[70] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Adv. Neural Inform. Process. Syst.*, 2017, pp. 3856–3866.

[71] R. LaLonde and U. Bagci, "Capsules for object segmentation," in *International conference on Medical Imaging with Deep Learning*, 2018.

[72] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3438–3446.

[73] 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.

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

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

[76] 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.

[77] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2010, pp. 49–56.

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

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[80] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE Int. Conf. Comput. Vis.*, 2021.

[81] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 1–10, 2022.

[82] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE Int. Conf. Comput. Vis.*, 2021.

[83] S. Chen, E. Xie, C. Ge, D. Liang, and P. Luo, "Cyclemlp: A mlp-like architecture for dense prediction," in *Int. Conf. Learn. Represent.*, 2022.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[85] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*, 2012, pp. 421–436.

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

[87] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *TIP*, vol. 29, pp. 3534–3545, 2020.

[88] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, 2018.

[89] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection," in *IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7254–7263.

[90] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2386–2395.

[91] 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.

[92] M.-M. Cheng and D.-P. Fan, "Structure-measure: A new way

to evaluate foreground maps," *IJCV*, vol. 129, no. 9, pp. 2622–2638, 2021.

[93] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function," *SCIENTIA SINICA Informationis*, 2021.

[94] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9141–9150.

[95] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *ECCV*. Springer, 2020, pp. 282–298.

[96] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *CVPR*, 2020, pp. 9799–9808.

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

[98] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1623–1632.

[99] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4039–4048.

[100] 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 *Int. Joint Conf. Artif. Intell.*, 2018, pp. 684–690.

[101] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1741–1750.

[102] 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.

[103] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[104] D.-P. Fan, J. Zhang, G. Xu, M.-M. Cheng, and L. Shao, "Salient objects in clutter," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[105] T. Hahn, M. Pyeon, and G. Kim, "Self-routing capsule networks," in *Adv. Neural Inform. Process. Syst.*, 2019, pp. 7658–7667.