DOI: 10.11992/tis.201812003
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190719.1524.004.html

基于双向消息链路卷积网络的显著性物体检测

申凯, 王晓峰, 杨亚东
(上海海事大学 信息工程学院, 上海 201306)

摘　要: 有效特征的提取和高效使用是显著性物体检测中极具挑战的任务之一。普通卷积神经网络很难兼顾提取有效特征和高效使用这些特征。本文提出双向消息链路卷积网络 (bidirectional message link convolution network, BML-CNN) 模型, 提取和融合有效特征信息用于显著性物体检测。首先, 利用注意力机制引导特征提取模块提取实体有效特征, 并以渐进方式选择整合多层次之间的上下文信息。然后使用带有跳过连接结构的网络与带门控函数的消息传递链路组成的双向信息链路, 将高层语义信息与浅层轮廓信息相融合。最后, 使用多尺度融合策略, 编码多层有效卷积特征, 以生成最终显著图。实验表明, BML-CNN 在不同指标下均获得最好的表现。

关键词: 显著性物体检测; 卷积神经网络; 注意力机制; 双向消息链路; 多尺度融合
中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 1673−4785(2019)06−1152−11

中文引用格式: 申凯, 王晓峰, 杨亚东. 基于双向消息链路卷积网络的显著性物体检测 [J]. 智能系统学报, 2019, 14(6): 1152–1162.
英文引用格式: SHEN Kai, WANG Xiaofeng, YANG Yadong. Salient object detection based on bidirectional message link convolution neural network[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1152–1162.

Salient object detection based on bidirectional message link convolution neural network

SHEN Kai, WANG Xiaofeng, YANG Yadong
(College Of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: The effective extraction and efficient utilization of features are among the most challenging tasks in salient object detection. The common convolutional neural network (CNN) can hardly reach a fine trade-off between effective feature extraction and efficient utilization. This paper proposes a bidirectional message link convolutional neural network (BML-CNN) model, which can extract and fuse effective features for salient object detection. First, the attention mechanism is used to guide the feature extraction module to extract the effective entity features, select, and integrate the multi-level context information in a progressive way. Second, the high-level semantic information is merged with shallow-profile information by a bidirectional message link, which is composed of a skip connection structure and a messaging link with a gating function. Finally, the saliency map can be generated by multi-scale fusion strategy, and effective features are encoded on several layers. The qualitative and quantitative experiments on six benchmark datasets show that the BML-CNN reaches the state-of-the-art performance under different indexes.

Keywords: salient object detection; convolutional neural network; attention mechanism; bidirectional message link; multi-scale fusion

视觉显著性是用来刻画图像中的部分区域, 这些区域相对于它们的临近区域更为突出。显著性模型可分为基于数据驱动的自底向上模型[1] 和基于任务驱动的自顶向下模型[2]。Itti 等[3] 提出的 ITTI 模型模拟生物视觉注意力机制用于显著性检测。Liu 等[4] 将显著性检测定义为二元分割问题

收稿日期: 2018−12−04.　**网络出版日期**: 2019−07−19.
基金项目: 国家自然科学基金项目 (61872231, 61703267); 上海海事大学研究生创新基金项目 (2017ycx083).
通信作者: 王晓峰. E-mail: xfwang@shmtu.edu.cn.

引发了显著性检测模型的热潮。基于卷积神经网络[5-7]的显著性检测方法消除了对手工特征的需求, 逐渐成为显著性检测的主流方向。显著性物体检测用于突出图像中最重要的部分, 常作为图像预处理步骤用于计算机视觉任务中, 包括图像分割[8-10]、视觉跟踪[11]、场景分类[12]、物体检测[13-15]、图像检索[16-18]、图像识别[19]等。

基于深度卷积神经网络, 特别是全卷积神经网络 (FCN), 已经在语义分割[20]、姿态估计[21]和对象提取[22]等标记任务中表现出优异的性能。同时也推动了尝试使用 FCN 解决显著性物体检测中显著性物体定位问题, 虽然这些模型[5-6, 23-24]在预测物体显著性的任务中有出色的高层语义提取能力, 但是显著图缺少精确的边界细节, 显著图无法保留精确的对象边界信息。这促使很多研究人员利用不同层级的特征的非线性组合进行显著性检测。Xiao 等[25]建议提取不同级别的显著图, 并将其进行非线性融合得到显著图, 使其获取高级语义信息的同时兼顾低级空间信息。Hou 等[26]建议在多个侧输出层之间添加短连接, 用以组合不同级别的特征。Zhang 等[27]提出通过低级别的特征与高级别的特征进行聚合, 生成多级特征。Jin 等[28]提出使用循环神经网络的方式将高级语义信息和低级空间信息相互传递, 生成显著图。Chen 等[29]提出使用空间注意机制与通道注意力机制捕捉图片高级语义信息, 但依旧存在边界信息缺失的现象, 且对于背景抑制、实体镜像问题的处理还需要进一步提高。

为解决上述问题, 本文提出了一种基于双向消息链路卷积网络的显著性物体检测方法。为了解决边界缺失问题使用设计一个具有跳过连接结构的上下文感知模块将高级语义与低级空间特征进行融合, 对于每一个侧输出采用了空洞卷积获取每一个侧输出的更多的上下文信息。为了准确地定位显著性物体的位置信息以及减少无关通道对显著物体的高级语义与空间信息的影响, 借助了空间注意力与通道注意力机制组成的注意力模块。为了更加有效地传递上下文语义信息, 借助具有门控的消息传递通道, 完成从高级特征到低级特征的传递。为了融合产生的多层特征信息, 借助多尺度融合策略生成物体显著性预测图。本文将提出的 BML-CNN 在 6 个数据集上与 13 种先进的显著性物体检测模型进行比较, 实验表明 BML-CNN 在不同的评价指标下均有最出色的表现, 此外, 模型的实时处理速度为 18 f/s。本文的贡献主要分为以下三个方面:

