软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(2):416−439 [doi: 10.13328/j.cnki.jos.005636]

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

视觉注意力检测综述*

王文冠, 沈建冰, 贾云得

(智能信息技术北京市重点实验室(北京理工大学),北京 100081)

通讯作者: 沈建冰, E-mail: shenjianbing@bit.edu.cn

摘 要: 人类能够迅速地选取视野中的关键部分,选择性地将视觉处理资源分配给这些视觉显著的区域.在计算机视觉领域,理解和模拟人类视觉系统的这种注意力机制,得到了学界的大力关注,并显示出了广阔的应用前景.近年来,随着计算能力的增强以及大规模显著性检测数据集的建立,深度学习技术逐渐成为视觉注意力机制计算和建模的主要手段.综述了视觉注意力检测的最新研究进展,包括人眼关注点检测和显著物体检测,并讨论了当前流行的视觉显著性检测数据集和常用的评估指标.对基于深度学习的工作进行了综述,也对之前代表性的非深度学习模型进行了讨论,同时,对这些模型在不同的数据集上的性能进行了详细评估.最后探讨了该领域的研究趋势和未来的发展方向.

关键词: 视觉注意力;视觉显著性;人眼关注点预测;显著物体检测

中图法分类号: TP391

中文引用格式: 王文冠,沈建冰,贾云得.视觉注意力检测综述.软件学报,2019,30(2):416−439. http://www.jos.org.cn/1000-9825/5636.htm

英文引用格式: Wang WG, Shen JB, Jia YD. Review of visual attention detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2):416−439 (in Chinese). http://www.jos.org.cn/1000-9825/5636.htm

Review of Visual Attention Detection

WANG Wen-Guan, SHEN Jian-Bing, JIA Yun-De

(Beijing Laboratory of Intelligent Information Technology (Beijing Institute of Technology), Beijing 100081, China)

Abstract: Humans have ability to quickly select a subset of the visual input and allocate processing resources to those visually important regions. In computer vision community, understanding and emulating such attention mechanism of the human visual system has attracted much attention from the researchers and shown a wide range of applications. More recently, with the ever increasing computational power and availability of large-scale saliency datasets, deep learning has become a popular tool for modeling visual attention. This review includes the recent advances in visual attention modeling, including fixation prediction and salient object detection. It also discusses popular visual attention benchmarks and various evaluation metrics. The emphasis of this review is both on the deep learning based studies and the represented non-deep learning models. Extensive experiments are also performed on various benchmarks for evaluating the performance of those visual attention models. In the end, the review highlights current research trends and provides insight into the future direction.

Key words: visual attention; visual saliency; eye fixation prediction; salient object detection

人类的视觉系统(human visual system)具有极强的感知和数据处理能力,有研究显示[1,2],每秒约有 10^8~10^9 字节的数据进入人眼.认知科学领域的研究表明[3,4],视觉注意力机制(visual attention mechanism)是人类视觉系

* 基金项目: 国家自然科学基金(61673062); 北京市自然科学基金(4182056)

Foundation item: National Natural Science Foundation of China (61673062); Natural Science Foundation of Beijing Municipality (4182056)

