智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于注意力机制的显著性目标检测方法

王凯诚， 鲁华祥， 龚国良， 陈刚

引用本文:
王凯诚, 鲁华祥, 龚国良, 等. 基于注意力机制的显著性目标检测方法[J]. 智能系统学报, 2020, 15(5): 956–963.
WANG Kaicheng, LU Huaxiang, GONG Guoliang, et al. Salient object detection method based on the attention mechanism[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(5): 956–963.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201903001

DOI: 10.11992/tis.201903001
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20200320.0957.008.html

基于注意力机制的显著性目标检测方法

王凯诚[1,2], 鲁华祥[1,3,4], 龚国良[1], 陈刚[1]
（1. 中国科学院 半导体研究所, 北京 100083; 2. 中国科学院大学 未来技术学院, 北京 100089; 3. 中国科学院 脑科学与智能技术卓越创新中心, 上海 200031; 4. 半导体神经网络智能感知与计算技术北京市重点实验室, 北京 100083）

摘　要: 针对目前主流的基于全卷积神经网络的显著性目标检测方法, 受限于卷积层感受野大小, 低层特征缺少全局性的信息, 而高层特征由于多次池化操作分辨率较低, 无法准确地预测目标边缘等细节的问题, 本文提出了基于注意力的显著性目标检测方法。在 ResNet-50 网络中加入注意力精炼模块, 利用训练样本的显著真值图对空间注意力进行有监督的学习, 使得不同像素位置的相关性更准确。通过深度融合多尺度的特征, 用低层特征优化高层特征, 精修网络的预测结果使其更加准确。在 DUT-OMRON 和 ECSSD 数据集上的测试结果显示, 本文方法能显著提升检测效果, F-measure 和平均绝对误差都优于其他同类方法。

关键词: 显著性目标检测; 深度学习; 全卷积神经网络; 视觉注意力; 多尺度特征; 图像处理; 人工智能; 计算机视觉
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)05−0956−08

中文引用格式: 王凯诚, 鲁华祥, 龚国良, 等. 基于注意力机制的显著性目标检测方法 [J]. 智能系统学报, 2020, 15(5): 956–963.
英文引用格式: WANG Kaicheng, LU Huaxiang, GONG Guoliang, et al. Salient object detection method based on the attention mechanism[J]. CAAI transactions on intelligent systems, 2020, 15(5): 956–963.

Salient object detection method based on the attention mechanism

WANG Kaicheng[1,2], LU Huaxiang[1,3,4], GONG Guoliang[1], CHEN Gang[1]
(1. Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China; 2. School of Future Technology, University of Chinese Academy of Sciences, Beijing 100089, China; 3. Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China; 4. Semiconductor Neural Network Intelligent Perception and Computing Technology Beijing Key Lab, Beijing 100083, China)

Abstract: Salient object detection simulates human visual mechanism. At present, the mainstream methods are based on fully convolutional neural networks. Limited by the receptive fields of convolution layers, low-level features lack a global description of images, whereas high-level features are too coarse to accurately segment details of objects, such as edges, because of multi-stage downsampling operations. To solve this problem, we propose a salient object detection method based on the attention mechanism. We introduce novel attention refinement modules. The ground-truth attention calculated from the training datasets is employed to supervise spatial attention. Through this method, the network learns more accurate position relevance between different pixels. In addition, to refine the output salient maps, we gradually combine the multi-scale features and optimize low-layer features with high-layer features. Sufficient experiments on DUT-OMRON and ECSSD datasets have demonstrated that the proposed method outperforms the others in terms of the value of the F measure and mean absolute error.

Keywords: salient object detection; deep learning; fully convolutional neural network; visual attention; multi-scale features; image processing; artificial intelligence; computer vision

收稿日期: 2019−03−02.　**网络出版日期:** 2020−03−20.
基金项目: 国家自然科学基金项目 (61701473); 中国科学院 STS 计划项目 (KFJ-STS-ZDTP-070); 北京市科技计划项目 (Z181100001518006); 中国科学院国防科技创新基金项目 (CXJJ-17-M152); 中国科学院战略性先导科技专项 (A 类)(XDA18040400).
通信作者: 龚国良. E-mail: gongmianjie@semi.ac.cn.