1) 使用由通道注意力与空间注意力组成的注意力模块来提取有效特征, 可赋予有效通道、有效卷积特征更高的权值, 减少背景对显著性物体预测的影响。

2) 提出具有跳过连接结构的上下文感知模块与带门控函数的消息链路组成的双向消息链路, 可在获取高级语义信息的同时, 保留完整的边界信息。

3) 借助多尺度融合策略将多级有效特征进行融合, 可在不同角度产生对显著性物体的预测, 并进一步融合不同尺度的信息生成具有完整边界的显著性物体预测图。

1　相关工作

本节将从 3 个方面介绍相关工作。首先, 描述特征传递在显著性检测中的应用。其次, 描述了注意力机制在各种视觉任务中的应用。最后, 介绍了多尺度融合在显著性物体检测任务中的应用。

1.1　特征传递

不同级别的特征传递是显著性物体检测任务中的一项重要工作, 也促使很多研究人员探讨更优异的特征传递策略。例如, Wang 等[6]提出了使用双卷积神经网络, 将局部超像素估计传递到高层卷积指导生成全局对象提议搜索的显著性物体检测。Jin 等[28]提出使用循环神经网络的方式将高级语义信息和低级空间信息相互传递, 生成显著图。Long 等[30]借助跳过连接的方法, 将高层语义添加到中间层, 已生成多分辨率, 多尺度的预测信息, 并由预测信息生成像素的预测结果。Zhao 等[19]通过融合全局和局部的上下文信息来预测每个超像素的显著度, 并依据每个超像素的显著度生成显著对象的显著图。Lee 等[23]提出将降低级空间信息与高级语义信息进行传递并编码, 并使用编码后的特征预测显著性图。Liu 等[24]建议使用分阶段检测物体的显著性, 第一阶段使用卷积神经网络提取全局结构特征, 并产生粗略估计, 第二阶段融合策略, 将本地上下文信息细化为显著图的细节, 并与第一阶段产生的粗略显著图进行相互传递并融合得到精确的显著性图。Wang 等[7]设计了全卷积神经网络 (FCN), 将粗略的显著性预测特征传递到高层, 并逐步指导显著性图的生成。上述方法在实现特征传递过程中并没有考虑到高层语义对低层轮廓提取的影响程度, 使得低层轮廓提取过于注重显著度高的位置, 从而导致显著度较低的边缘信息保留不足。

为控制高层语义对低层轮廓提取的影响程度, 提出使用带跳过连接结构与带门控函数组成的双向消息传递链路, 在实现高层语义信息与低

层轮廓信息相互传递的同时，能控制高层语义对低层轮廓提取的影响程度，达到高层语义有限指导低层轮廓的获取，低层轮廓信息为高层语义提供精确的空间信息。

1.2　注意力机制

视觉注意力机制是借鉴人类的视觉注意力机制，扫描全局图片获取需要关注的目标实体区域，并为这一区域投入更多的资源，可获取更为完善的关注目标的信息，而降低其他信息的影响。注意力机制在多个视觉任务中都有很出色的表现，例如，图像字幕[29, 31]、视觉问答[32-33]、目标识别[19]和图像分类[34]等。Xu 等[35]首先提出了使用“软”和“硬”注意力机制解决图像字幕。Wang 等[34]提出使用一种残差注意力机制来训练深度残差网络进行图像分类。Chen 等[29]提出了一种 SCA-CNN 网络，网络使用 CNN 结合了空间注意力机制与通道注意力机制赋予各个通道和空间位置不同的权重，提高目标响应并降低背景的干扰。

在显著性物体检测时，并不是所有通道的卷积特征对显著性物体的预测都具有同等重要性，个别通道会存在背景的卷积特征，且在同一通道中不同位置的卷积特征也对显著性物体的预测产生影响。为更有利地获取有效卷积特征，进一步消除背景对显著物体的影响，本文采用通道注意力机制为不同卷积通道赋予不同权值，以降低含背景卷积信息的通道对显著性物体预测的影响，另外引入空间注意力机制来为同一通道上不同位置的卷积特征赋予不同权值，以进一步消除背景的影响。将空间注意力与通道注意力串联组成注意力模块以实现对物体的初步关注，并以渐进的方式指导下一层关注的提取，逐步消除背景对显著性物体预测的影响。

1.3　多尺度特征融合

从一些可视化深度卷积神经网络的工作[9-10, 36-40]可以看出，不同层次的卷积特征是从不同的视角描述物体特征及其周围环境。高级语义有助于图像区域物体类别的识别，而低级视觉特征有助于保留空间细节，生成具有高分辨率的显著性图。然而如何有效地利用多尺度特征依然是一个值得探讨的问题。为此，已经有很多有价值的研究，例如，Li 等[5]通过使用先生成局部超像素估计，然后在多个 CNN 中提取多尺度特征来预测物体的显著性。Zhang 等[27]提出 Amulet 网络使用 RFC 生成多分辨率的预测信息，并使用 FS 进行多尺度融合，获得显著性的预测。Hariharan 等[22]提出使用 Hypercolumn 方法，不仅融合了来自多个中间层的卷积特征，还学习了密集特征分类器。Badrinarayanan 等[11]采用编码器−解码器网络，使用池化引导反卷积模块多级卷积特征。Ronneberger 等[2]提出 U-Net 网络，应用多个跳过连接结构来捕获上下文结构，并通过收缩路径和扩展路径融合多尺度卷积特征生成有精确定位的显著性预测图。

受到以上研究的启发，本文提出的方法中也使用跳过连接结构实现捕获上下文信息，并借助空洞卷积对上下文信息进一步提取，同时借助带门控的信息传递链路，实现高级语义与中间卷积特征的相互融合，这种融合是以阶段方式进行的，由跳过连接结构与带门控的信息传递链路组成了双向信息传递链路，有利于为显著性预测提供全面的信息。另一方面通过多尺度融合策略，为显著性预测提供不同视角的卷积特征，能够将低级的边缘感知特征与高级语义信息进行聚合，有助于保持对象边界。

2　算法模型

BML-CNN 模型使用含有注意力模块的特征提取模块来提取有效特征，借助双向消息链路实现高层语义信息与底层轮廓信息相互传递，融合上下文信息，最后使用多尺度融合策略，融合不同尺度的有效卷积特征，以实现物体的显著性预测。该模型具有出色的显著性预测能力，且边界保持较好。

