软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(2):953−968 [doi: 10.13328/j.cnki.jos.007443] [CSTR: 32375.14.jos.007443]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于边缘增强的宽解码器显著性目标检测方法*

童 彪, 宋晓宁, 华 阳, 张文杰, 吴小俊

(江南大学 人工智能与计算机学院, 江苏 无锡 214100)

通信作者: 宋晓宁, E-mail: x.song@jiangnan.edu.cn

摘 要: 当前, 显著性目标检测技术正在迅速发展, 但仍然存在一些问题亟待解决. 大多数现有的显著性目标检测方法在处理高分辨率图像任务时, 存在计算资源需求过高或者检测质量较差等问题. 其次, 许多现有算法采用的传统卷积操作缺乏针对性, 无法有效增强边缘细节特征, 导致边缘分割模糊不清. 为了在降低算力消耗的同时提高物体边缘分割质量, 并提升小尺度目标的检测性能, 提出了基于边缘增强的宽解码器显著性目标检测方法. 采用残差网络和 Swin Transformer 组合结构作为特征编码器, 以降低算力消耗. 并且将传统卷积替换为差分卷积模块, 通过多种不同类型的差分卷积并行使用, 从图像中提取了更加丰富的边缘信息. 设计了多尺度注意力模块, 对 4 层不同尺度特征进行注意力计算, 以更好地关注不同大小的目标. 此外, 采用含有大卷积核的多级宽解码器, 对融合特征进行长距离的上下文建模, 减少冗余信息, 进一步提升了网络的检测性能.

关键词: 显著性目标检测; 边缘增强; 差分卷积

中图法分类号: TP391

中文引用格式: 童彪, 宋晓宁, 华阳, 张文杰, 吴小俊. 基于边缘增强的宽解码器显著性目标检测方法. 软件学报, 2026, 37(2): 953–968. http://www.jos.org.cn/1000-9825/7443.htm

英文引用格式: Tong B, Song XN, Hua Y, Zhang WJ, Wu XJ. Salient Object Detection Method Based on Edge-enhanced Wide Decoder. Ruan Jian Xue Bao/Journal of Software, 2026, 37(2): 953–968 (in Chinese). http://www.jos.org.cn/1000-9825/7443.htm

Salient Object Detection Method Based on Edge-enhanced Wide Decoder

TONG Biao, SONG Xiao-Ning, HUA Yang, ZHANG Wen-Jie, WU Xiao-Jun

(School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214100, China)

Abstract: Salient object detection is developing rapidly. However, several critical challenges remain. Most existing methods struggle with high-resolution images due to either excessive computational demands or suboptimal detection quality. In addition, traditional convolutional operations commonly used in current algorithms lack targeted enhancement, resulting in inadequate edge detail extraction and blurred object boundaries. To address these limitations, this study proposes a salient object detection method based on an edge-enhanced wide decoder, which improves edge segmentation accuracy and enhances small-scale object detection while reducing computational overhead. A hybrid feature encoder combining a residual network and a Swin Transformer is employed to lower computational overhead. Traditional convolutions are replaced with a differential convolution module, where multiple types of differential convolutions are executed in parallel to extract richer edge information. A multi-scale attention module is incorporated to compute attention across four hierarchical feature layers, enabling better focus on objects of varying sizes. In addition, a multilevel wide decoder with large convolutional kernels is utilized to conduct long-range contextual modeling of fused features, effectively reducing redundant information and further boosting detection performance. Code will be released at https://github.com/wapitier/EEWDNet.

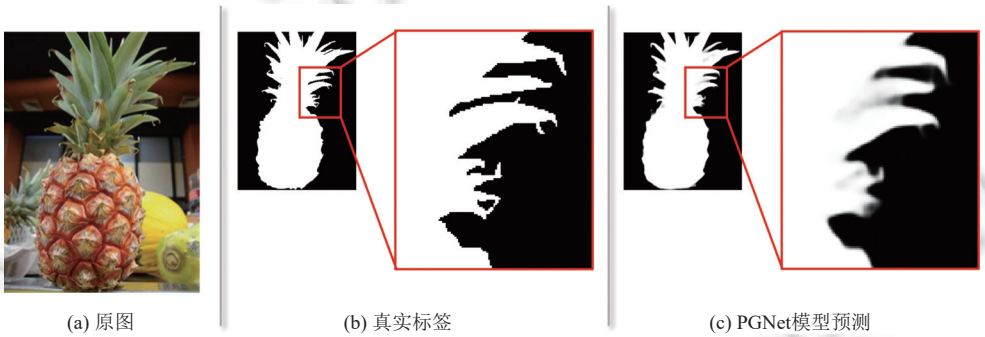
Key words: salient object detection; edge enhancement; difference convolution

* 基金项目: 国家重点研发计划 (2023YFF1105102, 2023YFF1105105); 国家社会科学基金 (21&ZD166); 国家自然科学基金 (61876072); 江苏省自然科学基金 (BK20221535)
收稿时间: 2024-04-14; 修改时间: 2024-06-21, 2025-02-20; 采用时间: 2025-04-04; jos 在线出版时间: 2025-11-13
CNKI 网络首发时间: 2025-11-14

随着信息技术的迅速发展, 智能终端设备的广泛应用使得图像和视频信息的获取变得更加便捷, 但也导致信息量呈指数级增长, 带来了处理大量冗余数据的挑战. 在这种背景下, 如何从海量信息中筛选出有效信息变得至关重要. 显著性目标检测具有以下 3 个方面的优势: 一是能够在图像或视频中有效分配计算资源, 使得对重要部分的处理更加集中和高效; 二是弥补了传统图像处理方法的不足, 能够提高检测的精度和效率; 三是通过定位和突出图像中最重要的目标, 所得到的检测结果能够符合人的理解认知, 使得识别更加直观和准确.

显著性目标检测[1]作为计算机视觉领域的研究热点, 具有多方面的研究价值. 在图像编辑中, 显著性目标检测可以用于自动裁剪或增强图像中的主要目标. 在自动驾驶领域, 显著性目标检测可以帮助车辆识别和跟踪道路上的行人、车辆等重要目标. 在医学影像分析中, 显著性目标检测可以用于辅助医生识别和定位疾病部位. 可以看出显著性目标检测技术具有广泛的应用前景, 对于提升图像处理、模式识别、智能交通等领域的智能化水平具有重要意义.

在研究初期, 传统的显著性目标检测方法往往针对低分辨率图像进行设计和优化, 主要目的是解决目标的定位问题. 然而随着摄像设备技术的进步, 高清摄像头的普及使得设备能够捕捉到更加清晰、细节丰富的图像. 这些高分辨率图像不仅能够提供更多的信息, 同时也为目标检测算法带来了更大的挑战. 在处理高分辨率图像时, 传统的显著性目标检测算法暴露出了一些不足之处. 例如, 部分算法将高分辨率图像下采样到固定小尺寸后进行识别, 在缩放过程中, 小尺度物体的有效像素变得更少, 使得模型对小目标的检测性能不佳. 此外, 针对大尺度物体, 传统算法无法很好地建模其全局信息, 导致在目标较大时难以进行准确的检测和定位. 另外, 高分辨率图像中的细节信息更加丰富, 传统算法在物体边缘分割和细节提取方面表现不佳, 经常出现边缘分割模糊、粘连和残缺等问题, 如图 1 所示, 导致检测结果的准确性和精度下降.



(a) 原图 (b) 真实标签 (c) PGNet模型预测

图 1 PGNet[2]模型的边缘分割缺陷

因此, 针对上述问题, 本文提出了基于边缘增强的宽解码器显著性目标检测方法. 在处理高分辨率图像时, 采用轻量化的残差网络[3]进行细节特征提取, 同时利用 Swin Transformer[4]处理下采样后的图像来提取语义信息. 这样一来, 既保证了计算量在合理可接受的范围内, 又能充分提取图像特征. 为确保细节特征与语义特征能正常融合互补, 引入了多尺度注意力模块, 将融合后的特征在不同尺度上重新修正, 抑制多余的背景噪声. 在解码器阶段, 采用内含大卷积核的多级宽解码器, 通过扩大原始感受野, 对特征进行更好的上下文建模, 从而提升模型的图像理解能力.

本文第 1 节介绍显著性目标检测的相关方法和研究现状. 第 2 节介绍本文提出的基于边缘增强的宽解码器显著性目标检测方法. 第 3 节通过对比实验验证所提模型的有效性. 第 4 节进行全文总结.

1 相关工作

1.1 显著性目标检测系统框架

常见的显著性目标检测系统框架如图 2 所示, 从图中可以看出, 显著性目标检测系统通常分为数据预处理、特征计算和结果输出这 3 个部分.

