基于多特征注意力循环网络的显著性检测

卢珊妹[1,2], 郭强[1,2)*], 王任[1,2], 张彩明[2,3]

[1] (山东财经大学计算机科学与技术学院 济南 250014)
[2] (山东省数字媒体技术重点实验室 济南 250014)
[3] (山东大学软件学院 济南 250010)
(guoqiang@sdufe.edu.cn)

摘 要: 特征表达是图像显著性检测的关键, 现有方法所提取的特征缺乏一定的可辨识性. 为此, 提出多尺度上下文特征提取机制和注意力循环机制来解决这一问题. 多尺度上下文特征提取机制通过空洞卷积增大高层特征的感受野来获取丰富的上下文语义特征, 并采用向量聚合策略对特征进行融合. 为增强融合特征的可辨识性, 利用注意力机制自适应地对卷积特征增加权重以区分每个像素的重要性, 使注意力集中于显著性区域, 并抑制背景中的干扰信息. 在此基础上, 采用循环网络能够有效地在空间位置上对卷积特征进行逐步细化, 进一步调整了显著性区域及其边缘, 从而生成精确的显著图. 该方法在 5 个常用的数据集上与 8 种相关方法进行了比较. 实验结果表明, 该方法不仅能够生成更加准确与完整的显著图, 而且其 MAE 和最大 *F*-measure 量化性能也有所提升.

关键词: 空洞卷积; 多尺度特征; 注意力机制; 循环网络; 显著性检测
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2020.18240

Salient Object Detection Using Multi-Scale Features with Attention Recurrent Mechanism

Lu Shanmei[1,2], Guo Qiang[1,2)*], Wang Ren[1,2], and Zhang Caiming[2,3]

[1] (*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan* 250014)
[2] (*Shandong Key Laboratory of Digital Media Technology, Jinan* 250014)
[3] (*Software College, Shandong University, Jinan* 250010)

Abstract: Feature representation is a key component to salient object detection. However, the features extracted by existing methods lack capability of discrimination for salient object detection. Hence, multi-scale context features extraction mechanism and attention recurrent mechanism are proposed to solve this problem. Specifically, the multi-scale context features extraction mechanism uses atrous convolution, which can expand receptive fields of high-level convolution features to obtain rich context sematic features, and adopts a vector aggregation strategy to fuse these features. In order to enhance the discriminative power of fused features, the proposed method adopts a attention mechanism to distinguish the importance of each pixel by adaptively increasing the weight of convolution features. The attention mechanism can focus on salient regions while suppressing the interference information in the background. Furthermore, a recurrent network is presented to gradually refine the convolution features in spatial position, and adjust salient regions to generate accurate saliency maps. The proposed method is compared with eight state-of-the-art methods on five benchmark datasets. Experimental results show that the method

收稿日期: 2020-02-19 ; 修回日期: 2020-09-07. **基金项目**: 国家自然科学基金(61873145, 61802229); 山东省自然科学省属高校优秀青年联合基金(ZR2017JL029); 山东省自然科学基金(ZR2018BF007); 山东省高等学校青创科技计划(2019KJN045); 山东省优势学科人才团队. **卢珊妹**(1996—), 女, 硕士研究生, 主要研究方向为显著性检测; **郭强**(1979—), 男, 博士, 教授, 硕士生导师, 论文通讯作者, 主要研究方向为计算机视觉、数字图像处理、高维数据分析; **王任**(1995—), 男, 硕士研究生, 主要研究方向为产品缺陷检测; **张彩明**(1955—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算机辅助几何设计、计算机图形学、医学图像处理.

can generate more accurate and complete saliency maps, and achieves good performance in both mean absolute error and maximum *F*-measure.

Key words: atrous convolution; multi-scale feature; attention mechanism; recurrent network; saliency detection

作为图像分析的重要预处理步骤, 显著性检测受到了许多研究人员的关注. 其目标是模拟人眼的视觉特征, 分辨出图像中最重要、最明显的目标区域, 并将其与背景区域分离. 该任务在手势识别[1]、视觉追踪[2]、图像检索[3]、语义分割[4]等多方面有着广泛的应用.

近年来, 深度学习已经成为机器学习中最受关注的工具, 并在计算机视觉中取得了突破性的进展. 由于深度学习网络能够提取图像的多层级特征和多尺度特征, 因此能够准确地捕获最显著性的区域. 基于深度学习的显著性检测通常是利用卷积神经网络(convolutional neural networks, CNN)较强的特征提取能力获取图像的特征, 并通过全连接层对特征进行加权平均来预测图像区域的显著性分数[5-6]. 然而, 全连接层不仅包含大量的参数, 使模型的计算复杂度增加, 还会使图像的空间信息丢失, 导致不准确的检测结果. 为了保持图像的空间信息并降低计算复杂度, 文献[7]用卷积层代替 CNN 中所有的全连接层, 提出了全卷积神经网络(fully convolutional networks, FCN). 此后, 各种基于 FCN 的显著性检测方法被提出[8-11]. 基于 FCN 的方法采用预训练的模型(VGG[12]或 ResNet[13]等)作为编码器, 提取图像的多层级特征. 在解码阶段, 通常会利用1×1的卷积层或 Softmax 层将所提取的低层特征和高层特征融合, 生成最终的显著性效果图.

基于 FCN 的显著性检测的关键是如何获取图像有效的卷积特征. 有效的卷积特征主要是指学习到的特征应包含丰富的上下文语义信息以及特征具备辨识性. 上下文语义信息能够捕捉不同区域之间相互作用的信息, 有利于提高预测的准确性. 在基于 FCN 方法提取的多层级特征中, 高分辨率的低层特征内含空间细节, 而低分辨率的高层特征包含丰富的语义信息. 由于受到感受野的限制, 仅利用 FCN 的多层级特征并不能获取丰富的上下文信息; 并且与高层特征相比, 低层特征主要反映细节信息, 其高层语义表示能力不足, 对显著性模型的性能没有明显的帮助. 因此, 本文利用 FCN 中高层卷积模块的多尺度特征来获取更大的感受野, 从而得到丰富的上下文信息, 以提高检测算法的性能. 有效卷积特征的辨识性是指所学特征更易于区分每个像素的显著性. 尽管多尺度特征包含丰富的上下文信息, 但是将多尺度特征直接聚合的方式使聚合特征对显著性区域和背景有着同等的关注度, 无法将两者准确地区分, 从而导致生成的显著图不够精细.

现有工作主要利用注意力机制使卷积特征具备辨识性[9,14]. 注意力机制是对人类大脑信号处理机制的一种模拟, 其可以从大量数据中快速且准确地捕捉最重要的信息. 得益于其特征选择能力, 注意力机制在计算机视觉领域中得到了广泛的应用. 在显著性检测中, 注意力机制通常通过计算出每个像素点的权重, 以区分像素点的显著程度, 这使注意力能够集中于显著性区域, 并过滤背景中的一些干扰信息, 从而实现显著性区域的精准定位. 然而, 这些方法都依赖于准确的注意力图. 当注意力图出现偏差时, 预测的显著图往往也会出现较大的误差. 为了纠正出现偏差的注意力图, 采用循环网络是一种可行的解决方法. 循环网络具有一定的记忆性, 在每个时间步长中将输入特征和隐藏状态特征输入到相同的结构, 从而捕捉输入特征的长期依赖. 同时, 循环网络能够抑制特征中的干扰信息, 实现对特征的细化. 本文将注意力机制嵌入循环网络中, 两者共同作用增强了模型的特征辨识能力, 不仅可以有效地抑制背景中的干扰信息, 而且以一种迭代的方式调整显著性区域, 达到了细化显著图的目的.