收稿时间: 2018-05-20; 修改时间: 2018-08-06; 采用时间: 2018-08-15

统具备如此惊人数据处理能力的重要基础:在处理视觉数据的初期,人类视觉系统会迅速将注意力集中在场景中的重要区域上,这一选择性感知机制极大地减少了人类视觉系统处理数据的数量,从而使人类在处理复杂的视觉信息时,能够抑制不重要的刺激,将有限的神经计算资源分配给场景中的关键部分,为更高层次的感知推理和更复杂的视觉处理任务(如物体识别[5]、场景分类[6]、视频理解[7]等),提供更易于处理且更相关的信息.从人类生理机理的角度而言,人类的视觉注意力机制基于视网膜的特殊生理结构:高分辨率的视网膜中央凹(central fovea)和较低分辨率的边缘视网膜(periphery).视网膜的中央凹区域集中了绝大多数的视锥细胞(cone cells),负责视力的高清成像.当人类关注某一物体时,通过转动眼球,将光线集中到中央凹,从而获取显著区域的更多细节而忽略其他不相关区域的信息.可见,人类视觉注意力机制引导视网膜的生理结构,完成对场景信息的选择性收集任务.而在计算机视觉领域,主要的研究问题在于怎样建立合适的计算模型来解释这种人类视觉注意力机制的潜在机理.在计算机视觉信息处理过程中引入注意力机制,不仅可以将有限的计算资源分配给重要的目标,而且能够产生出符合人类视觉认知要求的结果.因此,视觉注意力检测已经成为计算机视觉领域的研究热点,得到学界的大力关注.

人类视觉注意力机制研究起源于认知心理学(cognitive psychology)和神经科学(neuroscience),早期的代表工作可以追溯到 Koch 和 Ullman 的著作[8].Itti 等人[9]利用认知心理学中的经典理论——特征整合理论(feature integration theory,简称 FIT)[10]和指向搜索模型(guided search model)[11]提出了早期的人类视觉注意力机制的计算模型,并将人类视觉显著性检测研究引入了计算机视觉领域,该任务也被称为人眼关注点检测(human eye fixation prediction).在 Itti 等人的工作后,学界提出了大量的视觉显著性计算模型,这些模型对人眼在场景中某一个位置停留的可能性进行预测.随着计算机视觉的进一步发展,针对目标物体级别的理解显得尤为重要,在此背景下,视觉显著性检测出现了另一个重要的分支——显著物体检测(salient object detection).这一分支的早期研究有 Liu 等人[12]和 Achanta 等人[13]的工作,强调对场景中显著目标整体的准确预测并且获取清晰的显著物体边界,为物体级别的视觉任务(如目标检测[14]、目标备选提取[15]、视频摘要[16]、基于内容感知的图像裁剪[17,18]、目标跟踪[19]等)提供更直接更有效的信息.

与同类文献相比,本文的主要贡献如下.

(1) 对视觉注意力检测在近年来的代表性方法进行了系统和全面的研究,并根据输入数据的不同,将上述模型进一步划分为动态视频和静态图像的视觉显著性检测模型.

(2) 对近年来基于深度学习的视觉注意力计算模型进行了研究和分析,对它们的典型网络结构进行了阐述和分类.

(3) 对人眼关注点检测以及显著物体检测领域的代表性实验数据集、算法的性能评估指标进行了讨论和总结.

(4) 对经典的人眼关注点检测和显著物体检测模型,在静态及动态场景下的性能进行了定量分析,并探讨了视觉注意力检测领域未来的发展趋势.

本文第 1 节对人眼关注点检测模型进行综述.第 2 节介绍显著物体检测领域的代表性工作、主要假设以及基于深度学习的算法.第 3 节介绍人眼关注点检测和显著物体检测领域常用的数据集.第 4 节介绍人眼关注点检测和显著物体检测领域用于算法性能评估的指标.第 5 节针对当前经典的人眼关注点检测模型以及显著物体检测模型,在静态及动态场景下的性能进行定量评估.第 6 节对视觉注意力检测这一研究领域未来的发展趋势进行展望.

1 人眼关注点检测模型

人眼关注点检测是指通过数学建模的方式模拟人类视觉注意系统的机能,对图像或视频中不同位置受到视觉关注的概率进行计算,通过与真实的人类眼动数据相比对,能够对模型预测的视觉显著性结果进行量化评估.设有 K 个观测对象注视了 N 张图像 $\{I_n\}_{n=1}^{N},\{p_n^k\}_{k=1}^{K}$ 为观测对象在观测第 n 张图像时的眼动数据(人眼关注点位置),人眼关注点检测任务可以定义为:找到一个刺激-注意力变换函数(stimuli-saliency mapping function)$f\in F$.