作为计算机视觉领域基础的任务之一, 显著性目标检测模拟人类能够分辨不同目标重要程度的视觉机制, 提取出图像或视频中感兴趣的信息。利用这一视觉机制可以帮助找到图片中代表

场景语义的重要目标或区域, 而被广泛应用于图像理解[1]、场景解析[2]、目标追踪[3] 等各类计算机视觉任务。显著性目标检测任务一般包含 2 个步骤: 1) 检测出图像或视频中的显著目标; 2) 准确地分割出目标的区域。优秀的模型应当能够准确地检测和分割出显著目标, 以最大化地保留原始图片信息。同时, 作为其他算法的预处理过程, 显著性目标检测模型的复杂度需要尽可能的低, 从而提高计算效率。

在显著性目标检测任务的发展历史中, 首先被提出的是自下而上的计算模型[4-6], 将任务定义为寻找图像中的视觉注视点。人的注意力通常会被对比度、颜色和边缘等特征吸引, 因此基于中心-周围机制就可以由低级特征计算出视觉注视点。这些方法通常包括 3 个步骤: 1) 提取多种低级视觉特征; 2) 计算显著图; 3) 检测少数关键点, 从中找出视觉注视点。由于缺乏对图片全文信息的利用, 单纯依靠低级特征难以在复杂图片中检测出显著目标。为了改进自下而上的计算模型的不足, 科研人员提出基于学习的检测方法, 引入高级视觉特征如语义一致性来判断目标的显著性。Liu 等[7] 将显著性目标检测任务定义为二元检测问题, 使用类别边框来标记显著目标, 建立了第一个大规模图像数据集。对于基于学习的检测方法, 算法的关键在于如何从数据中有效地学习出显著目标与背景之间的语义关系。由于提取高级和多尺度特征的能力, 卷积神经网络 (convolutional neural network, CNN)[8] 消除了对手工提取特征的依赖, 近年来被广泛应用于各种计算机视觉任务中。基于 CNN 的显著性目标检测方法相比传统方法具有更加优异的性能, 已成为了主流研究方向。Long 等[9] 提出全卷积神经网络模型 (fully convolutional network, FCN), 端到端地解决了语义分割问题。如果将显著性目标检测视为二元密集预测任务, 则可以使用 FCN 在数据集上端到端地训练, 标记图片中每个像素点的显著性预测结果。如何进一步地提高 CNN 的网络性能, 就成为了显著性目标检测领域的研究热点所在。

CNN 通过堆叠卷积层来对图像进行多层级的视觉抽象, 较低的卷积层提取例如边缘、角等低级特征, 较高的卷积层提取高级语义模式。各级特征图包含了空间上和通道上的信息, 但由于卷积核的尺寸限制, 往往只有局部的信息得到利用。此外, 多级池化操作导致提取特征图分辨率较低, 缺乏局部区域的细节特征, 无法准确地分割显著目标边缘等细节。针对上述问题, 本文提出了一种基于注意力机制的显著性目标检测方法, 引入注意力精炼模块来改进网络结构, 提高网络对特征图中空间与通道上信息的利用能力, 并对低层与高层特征进行融合。在主流的显著性目标检测数据集上进行对比, 结果表明本文方法效果优于同类方法。

1　相关工作

1.1　显著性目标检测网络

得益于多层级和多尺度特征的提取能力, CNN 无需任何先验知识即可检测出显著目标, 能够更好地定位目标的边缘实现精准分割。He 等[10] 使用一维卷积和池化层学习基于区域的超像素特征, 计算出各个区域的显著性。Wang 等[11] 设计 2 个子网络来单独对低级与高级特征编码, 并对 2 个子网络提取的特征进行拼接, 使用 2 层的多层感知器来判断图像中每个区域的显著性。Wang 等[12] 设计了基于图像掩模的多尺度 R-CNN 框架计算各个区域的显著性, 使用低级特征对照和背景先验来补充语义特征, 同时利用超像素信息来精修目标边缘分割结果。上述几种方法高度依赖分割级的区域信息例如超像素、感兴趣区 (region of interest) 等, 使用图像分类网络来决定每个像素块的显著性。同时由于网络存在最后的全连接层, 空间信息和高级语义特征无法得到有效利用, 导致全局信息损失。而基于 FCN 的方法则通过像素级的操作避免了使用全连接层所带来的不足, 克服了显著目标边缘检测模糊与不准确的问题。Wang 等[13] 设计了递归全卷积网络结构, 并引入低级参照特征和中心先验知识作为显著性先验图到训练和预测过程中。Liu 等[14] 提出了 DHSNet, 在 GV-CNN 网络输出的全局显著图后, 加入分层递归网络来精修显著图。为了避免多级池化造成特征图分辨率较低的问题, Wang 等[15] 设计金字塔池化模块和多阶段提炼机制来改进网络结构。这些方法改进了传统 FCN 方法的不足, 考虑了全局信息和局部信息的整合, 逐层级利用特征图来精修目标边缘等细节。这说明利用像素间的相关信息, 对不同尺度的特征进行融合能有效地改善网络性能, 取得更好的预测结果。