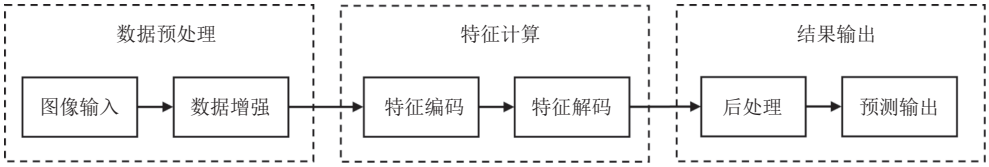


图 2　显著性目标检测系统框架

显著性目标检测系统的输入是一张或一系列图像, 这些图像可能来自摄像头、视频文件或其他图像源. 这些图像经过统一的缩放、旋转、翻转和归一化等数据增强[5]操作之后, 系统才可以使用显著性目标检测算法进行特征计算.

特征计算的过程通常包含特征编码和特征解码两个步骤. 作为显著性目标检测的第 2 步, 其目的是将输入图像转换为更高层次、更抽象的特征表示. 这些特征可以是图像的低层特征 (如颜色、纹理等), 或者是高层语义特征 (如形状、物体边界等). 特征编码器通过学习数据的有效表示, 可以帮助模型更好地理解输入数据的结构和语义信息, 从而提高模型在各种任务上的性能.

在特征编码之后, 网络中的特征解码器负责对提取的特征进行处理, 将编码器提取的高级抽象特征重新映射到原始数据的空间, 再采用基于图像对比度、颜色差异、空间先验等方法来计算每个像素的显著性得分或显著性预测. 通过特征解码器的重建过程, 模型可以学习到输入数据的有效表示, 并且能够保留数据的重要信息. 在获得图像的显著性得分之后, 通常会使用边缘细化、平滑、阈值设置等方法对显著性预测图进行后处理, 以进一步提高检测的准确性和鲁棒性. 最后, 系统将对显著性预测图进行阈值处理和区域分割, 以提取出显著性目标的位置信息, 从而确定图像中的显著性目标的位置.

1.2　Swin Transformer 网络

随着深度学习的不断发展, Transformer[6]在自然语言处理领域的成功应用吸引了在计算机视觉领域深耕的研究人员, Dosovitskiy 等人[7]于 2020 年提出了一种基于 Transformer 架构的视觉感知模型 Vision Transformer (ViT). ViT 的主要思想是将图像分割成固定大小的图像块 (patch), 然后将这些图像块转换成序列形式, 并输入到 Transformer 编码器中进行处理.

然而, 当 ViT 网络的输入图像尺寸较大时, 切分出的图像块数量会随之成平方倍地增加, 导致模型参数的数量呈现出平方级别的增长, 使得模型的空间复杂度过高. 为了解决 ViT 模型存在的问题, Liu 等人[4]于 2021 年提出一种新型网络 Swin Transformer, 该模型使用了一种新的分块策略和层次化的结构, 能够有效地处理大尺寸图像, 并在多个计算机视觉任务中取得了优异的性能. 首先, Swin Transformer 引入了跨阶段的分块策略, 将输入的大尺寸图像分解为一系列具有相同大小的小图像块, 但与传统的固定大小的图像块不同, Swin Transformer 采用了一种动态的分块策略, 即跨阶段分块. 该策略允许图像中不同位置的图像块之间存在重叠, 并且不固定于特定的图像块大小. 通过重叠区域, Swin Transformer 模型能够在不同图像块之间共享信息, 并允许图像中不同位置的信息得到充分的利用.

如图 3 所示, 原始的大尺寸图像会通过 Patch partition 模块被划分成一系列大小相同的图像块. 每个图像块会经过 Linear embedding 层, 将其映射到一个高维的向量空间中, 以便进行后续的处理. 这个过程通常包括一个线性变换操作, 将每个图像块中的像素值转换成一个特征向量. 经过 Linear embedding 层处理的特征向量将被送入 Swin Transformer block 层进行处理. Swin Transformer block 是 Swin Transformer 中的核心模块, 它由多个 Transformer 层组成. 在这一阶段, 模型会利用自注意力机制来捕捉图像中不同位置之间的关系, 并通过 MLP 层来实现特征的非线性变换和整合. 最后, 在 Swin Transformer 的最后一个阶段, 经过 Swin Transformer block 处理的特征向量将通过 Patch merging 层进行整合. Patch merging 层会将不同图像块之间的特征向量进行合并, 以得到整个图像的全局特征表示. 如图 3 右侧所示, W-MSA 结构首先使用大小相同的窗口对输入特征进行分割, 获得数个特征集合, 每个特征向量只需要与自身集合内部的其他成员做注意力计算. W-MSA 结构允许模型在不同窗口之间

进行交互和整合, 以更好地捕捉图像的全局信息. 对于每个窗口内部的特征向量, 首先计算出其内部的注意力矩阵, 接着将不同窗口之间的注意力矩阵进行加权求和, 得到最终的注意力权重. 最后, 利用注意力权重对特征向量进行加权求和, 得到每个窗口的输出特征向量. 而 SW-MSA 结构在 W-MSA 结构的基础上进行了改进, 将每个窗口的特征向量进行平移操作, 从而使得相邻窗口之间的特征能够重叠和共享, 增强了不同窗口之间的信息交互和整合能力.

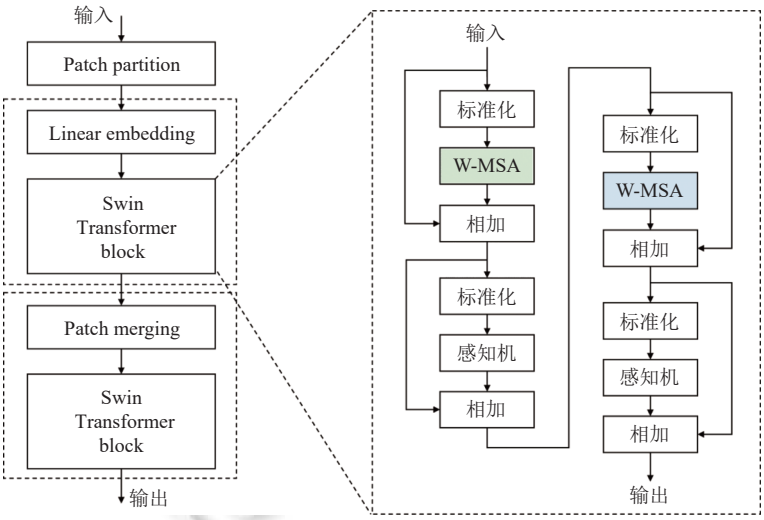


图 3 Swin Transformer 网络结构

1.3 研究难点与挑战

随着互联网技术的不断发展, 显著性目标检测作为计算机视觉上游任务中的关键一环, 受到了许多学者的关注和研究, 显著性目标检测算法的精度也在不断提高. 但是, 目前显著性目标检测领域仍然存在许多研究难点和挑战, 主要表现在以下两个方面. 一是针对高分辨率图像的有效检测. 现有大多数显著性目标检测算法在处理高分辨率图像时, 为了避免出现计算量过高的问题, 选择将图像压缩到统一的低分辨率尺寸后再输入网络. 然而在图像下采样的过程中, 会丢失大量的细节信息, 从而导致最终预测的粒度过粗, 降低了预测分割的质量. 因此, 如何在保持计算效率的同时, 实现对高分辨率图像的精确显著性目标检测成为当前的研究重点. 二是面向复杂场景下的精细像素分割. 在现实世界中, 图像往往会包含复杂的背景, 其中可能包含大量干扰信息, 使得显著目标的检测变得更加困难. 这些复杂的背景可能会掩盖目标, 导致显著性目标检测的精度下降, 例如对人体进行检测时头发丝的分割. 这些细节部分的分割对于显著性目标检测至关重要, 但由于其复杂性, 使得分割过程变得困难. 如何有效地识别和分割这些细微的像素是当前研究中的一个重要挑战.

2 本文方法

本节主要介绍提出的基于边缘增强的宽解码器显著性目标检测方法, 整体架构由残差网络 ResNet、Swin Transformer、差分卷积模块、多尺度注意力模块和多级宽解码器模块组成. 后续将首先详细介绍差分卷积模块, 接着介绍多尺度注意力模块和多级宽解码器模块, 最后介绍本网络中使用到的损失函数.