该函数通过最小化人眼关注点预测误差得到,如式(1)所示.

$$\sum_{k=1}^{K}\sum_{n=1}^{N} m(f(I_n), p_n^k) \tag{1}$$

这里,$m \in M$ 被定义为一种人眼关注点真值与显著性预测的距离度量(参见第 4.1 节).

这一领域早期的代表性工作是 Koch 等人[8]于 1985 年提出的视觉选择性注意理论.他们在对灵长类动物和人类视觉系统进行研究的基础上,提出了视觉注意力分配过程中的 3 个要素.

(1) 图像中的一些基本底层特征:颜色、朝向、运动方向和差异;

(2) 视觉选择性注意机制的一个重要功能,是使不同图像之间的信息变成一个连贯的整体;

(3) WTA 机制,即赢者取全(winner-take-all)的竞争机制,在视觉注意过程中,先选择最明显的目标,然后选择次明显目标.

1998 年,Itti 等人[9]基于 Koch 等人的理论以及认知心理学经典的特征整合理论[10]、指向搜索模型[11],提出了首个视觉显著性的计算模型,其算法流程主要含有 3 个步骤:提取颜色、亮度和朝向这 3 种初级视觉特征;在多尺度下使用中央-周围对比度(center-surrounding contrast)计算 3 种体现显著性的特征图(显著特征提取);对特征图进行归一化处理,然后进行特征图的合成(特征融合),运用 WTA 机制标注出图像中的显著目标.该算法对后来计算机视觉领域中视觉显著性计算模型的研究产生了重要影响,尤其是在深度学习技术得到大规模运用之前,主流的显著性检测算法都采用了类似的框架.

1.1 静态场景中的人眼关注点检测模型

在 Itti 的工作之后,计算机视觉领域出现了大量关于人眼注意点检测的工作,这些工作主要关注静态图像中的视觉显著性检测.根据这些模型所采用的人类视觉注意力机制的作用机理,可以将其划分为两种:自底向上(bottom-up)的模型和自顶向下(top-down)的模型.

自底向上的模型[20−25]受数据的驱动,典型的例子是人类在自由观看(free-viewing)模式下分配视觉注意力的情形.这类模型主要利用图像中的颜色、亮度、边缘等特征,考虑像素与周围领域在特征上的差异,计算该像素的显著性.Itti 等人在 1998 年的工作[9]就是这一类模型的典型代表.中央-周围(center-surround)原理是自底向上模型使用最多的理论,相关研究结果表明,视觉神经元往往只针对一个较小的中心区域敏感,如果在中心的周围区域也产生刺激,那么这个刺激会抑制中心区域对视觉神经元的刺激,这意味着视觉神经元对局部空间的不连续性较为敏感,容易注意到那些与局部周围邻域对比较为明显的位置,这也是视网膜、外侧膝状体和底层视觉皮层的工作原理.为了检测局部中心与周围邻域间的对比度(contrast),相关工作往往在不同的尺度上采用不同的特征进行计算,得到的差异度被作为估计最终显著性结果的依据.

自顶向下的模型[26−28]主要受任务驱动,受到人类主观意识的影响,包括先验性知识、当前的目标或对未来的预期.例如在等待客人时,人的注意力会集中在门的位置;或者在监控场景下,场景中的人往往更能引起监控者的注意.自顶向下的模型需要考虑高层的先验信息,例如人脸、车辆等,因此在基于特定任务数据上使用机器学习算法进行建模的方式,成为这类工作的主流.由于在自顶向下的注意力机制中,人类个体的情感、意志等主观因素难以控制,绝大部分人眼关注点检测算法都属于自底向上的模型.自底向上和自顶向下模型是基于不同的视觉注意力机制,根本机理不同.从人类认知学角度而言,自底向上模型主要是研究人类注意力机制的早期机制,数据驱动和任务、人的主观情感无关;而自顶向下的模型综合人类复杂推理和认知过程,和人类的心理活动、当下的主观情感相关.而自底向上的注意力机制是人在放松状态下,不加思考地自由观看场景时的视觉选择性特性,和人的主观个人因素关联较少,因此在计算机视觉领域,主要研究重点在自底向上的注意力机制,因为外部变量可控,内部变量影响少,而自顶向下模型主要在认知心理学相关领域有较多研究.