1.2　注意力机制

为了合理利用有限的视觉信息处理能力, 人类需要选择整个视觉区域中特定的部分来集中关

注，这一机制称为注意力机制[16]。视觉的注意力机制在人类处理复杂的场景信息时起到了至关重要的作用。当面对一个复杂场景时，人类首先快速地浏览场景中的全部内容，利用全局的空间信息寻找场景中最重要的部分。然后集中注意力针对这一部分进行深度感知，实现对场景视觉结构的理解。这一机制可以引入到卷积层对输入信息进行编码输出特征图的过程中。如果卷积层的输入图像尺寸为 $H\times W\times 3$，输出为 $H'\times W'\times C'$。那么图片中空间位置的相对信息存储在 $H'\times W'$ 维度上，而某一像素所包含的信息则存储在通道 C' 维度上。Fu 等[17]使用 APN 结构，从整个图片出发，迭代地生成子区域，并对子区域进行预测和整合，从而得到整张图片的预测。Chen 等[18]通过堆叠空间注意力模块来整合图像的区域特征，用于视觉问答任务中，取得了较好的效果。Hu 等[19]提出了压缩-激活模块，使用全局池化来获得通道上的注意力，提高了网络提取并利用特征的能力。从上述方法中可以看到，在神经网络中引入注意力机制可以更好地引导网络关注重点区域，提高网络对信息的提取和利用能力。

2　网络结构

本文在 ResNet-50 网络[20]的基础上，加入注意力精炼模块 (attention refinement module, ARM)，网络结构如图 1 所示。将 4 个注意力精炼模块逐级相连，上一级的输出与 ResNet-50 的特征图拼接，作为下一级模块的输入。通过对特征图的多级精炼，融合高级和低级特征精修网络预测显著图的细节。图 1 中的实线箭头代表着模块输出作为下一级的输入，虚线箭头代表模块的输出用于计算损失训练网络。

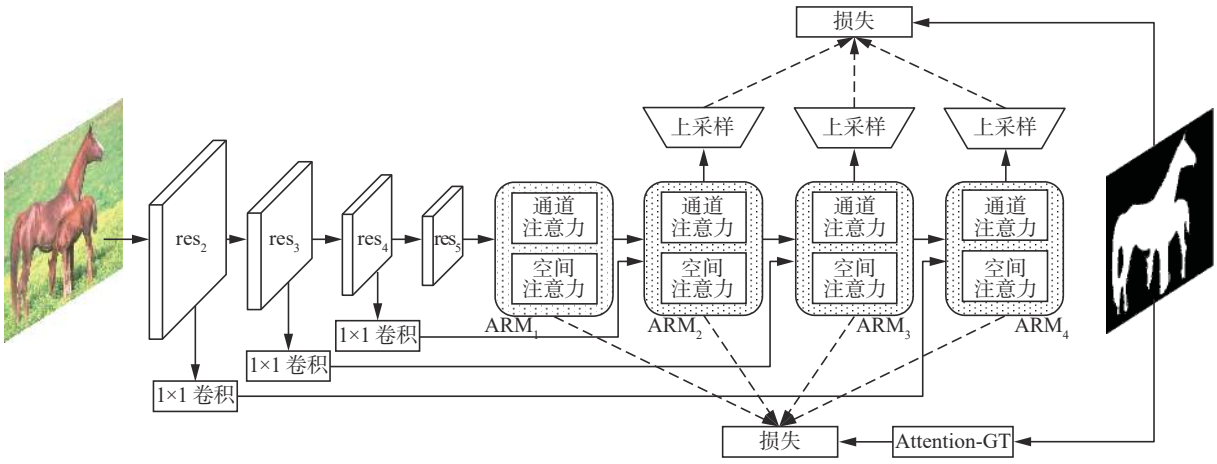

图 1　网络结构

Fig. 1　Network structure

2.1　注意力精炼模块

图 2 为本文的注意力精炼模块，用于处理 ResNet-50 提取的特征图。注意力精炼模块能够引导网络针对性地处理各个通道上的信息，集中关注空间上更有意义的部分。在网络较低的卷积层中，受限于感受野大小，往往无法利用感受野之外的上下文信息。因此使用全局均匀池化 (global average pooling) 来得到各个通道上的统计量 z，将全局空间信息压缩到一个通道表示器中。如图 2 中所示，如果输入特征图 $\boldsymbol{x}$ 尺寸为 $H\times W\times C$，通道统计量 z 在第 c 个通道上的值 z_c 为