2.1　通道注意力与空间注意力

通道注意力机制是调整特征通道对目标影响程度的方式，为有效的通道赋予更高的权重使其能对显著性对象有更高的响应，降低无效通道的权重使其能够降低对显著性对象预测的干扰。

将卷积特征用 $I \in \mathbf{R}^{W\times H\times C}$ 表示，其中 $W\times H\times C$ 表示卷积特征 I 的维度，用 $F=\{f_1, f_2, \cdots, f_C\}$ 表示卷积特征 I 上的通道，其中 $f_i \in \mathbf{R}^{W\times H}$，$i \in \{1,2,\cdots,C\}$ 表示卷积特征 I 上的第 i 个通道，W 表示宽，H 表示高，C 表示通道总数。用 $\boldsymbol{s} \in \mathbf{R}^C$ 表示通道权重向量，本文设计一个卷积层来学习每个通道的权值特征：

$$g = W_C * F + b_C \tag{1}$$

式中：W_C 表示卷积滤波器；b_C 表示卷积偏差。使用 Softmax 激活函数获得最终的通道注意力向量 $\boldsymbol{a}_C = \{a_C(1), a_C(2), \cdots, a_C(C)\}$：

$$a_C(i) = \mathrm{Softmax}(g(i)) = \frac{\exp(g(i))}{\sum_{i=1}^{C}\exp(g(i))} \tag{2}$$

$$\sum_{i=1}^{C} a_C(i) = 1 \tag{3}$$

空间注意力机制直接使用卷积特征预测显著

性往往可能由于非显著性区域所造成的噪音导致次优结果。空间注意力机制通过对每一个区域进行评估，为每一个区域赋予不同的权值，使得模型能够更加关注有助于显著性预测的有效信息。空间注意力机制可以突出显著性对象，减少背景区域的干扰。

使用 $I \in \mathbf{R}^{W\times H\times C}$ 表示卷积特征，使用 $L=\{(x,y)|x=1,2,\cdots,W;y=1,2,\cdots,H\}$ 表示卷积特征上空间位置，其中 (x,y) 表示空间上点的坐标。本文设计了一个卷积层来计算空间注意力特征图：

$$m = W_S * I + b_S \tag{4}$$

式中：$m \in \mathrm{R}^{W\times H}$ 是包含所有通道的信息；W_S 表示卷积滤波器；b_S 表示卷积偏差。使用 Softmax 激活函数获取每一个位置上的空间注意力权重。

$$a_S(l) = \mathrm{Softmax}(m(l)) = \frac{\exp(m(l))}{\sum_{l\in L}\exp(m(l))} \tag{5}$$

$$\sum_{l\in L} a_S(l) = 1 \tag{6}$$

式中：$m(l)$ 表示空间注意力特征图 m 中第 l 个点，其中 $l \in L$；$a_S(l)$ 表示第 l 个点的权值。令 $a_S=\{a_S(1),a_S(2),\cdots,a_S(W\times H)\}$ 为空间关注图。

注意力模块使用通道注意力模块与空间注意力模块串联成注意力模块，结构如图 1 所示。将注意力模块添加到带跳过连接的上下文感知模块，可从不同方向上减少背景区域的干扰，提高对显著性物体的预测，并精确的保留边界信息。

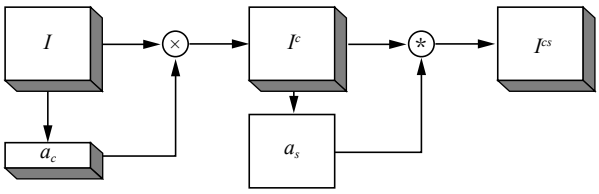


图 1　注意力模块模型

Fig. 1　Attention module model

使用 $I \in \mathbf{R}^{W\times H\times C}$ 表示输入注意力模块的卷积特征前半阶段为通道注意力机制，后半段为空间注意力机制。令 I^C 为经过通道注意力模块输出的卷积特征：

$$I^C(i) = I(i)\times a_C(i) \tag{7}$$

式中：$a_C(i)$ 表示第 i 层通道的通道注意力向量第 i 维参数，其中 $i \in \{1,2,\cdots,C\}$。将得到的卷积特征输入到空间注意力模块中得到 I^{CS}：

$$I^{CS} = a_S * I^C \tag{8}$$

式中*表示 Hadamard 矩阵乘积运算。得到的 I^{CS} 是通过注意力模块的带权卷积特征，模型使用 I^{CS} 指导下一层卷积对显著性物体特征的提取。

图 1 中，左半边为通道注意力模块和式，右半边为空间注意力模块和式，其中 I 为输入和式，a_C表示通道注意力向量和式由式 (2) 和式 (3) 计算得到。I^C 表示通道注意力模块的输出和式也是空间注意力模块的输入和式由式 (7) 计算得到。a_S 表示空间注意力权重和式，由式 (5) 和式 (6) 计算得到。I^S 表示空间注意力模块的输出和式，也是本文注意力模块的输出和式由式 (8) 计算得到。多层卷积之间添加注意力模块和式实现渐进式的注意力引导。每一层的注意力信息可指导下一层的训练，以自适应的方式生成新的注意项和式使得上下文信息实现由粗至简的细化过程。

2.2　双向消息链路

双向消息链路由带有跳过连接结构的上下文感知网络与带有门控函数的信息传递链路组成。带有跳过结构连接结构的上下文感知网络用来提取高级的语义信息，而带有门控函数的信息传递链路将高进语义信息和中间卷积特征指导低级空间信息提取。使得送入多尺度融合模块的不同尺度的特征图均具有完整的空间信息和语义信息，为最终的融合提供有效、可靠的输入源。

如图 2 所示，带有跳过连接结构的上下文传递模块，“Conv5”是对原始图片的特征提取，使用跳过连接结构将原始图片，与语义特征一起作为新的卷积层的输入，实现上下文传递，并使用后续的卷积将低级空间特征与高级语义相融合，使得显著性特征具有比较完备的边界信息和高级语义信息。另外，注意力机制的加入减少了背景对显著性物体预测的影响。