从 Itti 等人的工作[9]开始,传统的人眼关注点检测模型的计算框架主要是基于 Treisman 和 Gelade 的经典特征融合理论[10].该理论通过对人类视觉系统的研究,描述了不同视觉特征的融合,能够对人类的视觉注意力机制产生引导作用.基于这一理论,传统的人眼关注点检测模型主要包含 3 个步骤:(1) 显著性特征提取;(2) 基于显著特征的显著性图推断;(3) 不同特征的显著性图融合.在显著性特征提取阶段,首先检测不同的底层显著性特

征,如颜色、纹理等.在显著性推断阶段,根据中央-周围理论,计算中央区域与不同尺度上的周围区域的差异,如考虑局部邻域[9,23,27],或更大范围的全局邻域[20,22,24].由于上述过程同时使用不同的特征对显著性进行推断,因此在最后一步,需要融合不同特征得到的显著性图,这一融合过程可以基于不同的计算方式,如通过手工定义的线性组合权重[9],或通过支持向量机(support vector machine,简称 SVM)训练得到组合权重[25].

近年来,随着深度学习技术在计算机视觉领域的兴起,基于深度神经网络的显著性检测模型[29−36]已经成为主流.这些模型利用大规模的眼动数据集[37]以及深度学习技术的强大学习能力,达到了远好于传统模型的性能.在著名的公共人眼关注点检测数据集 MIT300[38]上,排名前 10 的显著性检测模型均使用了深度学习技术,其中,eDN 模型[29]是利用深度神经网络来对视觉注意机制进行建模的早期代表性工作,此后相继提出了 DeepFix 模型[30]、SALICON 模型[31]、Mr-CNN 模型[32]、Shallow and Deep 模型[33]、attentive LSTM 模型[34]、DVA 模型[35].这些工作的研究思路主要是探讨更复杂更有效的网络结构.Jetle 等人[36]测试了多个损失函数,这些损失函数主要基于概率理论的距离测度,实验结果表明,基于 Bhattacharyya 距离测度的损失函数能够给出最好的训练效果.

1.2 动态场景中的人眼关注点检测模型

在显著性检测领域中,有很多工作研究了如何模拟人类在观看图像时的视觉注意力机制,但关于动态场景下人类如何分配视觉注意力的研究相对较少,动态视觉注意力机制在人类日常行为中却更为普遍且更为重要.与静态的视觉注意力检测相比,动态视频中的运动信息为人眼关注点检测提供了很强的引导,然而,背景区域中的运动同样也会产生强烈的干扰,此外,光流模型计算运动信息时产生的计算误差也会给动态显著性检测带来很大的负面影响.

早期的动态人眼关注点检测的研究工作[39−46]主要为自底向上的模型,这些模型通过将静态显著性特征和时间域信息(如光流场、时域差分等)相结合,检测动态场景下的视觉注意力,其中大部分工作[39−41]都可被看作是已有静态显著性模型的基础上考虑运动信息后的扩展.例如,Gao 等人[39]通过在图像显著性检测模型[47]中添加额外的运动信息,来计算视频上的显著性.类似的,Mahadevan 等人[40]利用文献[47]中的模型,将中心-周围对比度显著性与动态纹理特征相结合;Guo 等人[48]采用傅里叶变换的相位谱(phase spectrum of the Fourier transform)计算动态显著性;Seo 等人[41]利用局部回归算子(local regression kernel)计算视频中像素或超体素和周围区域的相似性;Rahtu 等人[49]利用统计模型和局部特征(如光照、颜色和运动信息)上的对比度来计算视频显著性.这些模型严重依赖于特征工程,因而模型的性能受到了手工设计特征的限制.

