基于语义引导特征聚合的显著性目标检测网络

王正文 [1] 宋慧慧 [1] 樊佳庆 [1] 刘青山 [1]

摘 要 在显著性目标检测网络的设计中, U 型结构使用广泛. 但是在 U 型结构显著性检测方法中, 普遍存在空间位置细节丢失和边缘难以细化的问题, 针对这些问题, 提出一种基于语义信息引导特征聚合的显著性目标检测网络, 通过高效的特征聚合来获得精细的显著性图. 该网络由混合注意力模块 (Mixing attention module, MAM)、增大感受野模块 (Enlarged receptive field module, ERFM) 和多层次聚合模块 (Multi-level aggregation module, MLAM) 三个部分组成. 首先, 利用增大感受野模块处理特征提取网络提取出的低层特征, 使其在保留原有边缘细节的同时增大感受野, 以获得更加丰富的空间上/下文信息; 然后, 利用混合注意力模块处理特征提取网络的最后一层特征, 以增强其表征力, 并作为解码过程中的语义指导, 不断指导特征聚合; 最后, 多层次聚合模块对来自不同层次的特征进行有效聚合, 得到最终精细的显著性图. 在 6 个基准数据集上进行了实验, 结果验证了该方法能够有效地定位显著特征, 并且对边缘细节的细化也很有效.

关键词 显著性目标检测, 混合注意力, 多层次融合, 深度学习

引用格式 王正文, 宋慧慧, 樊佳庆, 刘青山. 基于语义引导特征聚合的显著性目标检测网络. 自动化学报, 2023, **49**(11): 2386−2395

DOI 10.16383/j.aas.c210425

Semantic Guided Feature Aggregation Network for Salient Object Detection

WANG Zheng-Wen[1] SONG Hui-Hui[1] FAN Jia-Qing[1] LIU Qing-Shan[1]

Abstract In the field of the salient object detection, the U-shaped structure has attracted much attention. However, the spatial and boundary details are ignored in the U-shaped salient object detection task. To address the issues, this paper proposes a salient object detection network that uses semantic information to guide feature aggregation, mainly including three modules: Mixing attention module (MAM), enlarged receptive field module (ERFM) and multi-level aggregation module (MLAM). Specifically, utilizes ERFM to process the low-level features, so that it can increase the receptive field while retaining the original edge details to obtain richer spatial context information. Then, utilize the MAM to process the high-level features to enhance its semantic representation, which guides the feature aggregation in the decoding process. Finally, MLAM is employed to effectively combine the multi-level features for the predicted salient map. This paper conducts extensive experiments on 6 benchmarks, and the results have proved that the method can effectively locate salient object and refine edge details.

Key words Salient object detection, mixing attention, multi-level aggregation, deep learning

Citation Wang Zheng-Wen, Song Hui-Hui, Fan Jia-Qing, Liu Qing-Shan. Semantic guided feature aggregation network for salient object detection. *Acta Automatica Sinica*, 2023, **49**(11): 2386−2395

显著性目标检测[1−5] 能够对图像中在视觉上最与众不同的对象或区域进行检测与识别. 目前, 显著性目标检测已经成功地作为许多计算机视觉领域任务的预处理过程, 包括目标跟踪[6]、物体识别[7]、语义分割[8] 等.

传统方法[9−10] 大多依靠颜色、纹理等手工特征或者启发式先验来捕获图像局部细节和全局上/下文. Goferman 等[9] 提出一种基于上/下文感知的方法, 对目标周围的不同区域均进行检测, 并最终基于四个心理学原理简单生成了显著性图. Yan 等[10] 设计了一个分层模型, 能够对显著信息进行层次分析, 并将不同层次的输出进行组合得到最终结果. 尽管上述算法取得了一定的成功, 但是由于缺乏高级语义信息的参与, 在复杂场景中检测显著物体的能力受到了很大限制.

近年来, 卷积神经网络得到快速发展. 例如文献 [11−13] 的卷积神经网络所具备的金字塔结构,

收稿日期 2021-05-17 录用日期 2021-10-18
Manuscript received May 17, 2021; accepted October 18, 2021
国家自然科学基金 (61872189, 61532009), 江苏省自然科学基金 (BK20191397), 江苏省"六大人才"高峰项目 (XYDXX-015) 资助
Supported by National Natural Science Foundation of China (61872189, 61532009), Natural Science Foundation of Jiangsu Province (BK20191397), and "Six Talent Peaks" Project of Jiangsu Province (XYDXX-015)
本文责任编委 陈龙
Recommended by Associate Editor CHEN Long
1. 南京信息工程大学江苏省大气环境与装备技术协同创新中心 江苏省大数据分析技术重点实验室 南京 210044
1. Jiangsu Key Laboratory of Big Data Analysis Technology, Collaborative InnovationCenter on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044

能够在较浅层拥有丰富的低层边缘细节特征, 而较深层则包含了更多语义信息, 更擅长定位显著物体的确切位置. 基于上述先验, 大量基于卷积神经网络的深度模型被相继提出. Hou 等[11] 对编码过程中每个阶段都引入了跳跃连接, 对特征图进行多层次多角度的聚合连接, 输出精确的结果. Li 等[14] 将粗纹理的显著图作为前景信息, 将图像边界的超像素值作为背景信息, 并将两者结合, 得到最终的结果. Qin 等[15] 设计了一种嵌套的 U 型结构, 融合了不同感受野大小的特征, 能够捕捉更多的上/下文信息. 在这些方法中, U 型结构由于能够通过在基础的分类网络上建立自上而下的路径来构建丰富的特征图, 而受到了最多的关注.

尽管上述方法相对于传统方法已经取得了很大进步, 但是还有很大改进空间. 首先, 在 U 型结构的解码过程中, 高层语义信息逐渐传递到较浅层, 虽然较浅层获得了显著物体的语义信息, 但是位置信息同时也被稀释, 造成最终输出的预测图中并不是当前图像中最显著部分, 丢失了显著物体准确的空间定位; 其次, 低层特征拥有丰富的边界信息, 但是由于在网络的较浅层, 无法获得较大感受野, 此时如果只是简单地融合高层特征与低层特征, 是无法精确地捕捉图片中显著物体边界的, 尤其是小目标. 因此, 本文考虑在增大低层特征感受野, 提高其表征力后, 将其送入到高效的特征聚合模块中, 以此来细化显著物体的边缘.