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \tag{1}$$

式中 $x_c(i,j)$ 为在第 c 个通道上 (i,j) 位置的特征值。这一操作整合特征图上各个通道中的信息。

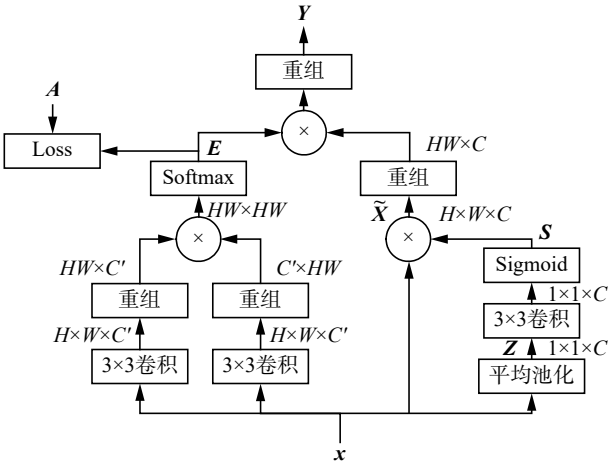

图 2　注意力精炼模块示意

Fig. 2　Attention refinement module

对特征图的通道信息整合后，需要基于通道统计量 z 对原始特征图进行精炼。这一操作应当

满足以下 2 个条件: 1) 是连续可导的非线性变换, 保证可以使用反向传播方法进行训练; 2) 能够学习各个通道之间非互斥的关系, 允许同时强调多个通道上的信息。在此, 与文献 [19] 相同, 使用 1×1 卷积层和 Sigmoid 激活函数来对通道表示器 z 进行非线性映射, 将值限定在 [0,1] 之间, 获得通道注意力向量 $\boldsymbol{s}\in \mathrm{R}^{1\times 1\times C}$ 为

$$\boldsymbol{s}=\sigma(\boldsymbol{W}_1 z) \tag{2}$$

式中: $\boldsymbol{W}_1$ 为 1×1 卷积层的权重; 函数 σ 代表 Sigmoid 函数。与文献 [19] 不同的是, 为了减少模块的参数量, 不使用瓶颈结构 (bottleneck) 对特征通道进行压缩。得到通道注意力向量 $\boldsymbol{s}$ 后, 将 $\boldsymbol{s}$ 与原始的输入特征图 $\boldsymbol{x}$ 在每个通道上进行相乘, 得到通道注意力精炼后的特征图 $\tilde{\boldsymbol{x}}_c$ 为

$$\tilde{\boldsymbol{x}}_c = \boldsymbol{s}_c \boldsymbol{x}_c \tag{3}$$

式中: $\boldsymbol{s}_c$ 和 $\boldsymbol{x}_c\in \mathbf{R}^{H\times W}$ 分别为 $\boldsymbol{s}$ 和 $\boldsymbol{x}$ 在第 c 个通道上的值。

进一步, 参考文献 [20] 中非局部操作 (non-local operation), 考虑特征图上其他像素对某像素的影响, 以此得到空间注意力特征。如果输入特征图 $\boldsymbol{x}$ 尺寸为 $H\times W\times C$, 对于输出 $\boldsymbol{Y}$ 上某一位置的特征向量 $\boldsymbol{y}_i\in \mathbf{R}^{1\times 1\times C}$, 非局部操作定义为

$$\boldsymbol{y}_i = \frac{1}{C(x)}\sum_{\forall j} f(x_i, x_j) g(x_j) \tag{4}$$

式中: $\boldsymbol{x}_i\in \boldsymbol{R}^{1\times 1\times C}$ 是该位置的输入特征向量; $\boldsymbol{x}_j\in \boldsymbol{R}^{1\times 1\times C}$ 是其他位置的输入特征向量; C 是归一化函数; f 是 2 个位置之间的相关函数; 函数 g 计算 $\boldsymbol{x}_j$ 的非线性映射。在这个操作中, 输出考虑了空间中所有位置对该位置特征向量的影响, 这与卷积操作只考虑该点邻域对其的影响不同。本文方法与文献 [20] 不同的是, 使用训练样本的真值图 (ground truth) 来有监督地学习不同位置之间的相关矩阵 $\boldsymbol{E}\in \mathbf{R}^{HW\times HW}$, 与经过通道注意力提炼后的特征图 $\tilde{\boldsymbol{x}}_c$ 进行矩阵乘法, 得到空间注意力精炼后的最终模块输出 $\boldsymbol{Y}$。根据训练样本的真值图, 将其缩放到 $H\times W$,就可以计算出对应的注意力真值图 $\boldsymbol{A}\in \mathbf{R}^{HW\times HW}$(attention-GT)。如果某两个像素同时都属于显著目标所在区域, 那么它们的相关度就为 1, 否则是 0。根据定义, 可以得到 $\boldsymbol{A}$ 为