为了使用有效的卷积特征来提升显著性检测的性能, 本文提出了一种基于多特征注意力循环网络的显著性检测模型. 该模型以 FCN 为基础, 用于提取图像的多层级特征. 为保证高效的计算, 本文利用不同感受野的空洞卷积层仅提取其后 3 个卷积模块的多尺度特征, 从而获取丰富的上下文语义信息. 在此基础上, 进一步采用注意力循环机制使多尺度特征具有辨识性. 注意力循环机制不仅能聚焦于图像的显著性区域, 而且能够对显著性区域做出适当的调整, 并进一步细化其边缘, 极大地提高了模型的性能.

1 相关工作

在过去的 20 多年中, 研究者提出了大量的显著性检测算法. 传统的显著性检测方法通常利用图像的低级特征, 如颜色、强度、对比度和梯度等特征来预测图像的显著性[15-16]. 虽然这类方法能够检测图像中主要的显著性物体, 但是在处理复杂场景图像时效果不够理想.

深度学习的兴起极大地推动了计算机视觉的发展, 促使显著性检测取得了很大的突破. 早期的基于深度学习的方法是利用 CNN 进行显著性检测. 其主要通过预测每个图像块(超像素或建议框)的显著性分数来得到显著图. Wang 等[17]采用 2 个深度神经网络(deep neural networks, DNN) DNN-L 和 DNN-G 分别进行局部估计和全局搜索, 并结合建议框共同预测图像的显著性. Li 等[6]通过计算每个超像素的显著性值来提取图像的上下文特征. 虽然这类基于 CNN 的方法能够通过卷积实现多尺度特征提取, 但并不能充分利用高层语义信息, 而且空间信息不能传播到最后的全连接层中, 从而造成全局信息的丢失. 同时, 全连接层的存在会增大整个网络的计算复杂度.

与基于 CNN 的模型利用图像块进行操作不同, FCN 考虑像素级别的操作来克服由全连接层引起的问题[18]. 基于 FCN 的模型移除了 CNN 中的全连接层, 低层趋向于编码更多的细节信息, 高层更偏向于提取全局语义信息. 然而单纯的高层语义信息缺乏图像的空间细节信息, 预测的显著性图不足以保持精细的区域边界. 为了更好地保持图像的空间信息, 研究者开始采用多尺度卷积特征进行显著性检测. Zhang 等[19]将低层和高层特征图集成来获取多尺度特征. 该方法只是简单地将多层级特征进行融合, 而不区分它们的重要性, 从而造成了信息的大量冗余. 文献[9]通过在多个 side-output 层之间加入短连接, 将高层语义信息传输到低层, 以得到多尺度特征, 但其所获得的上下文信息比较有限. Wang 等[20]采用金字塔注意模块, 通过考虑显著性的多尺度注意来增强模型的显著性表示能力. 与上述工作不同, 本文采用空洞卷积来增大特征图的感受野, 以获取信息丰富的多尺度上下文特征. 同时, 对于多尺度特征中存在一定的背景干扰信息, 仅利用多尺度上下文特征并不能保证准确的显著性预测, 需要一个机制过滤其中的干扰信息.

注意力机制能够在一定程度上缓解上述问题, 其可以抑制背景中的非显著性线索, 从而突出显著性区域. 已有研究工作证明, 注意力机制可使显著性检测模型的性能得到提升. 文献[21]利用传统的空间注意力机制对多尺度特征赋予权重, 但该方法中的多尺度仅仅是指特征尺寸的不同, 并不能获取丰富的上下文信息. 与传统的空间注意力机制不同, Chen 等[10]运用一种反向空间注意机制以自上而下的方式来指导 side-output 的残差学习, 补充缺失的部分目标和细节, 以提高显著性目标的完整性. 为了进一步准确地预测显著图, Zhang 等[14]提出了一种将空间和通道注意力机制结合的层级注意力机制, 该机制可以选择性地结合多层级上下文信息, 有效地抑制干扰信息. 然而, 上述方法生成的注意力图可能会存在偏差, 从而影响显著图的准确性. 为此, 本文通过将注意力机制与循环网络结合, 利用循环网络来逐步校正带有偏差的注意力图, 进而细化显著性区域, 使预测的显著图更接近真值图.

循环网络具有记忆功能并且参数共享, 它的输出取决于当前的输入和记忆, 能够保持序列前后的依赖关系, 其可用于对特征中的信息进行过滤, 实现对显著图的逐步细化. 文献[22]在每个卷积层加入循环连接的深度监督网络, 并在网络中间层加入 side-output 层来监督特征的学习. 虽然该模型能够充分利用多尺度上下文信息, 但在处理对比度不高或显著性对象较小的图像时性能有待进一步提高. Kuen 等[23]提出了循环注意力卷积-反卷积网络(recurrent attentional convolutional-deconvolution networks, RACDNN). 在 RACDNN 的每个时间步长中, 通过空间变换选择输入图像的一个子区域作为空间注意力. 随后这个子区域作为下一个时间步长的输入, 从而生成局部的显著图, 用于细化预测图所对应的区域. 尽管注意力机制可以迫使模型聚焦每一个时间步长的子区域, 但是子区域的重叠会导致显著性信息的冗余, 生成有背景干扰的显著图. 此外, 长短期记忆网络(long short-term memory, LSTM)也可以用于构建显著性注意力模型. 文献[24]基于 LSTM 提出了一种人眼注视点预测模型, 将注意力机制和卷积 LSTM 相结合, 在迭代的过程中聚焦到相应的空间位置, 细化了显著性特征, 能够较好地预测人眼注视点. 直觉上, 该模型也可用于定位图像中的显著性区域. 由于预测人眼注视点不需要精细的边缘, 而图像显著性目标位置通常是多变的, 这使得直接将该模型应用于显著性检测仅能得到粗略的显著图, 边缘不够准确. 为此, 本文结合前述的多尺度上下文融合特征和注意力循环机制, 提出了一种基于

多特征的注意力循环网络进行显著性检测，以实现更加准确的显著性检测.

2　本文方法

本文提出的基于多特征注意力循环网络的显著性检测模型包含 2 个组成部分：多尺度上下文特征提取模块以及注意力循环模块. 整体网络流程图如图 1 所示. 在多尺度上下文特征提取阶段，为保持图像的空间信息，本文使用基于 VGG-16 的 FCN 结构(即移除全连接层)作为预训练模型. 首先将图像输入到 FCN 中提取显著性区域的多层级特征；然后利用感受野不同的空洞卷积层提取最后 3 个卷积模块的多尺度特征；再将这些特征聚合并使用卷积操作实现特征融合，获取图像的上下文语义信息，从而准确地捕获显著性区域；最后利用由空间注意力机制和循环神经网络(recurrent neural network, RNN)组成的注意力循环模块，进一步细化融合后的多尺度上下文特征，以突出显著性区域并使显著性区域的边缘更加精细.