针对上述问题, 本文研究了如何在 U 型结构中通过高效的特征融合解决这些问题. 本文主要贡献包括以下 3 个方面: 1) 混合注意力模块 (Mixing attention module, MAM) 对来自第 5 个残差层的特征利用注意力机制进行显著性增强, 得到更加关注显著物体的语义特征, 同时为了解决解码过程中显著物体位置信息被不断稀释的问题, 将其作为整个解码过程中的语义指导, 不断指导解码过程中的特征聚合, 生成更加具体的显著性图. 2) 增大感受野模块 (Enlarged receptive field module, ERFM) 可以对来自低层的特征进行处理. 低层特征的边缘细节相当丰富, 但受限于感受野, 无法获得更加全局的信息. 因此, 考虑加入 ERFM, 可以在保留原有边缘细节的同时, 获得更大的感受野, 增强语义信息. 3) 多层次聚合模块 (Multi-level aggregation module, MLAM) 是对来自经过上述 2 个模块生成特征进行高效聚合, 以级联方式不断提取特征中的显著部分, 细化显著物体的边缘细节, 生成最终的显著图. 具体结构如图 1 所示.

1　相关工作

近年来, 大量基于全卷积神经网络深度模型[16−17] 被相继提出, 受益于全卷积神经网络强大的特征提取能力, 基于深度学习的方法已经在性能方面超越了大多数基于手工特征的传统方法. 文献 [18] 详细总结了传统方法. 本文主要讨论基于深度学习的显著性目标检测算法.

1.1　基于注意力机制的显著性目标检测

注意力机制具有很强的特征选择能力, 能够将特征信息进行深度整合, 使得网络更加去关注所需的语义信息. 根据加工域的不同, 注意力机制可以分为空间域注意力和通道域注意力两类, 其中空间注意力模块旨在捕获特征图中任意两个空间位置之间的空间依赖性, 通道注意力模块旨在捕获任意两个通道之间的通道依赖性. 因此, 许多学者利用注意力机制进行显著性目标检测. Zhang 等[19] 提出一种渐近注意力引导网络的显著性方法, 在解码阶段, 级联多个注意力模块渐近地生成最终结果. Zhao 等[20] 考虑到不同层次的特征所具备的信息并不相同, 因此, 对来自不同层次的特征, 分别设计了不同角度的注意力模块, 并对多个结果进行融合, 得到最终的结果. Chen 等[21] 提出一种反向注意网络, 将粗糙

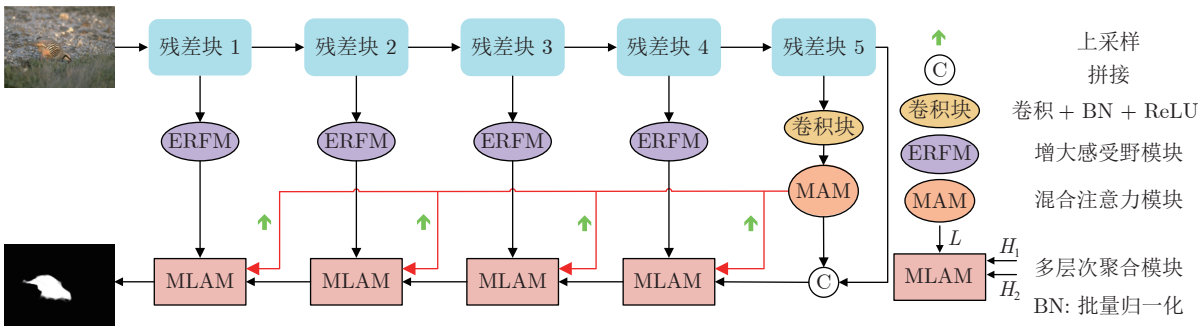


图 1　网络结构图

Fig. 1　Network structure diagram

的预测图反馈到中间特征层, 希望网络可以补全缺失的显著部分. Wang 等[22] 设计了一个金字塔注意力模块, 通过考虑多尺度注意力来增强显著特征的表征力. 上述方法都是对注意力机制的有效使用, 本文方法需要生成更加关注显著物体语义信息的高层特征, 利用注意力机制可以取得很好效果.

1.2 基于特征聚合的显著性目标检测

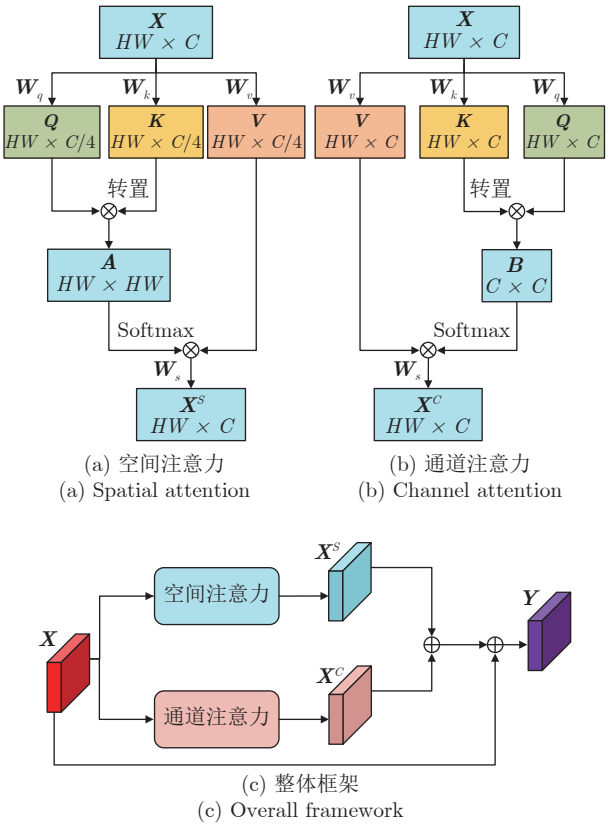
大多数对特征进行聚合的方法都是采用编码–解码的框架, 其中编码器用于提取多尺度特征, 解码器用于聚合特征以生成不同级别的上/下文信息. Wu 等[23] 对深层特征进行优化, 提高其表征力, 并利用双分支结构对特征进行聚合, 生成细化后的结果. Deng 等[24] 设计一种循环策略, 不断聚合来自不同层次的特征, 对网络进行细化, 增强显著信息. Wang 等[25] 提出一个特征打磨模块, 通过重复使用该模块, 对特征不断细化, 聚合来自不同层次的特征, 得到最终结果. 上述方法都探索了高效的特征聚合方法, 虽然有一定效果, 但是对于空间细节的捕捉仍然不够, 并且在解码过程中, 由于缺少高级语义的指导, 导致预测出的显著物体位置出现了偏移. 本文针对上述问题, 设计了多层次聚合模块, 使其能够在高级语义的指导下, 精确地定位显著物体, 并且通过级联多个、多层次聚合模块, 可以实现对边缘细节的细化.