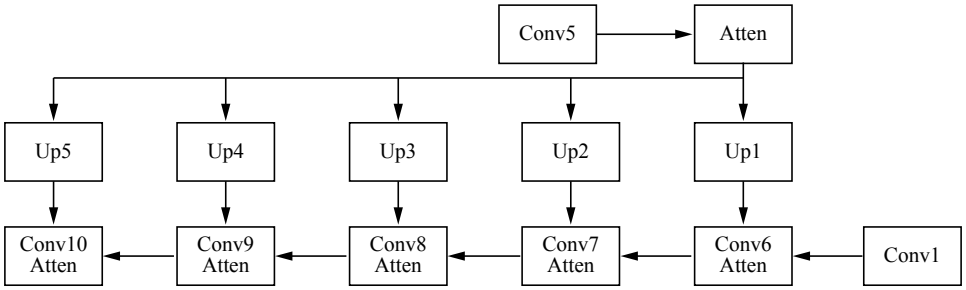


图 2　带跳过连接的上下文传递模块

Fig. 2　Context transfer module with skip connection

$$\text{att\_conv5} = \text{Atten}(\text{Conv5}) \tag{9}$$

$$\text{Up}_i = \text{Up}(\text{att\_conv5}, u_i) \tag{10}$$

其中 att_conv5 为“Conv5”通过注意力模块 Atten 的输出，具体计算方法由 2.1 节给出。Up_i, $i \in \{1,2,3,4,5\}$表示图 2 中上采样的输出，u_i 为大小分别为 $\{16\times16, 8\times8, 4\times4, 2\times2, 1\times1\}$ 的上采样内核。

$$\text{conv}_i = \text{Conv}(\text{Concat}(\text{Up}_{i-5}, \text{conv}_{i-1}), K) \tag{11}$$

$$\text{at}_i = \text{Atten}(\text{conv}_i) \tag{12}$$

式中：K 表示大小为 3×3 的卷积核，Concat 表示通道连接，Up_{i-5} 由式 (9) 和式 (10) 计算得到。式 (11) 中卷积的激活函数均为 Relu。at_i 表示 conv_i 通过注意力模块的输出。

如图 3 所示，本文使用带门控函数的信息传递链路将高级语义信息与中间层卷积特征相融合，因为并不是所有的中间层都对物体显著性的预测是有帮助的，所以借助门控函数产生 [0−1] 的权值向量，控制高层卷积特征对低级卷积特征的影响程度，从而每一层都是由上一层加权并与本层特征融合的结果，使得每一层都有在上一层高级语义的指导下选择本层的空间特征，从而产生不同级别、不同尺度、不同视角的显著性预测先验信息，为进一步的多尺度融合提供比较全面的特征信息。

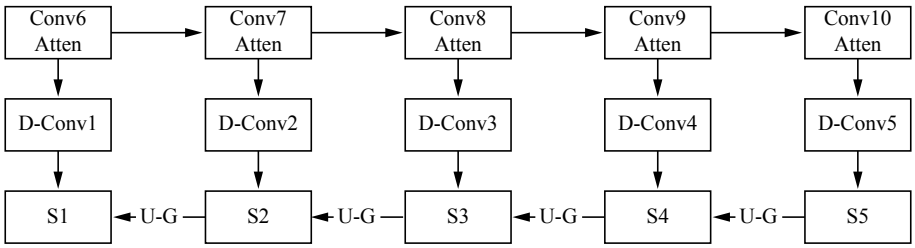


图 3　带门控函数的消息传递模块

Fig. 3　A messaging module with gated functions

$$\text{dc}_{ij} = \text{Conv}(\text{at}_i, K, D_j) \tag{13}$$

$$\text{sd}_i = \text{Concat}(\text{dc}_{i1}, \text{dc}_{i2}, \text{dc}_{i3}, \text{dc}_{i4}) \tag{14}$$

式中：dc_{ij}, $i \in \{1,2,3,4,5\}$, $j \in \{1,2,3,4\}$ 表示空洞卷积的输出；卷积核 K 的大小均为 3×3；D_j 表示大小分别为 1、3、5、7 的 dilation rate；sd_i 表示融合空洞卷积的输出，$i \in \{1,2,3,4,5\}$。

$$M_i = G\left(S_{i+1}, K^{i1}\right) \times \text{Conv}\left(S_{i+1}, K^{i2}\right) \tag{15}$$

$$G\left(S_{i+1}, K^{i1}\right) = \text{Sigmoid}\left(\text{Conv}\left(S_{i+1}, K^{i1}\right)\right) \tag{16}$$

$$S_i = \text{Conv}\left(\text{Concat}(M_i, \text{sd}_i), K^i\right) \tag{17}$$

式中：门控函数由 G 表示；K^i、K^{i1} 和 K^{i2} 均表示大小为 3×3 的卷积核；S_i 则表示双向消息链路的侧输出。

2.3　多尺度特征融合策略

本文提出的多尺度特征融合策略是将双消息链路的侧输出 S_i, $i \in \{1,2,3,4,5\}$ 进行有效融合。首先对 6 个侧输出进行上采样操作得到分层映射 Sm_i，它将用于对尺度特征融合的输入。

$$\text{Sm}_i = \text{Up}(S_i, u_i) \tag{18}$$

其中，Up 表示上采样操作；u_i 分别表示大小为 $\{1\times1, 2\times2, 4\times4, 8\times8, 16\times16\}$ 的采样内核。