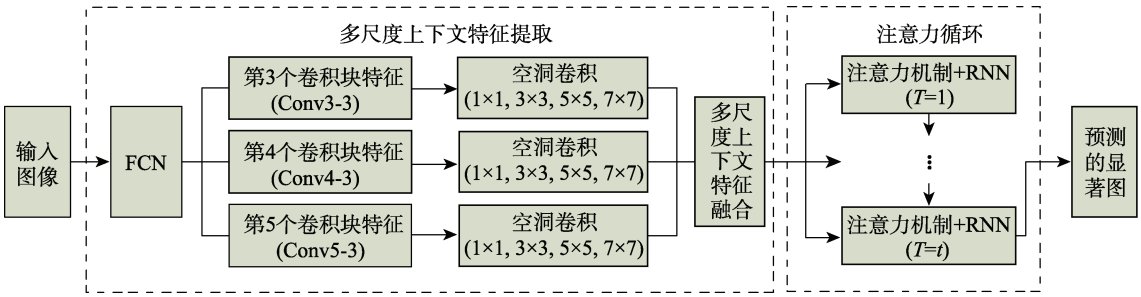


图 1　多特征注意力循环网络流程图

2.1　多尺度上下文特征提取

鉴于 FCN 简单的结构和强大的特征提取能力，本文选取基于 FCN 的网络作为模型的主干，用于提取图像的多层级特征. FCN 继承了 VGG-16 的 13 个卷积层，这些卷积层被划分为 5 组，每组后都紧接一个步长为 2 的最大池化层. 多个池化层会使得显著性预测图的输出尺寸太小，细节信息丢失太多，导致解码操作很难将其恢复，从而降低了预测的准确性. 为了解决这个问题，本文移除了最后一个池化层，使得最后一层的卷积特征能够保持更多的细节信息.

由于不同图像中的显著性区域位置、形状和尺寸都有很大的不同，使得显著性区域与背景之间的相互联系对显著性检测具有重要影响. 图像中的上下文信息能够直接捕获各个区域的相互联系，因此学习丰富的上下文语义特征是非常必要的. FCN 模型的高层能够提取丰富的语义特征，但是它缺乏提取上下文信息的能力，导致其提取的单尺度特征不能有效地预测复杂图像的显著性. 多尺度特征的提取能够捕获丰富的上下文语义信息，获取多尺度特征的策略主要有 2 种：(1) 采用在最后一个预测层上使用金字塔池化提取多尺度特征[25]，但大尺寸的池化层会导致重要信息的丢失；(2) 采用多尺度上下文感知网络[26]，通过堆叠一系列包含空洞卷积的模块来提取多尺度上下文信息. 与普通的卷积操作相比，空洞卷积能够增大特征的感受野而不增加模型的计算量. 受此启发，本文使用空洞卷积进行多尺度上下文特征提取来学习图像的上下文信息，如图 2 所示. 图中的前 5 个卷积模块表示 FCN 结构，后面的 3 个模块分别表示从

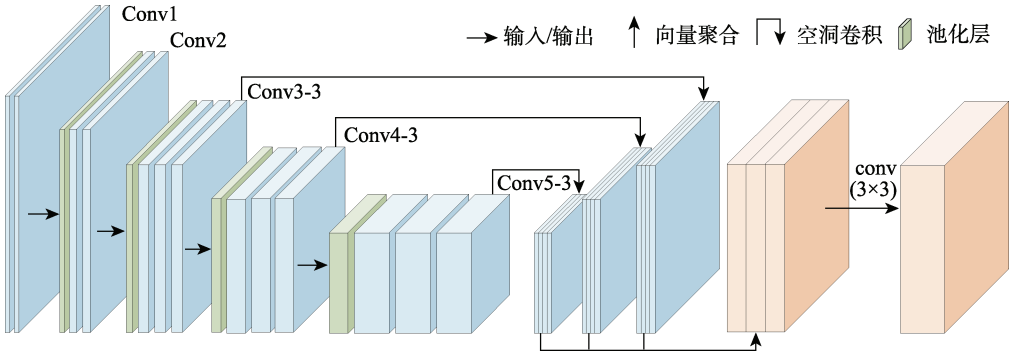


图 2　特征提取机制

最后 3 个卷积块提取的多尺度上下文特征；将其聚合后，使用 3×3 的卷积层可以得到多尺度上下文融合特征.

由于 FCN 提取的浅层特征所包含的语义信息较为有限，对于显著性预测性能没有明显的帮助，并且浅层特征的分辨率较高，使得对其进行操作会增加整个模型的计算量，因此本文只提取 FCN 高层卷积模块的多尺度特征. 每个卷积模块进行多尺度特征提取时，卷积核尺寸为 3×3，通道数为 64，扩张率为 1, 3, 5 和 7，输出尺寸如表 1 所示.

表 1　多尺度特征提取

层名称	输出尺寸
Conv3-3	88×88×256
Conv4-3	44×44×256
Conv5-3	22×22×256

对于大小为 $h\times w$ 的输入图像 $\boldsymbol{I}$，先用 FCN 提取 5 个尺寸的特征图，表示为 $\boldsymbol{F}=\{\boldsymbol{f}_i,\ i=1,2,3,4,5\}$，其分辨率为 $\left[\frac{h}{2^{i-1}},\frac{w}{2^{i-1}}\right]$. 为了获取更深层次的语义信息，用 4 个不同的空洞卷积层分别提取后 3 个卷积模块的多尺度特征. 这 4 个空洞卷积层的卷积核大小都是 3×3，扩张率分别为 1，3，5 和 7. 扩张率越大，其获取的感受野越大，特征所能感受到的区域越大. 例如，3×3 大小的卷积核，当扩张率为 3 时，它能够获取 5×5 的感受野. 本文方法提取的多尺度特征能够利用不同的感受野感知各个区域之间的联系，因此它包含丰富的上下文语义信息. 每个卷积模块的多尺度上下文特征用 $\boldsymbol{F}^{\mathrm{m}}=\{\boldsymbol{f}_i^{\mathrm{m}},\ i=3,4,5\}$ 表示，采用聚合操作将 3 个卷积模块的多尺度特征整合到一起，再用卷积核大小为 3×3 的卷积操作将其融合，得到多尺度上下文融合特征为

$$\boldsymbol{F}_{\mathrm{fuse}}=\mathrm{Conv}(\mathrm{Cat}(\boldsymbol{f}_3^{\mathrm{m}},\mathrm{Up}(\boldsymbol{f}_4^{\mathrm{m}}),\mathrm{Up}(\mathrm{Up}(\boldsymbol{f}_5^{\mathrm{m}})))).$$

其中，$\mathrm{Up}(\cdot)$ 表示因子为 2 的上采样操作；$\mathrm{Cat}(\cdot)$ 表示聚合操作；$\mathrm{Conv}(\cdot)$ 是 3×3 的卷积操作. 利用该机制可以得到上下文信息丰富的多尺度融合特征，有利于更准确地进行显著性预测.

2.2　注意力循环机制

在显著性检测中，注意力机制和循环网络发挥着重要的作用. 注意力机制能够定位显著性区域，区分其与背景区域，并给予不同的关注度. 循环网络则能够进一步强化显著性区域，并细化其边缘. 多尺度融合特征虽然包含丰富的上下文信息，但是利用该特征直接进行显著性检测时，显著性区域会出现定位不准确或者不完整的情况. 为此，本文扩展了传统的 RNN 模型，在 RNN 的基础上加入空间注意力机制，使其能够处理图像的空间特征. 注意力循环机制是针对像素进行操作的，它促使模型将注意力集中于显著性对象，抑制多尺度融合特征中的一些干扰信息. 同时，该注意力循环机制不需要处理输入特征在时间上的依赖关系，而是利用 RNN 的序列属性以一种迭代的方式处理特征图，实现了由粗糙到精细的显著性预测.