$$A_{ij}=\begin{cases} 1, & i,j\in T \\ 0, & i,j\notin T \end{cases} \tag{5}$$

式中: (i,j) 为特征图上某个像素的坐标; T 为真值图中显著目标所在像素位置序号的集合。与文献 [20] 中方法相似, 使用 2 个 3×3 卷积层在输入特征图上提取 2 个尺寸为 $H\times W\times C'$ 的特征矩阵, 其中 $C'=C/2$, 将其尺寸分别重塑为 $HW\times C'$ 和 $C'\times HW$, 进行矩阵乘法操作和 Softmax 归一化后就能获得相关矩阵 $\boldsymbol{E}$。

通过上述操作, 使得网络能够根据输入 $\boldsymbol{x}$ 动态调整各个通道上和空间位置上的值, 兼顾了空间信息和特征本身的信息, 提高网络对特征的辨别力。

2.2　多尺度特征融合

为了整合低级和高级特征, 把注意力精炼模块与 ResNet-50 的输出特征图逐级相连, 网络结构如图 1 所示。使用的 ResNet-50 结构与文献 [21] 中一致, 其最后一级特征图 res_5 是输入图像的 32 倍下采样。将其输入到 ARM_1 模块, 得到输出精炼特征图 arm_1, 接着对 arm_1 进行 2 倍上采样, 使其尺寸与 res_4 特征图相同。

为了将低层特征与高层特征融合, 使用 res_4 特征图和 ARM_2 模块对 arm_1 特征图进一步精炼。首先使用 1×1 卷积层将 res_4 特征图通道数压缩到 256, 然后与 arm_1 进行拼接, 将其作为 ARM_2 模块的输入, 得到输出精炼特征图 arm_2。重复这一步骤, 进一步整合 res_3 和 res_2 特征图, 产生 arm_3 和 arm_4 精炼特征图。这样, 实现了低级和高级特征的融合, 使用底层特征优化富有语义信息的高层特征, 对预测显著图进行了多级精修。

2.3　损失函数计算

把相关矩阵 $\boldsymbol{E}$ 的训练定义为回归问题, 根据计算出的注意力真值图 $\boldsymbol{A}$, 使用均方误差损失函数来计算 Loss:

$$\ell_k = \sum_{i,j}(E_k(i,j)-A(i,j))^2 \tag{6}$$

式中, ℓ_k为计算出的均方损失; $E_k(i,j)$ 和 $A(i,j)$ 分别为第 k 模块的 $\boldsymbol{E}$ 和 $\boldsymbol{A}$ 在 (i,j) 位置上的值。这样, 实现了特征图上空间注意力的监督, 更加准确地刻画了不同位置像素间的相关性。

对 arm_2、arm_3、arm_4 精炼特征图分别进行 16 倍、8 倍和 4 倍上采样, 如图 1 中所示, 使用 Sigmoid 计算每级预测显著图 P_n 为

$$P_n = \mathrm{Sigmoid}(\mathrm{arm}_n) \tag{7}$$

选择交叉熵损失函数计算第 n 模块的 P_n 与真值图 G 之间的 Loss:

$$\begin{aligned}\ell_n = -\sum_{i,j}&[G_n(i,j)=1]\log(P_n(i,j))+\\ &[G_n(i,j)=0]\log(1-P_n(i,j))\end{aligned} \tag{8}$$

式中ℓ_n为计算出的交叉熵损失。将每个模块的相关矩阵 $\boldsymbol{E}$ 和每级预测显著图 P_n 的损失函数相加, 得到最终的 Loss 为

$$\mathrm{Loss} = \sum_{n=2}^{4}\ell_n + \sum_{k=1}^{4}\ell_k \tag{9}$$

这样深度监督的训练策略可以使得网络的训练过程更容易收敛，降低训练所需的时间。

3　实验与结果

3.1　实验数据集

目前主流的显著性目标检测数据集包括 MSRA10K[22]、DUT-OMRON[23] 和 ECSSD[24] 等。这些数据集都包含大量的图片，并且样本分布广泛 (动物、植物和生活物体等)，因此被当前显著性目标检测算法用于效果对比中。MSRA10K 数据集包含 10 000 张图片，本文选用其作为模型的训练集。DUT-OMRON 数据集包含 5 158 张拥有复杂背景的图片，考验模型检测内容复杂场景的能力。ECSSD 数据集包含 1 000 张不同尺寸目标的复杂图像。本文选择 DUT-OMRON 和 ECSSD 数据集作为测试集，对比本文模型与其他模型的效果。