目前,基于深度学习的人眼关注点检测模型非常少[50−52],主要原因是动态场景的眼动数据集的数量较少且规模普遍较小.其中,

- Bak 等人[50]使用了经典的双流网络架构(two-stream network),将提取静态表观特征的网络与提取运动特征的网络相结合.
- Jiang 等人[51]使用两层长短期记忆神经网络(long-short-term memory network),与用于检测似物性(objectness)、光流和静态表观特征的网络相结合.
- Wang 等人[52]提出了基于卷积长短期记忆神经网络(convolutional long short-term memory network)的动态人眼关注点检测模型.该模型通过加入静态注意力模块(attentive module),将动态和静态显著性特征的提取进一步解耦合,并充分利用现有的大规模静态眼动数据,对整个网络结构进行充分的训练;同时,该网络设计还避免了之前动态显著模型需要进行耗时的光流计算的缺陷,进一步提升了检测速度.

相对于之前基于手工特征的动态显著性计算模型而言,这些基于深度学习的工作取得了更好的性能,同时也证明了将神经网络用于解决该问题的潜在优势.

2 显著物体检测模型

与人眼关注点检测任务相比,显著物体检测任务的研究历史相对较短,且该任务是一个纯计算机视觉任务,Liu 等人[12]和 Achanta 等人[13]的研究是该领域的早期代表性工作.2007 年,Liu 等人[12]正式提出了显著物体检测任务,可以视为视觉注意力机制在物体分割任务上的延拓,提出的背景是计算机视觉领域从底层视觉处理任务

向高层视觉理解方向的深入,对物体级别的感知和描述成为相关研究的关键.Liu 等人使用了不同尺度下的对比度(multi-scale contrast)、中心-周围直方统计(center-surround histogram)以及颜色空间分布(color spatial-distribution)这 3 种显著性度量方式,之后,使用条件随机场(conditional random field)对这些显著性特征进行整合,同时也提出了第 1 个显著物体检测数据集,并引入了查准率(precision)、查全率(recall)、F-值(F-measure)这 3 个重要的评估指标.2009 年,Achanta 等人[13]在 Liu 等人工作的基础上,提出了在频率域(frequency domain)上对显著物体进行快速检测的方法,该工作给出了查准率-查全率曲线(precision-recall curve),并进一步优化了 F-值的定义,这两种评估指标成为日后显著物体检测领域最常用的评估指标.Liu 等人[12]和 Achanta 等人[13]的研究为显著物体检测这一方向上的后续工作奠定了基础.

设有 N 张图像 $\{I_n\}_{n=1}^N$ 和相应的显著物体真值标定 $\{S_n\}_{n=1}^N$,这里,$S_n \in \{0,1\}^{W\times H}$ 为第 n 张图像的显著性二值化标定,基于以上定义,显著物体检测任务可以定义为:找到一个图像-显著物体预测函数 $f \in F$,该函数可以通过最小化显著物体预测误差得到,如式(2)所示.

$$\sum_{n=1}^{N} m(f(I_n), S_n) \tag{2}$$

这里,$m \in M$ 被定义为一种显著物体真值标定与显著物体预测的距离度量(参见第 4.2 节).显著物体真值 $\{S_n\}_{n=1}^N$ 可以通过观测对象的眼动数据进行标定,这表示显著物体检测和人眼关注点检测两者间存在着密切的相关性.