2.1 网络结构

基于边缘增强的宽解码器显著性目标检测方法的整体架构如图 4 所示. 受到 PGNet[2]模型的启发, 针对高分辨率显著性目标检测任务, 如果单独采用规模较小的模型, 例如残差网络, 则会面临着模型编解码能力不足的限制, 从而无法实现高质量的检测分割结果. 而采用规模稍大的模型, 如 Vision Transformer, 则会带来巨大的算力消

耗. 因此, 本文使用双流主干网络结构来解决上述矛盾. 具体而言, 采用残差网络 ResNet 来进行高分辨率图像的细节特征提取, 随后对图像进行下采样, 并借助 Swin Transformer 网络提取全局语义特征. 通过这种方式, 既保证了模型具有高质量的检测性能, 又避免了过高的计算开销.

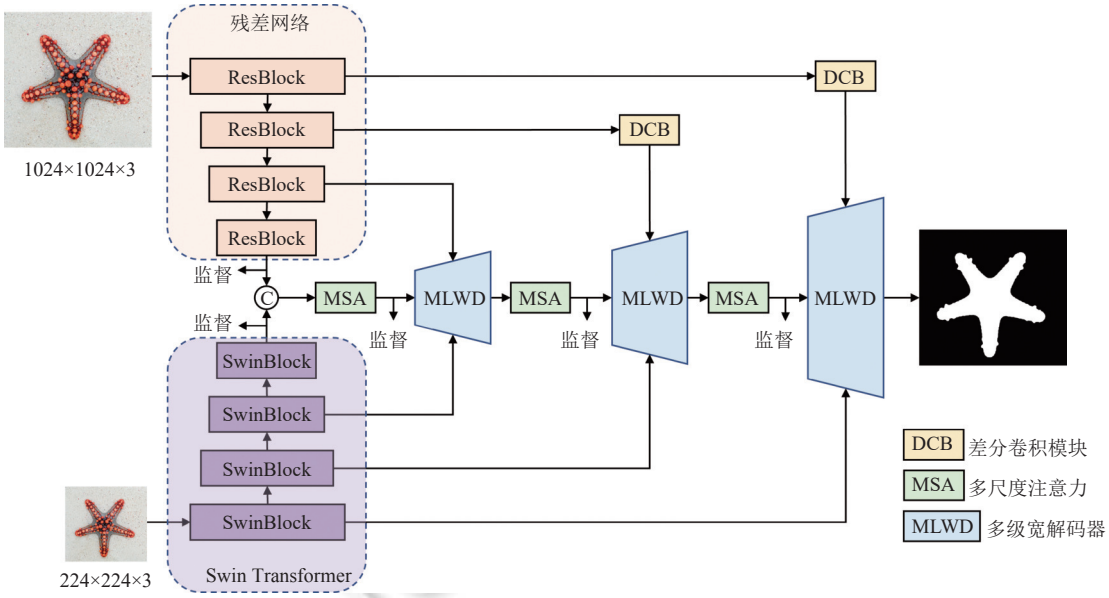


图 4 模型整体框架

为了提高不同尺度目标的检测能力, 这里选择残差网络的 4 层特征进行输出, 表示为 $\{R_i|i=2,3,4,5\}$. 得益于卷积的独特计算机制, 前两层的浅层特征 R_2 和 R_3 包含丰富的图像细节信息, 如纹理、颜色和边缘形状等信息, 在这里通过引入差分卷积操作, 能够在浅层特征中增加更丰富的边缘信息, 使得网络能够更准确地捕捉物体的边界和轮廓, 从而提高了最终模型的分割质量. 这种方法不仅可以有效地改善图像分割任务中的边缘检测效果, 还可以提升整体的分割性能, 为更精准的目标识别和定位提供了有力支持.

同样地, 选择 Swin Transformer 网络的后 4 层特征 $\{S_i|i=2,3,4,5\}$ 进行尺寸调整后输出. 为了融合残差网络和 Swin Transformer 网络输出的对应层相同维度的特征, 需要解决两者之间特征类型差异的问题. 残差网络更倾向于提取细节信息, 而 Swin Transformer 则更擅长提取语义信息. 为此, 引入多尺度注意力模块, 通过学习不同尺度下的特征关系, 对两种特征进行重新对齐和修正. 这样做可以实现细节信息与语义信息的互补, 并抑制其中多余的背景噪声, 提高最终模型的性能.

经过多尺度注意力模块的精修后, 特征被输入到多级宽解码器模块进行特征解码. 与传统的显著性目标检测方法中的经典解码范式不同, 多级宽解码器模块在逐级上采样的过程中加入了大尺寸卷积核. 这些大卷积核能够获得更大范围的原始感受野, 从而捕获更加全面的上下文信息. 因此, 网络能够更有效地提取抽象的高级特征, 从而使得其能够更好地理解图像中的语义内容.

与 PGNet 方法不同的是, 本文虽然采用了相同的编码器结构, 但在解码器结构上有显著区别. PGNet 解码器通过对交叉注意力矩阵进行显式监督来完成特定任务, 利用设计的 CMGM 模块接受编码器的输入特征并计算交叉注意力矩阵. 而本文的解码器则主要依靠多尺度注意力和多级宽解码器, 深入关注编码器输出的丰富细节特征信息. 通过计算不同层之间的融合特征信息, 不断修正语义细节和相关联的信息部分, 有效抑制背景噪声.

此外, 在模型的训练阶段, 采用了逐层监督方法. 当特征流处于不同尺度时, 将真实标注按照相同大小进行放缩, 然后通过对特征进行卷积并应用 Sigmoid[8]激活函数进行计算, 将特征值转换为每个像素对应前景或背景的分类概率. 最后, 通过计算预测图与标注之间的损失来产生梯度. 而逐层监督通过在网络的中间层添加监督信号, 可

以确保梯度的有效传播, 有助于减轻梯度消失或爆炸的问题, 可以加速网络的收敛速度. 并且逐层监督通过及时纠正特征的分布, 有效避免特征漂移问题, 提高了模型的泛化能力.

2.2 差分卷积模块

在计算机视觉领域的一些任务, 如图像分类[9]和目标检测[10], 经典卷积方法通常能够提取出图像中目标的颜色、纹理等基础特征, 这些特征已经足够辅助方法完成目标分类或定位. 然而, 在显著性目标检测任务中, 情况有所不同. 传统卷积提取到的特征往往不足以满足显著性目标检测的要求. 这类分割任务期望得到高质量、逐像素的分割结果, 而仅使用经典卷积方式常常会导致物体边缘分割模糊等问题.

针对以上问题, 本文提出了差分卷积模块. 差分卷积模块能够在特征提取过程中突出物体轮廓, 使得网络能够更好地捕获边缘信息, 从而有效地改善显著性目标的边缘分割效果.

在早期的边缘检测[11]任务中, 研究人员通常依赖于先验知识来设计边缘检测算子[12]. 这些传统的边缘算子主要基于梯度计算, 通过显式地计算像素之间的差异来捕获用于边缘检测的关键梯度信息, 常用的边缘检测算子如图 5 所示. 然而这些手工设计的边缘算子或者基于学习的边缘检测算法, 由于其浅层结构的局限性, 往往难以充分表达图像的复杂特征. 相比之下, 在深度学习中, 卷积操作被广泛应用于学习丰富且分层的图像表示. 普通的卷积核被用来探测局部图像模式, 但是这些卷积核的权重是从随机初始化开始优化的, 它们并没有直接编码梯度信息. 这导致它们在关注边缘相关特征方面的困难.

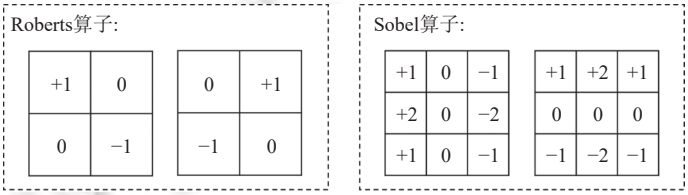


图 5 经典的边缘检测算子

差分卷积[13]的出现缓解了传统边缘检测算子存在的问题, 差分卷积的核心思想是通过计算像素点与其相邻像素之间的差值来突出边缘. 与传统的卷积核不同, 差分卷积采用特殊的差分核, 这个核对每个像素点与周围像素进行差分运算, 从而得到一个新的特征表示. 像素间的差分可以表示为:

$$D(x,y)=I(x+1,y)-I(x,y) \tag{1}$$

其中, $D(x,y)$ 表示图像在位置 (x,y) 处的像素差分值, $I(x,y)$ 表示图像在位置 (x,y) 处的像素值. 观察到差分卷积在边缘检测领域的显著贡献, 受像素差分卷积 (pixel difference convolution, PDC) 方法[14]的启发, 对其中的差分卷积结构进行了调整和改进. 本文提出的差分卷积模块具体结构如图 6 所示.