2 本文方法

如图 1 所示, 本文建立一个编码–解码结构. 首先, 选用 ResNeXt101 作为特征提取器, 提取图片的各层特征; 其次, 利用 MAM 生成一个全局语义特征, 来引导解码过程, 通过上采样、卷积和元素累加等操作, 将全局语义融合到解码器的各层特征中; 接着, 编码过程中生成的各级特征通过 ERFM 后, 生成具备更多边界信息的低层特征; 最后, 将各级特征一起送入 MLAM 进行特征的有效聚合, 通过级联方式生成最终的显著性图.

2.1 混合注意力模块

图片送入网络中, 经过编码后, 会生成一系列具备不同信息的特征. 最高层的特征具备最强的语义表征能力, 并且在解码过程中, 逐渐与低层特征进行融合, 最终得到显著图. 但是, 直接将这种语义信息进行解码融合, 会造成许多显著性细节的丢失, 原因在于高层特征的不同通道和不同空间位置对显著性计算的贡献是不同的. 具体地, 不同通道对同一对象会有不同响应, 而同一通道的不同空间位置也会包含不同的对象. 受文献 [26] 启发, 本文设计了混合注意力模块, 该模块分为通道注意力机制和空间注意力机制两部分, 用来捕捉不同通道和不同空间位置中最显著的部分, 利用这些最显著的语义信息, 对高层特征进行有效增强, 得到更具鲁棒性的全局语义特征. MAM 模块结构见图 2.



(a) 空间注意力 (a) Spatial attention

(b) 通道注意力 (b) Channel attention

(c) 整体框架 (c) Overall framework

图 2 混合注意力模块

Fig. 2 Mixing attention module

2.1.1 空间注意力机制

对于从残差块 5 中提取的高层特征, 首先, 将其宽、高维度展开成一维向量并进行转置, 得到二维矩阵 $\boldsymbol{X} \in \mathbf{R}^{HW \times C}$, C 是该特征的通道数, H 和 W 分别是高和宽, HW 为高与宽相乘的数量. 然后, 经过 3 个并行的全连接层 $\boldsymbol{W}_q$、$\boldsymbol{W}_k$ 和 $\boldsymbol{W}_v$ 对通道进行降维, 分别得到 $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q$、$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k$、$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v$ 三个矩阵. 接着, 利用 $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ 得到相关性矩阵, 其中, $\boldsymbol{A}_{ij}$ 代表 $\boldsymbol{Q}$ 中第 i 行与 $\boldsymbol{K}$ 中第 j 行的内积, 即两个不同空间位置处向量的相关性. 并且对相关性矩阵 $\boldsymbol{A}$ 的每一行利用 Softmax 函数进行归一化, 约束到 (0, 1) 内. 最后, 将相关性矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{V}$ 相乘, 并且经过一个全连接层 $\boldsymbol{W}_s$ 对通道维度进行恢复, 得到空间显著性增强后的特征图 $\boldsymbol{X}^S = \boldsymbol{A}\boldsymbol{V}\boldsymbol{W}_s$, 最终的特征表达式为:

$$\boldsymbol{X}^{S}=\sigma\left(\boldsymbol{X}\boldsymbol{W}_{q}(\boldsymbol{X}\boldsymbol{W}_{k})^{\mathrm{T}}\right)\boldsymbol{X}\boldsymbol{W}_{v}\boldsymbol{W}_{s} \tag{1}$$

式中，$\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbf{R}^{C\times C/4}$，$\boldsymbol{W}_s \in \mathbf{R}^{(C/4)\times C}$，$\sigma(\cdot)$ 为 Softmax 函数.

2.1.2　通道注意力机制

通道维度的操作与上述类似, 也是对残差块 5 提取的特征先沿着宽、高维度展开成一维向量并转置, 得到 $\boldsymbol{X}\in\mathbf{R}^{HW\times C}$ 经过三个全连接层, 输出 $\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_q$, $\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_k$, $\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_v$. 考虑到降维会带来过多的信息损失, 因此本文算法没有对通道进行降维. 然后, 通过 $\boldsymbol{B}=\boldsymbol{K}^{\mathrm{T}}\boldsymbol{Q}$ 得到相关性矩阵, 其中 $\boldsymbol{B}_{ij}$ 代表了 $\boldsymbol{K}$ 中第 i 列与 $\boldsymbol{Q}$ 中第 j 列的内积, 即两个不同通道向量的相关性. 同样, 需要对相关性矩阵 $\boldsymbol{B}$ 的每一列利用 Softmax 函数进行归一化, 约束到 (0, 1) 内. 最后, 将 $\boldsymbol{V}$ 与 $\boldsymbol{B}$ 相乘且经过一个全连接层 $\boldsymbol{W}_s$, 得到通道显著性增强后的特征图 $\boldsymbol{X}^{\boldsymbol{C}}=\boldsymbol{V}\boldsymbol{B}\boldsymbol{W}_s$, 最终的特征表达式为:

$$\boldsymbol{X}^{\boldsymbol{C}}=\boldsymbol{X}\boldsymbol{W}_{v}\sigma\left((\boldsymbol{X}\boldsymbol{W}_{k})^{\mathrm{T}}\boldsymbol{X}\boldsymbol{W}_{q}\right)\boldsymbol{W}_{s} \tag{2}$$

式中, $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v, \boldsymbol{W}_s \in \mathbf{R}^{C\times C}$. 最后合并这两个分支的输出. 考虑到残差结构的影响, 本文将合并后的特征与输入 $\boldsymbol{X}$ 进行相加, 生成最终特征图 $\boldsymbol{Y}\in\mathbf{R}^{HW\times C}$:

$$\boldsymbol{Y}=\boldsymbol{X}^{C}\oplus\boldsymbol{X}^{S}\oplus\boldsymbol{X} \tag{3}$$

式中, "$\oplus$"表示元素级的特征图相加. $\boldsymbol{Y}$ 在经过转置并且将维度展开恢复后, 送入到后续的模块中.