如图 4 所示，将式 (18) 计算得到的 5 个分层特征映射 Sm_i 输入到特征融合策略，生成最终的显著性预测图。

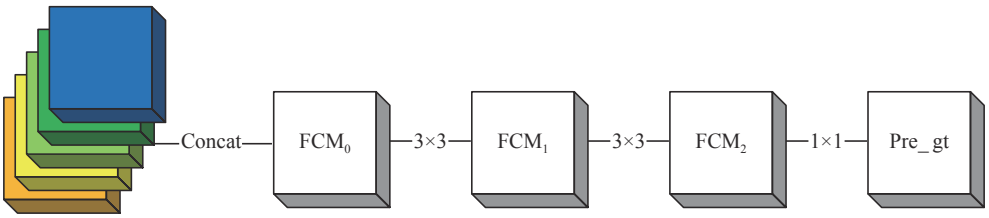


图 4　多尺度融合策略

Fig. 4　Multi-scale fusion strategy

$$\text{FCM}_0 = \text{Concat}(\text{Sm}_1, \text{Sm}_2, \text{Sm}_3, \text{Sm}_4, \text{Sm}_5) \tag{19}$$

$$\text{FCM}_1 = \text{Conv}(\text{FCM}_0, K_1) \tag{20}$$

$$\text{FCM}_2 = \text{Conv}(\text{FCM}_1, K_2) \tag{21}$$

$$\text{pre\_gt} = \text{Conv}(\text{FCM}_2, K_3) \tag{22}$$

式中：K_1、K_2 和 K_3 分别表示大小为 3×3、3×3、1×1 的卷积核；激活函数分别为 Relu、Relu、Sigmoid；pre_gt 为模型最终的输出，也是物体的显著性预测图。

BML-CNN 模型结构由图 5 给出，通过带有注意力机制的基础特征提取层，提取有效高级语义信息，并结合带有跳过连接结构的上下文传递与带门控的消息传递链路组成的双向信息传递模型

完成有效消息的双向传递, 使各层具有不同角度语义信息的同时保留完整的边界信息。最后使用特征融合策略将各层卷积特征相融合, 生成最终的显著性物体预测图。

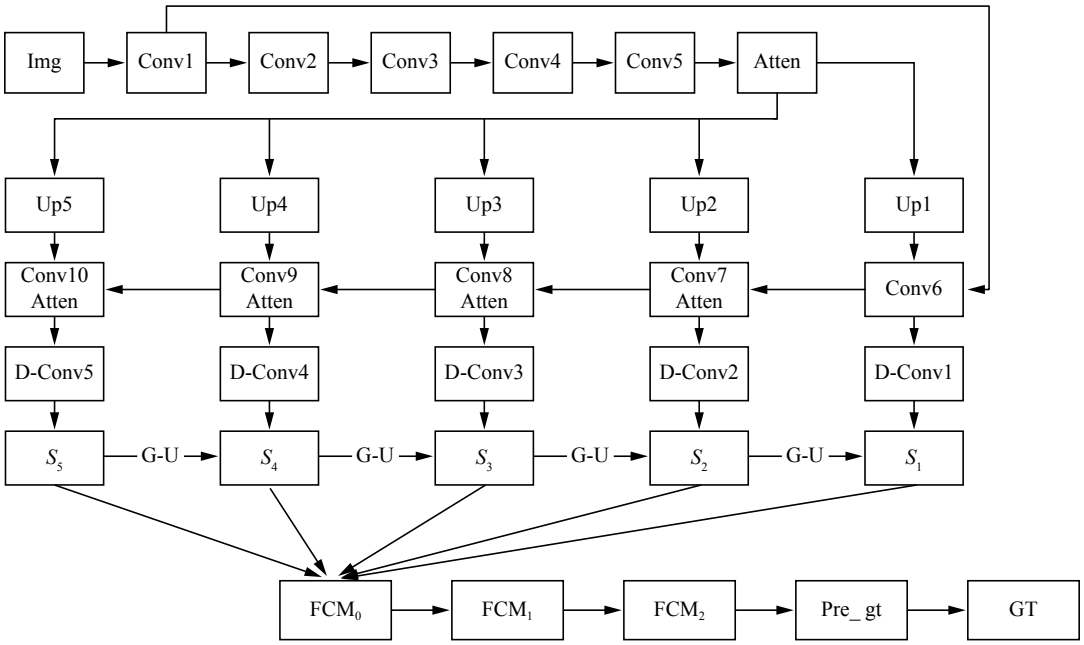


图 5　双向消息链路卷积网络的结构图

Fig. 5　Structure diagram of bidirectional message link convolution network

3　实验

3.1　实验设置

数据集: 该模型使用 DUTS-TR 数据集[19] 作为训练集, 数据集包括 10 553 张图片, 为了使模型获得更好的训练效果, 使用了数据增强策略生成了 63 318 张图片作为训练图片。为了评估模型, 本文使用了以下 6 个标准数据集作为模型先进性验证: DUTS-TE 数据集[19], 该数据集具有 5 019 个具有高像素注释的测试数据集。DUT-OMRON 数据集[41], 该数据集有 5 168 个高质量的图像, 数据集中的图像具有一个或多个显著性对象和相对复杂的背景。ECSSD 数据集[31], 该数据集具有 1 000 个图像, 在语义上具有比较复杂的分割结构。HKU-IS 数据集[19], 该数据集具有 4 447 幅图片, 具有多个不相连的显著性对象。PASCAL-S 数据集[42], 该数据集是从 PASCAL VOC 数据集[43] 中挑选的, 具有 850 张自然图像。

实现细节: 本文提出的算法使用 Keras 实现, 前 13 层卷积使用 VGG-16 预训练参数进行初始化, 其他权值的初始化采用 Xavier[44], 初始学习率为 10^{-5}, 权重衰减为 0.000 5, 输入图片大小为 256×256, 训练集采用 DUTS-TR, 并使用数据增强。在训练模型时模型共进行了 150 次迭代, 训练用时 29 个小时。对本文模型进行测试时, 实时处理速度为 18 f/s。源代码可在: https://github.com/yshenkai/SOD 中下载。

评价指标: 为了评估本文模型, 借助了 PR 曲线、F-measure 值和平均绝对误差 MAE 3 个指标将本文模型与其他的 13 个先进的模型相比较。

$$F_\beta = \frac{(1+\beta^2)\times \mathrm{Precision}\times \mathrm{Recall}}{\beta^2\times \mathrm{Precision}\times \mathrm{Recall}} \tag{23}$$