3.2　网络训练

本文选择“MXNET”深度学习框架，在 2 块 Titan Xp Pascal GPU 上进行网络的训练和测试。首先将 MSRA10K 训练集中的图片填充到 416×416 的大小，然后进行原尺寸 2/3 到 3/2 之间的随机尺度放缩。最后将图片随机裁剪到 416×416 像素，减去像素均值，送入网络中进行训练。

本文使用在 ImageNet 数据集[25] 上训练的 ResNet-50 模型作为网络的初始权重，再在 MSRA10K 数据集上进行微调 (finetune)。本文选择带有冲量 (momentum) 的随机梯度下降 (stochastic gradient descent) 作为梯度更新算法，使用文献 [26] 中的“Poly”学习率调整策略。mini-batch 的值设置为 16，训练共 45 000 个迭代。基准学习率设置为 0.01，下降率指数为 0.9。设置冲量值为 0.9，权重衰减 0.000 1。

3.3　评价标准

本文使用平均绝对误差 (MAE) 和 F-measure 值来评价算法在数据集上的测试结果。MAE 定义为输出预测结果 P 与二元真实值 G 在每个像素上错误率平均值，为

$$\mathrm{MAE}=\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}|P(x,y)-G(x,y)| \tag{10}$$

式中 W 和 H 分别为预测结果 P 的宽和高。MAE 越低意味着网络预测的准确率越高。

F-measure 是对算法的综合评价指标：

$$F_\beta=\frac{(1+\beta^2)\cdot P\cdot R}{\beta^2\cdot P+R} \tag{11}$$

式中：P 和 R 分别为平均准确率和平均召回率；β 设置为 0.3。为了强调准确率的重要性。F-measure 指标越高意味着网络预测的效果越好。与文献 [27] 相同，本文使用不固定的阈值来计算这准确率和召回率，设置该阈值为显著图平均值的 2 倍。有了 MAE 和 F-measure 这 2 个评价指标，就可以将本文方法与其他主流方法进行对比。

3.4　实验结果与性能对比

对比本文方法和其他 11 种主流的显著性目标检测方法，包括 SRM[15]、DRFI[28]、BL[29]、LEGS[11]、MDF[30]、MCDL[31]、DS[32]、DHS[14]、ELD[33]、DCL[34]、KSR[35] 和 RFCN[13]。使用作者文章中方法和参数设置训练网络，测试得到显著图或者直接使用其提供的显著图，再根据显著图计算在不同测试集上的结果，如表 1 所示。

表 1　本文方法与其他方法效果的对比

Table 1　Comparison between our method and the others on performance

方法	DUT-OMRON		ECSSD	
	F-measure	MAE	F-measure	MAE
SRM	0.707	0.069	0.892	0.056
DHS	—	—	0.871	0.063
RFCN	0.627	0.111	0.834	0.109
DCL	0.684	0.157	0.827	0.151
ELD	0.611	0.092	0.810	0.082
DS	0.603	0.120	0.821	0.124
MDF	0.644	0.092	0.805	0.108
MCDL	0.625	0.089	0.796	0.102
LEGS	0.592	0.133	0.785	0.119
BL	0.499	0.239	0.684	0.217
DRFI	0.550	0.138	0.733	0.166
本文方法	**0.720**	**0.064**	**0.906**	**0.049**

从表 1 可以看到，本文方法表现好过当前主流的显著性检测方法。对于 F-measure 指标，本文在 DUT-OMRON 和 ECSSD 数据集上分别超过表现最佳的 SRM 方法 0.013 和 0.014。同样，对于 MAE 指标，本文分别超过 SRM 方法 0.005 和 0.007。

为了研究注意力精炼模块的效果，在 ECSSD 数据集上测试了网络处于不同多尺度特征融合阶段时的表现，结果如表 2 所示。表 2 中 Baseline 代表直接使用 res_5 特征图进行 32 倍上采样得到显著图。可以看到，随着多尺度特征的融合，注意力精炼模块大幅度地提升了网络的表现，F-measure 指标提高了 0.057，MAE 指标降低了接近一半到达了 0.049。如图 3 所示，在 ECSSD 数据集上的预测结果图说明网络能够成功检测出植物、动物和人等显著目标，并对目标边缘进行准确分割。上述结果证明本文方法提高了网络对特征的利用能力，融合了多尺度的特征精修了预测结果，大幅度提高了网络的性能表现。