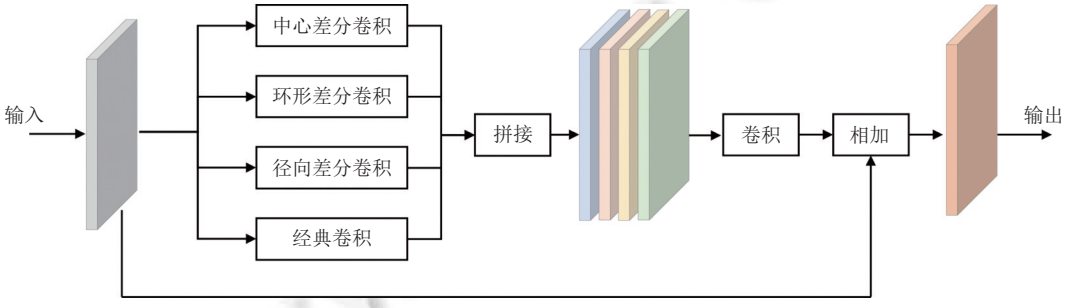


图 6 差分卷积模块结构

将残差网络流出的浅层特征 $\{R_i|i=2,3\}$ 输入差分卷积模块, 特征经过多分支的差分卷积区域, 分别使用中心差分、角度差分和径向差分对特征进行卷积, 通过不同类型的卷积方式提取不同种类的细节信息, 在这些信息中

重点加强边缘特征. 将 4 种特征在维度方向上进行拼接, 再使用卷积降维, 通过跳跃连接与原始特征相加后输出. 通过将差分特征与原始特征进行加权相加, 可以进一步增强图像的局部细节和边缘特征, 提高预测图像的视觉质量和对比度.

本文提出的差分卷积模块使用 3 种不同类型的差分卷积, 分别是中心差分卷积、环形差分卷积和径向差分卷积. 差分卷积的计算原理是基于各自预先设计的差分模块进行像素差异的计算, 3 种不同的差分模板如图 7 所示. 首先, 对于每个像素点, 差分模块会计算其与相邻像素之间的差值, 从而突出边缘特征的变化. 接下来, 得到的像素差值图像会与卷积内核的权重逐个相乘, 然后将所有相乘得到的结果进行求和. 最终, 这个求和结果会形成特征图中的数值, 反映了图像中不同位置的特征强度或变化程度. 计算过程可以用公式表示为:

$$y=\sum_{i=1}^{k\times k} w_i\cdot(x_i-x_i') \tag{2}$$

其中, x_i 和 x_i' 表示输入图像的像素, w_i 表示尺寸为 $k\times k$ 卷积核的权重. 采用并行结构叠加不同类型的差分卷积, 并通过卷积和短接结构重新调整这些不同类型细节信息在最终特征中的表达权重, 从而实现更符合显著性目标检测任务的差分卷积结构.

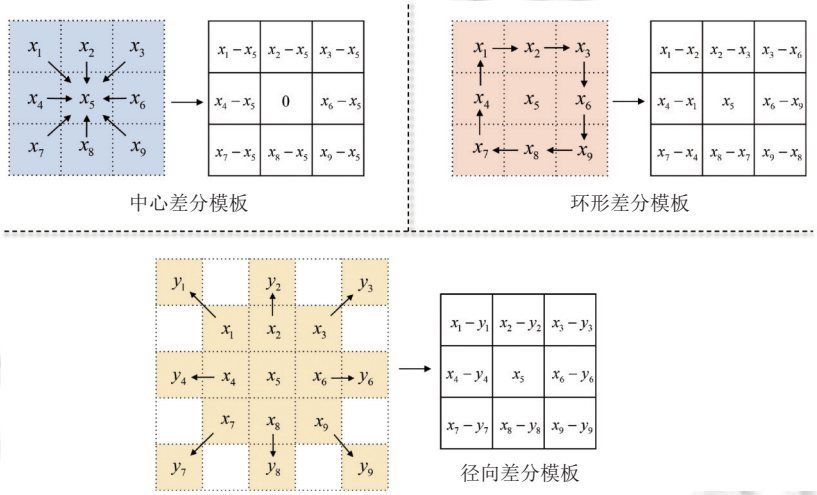


图 7　3 种类型的差分模板

2.3　多尺度注意力模块

注意力机制是 Vision Transformer 成功的重要因素之一. 传统的卷积操作在处理远距离像素之间的依赖关系时存在困难, 而自注意力机制能够通过学习到的权重动态地调整不同位置之间的关系, 使得模型能够更好地捕捉到图像中不同位置之间的长距离依赖关系. 本节提出多尺度注意力模块来计算不同分辨率下的注意力特征.

如图 8 所示, 输入特征首先经过连续的卷积池化操作, 获得了 4 层不同尺度的特征 $\{F_i|i=2,3,4,5\}$. 从最深层的特征开始, 逐层将相邻尺度的特征送入注意力计算模块. 在这个过程中, 低分辨率的深层特征被变换成二维向量键序列 (key), 而高分辨率的浅层特征被转换成二维向量查询序列 (query) 和值序列 (value). 通过注意力计算, 利用深层特征中的语义信息来指导浅层特征, 以保留显著目标的细节信息. 计算过程用公式表示为:

$$Attention\left(F_i^Q,F_j^K,F_i^V\right)=Softmax\left(F_i^Q\cdot F_j^{K^{\mathrm{T}}}\right)\cdot F_i^V \tag{3}$$

其中, F_i^Q 和 F_i^V 分别是浅层特征经过展平得到二维向量查询序列与值序列, F_j^K 是深层特征展平得到的键序列, 根据自注意力机制的类似原理, 首先计算查询序列和键序列的相似度, 然后使用 *Softmax* 激活函数对相似度进行归一化处理, 得到注意力权重. 接下来, 将注意力权重与值序列相乘, 得到加权的值序列. 最后, 将获得的注意力特征通过跳跃连接与原始特征相加后输出.

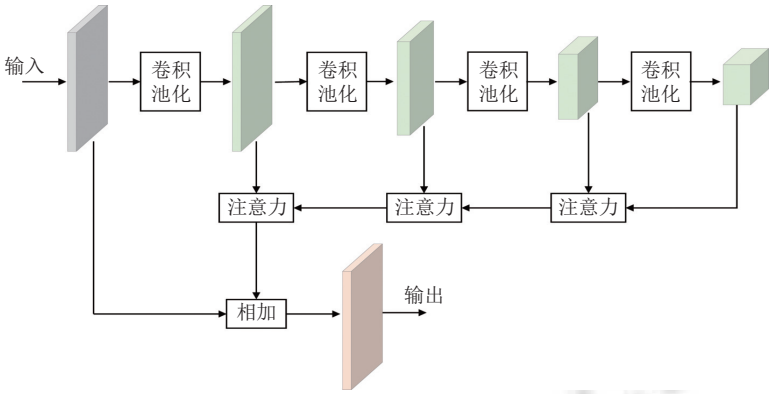


图 8　多尺度注意力模块结构

将输入的融合特征连续解耦, 可以制造特征的分层结构, 从而更好地捕获图像中的多层次信息. 在这个过程中利用注意力机制, 通过自身包含的全局语义信息来匹配并强化各层次特征中的细节信息. 逐层筛选和强化后, 可以得到与显著目标相关的特征信息, 同时抑制了浅层特征中的噪声. 这样的特征融合修正增强的过程能够有效提高模型对显著目标的检测性能, 使得模型能够更准确地定位和识别目标区域.

多尺度注意力模块利用不同尺度的特征进行融合, 使得高层特征能够补充低层特征的细节信息, 而低层特征则为高层特征提供上下文信息, 增强了特征表达的丰富性和准确性. 通过在不同网络层之间进行注意力计算, 将各层的特征信息进行整合. 这样不仅能够利用高层特征的抽象语义信息, 还能保留低层特征的细节信息. 从而能够有效提高显著性目标检测任务中的特征表达能力和检测精度. 该结构不仅能够捕捉到更多的细节信息, 还能够抑制背景噪声, 从而在复杂场景下实现更好的显著性目标检测效果.

2.4　多级宽解码器模块

在显著性目标检测方法中, 常见的模型架构之一是 U 型结构, 例如 U-Net[15]模型. 这类模型结构通常由编码器和解码器串联组成. 在编码器中, 采用卷积和下采样操作, 用以提取不同尺度的局部和全局信息, 从而逐步缩小特征尺寸. 然而, 在解码器中, 为了确保最终模型能够输出与原始图像相同尺寸的显著预测图, 通常需要进行上采样操作, 逐步将特征还原至初始大小. 然而, 如果仅仅使用常见的上采样方法[16], 例如双线性上采样, 由于其缺乏可学习性, 无法有效传递信息, 可能导致原始特征中的部分信息丢失. 这会导致特征在空间上的失真, 使得特征图被过度平滑, 从而丢失一些细微的细节信息, 不利于后续的预测输出.