其中令 $\beta^2 = 0.3$, Precision 表示准确率; Recall 表示召回率。使用 F_β 作为评价指标的目的在于消除 Precision 与 Recall 之间的矛盾, 可综合评价模型的优劣。除了 F_β 和 PR 曲线, 还计算了平均绝对误差 (MAE) 来测量预测的显著性图与真实显著图之间的差异。

$$\mathrm{AE} = \frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y)-G(x,y)| \tag{24}$$

式中: W、H 分别为输入图片的宽和高; $S(x,y)$ 表示在 (x,y) 点上的显著度预测; $G(x,y)$ 表示在该点真实显著度值。使用 MAE 作为评价指标的目的在于能够比较直观地反映预测值与真实值之间的偏差。在本文中 $W = H = 256$。

3.2　性能比较

本节使用了上述的评价指标将 BML-CNN 模型与其他 13 个先进模型 BL[45]、KSR[46]、DRFI[47]、LEGS[6]、MDF[5]、ELD[23]、DS[48]、MCDL[19]、DCL[49]、RFCN[7]、DHS[24]、UCF[40, 50] 和 Amulet[27] 进行比较, 同时为了实验的严谨性, 使用了作者推荐的参数

设计和其提供的源码或者直接利用作者提供的显著性图。

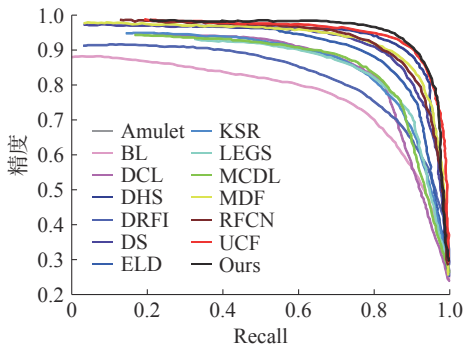
表 1 中 MAE 与 F-measure 是 14 个模型在 6 个标准数据集上计算而来，前三好的结果分别以红色、绿色和蓝色标注。可以看出本文所提出的模型在以上数据集中表现极为出色。

表 1　14 个模型的 MAE 和 F_β 对比

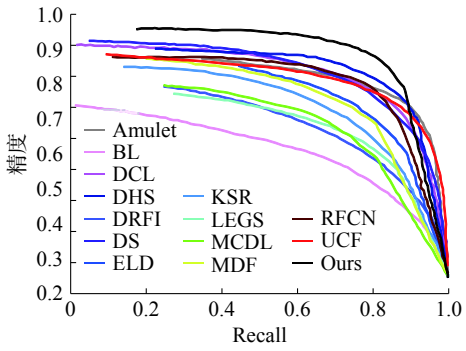
Table 1　MAE and F-measure were compared in 14 models

方法	DUTS-TE		DUT-OMRON		HKU-IS		THUR15K		ECSSD		PASCAL-S	
	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β
BL	0.238	0.409	0.239	0.499	0.207	0.660	0.219	0.530	0.217	0.684	0.249	0.574
KSR	0.121	0.602	0.131	0.591	0.120	0.747	0.123	0.604	0.135	0.782	0.157	0.704
DRFI	0.175	0.541	0.138	0.550	0.145	0.722	0.150	0.576	0.166	0.733	0.207	0.618
LEGS	0.138	0.585	0.133	0.592	0.119	0.723	0.125	0.607	0.119	0.785	0.155	0.697
MDF	0.100	0.673	0.092	0.644	—	—	0.109	0.636	0.108	0.805	0.146	0.709
ELD	0.093	0.628	0.092	0.611	0.074	0.706	0.098	0.634	0.082	0.810	0.123	0.718
DS	0.091	0.632	0.120	0.603	0.078	0.785	0.116	0.626	0.124	0.826	0.176	0.659
MCDL	0.105	0.594	0.089	0.625	0.092	0.757	0.103	0.620	0.102	0.796	0.145	0.691
DCL	0.149	0.714	0.157	0.684	0.136	0.853	0.161	0.676	0.151	0.827	0.181	0.714
RFCN	0.090	0.712	0.111	0.627	0.089	0.835	0.100	0.695	0.109	0.834	0.133	0.751
DHS	0.067	0.724	—	—	0.054	0.852	0.082	0.673	0.063	0.871	0.095	0.773
UCF	0.117	0.629	0.132	0.613	0.074	0.808	0.112	0.645	0.080	0.841	0.127	0.701
Amulet	0.085	0.678	0.098	0.647	0.052	0.839	0.094	0.670	0.061	0.869	0.100	0.763
Ours	0.063	0.758	0.070	0.732	0.049	0.872	0.071	0.731	0.063	0.882	0.090	0.803

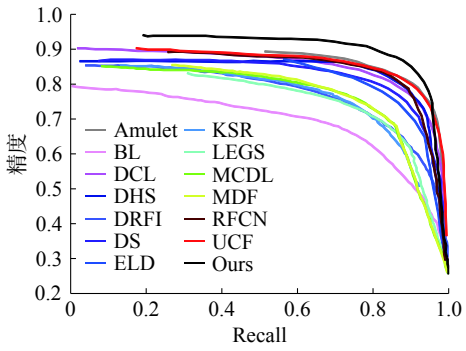
定量比较：由表 1 可以看出本文提出的模型 BML-CNN 在数据集 DUTS-TE、DUT-OMRON、HKU-IS、THUR15K、PASCAL-S 上 MAE 降低了 5.97%、21.35%、5.77%、13.41% 和 10%，在 F_β 指标上分别提高了 4.69%、7.02%、2.23%、8.62% 和 3.88%。在数据集 ECSSD 上 BML-CNN 比 Amulet 的 MAE 高了 3.28%，但 BML-CNN 却在 F_β 比 Amulet 高了 1.26%。由图 6 中 PR 曲线可以看出本文提出的 BML-CNN 模型在数据集 DUTS-TE、THUR15K、ECSSD 上，具有更高的召回曲线，表明在这 3 个数据集中，模型 BML-CNN 表现得比其他 13 个模型更加出色。



(a) ECSSD上的PR曲线比较



(b) HKU-IS上的PR曲线比较



(c) THUR15K上的PR曲线比较

图 6　14 种显著性检测方法在 ECSSD、HKU-IS 和 THUR15K 上的 PR 曲线比较

Fig. 6　PR curves of 14 saliency detection methods on DUTS-TE, THUR15K and ECSSD were compared