表 2　不同尺度融合阶段网络性能差异

Table 2　Performance of our network in different stages

阶段	ECSSD		上采样率
	F-measure	MAE	
Baseline	0.849	0.097	32×
arm_2	0.878	0.060	16×
arm_3	0.890	0.057	8×
arm_4	0.906	0.049	4×

图 3　在 ECSSD 数据集上的预测结果

Fig. 3　Predicted results on ECSSD dataset

为了分析本文提出的注意力精炼模块对通道注意力的整合效果, 对比了文献 [36] 中的通道注意力模块 (channel attention block) 与本文方法在 ECSSD 数据集上测试的结果, 结果如表 3 所示, 其中 CAB 表示使用通道注意力模块整合通道上的信息, ARM 表示使用本文的注意力精炼模块中的通道注意力对特征图进行精炼, 而不使用空间注意力。ARM+AS 表示使用训练样本真值图对只使用通道注意力的注意力精炼模块进行额外监督。结果可以发现, 本文方法相较文章中的方法, F-measure 指标提高了 0.027, MAE 指标降低了 0.011, 展示出更加优异的通道全局信息整合能力。

表 3　通道注意力精炼效果对比

Table 3　Comparison on channel attention refinement performance

使用方法	ECSSD	
	F-measure	MAE
Baseline	0.849	0.097
CAB	0.879	0.06
ARM	0.89	0.057
ARM+AS	0.906	0.049

进一步, 对比其他文章中上下文提取模块方法, 分析本文提出的注意力精炼模块对空间注意力特征的利用情况, 如表 4 所示。其中 DeepLab V2 指文献 [26] 中提出的使用全连接的条件随机场来整合空间信息, PSP-Net 为文献 [37] 中提出的特征金字塔方法, DeepLab V3[38] 使用空洞卷积提高网络的感受野, 提高网络对多尺度信息的获取能力。根据结果可以发现, 本文方法相较性能最好的 DeepLab V3 方法提高了 0.016 的 F-measure 指标, 降低了 0.008 的 MAE 指标, 取得了更加优异的空间注意力精炼效果。

表 4　空间注意力精炼效果对比

Table 4　Comparison on spatial attention refinement performance

使用方法	ECSSD	
	F-measure	MAE
DeepLab V2	0.866	0.083
PSP-Net	0.889	0.058
DeepLab V3	0.882	0.057
本文方法	0.906	0.049

4　结束语

本文提出了一种基于注意力机制的显著性目标检测方法, 设计注意力精炼模块融合通道和空间注意力, 使得网络能够根据输入特征图选取其中重要的信息。使用训练样本的真值图有监督地训练空间注意力, 提高了像素间相关关系的准确性。最后, 本文将注意力精炼模块逐级连接, 使用低级特征精修高级语义特征, 修正预测显著图细节, 实现了多尺度特征的融合。在 MSRA10K 数据集上训练模型后, 在 DUT-OMRON 和 ECSSD 数据集上进行测试, 并在 ECSSD 数据集上与其他主流通道和空间特征提取方法对比。实现结果表明, 与目前主流的显著性目标检测方法相比, 本文提出的方法能够更有效地精炼特征图上的通道和空间信息, 因此取得了更加优异的效果。

参考文献:

[1] ZHANG Fan, DU Bo, ZHANG Liangpei. Salien-cy-guided unsupervised feature learning for scene classi-fication[J]. IEEE transactions on geoscience and remote sensing, 2015, 53(4): 2175–2184.

[2] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE transactions on pattern analysis and machine intelligence, 1998,

20(11): 1254–1259.

[3] HONG S, YOU T, KWAK S, et al. Online tracking by learning discriminative saliency map with convolutional neural network[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France, 2015: 597−606.

[4] TREISMAN A M, GELADE G. A feature-integration theory of attention[J]. Cognitive psychology, 1980, 12(1): 97–136.

[5] KOCH C, ULLMAN S. Shifts in selective visual attention: towards the underlying neural circuitry[J]. Human neurobiology, 1985, 4(4): 219–227.

[6] WOLFE J M, CAVE K R, FRANZEL S L. Guided search: an alternative to the feature integration model for visual search[J]. Journal of experimental psychology: human perception and performance, 1989, 15(3): 419–433.

[7] LIU Tie, SUN Jian, ZHENG Nanning, et al. Learning to detect a salient object[C]//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1−8.

[8] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradi-ent-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[9] LONG J, SHELHAMER E, DARRELL T. Fully convo-lutional networks for semantic segmenta-tion[C]//Proceedings of 2015 IEEE Conference on Com-puter Vision and Pattern Recognition. Boston, USA, 2015: 3431−3440.