本文采用的注意力循环机制结构如图 3 所示. 该模块使用多尺度融合特征初始化隐藏层 $\boldsymbol{H}_0$，它与多尺度融合特征共同被几个相同模块组成的注意力循环机制迭代细化，得到最终的显著性效果图. 多尺度特征 $\boldsymbol{F}_{\mathrm{fuse}}$ 为该模块的输入. $\boldsymbol{F}_{\mathrm{fuse}}$ 在每个时间步长都需要重新输入注意力循环机制，对细化后的特征进行强化和补充. 它与上一步的隐藏状态 $\boldsymbol{H}_{t-1}$ 进行卷积，将结果输入到 tanh 激活函数，得到中间特征

$$\boldsymbol{X}_t=\boldsymbol{V}_{\mathrm{a}}*\tanh\left(\boldsymbol{W}_{\mathrm{a}}*\boldsymbol{F}_{\mathrm{fuse}}+\boldsymbol{U}_{\mathrm{a}}*\boldsymbol{H}_{t-1}+\boldsymbol{b}_{\mathrm{a}}\right).$$

其中，$\boldsymbol{V}_{\mathrm{a}}$，$\boldsymbol{W}_{\mathrm{a}}$，$\boldsymbol{U}_{\mathrm{a}}$ 均表示卷积核；$\boldsymbol{b}_{\mathrm{a}}$ 表示偏置项. 在 $\boldsymbol{X}_t$ 上对每个像素进行 Softmax 操作，得到一幅归一化的空间注意力图

$$A_t^{xy}=\mathrm{Softmax}\left(X_t^{xy}\right)=\frac{\exp\left(X_t^{xy}\right)}{\sum_x\sum_y\exp\left(X_t^{xy}\right)}.$$

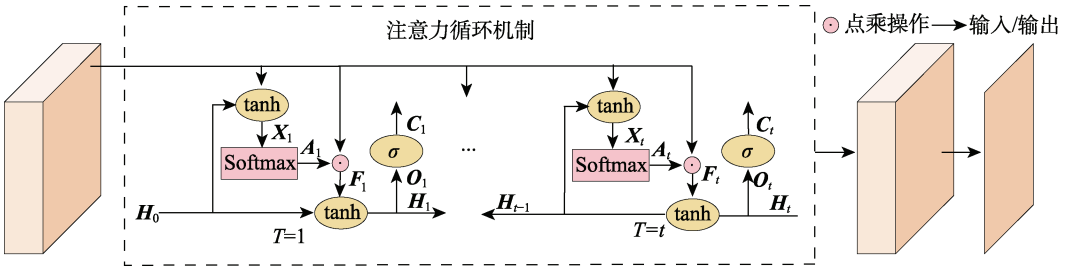


图 3　注意力循环机制结构

其中, A_t^{xy} 是在位置 (x,y) 上的注意力值. 对注意力图 $\boldsymbol{A}_t$ 与输入特征 $\boldsymbol{F}_{\text{fuse}}$ 的每一个通道进行点乘操作, 从而将注意力集中于显著性区域, 得到注意力特征图 $\boldsymbol{F}_t = \boldsymbol{A}_t \odot \boldsymbol{F}_{\text{fuse}}$.

将 $\boldsymbol{F}_t$ 输入到 RNN 中, 注意力特征 $\boldsymbol{F}_t$ 和上一步的隐藏状态 $\boldsymbol{H}_{t-1}$ 共同作用来更新隐藏层的状态

$$\boldsymbol{H}_t = \tanh\left(\boldsymbol{W}_{\text{h}} * \boldsymbol{F}_t + \boldsymbol{U}_{\text{h}} * \boldsymbol{H}_{t-1}\right).$$

其中, $\boldsymbol{W}_{\text{h}}$ 和 $\boldsymbol{U}_{\text{h}}$ 都表示卷积核. 最后用更新后的隐藏层状态 $\boldsymbol{H}_t$ 来更新输出 $\boldsymbol{O}_t$ 和记忆细胞 $\boldsymbol{C}_t$ 的状态, 即 $\boldsymbol{O}_t = \boldsymbol{V}_{\text{o}} * \boldsymbol{H}_t + \boldsymbol{b}_{\text{o}}$, $\boldsymbol{C}_t = \text{sigmoid}\left(\boldsymbol{O}_t\right)$. 其中, $\boldsymbol{V}_{\text{o}}$ 表示卷积核; $\boldsymbol{b}_{\text{o}}$ 表示偏置项.

需要指出的是, 当 RNN 中的时间步长过长, 在反向传播的过程中, 隐藏层之间的连乘操作变多, 会出现梯度消失的情况. 一种解决方法是利用门控制结构, 如 LSTM[27]和 GRU[28]等. 这种结构能让信息有选择地通过, 在一定程度上解决了 RNN 的梯度消失问题. 在本文的模块中, 由于时间步长较短($T=4$), 因此不必担心出现梯度消失的问题. 而且门控制结构的参数比较多, 计算量比较大. 为了保持模型的简洁, 采用了 RNN 结构进行迭代细化. 本文实验结果表明, RNN 足可以满足本文的要求, 具体讨论见第 3.5.4 节. 利用该注意力循环机制将注意力集中于图像的显著性区域, 抑制背景中的干扰信息, 同时, 逐步修改和细化由 FCN 产生的原始显著图, 实现准确的显著性预测.

2.3　损失函数

显著性检测在本质上是对每一个像素进行二分类的任务. 交叉熵损失是分类任务中常用损失函数, 它能够克服均方误差参数更新过慢的问题, 实现快速地收敛. 因此, 本文使用交叉熵损失以端到端的方式训练, 使预测的显著图与真值图更加接近. 该损失函数定义为

$$L = -\sum_{x,y} l_{x,y} \log\left(p_{x,y}\right) + \left(1 - l_{x,y}\right) \log\left(1 - p_{x,y}\right).$$

其中, $l_{x,y} \in \{x, y\}$ 是像素(x, y)的标签; $p_{x,y}$ 是像素(x, y)属于前景像素的概率.

3　实验及分析

3.1　数据集

本文方法在 PASCAL-S[29], SOD[30], HKU-IS[6], ECSSD[31]和 DUTS[32]这 5 个常用标准数据集上进行评估. PASCAL-S 数据集是从 PASCAL-VOC 2009 分割数据集的验证集中挑选出来的, 这个数据集共有 850 幅自然图像, 其中包含很多背景复杂以及存在多个物体的图像. SOD 数据集由 Berkeley 数据集中挑选出的 300 幅图像组成, 它包含各种显著性区域不明显的自然图像和低对比度图像, 是目前最有挑战性的数据集之一. HKU-IS 数据集包含 4447 幅图像, 其中有很多具有不连续显著性对象的图像. ECSSD 数据集中有 1000 幅各种复杂场景的图像. DUTS 数据集是一个包含 10553 幅训练图像和 5019 幅测试图像的大规模数据集, 这些图像不仅有复杂的背景, 而且显著性物体位置和尺寸都较多变. 因此, 本文使用 DUTS 数据集进行训练和测试, 其他 4 个数据集仅用于测试评估.