如图 7, 第 1 行表示输入的图片, 可以看出, 本文提出的 BML-CNN 模型在显著物体预测和边界保持中均优于现有的 Amulet 和 UCF 方法, 此外在处理含有倒影的图片 (例如图 7 中第 5 幅图片), BML-CNN 模型具有更高的鲁邦性。

定性比较: 在 HKU-IS 与 DUTS-TE 两个数据集上, 使用 13 个模型中表现比较出色的 Amulet 与 UCF 模型给出的显著性预测图进行比较, 为防止模型对特定显著性物体出现过拟合, 选取具有不同显著性物体来进行比较。如图 7, 从第一幅图片可以看出, 在动物显著性预测时本文模型比其他模型保留了更多实体的信息, 且边界保持较好, 从第 3 幅图片中可以看出, 注意力机制的应用消除了更多背景的影响, 实现了更准确的预测。从第 4 幅图中可以看出, 使用更有效的高层语义与底层轮廓传递策略, 可在显著性预测时保留更加完整的边界信息。从第 5 幅图片中可以看出, 注意力机制和高层语义与低层轮廓传递策略在处理镜像实体的问题中表现出更高的鲁棒性。

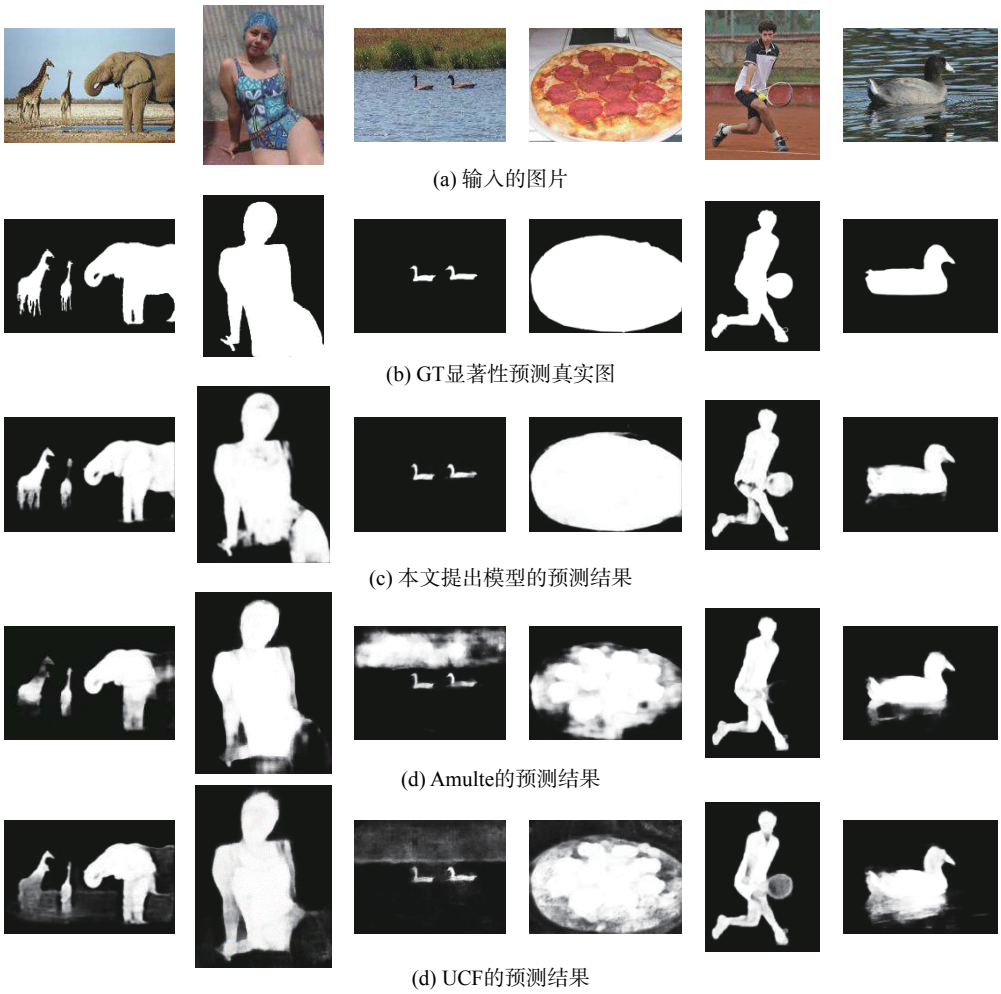


图 7　物体显著性预测图对比

Fig. 7　Object saliency prediction graph comparison

4　结束语

本文提出了结合注意力机制与多尺度融合的双向消息传递链路显著性目标检测算法, 首先通过带有注意力模块的特征提取层获取有效高层语义信息, 然后通过双向消息传递链路实现高层语义与底层轮廓的双向传递, 最后通过多尺度融合策略实现多层不同尺度的卷积特征的融合, 从而产生显著物体的预测图。与现有算法相比, BML-CNN 模型的性能在不同数据集上均获得较高的提升, 在边界保持、抑制背景噪声和镜像实体的处理等问题上都有最优异的表现。

模型虽然在复杂背景下的表现比较出色, 也有很好的边界保持。但是该模型对于镜像实体问题 (如倒影、镜中映像) 的处理尚未达到最优效果, 接下来可以针对镜像实体问题来优化模型。此外, 注意力模块与带门控函数的消息传递链路

的引入导致网络实时处理能力下降，如何同时提高预测效果和实时处理能力值得进一步研究。

参考文献：

[1] 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.

[2] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [J]. arXiv: 1505.04597, 2015.

[3] 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.

[4] LIU Tie, ZHENG Nanning, DING Wei, et al. Video attention: learning to detect a salient object sequence[C]/Proceedings of 200819th International Conference on Pattern Recognition. Tampa, USA, 2008: 1–4.

[5] LI Guanbin, YU Yizhou. Visual saliency based on multiscale deep features[J]. Computer science, 2015.