2.2　增大感受野模块

低层特征的边缘细节非常丰富, 但由于下采样的次数有限, 感受野相对受限, 无法捕捉全局的信息. 在解码过程中, 如果仅仅是简单利用低层特征, 虽然边缘的细节信息得到利用, 但并没有充分挖掘特征的空间细节. 受文献 [27] 启发, 本文设计如图 3 所示的增大感受野模块. 低层特征经过该模块后, 在保证边缘细节不丢失的前提下, 扩大了感受野, 具备了更多空间细节.

首先, 对于特征 $M\in\mathbf{R}^{C\times H\times W}$, 设计四个并行分支 $(l_i, i=1, 2, 3, 4)$, 其中 l_1 采用一个 1×1 卷积, 剩下的三个分支均采用 3×3 卷积, 并且对这三个分支设置不同的扩张率. 根据低层特征分辨率的不同设置不同的扩张率: 对于分辨率较低的特征设置较小的扩张率, 对于分辨率较高的特征设置较大的扩张率. 本文最大的扩张率设置为 $d=5, 8, 11$, 并随着特征图的缩小而不断缩小 (具体设置见第 3.6 节). 然后, 对四个分支输出进行通道维度拼接, 并利用一个 1×1 卷积得到融合后的特征.

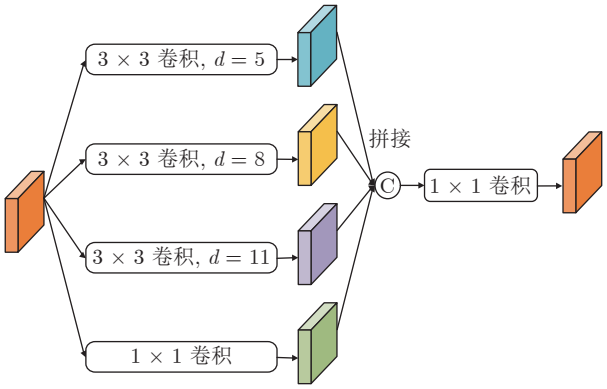


图 3　增大感受野模块

Fig. 3　Enlarged receptive field module

2.3　多层次聚合模块

在解码过程中, 高效利用每一层的特征尤为关键. 以前的研究只对高层特征与低层特征进行简单的拼接融合, 得到的结果非常粗糙. 因此, 本文设计了多层次聚合模块, 对来自不同层、不同空间尺度的特征进行有效聚合. 该模块的输入分为 MAM 生成的语义特征 H_1, 经过 ERFM 增强后的低层特征 L 和当前进行解码的特征 H_2 三个部分. 图 4 是多层次聚合模块示意图.

整个聚合过程分为 2 个阶段: 第 1 阶段是语义特征对当前解码特征的指导融合. 首先让 H_1 经过两个并行的 1×1 卷积, 第 1 个分支与 H_1 在通道维度上进行拼接融合后, 与第 2 个分支的结果相加完成第 1 次融合, 得到高层特征 H:

$$H=f_{conv}(f_{cat}(f_{conv}(H_1), H_2))\oplus f_{conv}(H_1) \tag{4}$$

式中, $f_{conv}(\cdot)$ 指卷积操作, $f_{cat}(\cdot)$ 指通道的拼接操作. 第 2 阶段是第 1 阶段融合得到的高层特征 H 与经过 ERFM 增强后的低层特征 L 的聚合. 此阶段分为自下而上和自上而下两个并行分支. 自下而上是 H 向 L 的聚合, 此阶段 L 不变, H 经过一次上采样和一个 1×1 卷积后与 L 进行通道维度的拼接, 得到聚合图 $X^{h\to l}$:

$$X^{h\to l}=f_{conv}(f_{cat}(L, f_{up}(H))) \tag{5}$$

式中, $f_{up}(\cdot)$ 指上采样操作. 自上而下是 L 向 H 的聚合, 此阶段 H 不变, L 首先经过一个并行的池化操作, 其中最大池化可以提取特征中响应值较大的信息即特征中所包含的显著信息, 平均池化可以得到特征的全局信息. 经过并行池化后, 特征 L 具备更强的表征力, 并且与 H 有相同的空间尺寸, 此时将其与特征 H 在通道维度上进行拼接, 并利用 1×1 卷积完成融合. 然后, 对其进行上采样, 得到最终的 $X^{l\to h}$:

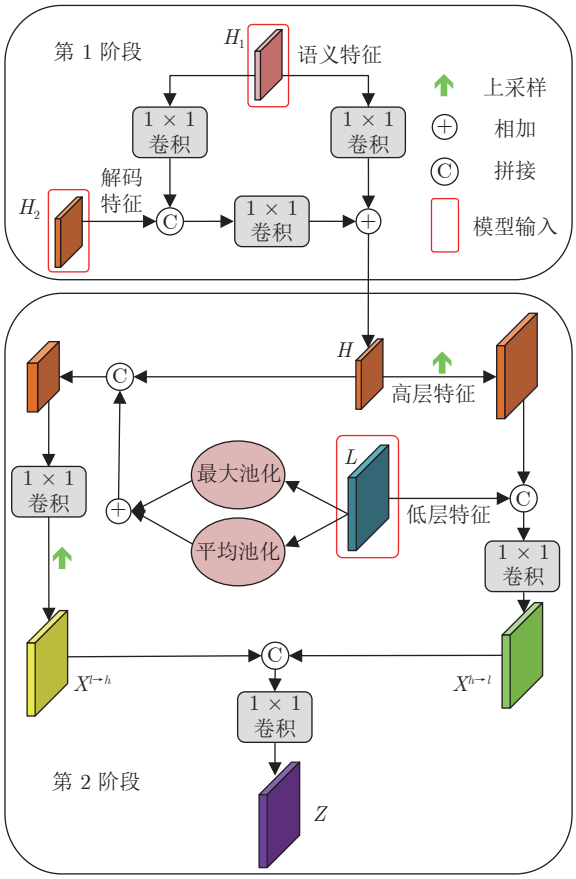


图 4 多层次聚合模块

Fig. 4 Multi-level aggregation module

$$X^{l\to h} = f_{up}(f_{conv}(f_{cat}(H,\ f_{avg}(L) + f_{\max}(L)))) \quad (6)$$