3.2　评估标准

本文使用 3 种评估方法来评价各种检测方法的性能, 包括: 查准率-查全率(precision-recall, PR)曲线, 最大 F-measure[33]和平均绝对误差(mean absolute error, MAE)[34]. 这些评估标准被广泛用于显著性检测[18].

(1) 查准率 P 和查全率 R 是通过给预测的显著图设定一个阈值来计算, 并将计算结果与相应的真值图进行比较. 计算 P 和 R 时, 首先应将显著图 S 转换为二值图 M. 假设原图像的真值图为 G. 查准率是指在预测的显著性区域中真值显著性像素所占的比例, 而查全率定义为检测到的显著性像素在真值区域中所占比例, 即

$$P = \frac{|M \cap G|}{|M|},$$

$$R = \frac{|M \cap G|}{G}.$$

利用从 0~255 的所有灰度级作为阈值对 S 进行二值化, 每个阈值都可以计算出一组查准率和查全率用于绘制 PR 曲线. 可以根据 PR 曲线下方的面积来评估方法的性能, 但更常用的是平衡点. 平衡点是查准率等于查全率时的取值, 该值越大, 证明该方法性能越高.

(2) F-measure 是一个总体性能评估指标, 它是对查准率和查全率进行加权计算, 即

$$F_\beta = \frac{\left(\left(1 + \beta^2\right) \times P \times R\right)}{\beta^2 \times P + R}.$$

其中, β^2 用于强调查准率的重要性, 按照文献

[31]的建议，其通常设置为 0.3. F-measure 值越大，方法性能越好.

(3) MAE 用于衡量预测的显著图和真值图之间的平均误差，即

$$\mathrm{MAE}=\frac{1}{w\times h}\sum_{x=1}^{w}\sum_{y=1}^{h}\left|S(x,y)-G(x,y)\right|.$$

其中，w 和 h 代表显著图的宽度和高度；S 和 G 分别表示预测的显著图和真值图. MAE 值越低，证明预测的显著性效果图与真值图之间的差距越小，方法性能越好.

3.3 实现细节

采用 PyTorch 框架实现本文方法，并使用 GTX 1080 GPU 进行加速. 在本文中用 DUTS 数据集的训练集训练提出的模型，所有训练和测试图像的尺寸都调整为 352×352 . FCN 的前 13 个卷积层由 VGG-16 网络初始化，其他卷积层则采用 PyTorch 的默认设置策略进行初始化. 本文构建的模型利用 Adam 优化器[35]进行训练，批量设置为 10，初始学习率设置为 10^{-4}，并且每迭代 50 次学习率会下降 10%. 共迭代 100 次，模型的训练时间大概需要 14 h.

3.4 与相关方法的性能比较

本文方法与 8 种相关基于深度学习的显著性检测方法进行比较，包括 MDF[6], DHS[11], Amulet[19], PAGR[14], RAS[10], DSS[9], CapSal[36]和 PAGE[20]. 其中，MDF, DHS, Amulet, DSS 中利用了多层级或多尺度特征；PAGR, RAS, CapSal, PAGE 都引入了注意力机制，PAGR 和 CapSal 中还包含循环机制. 为了公平地进行性能比较，本文采用作者公开的实现代码以及默认的参数实现这些模型，或者直接使用作者提供的显著图对这些方法进行评估. 在实验过程中，本文方法与所有对比方法均没有使用任何预处理和后处理步骤(例如 CRF[37]).

3.4.1 定量评估

将本文方法与 8 种显著性检测方法进行了性能比较，图 4 所示为各方法的 PR 曲线. 从中可以看到，本文方法与 PAGE 性能相当，除此之外，在所有数据集上，其平衡点都要高于其他方法，这保证了高的查准率和查全率，说明其性能要优于其他方法. 表 2 所示为各方法的 MAE 和 F-measure 值对比. 可以看出，本文方法在 DUTS, HKU-IS, PASCAL-S, SOD 数据集上的 F-measure 值与排第 2 位的方法相比，分别提高了 0.4%, 0.1%, 0.1%, 0.2%；在 DUTS, HKU-IS, SOD 数据集上的 MAE 值与排第 2 位的方法相比，分别降低了 3.8%, 2.7%, 1.8%. 这说明本文方法比其他先进方法的鲁棒性更强，更适用于复杂的显著性数据集.

3.4.2 定性评估

图 5 给出不同方法生成的显著图. 这些图像来自于 5 个数据集的测试图像. 可以看出，本文方法比其他方法的视觉效果更好，能够更加准确地检测显著性物体. 例如，第 3 幅图像中的显著性区域与背景的颜色对比度较小，本文方法可以准确地