2.1 图像显著物体检测模型

早期的图像显著物体检测模型[53−57]主要基于自底向上的方法,使用了不同的底层视觉特征,如颜色、边缘等,由于显著物体检测与人眼关注点检测任务关系密切,都是对人类视觉注意力机制的建模,因此早期的显著物体检测模型也借鉴了人类视觉注意力机制的一些基本理论,包括经典的对比度假设、中心-周围假设.比如,Liu 等人[12]和 Achanta 等人[13]都使用了这两种假设,Cheng 等人[53]也使用了类似的假设,他们考虑了局部和全局范围上的颜色对比度信息,算法简洁明了,得到了学界的广泛关注.此外,Yan 等人[55]提出通过对图像进行不同尺度的过分割,完成在不同尺度上表观一致的图像表达,并在不同尺度上对显著性特征进行提取和融合优化,来得到最终显著物体检测结果.视觉中心偏移(center bias)也是一个常用的基于人类注意力机制的假设[55].该假设基于这样的现象:人类在观测场景时,视觉系统具有向场景中央分配较高注意力权重的倾向.之后,流行的假设是背景先验假设(background prior),该假设在 2012 年由 Wei 等人[54]提出.与中心-周围假设和视觉中心偏移假设尝试定义“什么更有可能是显著区域”不同,该假设尝试定义“什么更有可能是背景”.该假设基于这样的观察:在大部分场景中,图像四周边缘的部分属于背景的概率较大.该假设可视为对视觉中心偏移假设的进一步发展,在深度学习技术得到大规模应用之前,背景先验假设是显著性检测领域最有效的假设,绝大多数性能优异的模型[58−62]都基于这一假设,这些工作主要关注如何进一步提高背景先验假设的准确度以及如何应用更先进的单分类器(one-class classifier).通过背景先验假设,相当于获取了一类(背景)样本,那么该问题可被视为只给出 1 类样本的单类分类(one-class classification)问题.例如,Jiang 等人[59]的工作可被视为基于可吸收随机游走算法的单分类器,Wei 等人[54]和 Zhang 等人[61]的工作则是通过不同的距离度量方式对样本进行分类.

随着深度学习技术在图像分类问题上取得巨大的成功,显著物体检测领域的研究重心也逐渐向基于深度学习的模型偏移.稍早期的工作(2015 年~2016 年)使用了深度学习特征作为更有效的显著性表达,并使用全卷积神经网络进行训练.例如,Zhao 等人[63]的工作使用深度神经网络预测图像超像素(superpixel)或目标物体备选(object proposal)的显著性值,从而将显著物体检测任务转换为对图像超像素或目标物体备选的分类问题(显著/不显著);Wang 等人[64]使用两个深度网络分别用于预测局部超像素和全局目标物体备选的显著性值;Li 等人[65]利用每个超像素在不同尺度上的深度学习特征,提取上下文信息(contextual information),然后,通过分类网络来对每个超像素是否显著进行分类;Lee 等人[66]将深度特征作为高层信息,将 Gabor 滤波响应、颜色直方统计等作为底层特征,融合不同层次的显著性信息后进行显著性预测.这类模型取得了较好的性能,但存在一些缺陷,比如,由于使用了基于全连层(fully connected layer)的分类网络,这类模型的参数量较大且损失了空间信息;同时,

由于需要对每一个超像素或目标物体备选进行显著/不显著分类,这类算法的计算代价较大.