式中, $f_{\max}(\cdot)$ 和 $f_{avg}(\cdot)$ 分别代表最大池化和平均池化操作. 最后, 对两个分支得到的聚合特征也进行一次聚合:

$$Z = f_{conv}(f_{cat}(X^{l\to h},\ X^{h\to l})) \quad (7)$$

3 实验设置与结果分析

3.1 训练细节

本文代码是在 Pytorch1.5.0 框架下完成, 并且使用 1 张 GeForce GTX2080Ti GPU 进行训练. 训练数据使用 DUTS[28] 数据集中 10 553 张图片. 使用 Adam[29] 优化器进行优化, 初始学习率设置为 1×10^4, 并且在每训练完成两个周期后衰减一半, 批量大小为 8. 使用 ResNeXt101 作为特征提取器提取各层特征, 并加载在 ImageNet 上预训练的分类权重, 作为初始权重. 为了减少过拟合的影响, 在训练阶段, 对图片进行了随机翻转和遮挡, 并将图片缩放到 320 × 320 像素后, 将其随机裁剪为 288 × 288 像素, 输入到网络中进行训练; 测试阶段, 仅将图片缩放到 288 × 288 像素后, 输入到网络中进行测试.

3.2 数据集

本文在 6 个基准数据集上进行实验, 包括 DUTS-TE[28]、DUT-OMRON[30]、ECSSD[31]、HKU-IS[32]、PASCAL-S[33] 和 SOD[34]. 其中, DUTS-TE 与训练集的 10 553 张图片同属一个数据集, 包含 5 019 张测试图片. DUT-OMRON 是最具有挑战性的数据集, 包含 5 188 张图片, 该数据集的难点在于背景非常复杂, 对网络预测显著目标有很大干扰作用. ECSSD 相对简单, 由 1 000 张图片组成, 其中显著目标形状与外观有很大差异. HKU-IS 包含 4 447 张图片, 其中包含多个具有不同类别或外观的显著物体. PASCAL-S 包含 850 张图片, 图片中物体之间会出现很大程度的重叠. SOD 只有 300 张图片, 但场景的复杂多变, 带来很大挑战.

3.3 评价指标

本文使用平均绝对误差 (Mean absolute error, MAE)、F_β (F-measure) 和 S_m (Structure measure) 作为评价指标.

1) MAE 计算预测的显著图与真实标签之间的差异:

$$\mathrm{MAE} = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|\boldsymbol{P}(x,\ y) - \boldsymbol{G}(x,\ y)| \quad (8)$$

式中, $\boldsymbol{P}$ 指预测的显著图, $\boldsymbol{G}$ 指真实标签值.

2) F_β 是一种经典且有效的测量指标, 通过对查准率 (Precision) 与查全率 (Recall) 设置不同的权重来计算:

$$F_\beta = \frac{(1+\beta^2) \times \mathrm{Recall} \times \mathrm{Precision}}{\beta^2 \times \mathrm{Precision} + \mathrm{Recall}} \quad (9)$$

式中, β^2 设置为 0.3.

3) S_m 用来考虑预测的显著图与真实标签之间的全局和局部的结构相似性, 该指标的详细介绍见文献 [35].

3.4 损失函数

本文使用标准的二元交叉熵损失作为训练的损失函数:

$$L_{bce} = -\sum_{(x,\ y)}[\boldsymbol{G}(x,\ y)\lg(\boldsymbol{P}(x,\ y)) + (1 - \boldsymbol{G}(x,\ y))\lg(1 - \boldsymbol{P}(x,\ y))] \quad (10)$$

3.5 与其他算法的性能比较

本文与最新 10 种基于深度学习的方法进行比

较, 包括 U2Net[15]、PAGR[19]、RAS[21]、CPD[23]、DGRL[36]、MLMS[37]、PoolNet[38]、AFNet[39]、BASNet[40] 和 ITSD[41]. 为了指标的公平性, 所有指标均在同一评测代码下进行评测, 并且所有用于评测的显著图均从作者发布的模型中得出.

3.5.1　定量分析

表 1、表 2 和表 3 分别列出了各算法的 F_β、MAE 和 S_m 评价指标结果. 本文方法在 3 项指标中均表现优异. 由表 1 和表 3 可以看出, 本文方法在指标 F_β和 S_m上大幅领先于其他方法, 即便是次优的 ITSD 算法, 在较难的数据集 DUT-OMRON 中, 本文也在 F_β 指标上领先其 0.003, S_m 指标领先其 0.007. 这主要得益于本文多层次聚合模块能够最大限度地保留显著物体的空间信息和边界细节. 对于表 2 中 MAE 指标, 本文方法也仅在相对较难的 3 个数据集上表现稍有不足, 但与第 1 名的差距是非常小的, 基本保持在 0.001 ~ 0.002 之间. 图 5 是各方法的查准率–查全率曲线图, 加粗实线是本文方法, 由图 5 可以看出, 本文算法性能的优越性.