[6] 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.

[7] WANG Linzhao, WANG Lijun, LU Huchuan, et al. Salient object detection with recurrent fully convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(7): 1734–1746.

[8] CHENG Mingming, ZHANG Guoxin, MITRA N, et al. Global contrast based salient region detection[C]//Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 409–416.

[9] JIANG Zhuolin, DAVIS L S. Submodular salient region detection[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2043–2050.

[10] JUNG C, KIM C. A unified spectral-domain approach for saliency detection and its application to automatic object segmentation[J]. IEEE transactions on image processing, 2012, 21(3): 1272–1283.

[11] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. Computer science, arXiv: 1511.00561, 2015.

[12] REN Zhixiang, GAO Shenghua, CHIA L T, et al. Region-based saliency detection and its application in object recognition[J]. IEEE transactions on circuits and systems for video technology, 2014, 24(5): 769–779.

[13] MU Nana, XU Xiaolong, ZHANG Xong, et al. Salient object detection using a covariance-based CNN model in low-contrast images[J]. Neural computing and applications, 2018, 29(8): 181–192.

[14] ZHOU Li, YANG Zhaohui, ZHOU Zongtan, et al. Salient region detection using diffusion process on a two-layer sparse graph[J]. IEEE transactions on image processing, 2017, 26(12): 5882–5894.

[15] LIU Tie, DUAN Haibin, SHANG Yuanyuan, et al. Automatic salient object sequence rebuilding for video segment analysis[J]. Science China information sciences, 2018, 61(1): 012205.

[16] ZHANG Jing, FENG Shengwei, LI Da, et al. Image retrieval using the extended salient region[J]. Information sciences, 2017, 399: 154–182.

[17] SINGH C, PREET KAUR K. A fast and efficient image retrieval system based on color and texture features[J]. Journal of visual communication and image representation, 2016, 41: 225–238.

[18] XU Gongwen, XU Lina, LI Xiaomei, et al. An image retrieval method based on visual dictionary and saliency region[J]. International journal of signal processing, image processing and pattern recognition, 2016, 9(7): 263–274.

[19] ZHAO Rui, OUYANG Wanli, LI Hongsheng, et al. Saliency detection by multi-context deep learning[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1265–1274.

[20] DAI Jifeng, HE Kaiming, LI Yi, et al. Instance-sensitive fully convolutional networks[J]. Computer science, 2016.

[21] YANG Wei, OUYANG Wanli, LI Hongsheng, et al. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3073–3082.

[22] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Hypercolumns for object segmentation and fine-grained localization[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 447–456.

[23] LEE G, TAI Y W, KIM J. Deep saliency with encoded low level distance map and high level features[C]//Pro-

ceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 660–668.

[24] 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.

[25] XIAO Fen, DENG Wenzheng, PENG Liangchan, et al. Multi-scale deep neural network for salient object detection[J]. IET image processing, 2018, 12(11): 2036–2041.

[26] HOU Qibin, CHENG Mingming, HU Xiaowei, et al. Deeply supervised salient object detection with short connections[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5300–5309.

[27] ZHANG Pingping, WANG Dong, LU Huchuan, et al. Amulet: aggregating multi-level convolutional features for salient object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 202–211.

[28] JIN Xiaojie, CHEN Yunpeng, FENG Jiashi, et al. Multi-path feedback recurrent neural network for scene parsing[J]. Computer science, arXiv: 1608.07706, 2016.

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

[30] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3431–3440.

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

[32] YANG Zichao, HE Xiaodong, GAO Jianfeng, et al. Stacked attention networks for image question answering[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 21–29.

[33] XU Huijuan, SAENKO K. Ask, attend and answer: exploring question-guided spatial attention for visual question answering[J]. Computer science, arXiv: 1511.05234, 2015.

[34] WANG Fei, JIANG Mengqing, QIAN Chen, et al. Residual attention network for image classification[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6450–6458.

[35] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[J]. Computer science, arXiv: 1502.03044, 2015.

[36] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[J]. Computer science, 2013.

[37] ZERIER M D, FERGUS R. Visualizing and understanding convolutional networks[J]. Computer science, 2013.

[38] MAHENDRAN A, VEDALDI A. Understanding deep image representations by inverting them[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5188–5196.

[39] WANG Lijun, OUYANG Wanli, WANG Xiaogang, et al. Visual tracking with fully convolutional networks[C]/Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3119–3127.

[40] ZHANG Pingping, WANG Dong, LU Huchuan, et al. Learning uncertain convolutional features for accurate saliency detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 212–221.

[41] YANG Chuan, ZHANG Lihe, LU Huchuan, et al. Saliency detection via graph-based manifold ranking[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3166–3173.

[42] LI Yin, HOU Xiaodi, KOCH C, et al. The secrets of salient object segmentation[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 280–287.

[43] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The Pascal visual object classes (VOC) challenge[J]. International journal of computer vision, 2010, 88(2): 303–338.

[44] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy, 2010: 249–256.

[45] 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.

[46] WANG Tiantian, ZHANG Lihe, LU Huchuan, et al. Kernelized subspace ranking for saliency detection[C]//Proceedings of the 14thEuropean Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 450–466.

[47] 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.

[48] 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.

[49] 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.

[50] ZHU Wangjiang, LIANG Shuang, WEI Yichen, et al. Saliency optimization from robust background detection[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2814–2821.

作者简介：

申凯，男，1996 年生，硕士研究生，主要研究方向为计算机视觉、图像处理与视觉问答。

王晓峰，男，1958 年生，教授，博士生导师，International Journal of Granular Computing, Rough Sets and Intelligent Systems (IJGCRSIS) 编委，中国人工智能学会机器学习专业委员会常务委员，中国人工智能学会智能交通专业委员会委员等。主要研究方向为人工智能、数据挖掘与知识发现。主持和参加国家 863 计划课题、国家自然科学基金重点课题各 1 项，主持国家合作项目 2 项、辽宁省自然科学基金 2 项，科研项目 30 余项。发表学术论文 70 余篇。

杨亚东，男，1990 年生，博士研究生，主要研究方向为计算机视觉、图像处理。