随着全卷积神经网络(fully convolutional neural network)的兴起,近年来(2016 年~2018 年),基于深度学习的显著物体检测工作都使用或改造了全卷积神经网络,进行像素级别的显著性预测.例如,Wang 等人[67]将深度学习技术与之前的显著性先验相结合,利用显著性先验获取初始的显著性估计,然后,使用循环神经网络(recurrent neural network)来对初始的显著性先验进行优化.有一些工作[68−72]受到像素级语义分割任务的启发,提出将不同神经网络层的特征相融合来进行显著物体检测.由于深度神经网络的较浅层网络能够保留较多较细粒度的底层视觉特征,而较深层的网络能够提取更高层的、语义级的特征,因而,融合不同神经网络层的特征既能保留原有的底层空间信息,又能获得高层语义信息.目前,基于深度学习技术的显著物体检测工作的主要研究重心是探索更有效、能保留更多空间细节的网络结构.例如,Zhang 等人[68]利用不同尺度输入得到了深度信息,Hou 等人[69]将每一层的深度神经网络特征都进行互连.除此之外,2018 年,Wang 等人[70]提出了通过视觉注意力先验来检测视觉显著物体的 ASNet 模型.该模型将视觉注意力作为对整个场景的高层次理解,通过较高层的神经网络层进行学习,显著物体检测任务则被视为更细粒度的、物体层面的显著性检测,由视觉注意力提供自顶向下地引导.ASNet 模型基于堆栈卷积长短期记忆神经网络,该网络特有的循环结构能够迭代地优化显著性检测结果.该工作为视觉注意力机制提供了更深层次的解读,揭示了显著物体检测和人眼关注点检测二者之间的关联性.就整体而言,基于深度学习的显著物体检测模型取得了远超传统模型的性能.

2.2 视频显著物体检测模型

早期的动态视觉显著性模型主要关注动态场景下的人眼关注点检测任务,针对视频显著物体检测的研究,可以追溯到 Liu 等人[73]和 Wang[74]等人的工作.2014 年,Liu 等人[73]提出在超像素级别上,利用运动和表观信息来检测视频中的显著物体整体.Wang 等人[74]提出了梯度流场(gradient flow field)和全局显著物体连续性假设,首先,利用目标的表观和运动的不连续性,计算光流梯度幅值和颜色梯度幅值,建立梯度流场来确定显著物体的初始位置,结合局部和全局显著性线索进一步优化;然后,利用视频显著物体在时空域上连续的假设,建立全局显著性优化方程,得到最终的时空域平滑的显著物体估计结果.该工作同时提出了第 1 个专门用于显著物体检测的数据集 ViSal,并将查准率-查全率曲线和 MAE 值这两个评估指标用于视频显著物体检测任务.

关于视频显著物体检测的工作逐年增多.2015 年,Wang 等人[75,76]进一步提出了基于测地距(geodesic distance)的视频显著物体检测算法,并将显著物体检测用于无监督的视频分割.该算法通过建立帧内和帧间图模型对视频帧内和帧间信息进行建模,并使用测地距在帧内和帧间图模型上对每个超像素的显著性进行度量,因为测地距能够较好地获取相应的结构化信息.Kim 等人[77]提出了基于重启动随机游走(random walk with restart,简称 RWR)的视频显著物体检测算法.该算法将空间域的显著性作为随机游走的重启动分布,利用空间特征建立随机游走的转移概率矩阵,将达到稳定状态时相应的概率分布作为最终的时空显著物体估计.2017 年,Liu 等人[78]通过对显著性检测结果在时间域上的迭代更新,进一步发展了他们之前的工作[73].此外,还有 Guo 等人[79]提出的基于目标物体备选(object proposal)的频显著物体检测模型;文献[80]提出的基于低秩相关性(low-rank coherency)的模型;文献[81]提出的利用时空显著性线索、局部约束以及似物性指标的算法;Li 等人[82]提出的基于栈式自动编码器(stacked autoencoder)的视频显著物体检测模型;Alshawi 等人[83]通过计算每个像素的不确定性(uncertainty)来提高视频显著物体的检测结果.

2017 年,Wang 等人[84]提出了基于全卷积神经网络的视频显著性物检测模型,这也是第 1 个基于深度学习的视频显著物体检测模型.该工作主要解决了两个关键问题:(1) 在缺乏充分训练样本的条件下,如何对深度学习模型进行训练;(2) 如何建立快速且准确的视频显著性检测模型.该模型包含了两个模块,分别用于学习空间域和时间域上的显著性信息.其中,动态显著性检测模块,显式地利用了静态显著性检测模块的静态显著性估计,直接生成时空显著性检测结果,并且避免了耗时的光流计算.同时,该工作中提出了一个重要的数据扩充技术,能够利用已有的标定好的图像数据集来合成大量的视频数据,从而使深度视频显著物体检测模型能够学习到丰富的显著性信息,并避免了在原来少量视频样本上的过拟合.通过利用合成的视频数据和真实的视频数据,该视频显著物体检测模型能够成功地学习到时间域和空间域的显著性信息,从而产生更准确的显著性检测结果

和达到更快的检测速度.