3.5.2　定性分析

图 6 是本文方法与其他 10 种方法的显著性图. 由图 6 可以看出, 本文方法对显著信息的捕捉明显更强. 在第 1 行中, 即便是指标与本文最接近的 ITSD 也将座椅当作显著物体, 但是在人类的视觉效果上, 明亮的灯与背景的区分度更大, 本文方法因为有全局语义指导特征聚合, 可以捕捉到壁灯的

表 1　不同方法的 F_β 指标结果比较

Table 1　Comparison of F_β values of different models

数据集	本文方法	PAGR	RAS	DGRL	CPD	MLMS	PoolNet	AFNet	BASNet	U2Net	ITSD
ECSSD	**0.951**	0.924	0.921	0.921	0.936	0.930	0.944	0.935	0.942	**0.951**	0.947
DUT-OMRON	**0.827**	0.771	0.786	0.774	0.794	0.793	0.808	0.797	0.805	0.823	0.824
PASCAL-S	**0.873**	0.847	0.837	0.844	0.866	0.858	0.869	0.868	0.854	0.859	0.871
HKU-IS	**0.937**	0.919	0.913	0.910	0.924	0.922	0.933	0.923	0.928	0.935	0.934
DUTS-TE	**0.888**	0.855	0.831	0.828	0.864	0.854	0.880	0.862	0.860	0.873	0.883
SOD	0.873	0.838	0.810	0.843	0.850	0.862	0.867	—	0.851	0.861	**0.880**

注: F_β 值越大越好, 加粗数字为最优结果, 加下划线数字为次优结果.

表 2　不同方法的 MAE 指标结果比较

Table 2　Comparison of MAE values of different models

数据集	本文方法	PAGR	RAS	DGRL	CPD	MLMS	PoolNet	AFNet	BASNet	U2Net	ITSD
ECSSD	**0.034**	0.064	0.056	0.043	0.040	0.038	0.039	0.042	0.037	**0.034**	0.035
DUT-OMRON	0.058	0.071	0.062	0.062	0.056	0.060	0.056	0.057	0.056	**0.054**	0.061
PASCAL-S	**0.065**	0.089	0.104	0.072	0.074	0.069	0.075	0.069	0.076	0.074	0.072
HKU-IS	0.032	0.047	0.045	0.036	0.033	0.034	0.033	0.036	0.032	**0.031**	**0.031**
DUTS-TE	0.042	0.053	0.060	0.049	0.043	0.045	**0.040**	0.046	0.047	0.044	0.041
SOD	**0.093**	0.145	0.124	0.103	0.112	0.106	0.100	—	0.114	0.108	0.095

注: MAE 值越小越好.

表 3　不同方法的 S_m 指标结果比较

Table 3　Comparison of S_m values of different models

数据集	本文方法	PAGR	RAS	DGRL	CPD	MLMS	PoolNet	AFNet	BASNet	U2Net	ITSD
ECSSD	**0.932**	0.889	0.893	0.906	0.915	0.911	0.921	0.914	0.916	0.928	0.925
DUT-OMRON	**0.847**	0.775	0.814	0.810	0.818	0.817	0.836	0.826	0.836	**0.847**	0.840
PASCAL-S	**0.865**	0.749	0.795	0.869	0.844	0.849	0.845	0.850	0.838	0.844	0.859
HKU-IS	**0.930**	0.887	0.887	0.897	0.904	0.901	0.917	0.905	0.909	0.916	0.917
DUTS-TE	**0.873**	0.838	0.839	0.842	0.867	0.856	0.883	0.866	0.853	0.861	0.872
SOD	0.808	0.720	0.764	0.771	0.771	0.780	0.795	—	0.772	0.786	**0.809**

注: S_m 值越大越好.

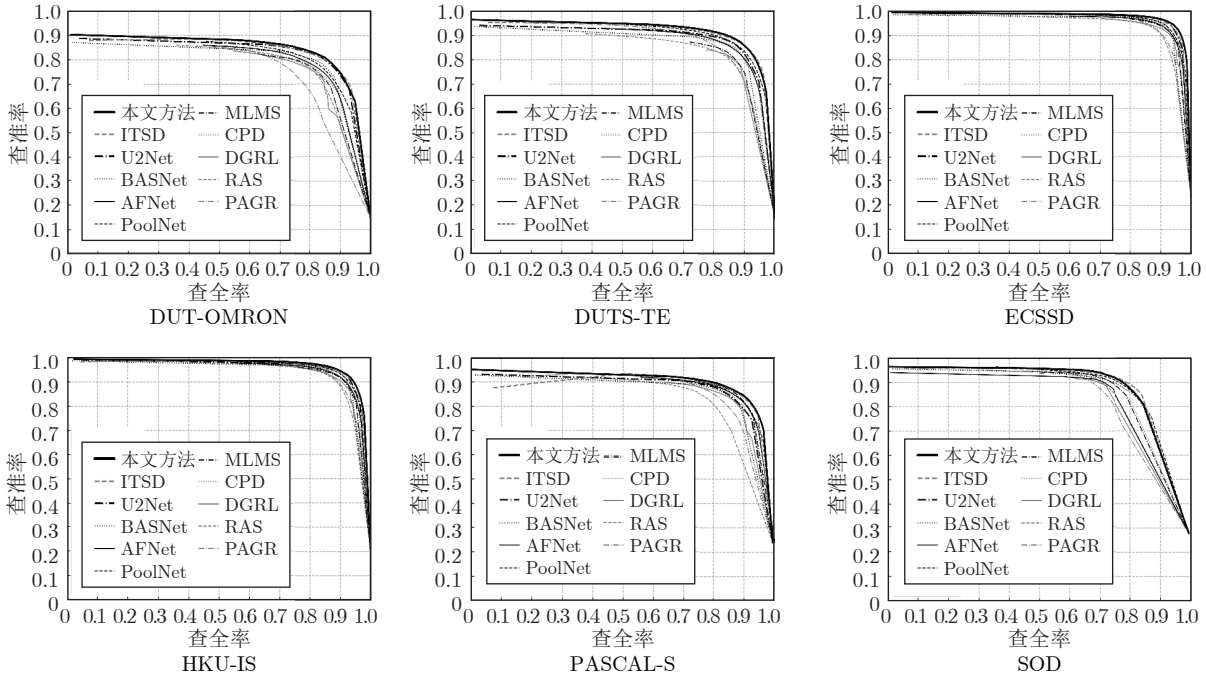


图 5 不同算法的查准率–查全率曲线示意图

Fig. 5 Comparison of precision–recall curves of different methods

图 6 不同算法的显著性图

Fig. 6 Salient maps of different methods

显著信息. 在第 2 行中, 围绳与人之间有很多交叉, 即便是当前性能较好的方法也并没有将目标完整地识别出来, 而本文方法由于对低层特征进行了感受野增强, 可以捕捉目标周围更多的上/下文信息, 能

够将全部目标识别出来, 但同时也存在围绳部分被识别为人的问题. 综上所述, 本文算法对于复杂背景下的物体边界并不能很好地细化. 但对于背景较简单的物体 (如第 5 行和最后 1 行), 本文均能很好地预测出边界轮廓.