为了解决这一问题, 一些方法选择在上采样过程中引入卷积层和归一化层, 以调整特征分布, 从而恢复或增强细节信息, 减缓连续上采样带来的影响. 然而, 这些卷积层通常使用小尺寸的内核, 由于感受野的限制, 无法有效处理较长距离的依赖信息. 这可能会导致模型在恢复细节信息的过程中, 丢失一些重要的上下文信息, 从而影响最终的预测结果. 因此, 本文提出带有宽卷积的解码器模块, 宽卷积模块结构如图 9 所示.

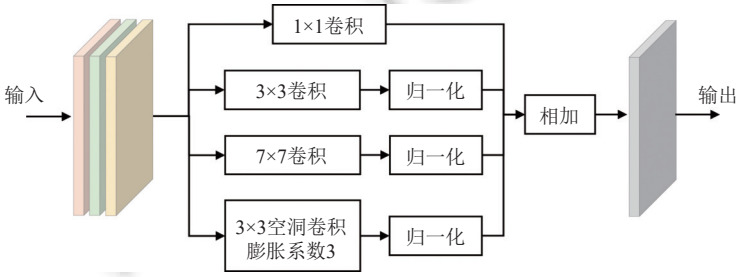


图 9　宽卷积模块结构

融合特征经过 4 条不同的卷积分支进行特征重整. 首先, 第 1 条分支中的 1×1 卷积主要用于降低特征通道维度, 以减少计算量和参数数量, 并且有助于提取更加抽象的特征. 其次, 第 2 条分支采用 3×3 卷积核, 用于捕捉小尺度目标的特征. 这是因为在处理宽卷积过程中, 往往容易丢失小尺度的信息, 因此使用 3×3 卷积有助于保留这些细微特征, 提高模型对小尺度目标的感知能力. 第 3 条分支采用 7×7 卷积核, 这种大卷积核[17]可以覆盖更广阔的空间范围, 从而能够捕获更多的局部和全局信息. 这对于特征提取过程中的上下文理解非常重要, 有助于增强模型对目标周围环境的感知和理解能力. 最后, 第 4 条分支采用 3×3 空洞卷积[18]. 与其他分支相比, 它在不增加模型参数量的同时, 可以辅助提取特征图中的远距离关联信息. 空洞卷积可以通过调整膨胀率来灵活调节感受野大小, 从而更好地捕获远距离的依赖关系, 提高模型对全局信息的感知能力. 最终, 将 4 条分支的特征相加并输出. 过程用公式表示为:

$$F_{b1} = Conv_{1\times1}(F_i) \tag{4}$$

$$F_{b2} = BN(Conv_{3\times3}(F_i)) \tag{5}$$

$$F_{b3} = BN(Conv_{7\times7}(F_i)) \tag{6}$$

$$F_{b4} = BN(Conv_{3\times3,d=3}(F_i)) \tag{7}$$

$$F = F_{b1} + F_{b2} + F_{b3} + F_{b4} \tag{8}$$

其中, F_i 表示输入特征, F_{bi} 表示第 i 分支的输出特征, $BN(\cdot)$ 表示批归一化 (batch normalization)[19]操作.

多级宽解码器模块的结构如图 10 所示, 该模块接收 3 个分支的特征流作为输入. 这 3 个分支分别来自编码器中的残差网络、Swin Transformer 网络以及上一级解码器的输出特征. 在接收到这 3 种特征后, 首先需要统一它们的维度. 为了实现这一点, 采用采样操作将 3 种特征放缩至相同的大小, 并在通道维度上进行特征拼接, 以便后续处理.

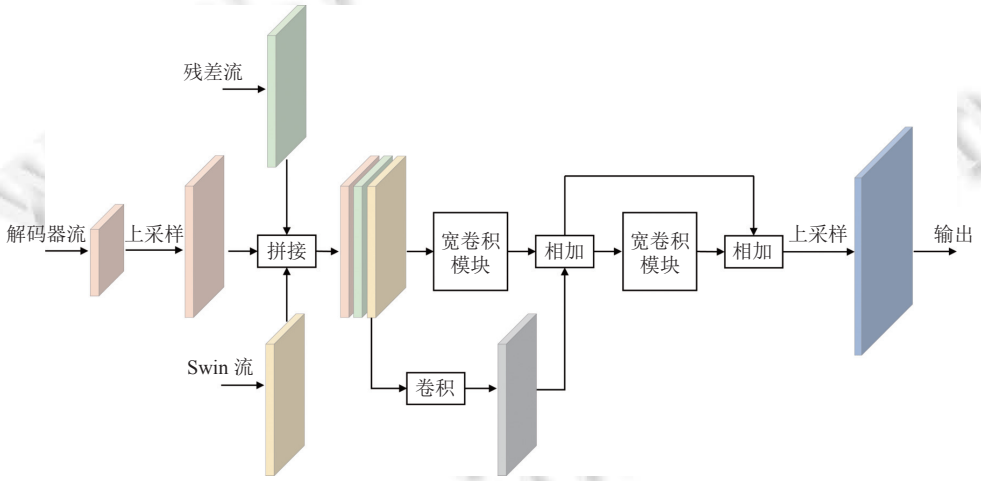


图 10　多级宽解码器模块结构

接下来, 通过宽卷积模块对拼接后的特征进行调整和压缩, 这一步骤旨在进一步整合和优化特征, 以确保模型能够更好地理解图像内容. 然后, 将调整后的特征与融合特征的普通卷积分支特征进行相加, 以获得更丰富和多样的特征表示. 得到的特征再次送入宽卷积模块, 以进一步提升特征的表达能力. 最后, 输出的特征通过跳跃连接与输入前的特征相加, 以保留原始信息, 并最终通过上采样操作输出. 通过这样的多级宽解码器模块设计, 能够充分利用不同分支提取的特征信息, 通过逐步整合和优化特征, 进一步提升模型在任务中的性能表现. 过程用公式表示为:

$$F_i^D = Concat\left(F_i^R, F_i^S, Up\left(F_{i-1}^D\right)\right) \tag{9}$$

$$F_i^D = WC\left(F_i^D\right) + Conv\left(F_i^D\right) \tag{10}$$

$$F_i^D = Up\left(WC\left(F_i^D\right) + F_i^D\right) \tag{11}$$

其中, F_i^R 表示残差网络第 i 层输出的特征, F_i^S 表示 Swin Transformer 网络第 i 层输出的特征, $Up(\cdot)$ 表示双线性插值上采样操作, F_i^D 表示第 i 层解码器输出的特征, $Concat(\cdot)$ 表示在通道方向上进行特征拼接操作, $WC(\cdot)$ 表示宽卷积, $Conv(\cdot)$ 表示普通卷积.

3 实验分析

3.1 数据集和评价指标

目前, 主流的显著性目标检测数据集包括两个低分辨率数据集, 分别是 DUTS[20]和 DUT-OMRON[21], 以及 3 个高分辨率数据集, 包括 HRSOD[22]、DAVIS-S[23]和 UHRSD[2]. 这 5 个数据集包含了显著性目标检测任务中常见情形, 覆盖了多种类别的显著物体, 同时也包含多种场景, 例如城市、森林、沙滩、室内等, 十分适合对显著性目标检测算法进行训练和评估.

评价指标能够提供客观的性能度量标准, 避免了主观评价的偏差, 使得不同研究方法的比较更加公平和准确. 在此次实验中, 选用 4 种不同的评价指标, 从不同角度去衡量显著性目标检测模型的性能, 分别是 F-measure 指标[24]、S-measure 指标[25]、E-measure 指标[26]和平均绝对误差 (mean absolute error, *MAE*)[27]. 其中, F-measure 是通过调和精确率 (*precision*) 和召回率 (*recall*) 的值来综合评价模型的性能; S-measure 指标可以衡量目标检测结果与真实标注之间的结构相似性; E-measure 指标从结构相似性和细节差异两个方面综合衡量模型性能; 平均绝对误差用于衡量模型预测值与真实值之间的平均绝对偏差.