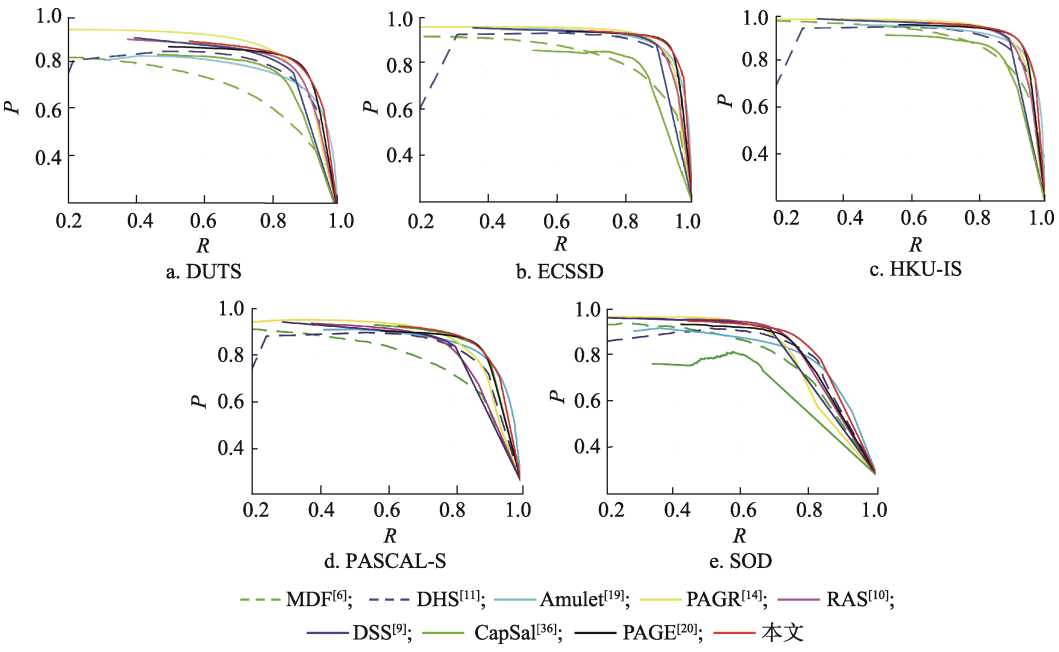
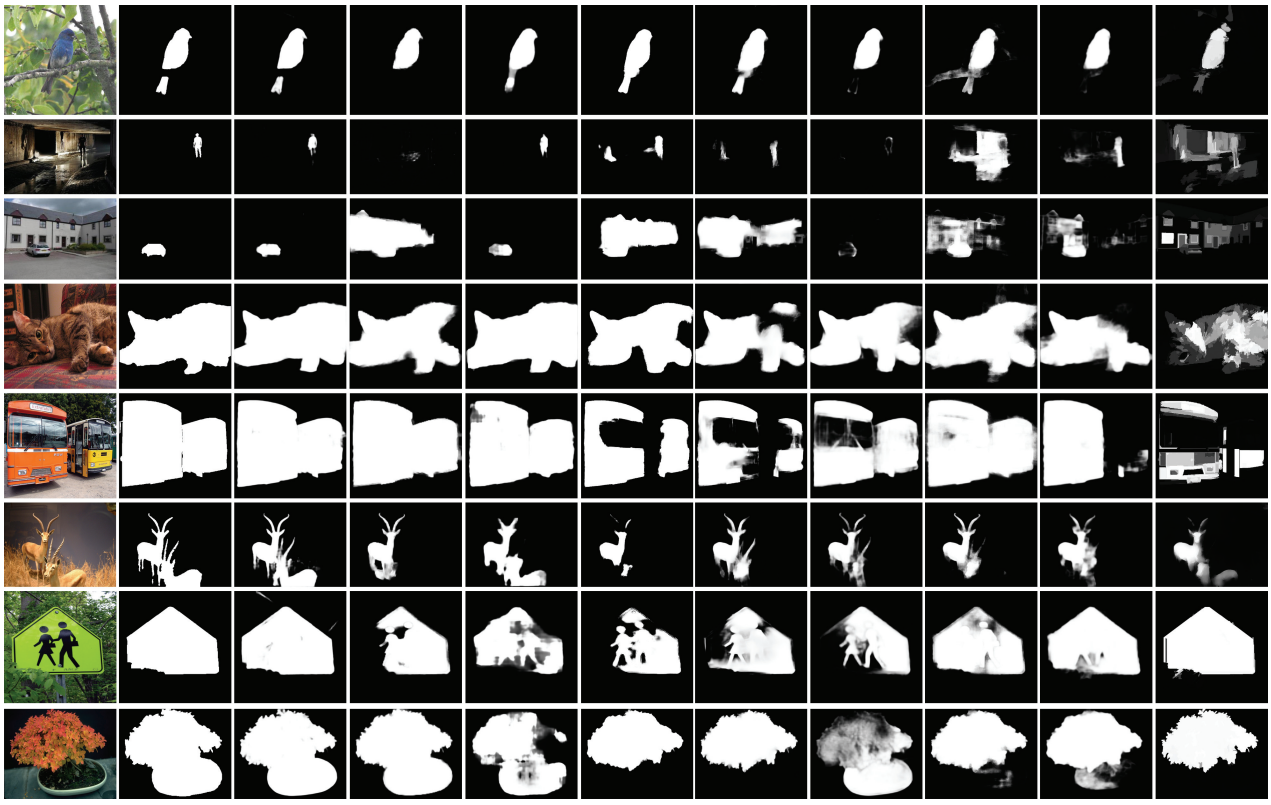


图 4 不同方法的 PR 曲线

表 2　不同方法的 MAE 和最大 *F*-measure 对比

方法	DUTS		ECSSD		HKU-IS		PASCAL-S		SOD	
	MAE	*F*-measure	MAE	*F*-measure	MAE	*F*-measure	MAE	*F*-measure	MAE	*F*-measure
MDF[6]	0.114	0.657	0.105	0.797	0.129	0.839	0.147	0.709	0.165	0.736
DHS[11]	0.067	0.776	0.060	0.893	0.053	0.875	0.094	0.800	0.129	0.790
Amulet[19]	0.085	0.751	0.059	0.905	0.051	0.887	0.099	0.815	0.142	0.773
PAGR[14]	0.055	0.817	0.061	0.904	0.048	0.897	0.094	0.815	0.147	0.761
RAS[10]	0.059	0.807	0.056	0.908	0.045	0.901	0.105	0.805	0.124	0.810
DSS[9]	0.056	0.796	0.052	0.906	0.040	0.901	0.098	0.809	0.124	0.802
CapSal[36]	0.061	0.772	0.077	0.813	0.057	0.842	**0.075**	0.830	0.148	0.669
PAGE[20]	0.052	0.815	**0.042**	**0.924**	0.037	0.907	0.078	0.836	0.111	0.796
本文	**0.050**	**0.818**	0.043	0.919	**0.036**	**0.908**	0.078	**0.837**	**0.109**	**0.812**

注. 粗体为最优结果.



a. 原图　b. 真值图　c. 本文　d. PAGE[20]　e. CapSal[36]　f. DSS[9]　g. RAS[10]　h. PAGR[14]　i. Amulet[19]　j. DHS[11]　k. MDF[6]

图 5　不同方法的视觉对比效果

定位显著性区域，而其他大多数方法检测出的显著图带有大量背景干扰；在第 8 幅图像中，大部分方法检测出的显著性区域不完整，本文方法不仅检测出完整的显著性区域，而且其边缘也比较精细. 无论是在显著性背景复杂(第 1, 2 行)、显著性区域和背景对比度不大(第 3, 4 行)的图像中，还是在多个显著性区域(第 5, 6 行)、显著性区域复杂(第 7, 8 行)的情况下，本文方法都可以准确地定位和突出整个显著性区域，并且具有良好的视觉效果.

3.5　消融分析

本文模型由 2 个部分组成，包含多尺度上下文特征提取机制和注意力循环机制. 为了验证各部分的有效性，在 DUTS 数据集上进行训练，在 DUTS 和 SOD 这 2 个具有挑战性的数据集上进行测试.

3.5.1　基本模型(FCN)

为了证明多特征注意力循环网络结构的有效性，将本文模型与仅使用 FCN 结构的模型进行比

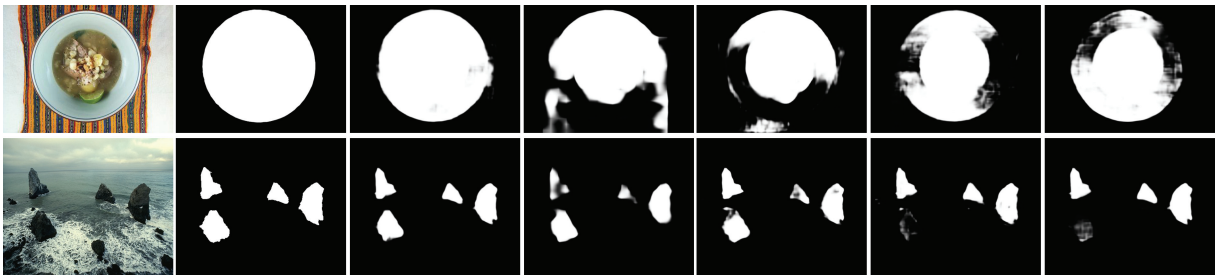
较. 基于 FCN 模型的视觉效果如图 6d 所示. 可以发现利用 FCN 模型生成的 2 幅图像中的显著性区域都不完整, 并且第 1 幅显著图中还含有大量的背景干扰. 而本文所提出的多特征注意力循环网络可以有效地解决该问题. 为了证明本文模型的有效性, 表 3 给出了对 FCN 模型的定量评估. 本文方法在 DUTS 和 SOD 数据集上的 MAE 值比 FCN 模型分别降低了 20.6%和 18.0%, 最大 F-measure 则分别提高了 8.5%和 7.1%. 可以看出, 本文模型可以有效地结合多尺度特征并且表现出良好的性能.