3.6　消融实验

表 4 是在数据集 ECSSD 上针对各模块的消融实验结果: 1) 混合注意力模块. 由表 4 第 3 行可知, 当缺少混合注意力模块时, MAE 指标上升了 0.008, 由此可见, 利用该模块生成的全局语义特征引导特征聚合, 能够大幅提升聚合性能; 2) 增大感受野模块. 由表 4 第 4 行可知, MAE 指标上升了 0.005, 主要是因为缺少了感受野增强, 没有充分提取低层特征的空间上/下文信息, 不利于细化边界; 3) 多层次融合模块. 由表 4 第 2 行可知, 当用简单的上采样和相加操作代替该模块时, MAE 上升了 0.011, 说明多层次融合模块聚合方式非常高效.

表 4　消融实验结果
Table 4　Results of ablation experiment

MAM	ERFM	MLAM	MAE/ F_β
—	—	✓	0.049/0.935
✓	✓	—	0.045/0.937
—	✓	✓	0.042/0.942
✓	—	✓	0.039/0.944
✓	✓	✓	**0.034/0.951**

注: MAE 值越小越好, 加粗字体为最优结果, "✓"为使用指定模块.

表 5 是对 ERFM 模块中, 不同扩张率设置的对比实验结果. 表 5 中不同设置组合从左向右依次对应不同分辨率的特征图 (见图 3), 即左边第 1 组扩张率对应分辨率最大的特征图, 最后 1 组扩张率对应分辨率最小的特征图. 表 5 的第 4 行是本文方法的设置. 由表 5 第 1 行可以看出, 当扩张率全部设置为 (1, 3, 5) 时, 与本文方法相比, MAE 上升了 0.005, 而随着本文方法对分辨率较高的特征图分配更大的扩张率时, MAE 的指标不断降低. 实验结果表明, 在本文方法中, 扩张率的选择是有效的.

表 6 是对 MLAM 模块第 2 阶段中, 自上而下和自下而上两个分支在数据集 ECSSD 上的消融实验结果. 由表 6 第 1 行可知, 当只使用自下而上分支时, 相比两个并行分支均使用时, MAE 上升了 0.007; 而只使用自上而下分支时, 上升了 0.006. 由此可见, 本文方法将两个分支并行使用的方式是有效的, 能够对精度有所提升.

表 5　ERFM 模块中, 不同扩张率设置的对比实验
Table 5　Comparative experiment of different dilation rate configurations in ERFM

扩张率的不同设置组合	MAE/ F_β
(1, 3, 5), (1, 3, 5), (1, 3, 5), (1, 3, 5)	0.039/0.946
(1, 3, 5), (1, 3, 5), (3, 5, 7), (1, 3, 5)	0.037/0.948
(1, 3, 5), (4, 6, 8), (3, 5, 7), (1, 3, 5)	0.036/0.950
(5, 8, 11), (4, 6, 8), (3, 5, 7), (1, 3, 5)	**0.034/0.951**

表 6　MLAM 模块中, 两个分支的消融实验
Table 6　Ablation experiment of two branches in MLAM

自下而上分支	自上而下分支	MAE/ F_β
✓	—	0.041/0.940
—	✓	0.040/0.946
✓	✓	**0.034/0.951**

表 7 是对 MAM 中, 注意力模块位置关系的消融实验结果. 前 2 行是将两个模块串联并考虑其先后位置, 第 3 行是两个模块并行即本文方法. 当通道注意力位置在前时, 与本文方法相比, MAE 上升了 0.002; 当空间注意力位置在前时, MAE 上升了 0.004. 该实验结果验证了本文将两个模块设置成并行的有效性.

表 7　MAM 模块中, 注意力模块位置关系的消融实验
Table 7　Ablation experiment on the position relationship of attention module in MAM

注意力模块之间的位置关系	MAE/ F_β
通道注意力在前	0.036/0.947
空间注意力在前	0.038/0.944
并行放置 (本文方法)	**0.034/0.951**

4　结束语

本文提出一种基于语义引导特征聚合的显著性目标检测算法, 主要包括混合注意力模块、增大感受野模块和多层次融合模块 3 个模块. MAM 能够生成更佳的语义特征, 用来指导解码过程中的特征融合, 使得聚合的特征能够更好地定位显著物体; ERFM 能够丰富低层特征所具备的上/下文信息, 并将增强后的特征输入到 MLAM 中; MLAM 利用 MAM 生成的语义信息, 对当前解码的特征和 ERFM 输出的低层特征进行指导融合, 并最终以级联方式逐步恢复边界细节, 生成最终的显著图. 本文与目前流行的 10 种算法在 6 个基准数据集上进行了实验比较, 由可视化图 6 可以看出, 本文算法能

够有效地保留显著物体的空间位置信息, 并且边缘也得到了很好细化. 实验结果也验证了本文算法具有领先性能.