3.2 实验设置

实验平台的操作系统为 Ubuntu 20.04.4 LTS, 配置 Python 3.8 环境, 基于 PyTorch 框架实现网络模型, 使用一张 RTX 2080Ti 显卡来加速训练. 网络结构方面按照前文所述设计依次搭建. 整个网络使用随机梯度下降法 (SGD)[28]进行端到端训练. 将两个编码器的最大学习率设置为 0.000 1, 解码器的最大学习率设置为 0.001. 训练过程中采用余弦退火 (cosine annealing) 学习率衰减[29]策略, Momentum 和权重衰减分别设置为 0.9 和 0.000 5. 批量化大小 (batch size) 设置为 4, 最大迭代次数 (epoch) 设置为 50.

本次实验采用两种不同的数据集组合方式来训练模型, 以探究其对模型性能的影响. 首先是单独使用低分辨率数据集 DUTS 进行训练, 这有助于评估模型在处理普通分辨率图像时的性能表现, 并验证其在常见场景下的适用性. 其次, 采用低分辨率数据集 DUTS 和高分辨率数据集 HRSOD 的组合进行训练. 这种组合训练方式可以使模型在更广泛的场景下学习特征, 提高其对不同分辨率图像的泛化能力. 在测试过程中, 需要对每个图像的大小进行统一调整, 将其调整为相同的大小. 这一步骤的目的是确保在输入到网络中之前, 所有图像都具有相同的尺寸, 以便模型能够统一处理, 减少因图像大小差异而引入的偏差.

测试实验分为两组. 一组在低分辨率数据集 DUTS 和 DUT-OMRON 上进行性能测试. 这组实验旨在评估模型在处理普通分辨率图像时的性能, 并验证其在常见场景下的泛化能力. 另一组实验则在高分辨率数据集 DAVIS-S、HRSOD 和 UHRSD 上测试模型在处理高分辨率图像时的性能表现. 通过这两组实验, 可以全面了解模型在不同分辨率图像上的表现, 为实际应用提供更全面的参考依据.

3.3 整体性能评估

为了验证基于边缘增强的宽解码器显著性目标检测方法的有效性, 选择了几种常用的对比网络, 包括 CPD[30]、SCRN[31]、DAS[32]、F^3Net[33]、GCPA[34]、ITSD[35]、LDF[36]、CTD[37]、PFS[38]、SelfReformer[39]和 PGNet[2], 这些对比方法代表了当前显著性目标检测领域的一些主流模型架构. 对所有网络进行了训练, 并将它们在 5 个公共数据集上的指标值进行了对比, 并给出具体结果. 其中, 标注“D”表示模型在 DUTS 数据集上进行训练, “DH”表示在 DUTS 和 HRSOD 数据集上混合训练, 粗体数字表示每个数据集上的最优性能值. $F_{\max}$ 表示多个 F-measure 值中的最大值, S_m 和 E_m 分别表示 S-measure 和 E-measure 指标的取值, 符号“↑”表示该指标值越大性能越好, 符号“↓”表示

该指标值越小性能越好, 下同.

实验结果如表 1 和表 2 所示. 表 2 中, Params 表示模型的参数量大小 (number of parameters), FLOPs 是指在输入图像尺寸为 1024×1024 时的模型计算量. 可以看出, 本文提出的基于边缘增强的宽解码器显著性目标检测方法在 5 个公共数据集上表现出了令人满意的性能. 值得注意的是, 该方法在 3 个高分辨率数据集上的指标相较于两个低分辨率数据集有着更为显著的提升, 这一现象突显了所提出方法中各模块的有效性.

表 1　不同方法在低分辨率数据集上的实验比较结果

方法	DUTS				DUT-OMRON			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
CPD	0.865	0.043	0.887	0.869	0.797	0.056	0.866	0.825
SCRN	0.888	0.040	0.888	0.885	0.811	0.056	0.863	0.837
DAS	0.895	0.034	0.908	0.894	0.827	0.050	0.877	0.845
F^3Net	0.891	0.035	0.902	0.888	0.813	0.053	0.871	0.838
GCPA	0.888	0.038	0.891	0.891	0.812	0.056	0.860	0.839
ITSD	0.883	0.041	0.895	0.885	0.821	0.061	0.863	0.840
LDF	0.898	0.034	0.910	0.892	0.820	0.051	0.873	0.838
CTD	0.897	0.034	0.909	0.893	0.826	0.052	0.875	0.844
PFS	0.896	0.036	0.902	0.892	0.823	0.055	0.875	0.842
SelfReformer	0.916	0.026	0.920	0.911	0.836	0.041	0.886	0.856
PGNet-D	0.917	0.027	0.922	0.911	0.835	0.045	0.887	0.855
PGNet-DH	0.919	0.028	0.925	0.912	0.835	0.046	0.887	0.858
Ours-D	0.927	0.024	0.930	0.914	0.844	0.039	0.908	0.858
Ours-DH	**0.935**	**0.020**	**0.939**	**0.916**	**0.848**	**0.037**	**0.911**	**0.861**

表 2　不同方法在高分辨率数据集上的实验比较结果

方法	Params (M)	FLOPs (G)	DAVIS-S				HRSOD				UHRSD			
			$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
CPD	47.85	149.42	0.871	0.029	0.921	0.893	0.867	0.041	0.891	0.881	0.894	0.055	0.884	0.878
SCRN	25.23	126.69	0.893	0.027	0.911	0.902	0.880	0.042	0.887	0.888	0.904	0.051	0.880	0.887
DAS	36.68	240.88	0.902	0.020	0.949	0.911	0.893	0.032	0.925	0.897	0.914	0.045	0.892	0.889
F^3Net	25.54	138.38	0.915	0.020	0.940	0.914	0.900	0.035	0.913	0.897	0.909	0.046	0.887	0.890
GCPA	67.06	555.53	0.922	0.020	0.934	0.929	0.889	0.036	0.898	0.898	0.912	0.047	0.886	0.896
ITSD	51.07	319.49	0.899	0.022	0.922	0.909	0.896	0.036	0.912	0.898	0.911	0.045	0.895	0.897
LDF	25.15	130.69	0.911	0.019	0.947	0.922	0.904	0.032	0.919	0.904	0.913	0.047	0.891	0.888
CTD	11.83	51.71	0.904	0.019	0.938	0.911	0.905	0.032	0.921	0.905	0.917	0.043	0.898	0.897
PFS	31.18	382.86	0.916	0.019	0.946	0.923	0.911	0.033	0.922	0.906	0.918	0.043	0.896	0.897
SelfReformer	91.59	106.61	0.932	0.016	0.954	0.918	0.924	0.026	0.945	0.900	0.924	0.040	0.926	0.888
PGNet-D	72.72	57.04	0.936	0.015	0.947	0.935	0.931	0.021	0.944	0.930	0.931	0.037	0.904	0.912
PGNet-DH	72.72	57.04	0.950	0.012	0.975	0.948	0.937	0.020	0.946	0.935	0.935	0.036	0.905	0.912
Ours-D	122.84	264.77	0.944	0.014	0.958	0.939	0.944	0.018	0.955	0.936	0.942	0.031	0.937	0.915
Ours-DH	122.84	264.77	**0.961**	**0.009**	**0.979**	**0.949**	**0.953**	**0.015**	**0.968**	**0.939**	**0.954**	**0.023**	**0.949**	**0.918**

这种优势主要源于模型在设计时更加注重细节信息的捕捉和长距离依赖关系的建模, 使其在特征提取和多尺度信息融合方面具备更丰富的表示能力. 然而, 这种设计也导致参数量相对较高. 为平衡计算开销, 本文方法采用双分支结构, 在不同分辨率下分别处理图像, 以综合考虑细节信息与语义信息的提取. 即使面对 1024×1024 的高分辨率输入, 计算量依然保持在对比方法的平均水平, 兼顾了高效性与性能表现.

此外, 差分卷积模块、多尺度注意力模块和多级宽解码器模块的联合作用, 使得模型能够学习到更为丰富的细节信息, 实现更精准的显著物体分割, 从而提升整体性能和分割质量. 实验结果表明, 所提出的方法在处理高分辨率图像时具有显著优势, 并且能够有效应对不同分辨率图像带来的挑战.

3.4 消融实验

为了验证网络的每个模块的必要性和有效性, 本文针对每个策略和模块分别在 5 个公共数据集上进行训练和测试, 具体的消融实验如下.

(1) 差分卷积模块的有效性

本节对差分卷积模块进行消融实验, 旨在探究其对整体模型性能的影响. 定义了 3 种不同的训练方法, 包括: 方法 1 不使用差分卷积模块, 方法 2 对残差网络的深层特征使用差分卷积模块, 方法 3 对残差网络的浅层特征使用差分卷积模块. 在这 3 组对照实验中, 其他训练参数均保持一致, 以确保实验结果的可比性. 随后, 在 DUTS 数据集上对这 3 种方法进行了 50 轮训练, 并分别在低分辨率和高分辨率数据集上进行性能测试. 这样设计实验的目的是观察不同位置引入差分卷积模块对模型性能的影响程度.