3.5.2 多尺度特征模型(FCN+MSF)

本文利用该机制捕获各种尺寸显著性区域的更多上下文信息. 为了验证多尺度上下文特征机制的有效性, 将只包含多尺度上下文特征提取机制的模型进行训练, 并对训练后模型进行测试. 图 6e 为该模型产生的显著图. 这幅图像中的显著性区域仍然不完整, 但是背景干扰明显减少. 表 3 中的定量评估显示, 该模型在这 2 个数据集上的 MAE 值比 FCN 模型分别降低了 12.7%和 4.5%, 最大 F-measure 则提高了 5.4%和 3.4%. 由此可见, 该机制可以提取更多的上下文信息, 对显著性检测更有效.

3.5.3 注意力循环模型(FCN+ARNN)

注意力循环机制不仅能够将注意力集中于显著性物体, 而且能够细化最终的显著图. 本文把只包含注意力循环机制的模型与基于 FCN 的模型进行比较. 注意力循环模型的视觉效果如图 6f 所示. 显著图中的显著区域不够完整, 区域内部不连续, 但是几乎没有背景干扰. 其在 2 个数据集上的 MAE 值与 FCN 模型相比, 分别降低了 14.3%和 9.8%, 最大 F-measure 分别提高了 5.0%和 5.8%. 实验结果表明, 注意力循环机制有助于显著性检测, 不仅能准确地定位显著性区域, 而且能够细化显著图.



a. 原图 b. 真值图 c. 本文 d. FCN e. FCN+MSF f. FCN+ARNN g. FCN+MSF+ALSTM

图 6 消融实验方法的视觉效果

表 3 消融实验方法的 MAE 和最大 F-measure

方法	DUTS		SOD	
	MAE	F-measure	MAE	F-measure
FCN	0.063	0.754	0.133	0.758
FCN+MSF	0.055	0.797	0.127	0.784
FCN+ARNN	0.054	0.792	0.120	0.802
FCN+MSF+ALSTM	0.049	0.824	0.199	0.796
FCN+MSF_5+ARNN	0.054	0.803	0.121	0.778
FCN+$MSF_{4,5}$+ARNN	0.051	0.811	0.118	0.798
FCN+MSF+ARNN (T=4, 本文)	0.050	0.818	0.109	0.812
FCN+MSF+ARNN(T=3)	0.050	0.820	0.118	0.798
FCN+MSF+ARNN(T=5)	0.053	0.815	0.121	0.793

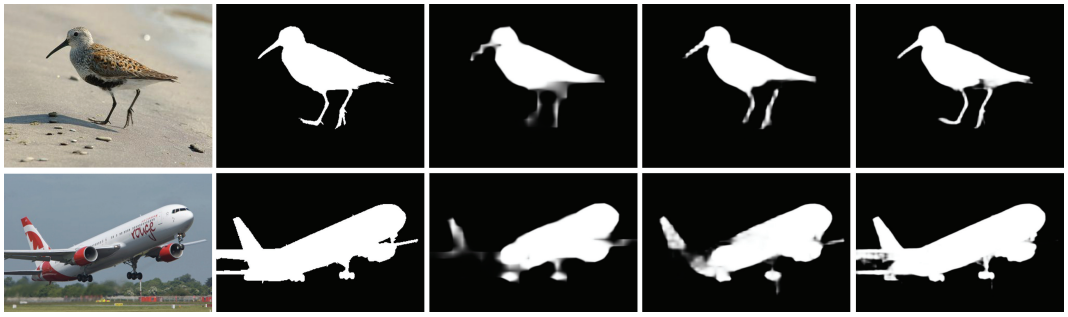
3.5.4 多特征注意力循环模型(FCN+MSF+ALSTM)

将基于 LSTM 的模型在 DUTS 数据集上训练, 并将测试的结果与本文模型进行比较. 视觉效果如图 6g 所示. 在具有复杂显著性区域的图像中, 基于 LSTM 方法不能完整地检测出显著性区域. 表 3 中的定量评估显示, 虽然基于 LSTM 的模型在 DUTS 数据集上性能稍好, 但是在 SOD 数据集上性能却远逊于基于 RNN 模型的性能. 同时, 考虑到 RNN 的简洁性和较低计算复杂度, 本文模型更适合采用 RNN 而非 LSTM.

3.5.5 多特征注意力循环模型(FCN+MSF+ARNN)

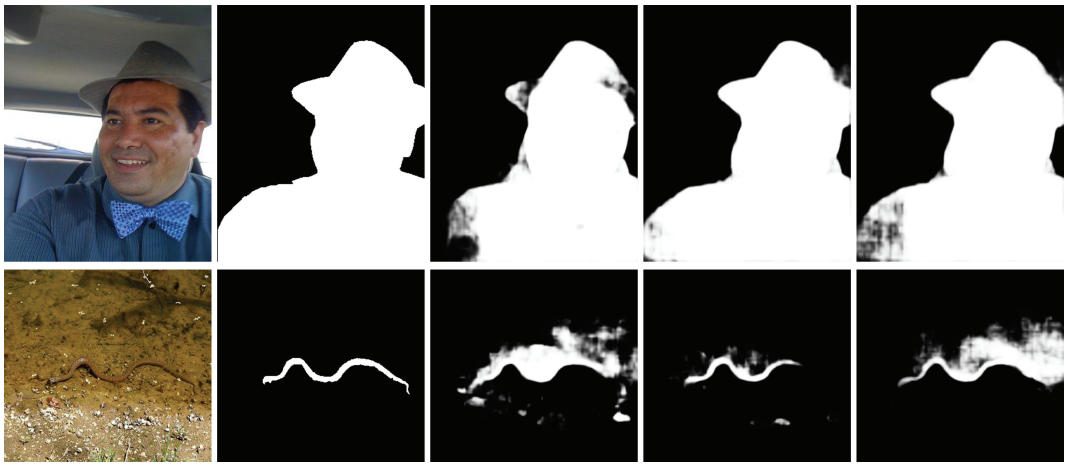
本文方法将多尺度特征与注意力循环特征结合. 考虑多尺度特征融合层数对模型的影响, 分别对融合后 3 层多尺度特征(FCN+MSF+ARNN 即本文方法), 融合后 2 层多尺度特征(FCN+$MSF_{4,5}$+ARNN)和只使用最后一层多尺度特征(FCN+MSF_5+ARNN)的模型进行实验. 视觉效果对比如图 7 所示, 融合后 3 层多尺度特征模型预测的显著目标细节完整性以及目标边缘连续性比其余 2 个模型更好. 表 3 展示了 3 个模型在 DUTS 和 SOD 数据集上的定量评估, 融合后 3 层多尺度特

征模型的 MAE 值和最大 F-measure 值是性能最好的. 因此，选择融合后 3 层多尺度特征作为本文模型的设置. 另外，注意力循环机制的视觉步长对显著性检测的性能也有一定的影响. 循环机制时间步长太短使性能提升不明显，太长则会导致模型计算复杂度大量增加. 因此，本文选择对循环机制的时间步长(T=3, 4, 5)进行比较. 表 3 的定量评估结果表明，在 DUTS 数据集上，$T=4$ 时模型的性能与 $T=3$ 相当，但比 $T=5$ 有所提高；而在 SOD 数据集上，$T=4$ 时的 2 项评价指标均优于 $T=3$ 和 $T=5$. 图 8 给出了不同时间步长的显著图比较，可以很明显地看出，当 $T=4$ 时，显著性区域最接近真值图，并且背景干扰最少，模型效果最好.