References

1 Donoser M, Urschler M, Hirzer M, Bischof H. Saliency driven total variation segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 817−824

2 Wei J, Wang S, Huang Q. F3Net: Fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: Spring, 2020. 12321−12328

3 Li Dong-Min, Li Jing, Liang Da-Chuan, Wang Chao. Multiple salient objects detection using multi-scale prior and deep features. *Acta Automatica Sinica*, 2019, **45**(11): 2058−2070
(李东民, 李静, 梁大川, 王超. 基于多尺度先验深度特征的多目标显著性检测方法. 自动化学报, 2019, **45**(11): 2058−2070)

4 Xu Wei, Tang Zhen-Min. Exploiting hierarchical prior estimation for salient object detection. *Acta Automatica Sinica*, 2015, **41**(4): 799−812
(徐威, 唐振民. 利用层次先验估计的显著性目标检测. 自动化学报, 2015, **41**(4): 799−812)

5 Yang Sai, Zhao Chun-Xia, Xu Wei. A novel salient object detection method using bag-of-features. *Acta Automatica Sinica*, 2016, **42**(8): 1259−1273
(杨赛, 赵春霞, 徐威. 一种基于词袋模型的新的显著性目标检测方法. 自动化学报, 2016, **42**(8): 1259−1273)

6 Hong S, You T, Kwak S, Han B. Online tracking by learning discriminative saliency map with convolutional neural network. In: Proceedings of the 32nd International Conference on Machine Learning. Miami, USA: IMLS, 2015. 597−606

7 Ren Z, Gao S, Chia L, Tsang I W. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014, **24**(5): 769−779

8 Wang X, You S, Li X, Ma H. Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1354−1362

9 Goferman S, Zelnik-Manor L, Tal A. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(10): 1915−1926

10 Yan Q, Xu L, Shi J, Jia J. Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. America, Portland: IEEE, 2013. 1155−1162

11 Hou Q, Cheng M M, Hu X, Borji A, Tu Z, Torr P. Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 5300−5309

12 Luo Z, Mishra A, Achkar A, Eichel J, Jodoin P M. Non-local deep features for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 6609−6617

13 Zhang P, Wang D, Lu H, Wang H, Ruan X. Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 202−211

14 Li X, Zhao L M, Wei L, Yang M, Wu F, Zhuang Y T, et al. DeepSaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, 2016, **25**(8): 3919−3930

15 Qin X, Zhang Z, Huang C, Dehghan M, Jagersand M. U2Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 2020, **106**: Article No. 107404

16 Pang Y, Zhao X, Zhang L, Lu H. Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 9413−9422

17 Wei J, Wang S, Wu Z, Su C, Huang Q, Tian Q. Label decoupling framework for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 13025−13034

18 Borji A, Cheng M M, Jiang H, Li J. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 2015, **24**(12): 5706−5722

19 Zhang X N, Wang T T, Qi J Q, Lu H C, Wang G. Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 714−722

20 Zhao T, Wu X Q. Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 3085−3094

21 Chen S, Tan X, Wang B, Hu X. Reverse attention for salient object detection. In: Proceedings of the IEEE Europeon Conference on Computer Vision. Munich, Germany: IEEE, 2018. 234−250

22 Wang W, Zhao S, Shen J, Hoi S C, Borji A. Salient object detection with pyramid attention and salient edges. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 1448−1457

23 Wu Z, Su L, Huang Q. Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 3907−3916

24 Deng Z J, Hu X W, Zhu L, Xu X M, Qin J, Han G Q, et al. R3Net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press, 2018. 684−690

25 Wang B, Chen Q, Zhou M, Zhang Z, Jin X, Gai K. Progressive feature polishing network for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: Springer, 2020.

26 Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 3146−3154

27 Zhao H S, Shi J P, Qi X J, Wang X G, Jia J Y. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 6230−6239

28 Wang L J, Lu H C, Wang Y F, Feng M Y, Wang D, Yin B C, et al. Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 3796−3805

29 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980, 2014.

30 Yang C, Zhang L, Lu H C, Ruan X, Yang M. Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013. 3166−3173

31 Li X H, Lu H C, Zhang L, Ruan X, Yang M. Saliency detection via dense and sparse reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 2976−2983

32 Li G B, Yu Y Z. Visual saliency based on multi-scale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 5455−5463

33 Li Y, Hou X, Koch C, Rehg J M, Yuille A L. The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 280−287

34 Movahedi V, Elder J H. Design and perceptual validation of performance measures for salient object segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, USA: IEEE, 2010. 49−56

35 Fan D P, Cheng M M, Liu Y, Li T, Borji A. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 4548−4557

36 Wang T T, Zhang L, Wang S, Lu H C, Yang G, Ruan Y, et al. Detect globally, refine locally: A novel approach to saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3127−3135

37 Wu R, Feng M, Guan W, Wang D, Lu H, Ding E. A mutual learning method for salient object detection with intertwined multi-supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 8150−8159

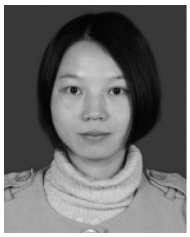
38 Liu J J, Hou Q, Cheng M M, Feng J, Jiang J. PoolNet: A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 3917−3926

39 Feng M Y, Lu H C, Ding E. Attentive feedback network for boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 1623−1632

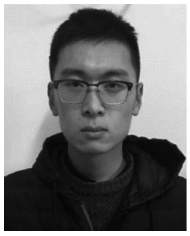
40 Qin X B, Zhang Z C, Huang C Y, Gao C, Dehghan M, Jagersand M. BASNet: Boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE, 2019. 7479−7489

41 Zhou H, Xie X, Lai J H, Chen Z, Yang L. Interactive two-stream decoder for accurate and fast saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 9141−9150

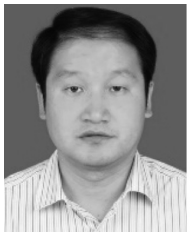
王正文 南京信息工程大学自动化学院硕士研究生. 主要研究方向为显著性目标检测, 深度学习.
E-mail: 20191223064@nuist.edu.cn
(**WANG Zheng-Wen** Master student at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers salient object detection and deep learning.)



宋慧慧 南京信息工程大学自动化学院教授. 主要研究方向为视频目标分割, 图像超分. 本文通信作者.
E-mail: songhuihui@nuist.edu.cn
(**SONG Hui-Hui** Professor at the School of Automation, Nanjing University of Information Science and Technology. Her research interest covers video object segmentation and image super-resolution. Corresponding author of this paper.)



樊佳庆 南京信息工程大学自动化学院硕士研究生. 主要研究方向为视频目标分割. E-mail: jqfan@nuaa.edu.cn
(**FAN Jia-Qing** Master student at the School of Automation, Nanjing University of Information Science and Technology. His main research interest is video object segmentation.)



刘青山 南京信息工程大学自动化学院教授. 主要研究方向为视频内容分析与理解. E-mail: qsliu@nuist.edu.cn
(**LIU Qing-Shan** Professor at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video content analysis and understanding.)