表 3 和表 4 的结果显示, 与方法 1 相比, 方法 2 和方法 3 使用了差分卷积模块, 表现更出色. 特别是在高清数据集上, 性能提升更为显著, 这说明随着图像清晰度的提高, 差分卷积模块能更好地突出物体的边缘细节, 从而提升分割质量. 相比于在网络深层添加差分卷积模块的方法 2, 方法 3 在浅层位置引入了这一模块, 且性能提升更为显著. 这是因为浅层特征包含了大部分细节信息, 而深层特征主要包含上下文语义信息. 因此, 从效率和收益的角度考虑, 在浅层位置添加差分卷积模块是更优的选择.

表 3 不同方法在低分辨率数据集上的实验比较结果

方法	DUTS				DUT-OMRON			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
方法1	0.929	0.024	0.929	0.820	0.841	0.044	0.889	0.855
方法2	0.932	0.023	0.937	0.914	0.843	0.041	0.894	0.856
方法3	**0.935**	**0.020**	**0.939**	**0.916**	**0.848**	**0.037**	**0.911**	**0.861**

表 4 不同方法在高分辨率数据集上的实验比较结果

方法	DAVIS-S				HRSOD				UHRSD			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
方法1	0.919	0.021	0.941	0.923	0.911	0.033	0.923	0.909	0.919	0.042	0.897	0.903
方法2	0.935	0.016	0.949	0.935	0.934	0.020	0.948	0.930	0.935	0.036	0.913	0.911
方法3	**0.944**	**0.014**	**0.958**	**0.939**	**0.944**	**0.018**	**0.955**	**0.936**	**0.942**	**0.031**	**0.937**	**0.915**

(2) 多尺度注意力模块的有效性

在讨论多尺度注意力模块对整体模型性能的影响时, 设置了两组对照实验. 在方法 1 中, 不使用多尺度注意力模块, 而在方法 2 中则添加了该模块. 这两组实验保持了其他结构和训练参数不变, 以确保能够准确评估多尺度注意力融合增强模块对模型性能的独立影响.

实验结果如表 5 和表 6 所示, 添加多尺度注意力模块的方法 2 相较于方法 1 在显著性目标检测任务中取得了显著的性能提升. 这表明多尺度注意力模块在模型中的引入, 对提升性能发挥了重要作用. 这种模块能够通过在不同尺度中计算注意力, 有效地将语义信息引导到细节信息中, 并抑制多余的噪声信息. 这样的处理方式有助于强化特征中的细节信息, 修正融合特征中的潜在冲突, 从而改善了模型对特征的表征, 进而提升了性能表现. 因此, 多尺度注意力模块能够有效地实现细节信息与语义信息的互补, 为模型性能的提升做出了重要贡献.

(3) 多级宽解码器模块的有效性

为了验证设计的多级宽解码器模块和逐层监督策略的有效性, 分别设置了 4 组对照网络. 在方法 1 中, 不使用多级宽解码器模块, 并且不进行逐层监督; 在方法 2 中, 虽然不使用多级宽解码器模块, 但进行了逐层监督; 在方法 3 中, 使用了多级宽解码器模块, 但没有逐层监督; 而在方法 4 中, 既使用了多级宽解码器模块, 又进行了逐层监督. 这 4 组实验旨在系统地研究多级宽解码器模块和逐层监督对模型性能的影响.

表 5 不同方法在低分辨率数据集上的实验比较结果

方法	DUTS				DUT-OMRON			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
方法1	0.930	0.021	0.931	0.911	0.847	0.041	0.895	0.854
方法2	**0.935**	**0.020**	**0.939**	**0.916**	**0.848**	**0.037**	**0.911**	**0.861**

表 6 不同方法在高分辨率数据集上的实验比较结果

方法	DAVIS-S				HRSOD				UHRSD			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
方法1	0.931	0.019	0.942	0.932	0.931	0.022	0.943	0.931	0.930	0.038	0.911	0.912
方法2	**0.944**	**0.014**	**0.958**	**0.939**	**0.944**	**0.018**	**0.955**	**0.936**	**0.942**	**0.031**	**0.937**	**0.915**

实验结果如表 7 和表 8 所示, 通过对比 4 种不同方法的性能表现, 可以得出一些结论. 相比于不使用多级宽解码器模块的方法, 使用该模块能够显著提升模型的性能; 同时, 逐层辅助监督也能够对模型的性能产生积极的影响. 进一步分析发现, 将多级宽解码器模块与逐层辅助监督相结合的方法 4 在性能上取得了最佳的表现, 证明了这两种技术的互补性和有效性.

表 7 不同方法在低分辨率数据集上的实验比较结果

方法	DUTS				DUT-OMRON			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
方法1	0.926	0.031	0.925	0.909	0.828	0.046	0.880	0.847
方法2	0.928	0.022	0.928	0.914	0.835	0.041	0.894	0.856
方法3	0.933	0.020	0.931	0.916	0.840	0.040	0.906	0.856
方法4	**0.935**	**0.020**	**0.939**	**0.916**	**0.848**	**0.037**	**0.911**	**0.861**

表 8 不同方法在高分辨率数据集上的实验比较结果

方法	DAVIS-S				HRSOD				UHRSD			
	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{\max}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
方法1	0.921	0.028	0.932	0.928	0.926	0.031	0.929	0.922	0.928	0.045	0.924	0.903
方法2	0.937	0.019	0.944	0.934	0.935	0.023	0.941	0.931	0.937	0.035	0.932	0.912
方法3	0.939	0.018	0.948	0.935	0.937	0.022	0.944	0.931	0.936	0.036	0.929	0.911
方法4	**0.944**	**0.014**	**0.958**	**0.939**	**0.944**	**0.018**	**0.955**	**0.936**	**0.942**	**0.031**	**0.937**	**0.915**

3.5 结果可视化

为了展示模型在高分辨率数据集上的优越性, 图 11 将本节方法和其他方法的预测图像进行可视化对比. 图中第 1 列是数据集中的原图像, 第 2 列是图像所对应的真实标注, 第 3 列是本节方法训练 50 轮后的预测结果, 第 4 列至最后一列是其他对比方法.

从图 11 中的分割结果可以观察到不同场景下的目标检测情况. 首先, 在图中第 1 行所示的复杂场景中, 对比方法均未能成功检测出小尺度目标, 而本文方法成功检测并分割出了人物, 但不足之处在于错误地将人物的影子也进行了分割. 其次, 在第 2、3、4 行的场景中, 对比方法未能正确理解显著目标, 例如第 2 行的白色楼梯和第 3 行飞机前的人物. 然而, 从第 5、6、7 行可以看出, 本文方法在分割具有复杂结构的物体时表现优异, 例如第 5 行中叶子的轮廓和第 6 行中栅栏之间的缝隙. 相比之下, 对比方法存在无法分割、多分割、部分分割和边缘模糊等问题, 这些问题严重影响了检测质量. 这表明本文提出的差分卷积模块有效提升了边缘特征的表达能力, 而多尺度注意力模块和多级宽解码器模块有助于模型有效地关注不同尺度的目标, 从而提升了模型的物体理解能力和检测准确性.