a. 原图　b. 真值图　c. FCN+MSF$_5$+ARNN　d. FCN+MSF$_{4,5}$+ARNN　e. 本文

图 7　融合不同层数多尺度特征的显著图比较



a. 原图　b. 真值图　c. $T=3$　d. $T=4$　e. $T=5$

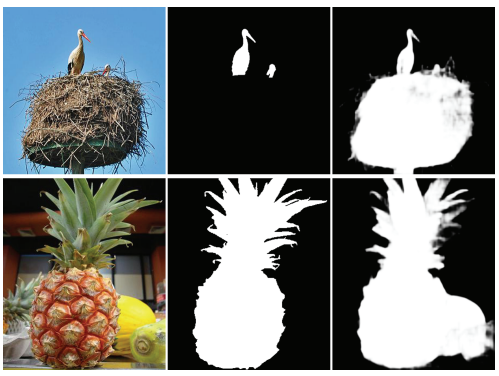
图 8　不同时间步长的显著图比较

3.6　局限性

得益于多尺度上下文特征提取机制和注意力循环机制的有效性，本文方法可以准确地检测大部分显著性区域. 但是，当图像中存在多个目标且语义信息不明确导致语义显著性区域存在歧义时，本文方法的检测结果不够理想. 比如，图 9a 中第 1 幅图像的鸟和鸟巢，第 2 幅图像的菠萝和它背后的水果都存在语义歧义，这导致本文方法生成的显著图与真值图相比误差较大.

4　结　语

为获取有效的卷积特征进行显著性检测，本文提出了一种基于多特征的注意力循环网络. 该网络利用由 4 个不同感受野的空洞卷积组成的多



a. 原图　b. 真值图　c. 本文

图 9　本文方法的失败案例

尺度上下文特征提取机制来获取高层特征的多尺度上下文特征. 在此基础上, 本文进一步采用注意力循环机制增强特征的辨识性并细化显著性区域, 从而生成准确的显著图. 本文方法在常用的公开数据集上与其他方法进行比较, 实验结果表明该方法整体性能优于新近提出的方法.

参考文献(References):

[1] Pisharady P K, Vadakkepat P, Loh A P. Attention based detection and recognition of hand postures against complex backgrounds[J]. International Journal of Computer Vision, 2013, 101(3): 403-419

[2] 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. New York: International Machine Learning Society, 2015: 597-606

[3] Wang Ruixia, Peng Guohua. An image retrieval method with sparse coding based on Riemannian manifold[J]. Acta Automatica Sinica, 2017, 43(5): 778-788(in Chinese)
(王瑞霞, 彭国华. 基于黎曼流形稀疏编码的图像检索算法[J]. 自动化学报, 2017, 43(5): 778-788)

[4] Liu Z, Shi R, Shen L Q, *et al*. Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut[J]. IEEE Transactions on Multimedia, 2012, 14(4): 1275-1289

[5] Lee G, Tai Y W, Kim J. Deep saliency with encoded low level distance map and high level features[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 660-668

[6] Li G B, Yu Y Z. Visual saliency based on multiscale deep features[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 5455-5463

[7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[8] Yu Chunyan, Xu Xiaodan, Zhong Shijun. Saliency region detection based on deconvolutional and skip nested module[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(11): 2150-2158(in Chinese)
(余春艳, 徐小丹, 钟诗俊. 融合去卷积与跳跃嵌套结构的显著性区域检测[J]. 计算机辅助设计与图形学学报, 2018, 30(11): 2150-2158)

[9] Hou Q B, Cheng M M, Hu X W, *et al*. Deeply supervised salient object detection with short connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 815-828

[10] Chen S H, Tan X L, Wang B, *et al*. Reverse attention for salient object detection[C]//Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 236-252

[11] Liu N, Han J W. DHSNet: deep hierarchical saliency network for salient object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 678-686

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2020-02-19]. https://arxiv.org/abs/1409.1556

[13] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[14] Zhang X N, Wang T T, Qi J Q, *et al*. Progressive attention guided recurrent network for salient object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 714-722

[15] Cheng M M, Mitra N J, Huang X L, *et al*. Global contrast based salient region detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 569-582

[16] Qian Sheng, Chen Zonghai, Lin Mingqiang, *et al*. Saliency detection based on conditional random field and image segmentation[J]. Acta Automatica Sinica, 2015, 41(4): 711-724(in Chinese)
(钱生, 陈宗海, 林名强, 等. 基于条件随机场和图像分割的显著性检测[J]. 自动化学报, 2015, 41(4): 711-724)

[17] Wang L J, Lu H C, Ruan X, *et al*. Deep networks for saliency detection via local estimation and global search[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3183-3192

[18] Borji A, Cheng M M, Hou Q B, *et al*. Salient object detection: A survey[J]. Computational Visual Media, 2019, 5(2): 117-150

[19] Zhang P P, Wang D, Lu H C, *et al*. Amulet: aggregating multi-level convolutional features for salient object detection[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 202-211

[20] Wang W G, Zhao S Y, Shen J B, *et al*. Salient object detection with pyramid attention and salient edges[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1448-1457

[21] Li G B, Xie Y, Lin L, *et al*. Instance-level salient object segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 247-256

[22] Tang Y B, Wu X Q, Bu W. Deeply-supervised recurrent convolutional neural network for saliency detection[C] //Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 397-401

[23] Kuen J, Wang Z H, Wang G. Recurrent attentional networks for saliency detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3668-3677

[24] Cornia M, Baraldi L, Serra G, *et al*. Predicting human eye fixations via an LSTM-based saliency attentive model[J]. IEEE Transactions on Image Processing, 2018, 27(10): 5142-5154

[25] Wang T T, Borji A, Zhang L, *et al*. A stagewise refinement

model for detecting salient objects in images[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4039-4048

[26] Li D W, Chen X T, Zhang Z, *et al*. Learning deep context-aware features over body and latent parts for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 7398-7407

[27] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780

[28] Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation[OL]. [2020-02-19]. https://arxiv.org/abs/1406.1078

[29] Li Y, Hou X D, Koch C, *et al*. The secrets of salient object segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 280-287

[30] Movahedi V, Elder J H. Design and perceptual validation of performance measures for salient object segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2010: 49-56

[31] Yan Q, Xu L, Shi J P, *et al*. Hierarchical saliency detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2013: 1155-1162

[32] Wang L J, Lu H C, Wang Y F, *et al*. Learning to detect salient objects with image-level supervision[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3796-3805

[33] Achanta R, Hemami S, Estrada F, *et al*. Frequency-tuned salient region detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 1597-1604

[34] Perazzi F, Krähenbühl P, Pritch Y, *et al*. Saliency filters: Contrast based filtering for salient region detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 733-740

[35] Kingma D P, Ba J. Adam: a method for stochastic optimization[OL]. [2020-02-19]. https://arxiv.org/abs/1412.6980

[36] Zhang L, Zhang J M, Lin Z, *et al*. CapSal: leveraging captioning to boost semantics for salient object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 6017-6026

[37] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C] //Proceedings of the 24th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2011: 109-117