[10] HE Shengfeng, LAU R W, LIU Wenxi, et al. Supercnn: a superpixelwise convolutional neural network for salient object detection[J]. International journal of computer vision, 2015, 115(3): 330–344.

[11] WANG Lijun, LU Huchuan, RUAN Xiang, et al. Deep networks for saliency detection via local estimation and global search[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3183−3192.

[12] WANG Xiang, MA Huimin, CHEN Xiaozhi. Salient object detection via fast R-CNN and low-level cues[C]//Proceedings of 2016 IEEE International Con-ference on Image Processing (ICIP). Phoenix, USA, 2016: 1042−1046.

[13] WANG Linzhao, WANG Lijun, LU Huchuan, et al. Sa-liency detection with recurrent fully convolutional networks[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands, 2016: 825−841.

[14] LIU Nian, HAN Junwei. Dhsnet: deep hierarchical saliency network for salient object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 678−686.

[15] WANG Tiantian, BORJI A, ZHANG Lihe, et al. A stage-wise refinement model for detecting salient objects in images[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4039−4048.

[16] LAROCHELLE H, HINTON G. Learning to combine foveal glimpses with a third-order Boltzmann machine[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 1243−1251.

[17] FU Jianlong, ZHENG Heliang, MEI Tao. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 4476−4484.

[18] CHEN Long, ZHANG Hanwang, XIAO Jun, et al. SCA-CNN: spatial and channel-wise attention in convo-lutional networks for image captioning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 6298−6306.

[19] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132−7141.

[20] WANG Xiaolong, GIRSHICK R B, GUPTA A, et al. Non-local neural networks[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7794−7803.

[21] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.

[22] CHENG Mingming, MITRA N J, HUANG Xiaolei, et al. Global contrast based salient region detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(3): 569–582.

[23] YANG Chuan, ZHANG Lihe, LU Huchuan, et al. Sali-ency detection via graph-based manifold rank-ing[C]//Proceedings of 2013 IEEE Conference on Com-puter Vision and Pattern Recognition. Portland, USA, 2013: 3166−3173.

[24] YAN Qiong, XU Li, SHI Jianping, et al. Hierarchical saliency detection[C]//Proceedings of 2013 IEEE Con-ference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 1155−1162.

[25] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pat-

tern Recognition. Miami, USA, 2009: 248−255.

[26] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(4): 834–848.

[27] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1597−1604.

[28] JIANG Huaizu, WANG Jingdong, YUAN Zejian, et al. Salient object detection: a discriminative regional feature integration approach[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2083−2090.

[29] TONG Na, LU Huchuan, RUAN Xiang, et al. Salient object detection via bootstrap learning[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1884−1892.

[30] LI Guanbin, YU Yizhou. Visual saliency based on multiscale deep features[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5455−5463.

[31] ZHAO Rui, OUYANG Wanli, LI Hongsheng, et al. Saliency detection by multi-context deep learn-ing[C]//Proceedings of 2015 IEEE Conference on Com-puter Vision and Pattern Recognition. Boston, USA, 2015: 1265−1274.

[32] LI Xi, ZHAO Liming, WEI Lina, et al. DeepSaliency: multi-task deep neural network model for salient object detection[J]. IEEE transactions on image processing, 2016, 25(8): 3919–3930.

[33] LEE G, TAI Y, KIM J. Deep saliency with encoded low level distance map and high level features[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 660−668.

[34] LI Guanbin, YU Yizhou. Deep contrast learning for salient object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 478−487.

[35] WANG Tiantian, ZHANG Lihe, LU Huchuan, et al. Kernelized subspace ranking for saliency detec-tion[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands, 2016: 450−466.

[36] YU Changqian, WANG Jingbo, PENG Chao, et al. Learning a discriminative feature network for semantic segmentation[C]//Proceedings of 2018 IEEE/CVF Con-ference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1857−1866.

[37] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6230−6239.

[38] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv: 1706.05587, 2017.

作者简介：

王凯诚，硕士研究生，主要研究方向为神经网络芯片、机器学习。

鲁华祥，研究员，博士生导师，主要研究方向为类神经计算芯片、类脑神经计算技术和应用系统、信息与信号处理。出版专著 1 部，授权发明专利 10 项。发表学术论文 40 余篇。

龚国良，副研究员，主要研究方向为智能算法与类脑计算系统、图像处理芯片、AI 芯片、神经网络算法及其应用研究。授权发明专利 4 项。发表学术论文 6 篇。