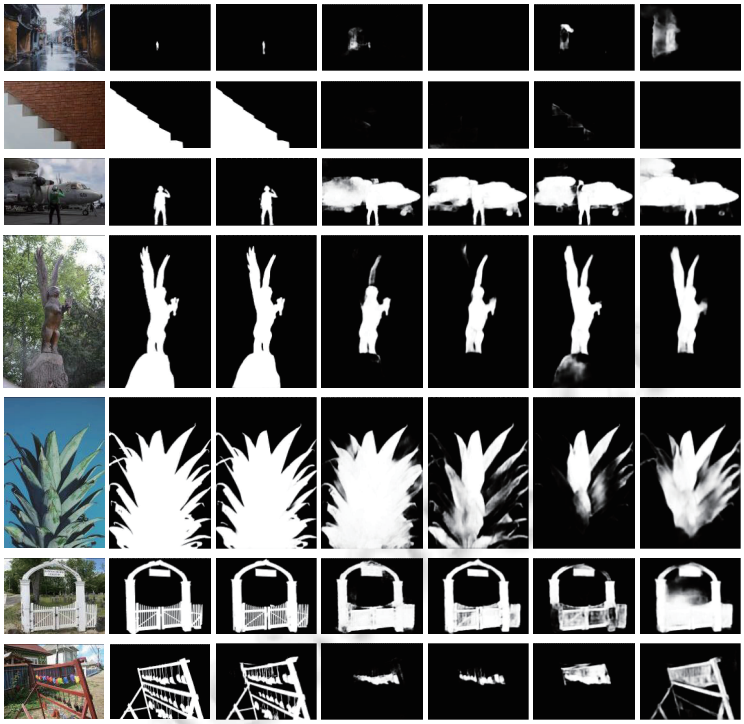


图 11 不同模型的分割结果图

4 总 结

本文针对许多显著性目标检测模型存在小尺度物体分割困难和物体边缘分割质量差的问题提出了基于边缘增强的宽解码器显著性目标检测方法, 使用融合差分卷积模块的残差网络来重点关注物体边缘特征, 结合另一条分支的 Swin Transformer 网络提取的丰富全局语义特征. 在融合特征的过程中加入多尺度注意力模块, 综合学习不同类型特征中的不同尺度图像特征, 在类型和尺度上完成信息互补, 抑制背景噪声特征, 提高模型泛化能力. 在解码器中应用多级宽解码模块, 利用大卷积核扩大原始感受野, 更好地进行上下文建模, 进一步提升网络性能. 在 5 个公共基准数据集上, 模型取得出色表现. 本文还通过消融实验分析每个模块的有效性. 在后续研究工作中, 将进一步研究如何在不提高计算成本的同时从高分辨率图像中获得清晰的显著目标分割.

References

[1] Cong RM, Zhang C, Xu M, Liu HY, Zhao Y. Research progress of RGB-D salient object detection in deep learning era. Ruan Jian Xue Bao/Journal of Software, 2023, 34(4): 1711–1731 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6700.htm [doi: 10.13328/j.cnki.jos.006700]

[2] Xie CX, Xia CQ, Ma MC, Zhao ZR, Chen XW, Li J. Pyramid grafting network for one-stage high resolution saliency detection. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11707–11716. [doi: 10.1109/CVPR52688.2022.01142]

[3] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[4] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

[5] Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. In: Proc. of the 7th Int'l Conf. on Document Analysis and Recognition. Edinburgh: IEEE, 2003. 958–963. [doi: 10.1109/ICDAR.2003.1227801]

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. 2021.

[8] Kyurkchiev N, Svetoslav M. Sigmoid Functions: Some Approximation and Modelling Aspects. Saarbrucken: LAP LAMBERT Academic Publishing, 2015. [doi: 10.11145/j.bmc.2015.03.081]

[9] Chen LY, Li SB, Bai Q, Yang J, Jiang SL, Miao YM. Review of image classification algorithms based on convolutional neural networks. Remote Sensing, 2021, 13(22): 4712. [doi: 10.3390/rs13224712]

[10] Zou ZX, Chen KY, Shi ZW, Guo YH, Ye JP. Object detection in 20 years: A survey. Proc. of the IEEE, 2023, 111(3): 257–276. [doi: 10.1109/JPROC.2023.3238524]

[11] Dharampal, Mutneja V. Methods of image edge detection: A review. Journal of Electrical & Electronic Systems, 2015, 4(2): 1000150. [doi: 10.4172/2332-0796.1000150]

[12] Prewitt JMS. Object enhancement and extraction. Picture Processing and Psychopictorics, 1970, 10: 15–19.

[13] Yu ZT, Zhao CX, Wang ZZ, Qin YX, Su Z, Li XB, Zhou F, Zhao GY. Searching central difference convolutional networks for face anti-spoofing. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5294–5304. [doi: 10.1109/CVPR42600.2020.00534]

[14] Su Z, Liu WZ, Yu ZT, Hu DW, Liao Q, Tian Q, Pietikäinen M, Liu L. Pixel difference networks for efficient edge detection. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 5097–5107. [doi: 10.1109/ICCV48922.2021.00507]

[15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI 2015). Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[16] Tian Z, He T, Shen CH, Yan YL. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3121–3130. [doi: 10.1109/CVPR.2019.00324]

[17] Ding XH, Zhang XY, Han JG, Ding GG. Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11953–11965. [doi: 10.1109/CVPR52688.2022.01166]

[18] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[19] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. of the 32nd Int'l Conf. on Machine Learning, Vol. 37. 2015. 448–456. [doi: 10.5555/3045118.3045167]

[20] Wang LJ, Lu HC, Wang YF, Feng MY, Wang D, Yin BC, Ruan X. Learning to detect salient objects with image-level supervision. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3796–3805. [doi: 10.1109/CVPR.2017.404]

[21] Yang C, Zhang LH, Lu HC, Ruan X, Yang MH. Saliency detection via graph-based manifold ranking. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 3166–3173. [doi: 10.1109/CVPR.2013.407]

[22] Zeng Y, Zhang PP, Lin Z, Zhang JM, Lu HC. Towards high-resolution salient object detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7233–7242. [doi: 10.1109/ICCV.2019.00733]

[23] Perazzi F, Wang O, Gross M, Sorkine-Hornung A. Fully connected object proposals for video segmentation. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 3227–3234. [doi: 10.1109/ICCV.2015.369]

[24] Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Proc. of the 19th Australian Joint Conf. on Artificial Intelligence. Hobart: Springer, 2006. 1015–1021. [doi: 10.1007/11941439_114]

[25] Fan DP, Cheng MM, Liu Y, Li T, Borji A. Structure-measure: A new way to evaluate foreground maps. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 4558–4567. [doi: 10.1109/ICCV.2017.487]

[26] Fan DP, Gong C, Cao Y, Ren B, Cheng MM, Borji A. Enhanced-alignment measure for binary foreground map evaluation. In: Proc. of

the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 698–704. [doi: 10.5555/3304415.3304515]

[27] Hodson TO. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development, 2022, 15(14): 5481–5487. [doi: 10.5194/gmd-15-5481-2022]

[28] Ruder S. An overview of gradient descent optimization algorithms. arXiv:1609.04747, 2016.

[29] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[30] Wu Z, Su L, Huang QM. Cascaded partial decoder for fast and accurate salient object detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3902–3911. [doi: 10.1109/CVPR.2019.00403]

[31] Wu Z, Su L, Huang QM. Stacked cross refinement network for edge-aware salient object detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7263–7272. [doi: 10.1109/ICCV.2019.00736]

[32] Zhao JW, Zhao YF, Li J, Chen XW. Is depth really necessary for salient object detection? In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1745–1754. [doi: 10.1145/3394171.3413855]

[33] Wei J, Wang SH, Huang QM. F³Net: Fusion, feedback and focus for salient object detection. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 12321–12328. [doi: 10.1609/aaai.v34i07.6916]

[34] Chen ZY, Xu QQ, Cong RM, Huang QM. Global context-aware progressive aggregation network for salient object detection. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 10599–10606. [doi: 10.1609/aaai.v34i07.6633]

[35] Zhou HJ, Xie XH, Lai JH, Chen ZX, Yang LX. Interactive two-stream decoder for accurate and fast saliency detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9138–9147. [doi: 10.1109/CVPR42600.2020.00916]

[36] Wei J, Wang SH, Wu Z, Su C, Huang QM, Tian Q. Label decoupling framework for salient object detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13022–13031. [doi: 10.1109/CVPR42600.2020.01304]

[37] Zhao ZR, Xia CQ, Xie CX, Li J. Complementary trilateral decoder for fast and accurate salient object detection. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 4967–4975. [doi: 10.1145/3474085.3475494]

[38] Ma MC, Xia CQ, Li J. Pyramidal feature shrinking for salient object detection. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 2311–2318. [doi: 10.1609/aaai.v35i3.16331]

[39] Yun YK, Lin WS. Towards a complete and detail-preserved salient object detection. IEEE Trans. on Multimedia, 2024, 26: 4667–4680. [doi: 10.1109/TMM.2023.3325731]

附中文参考文献

[1] 丛润民, 张晨, 徐迈, 刘鸿羽, 赵耀. 深度学习时代下的 RGB-D 显著性目标检测研究进展. 软件学报, 2023, 34(4): 1711–1731. http://www.jos.org.cn/1000-9825/6700.htm [doi: 10.13328/j.cnki.jos.006700]

作者简介

童彪, 硕士生, 主要研究领域为人工智能, 显著性目标检测.

宋晓宁, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为模式识别, 计算机视觉, 生物信息学, 机器学习.

华阳, 博士生, 主要研究领域为生物信息学, 计算机视觉, 深度学习.

张文杰, 博士生, 主要研究领域为自然语言处理, 信息抽取.

吴小俊, 博士, 教授, 博士生导师, 主要研究领域为模式识别, 计算机视觉, 生物信息学, 机器学习.