3 视觉显著性检测数据集

本节主要介绍图像人眼关注点检测、视频人眼关注点检测、图像显著物体检测以及视频显著物体检测领域的代表性数据集.

3.1 图像眼动数据集

常用的静态眼动数据集有 MIT300[38]、MIT1003[25]、TORONTO[22]、PASCAL-S[85]、SALICON[37]以及 DUT-OMRON[58].

(1) MIT300 数据集

2012 年,麻省理工的 Judd 等人建立了 MIT300 数据集[38].该数据集包含了 300 张自然图像以及 39 名观测者的眼动数据,是图像人眼关注点检测领域影响力最大、使用最广泛的数据集.该数据集得以广泛应用的原因是:数据分布较为合理且具有一定的难度;建立较早,影响力较大;人眼关注点的真值标定不公开,从而防止了模型在该数据集上的过拟合;发布了相关评估实验的代码,且评估结果详实充分.

(2) MIT1003 数据集

MIT1003 数据集[25]也是由麻省理工的 Judd 等人建立的.该数据集包含了从 Flikr 和 LabelMe 网站得到的 1 003 张图像,其中 779 张为风景像,228 张为肖像,并公开了 15 名观测者的眼动数据.同时,眼动数据的记录过程还考虑了记忆机制:每个观测者被要求在 100 张图像中指出哪一张是先前看到的.MIT1003 数据集可以作为 MIT300 数据集的补充,即在 MIT1003 数据集上训练基于机器学习的注意力模型,然后,以 MIT300 数据集作为测试集进行性能评估.

(3) TORONTO 数据集

TORONTO 数据集[22]于 2006 年由约克大学的 Bruce 等人建立,是计算机视觉领域提出最早、使用最广的数据集之一.它包括了 120 张分辨率为 511×681 的彩色图像.这些图像属于室内和室外场景,一共记录了 20 名观测者的眼动数据.在每名观测者眼动数据的采集过程中,每张图像呈现 3s,图像之间插入为时 2s 的灰度图像作为间隔.

(4) PASCAL-S 数据集

PASCAL-S 数据集[85]于 2014 年由乔治亚理工学院的 Li 等人建立.该数据集使用了 PASCAL VOC 2010[86]数据集验证集的 850 张图像,并公布了 8 名观测者在 2s 内、自由观看模式下观测图像得到的眼动数据.

(5) SALICON 数据集

SALICON 数据集[37]是 2015 年由新加坡国立大学的 Jiang 等人建立的.该数据集包含了 20 000 张选自 Microsoft COCO 数据集[87]的图像,是迄今为止图像人眼关注点检测领域规模最大的数据集.但是该数据集没有使用眼动仪录制眼动数据,而是利用了亚马逊众筹标记平台(Amazon Mechanical Turk,简称 AMT),让标注者用鼠标点击自己关注的位置.Jiang 等人强调了用鼠标记录的眼动数据与眼动仪记录的实际数据高度接近,但是 Tavakoli 等人[9]指出,眼动仪记录的真实眼动数据和鼠标记录的眼动数据之间仍然存在着较大的区别,当分别利用不同方式记录的眼动数据作为训练样本训练模型时,不同的训练样本会对模型的最终性能产生不同的影响;同时,利用鼠标记录的眼动数据对模型的性能进行评估时,产生的评估结果以及模型性能的相对好坏也与在真实眼动数据上的测试结果不符.尽管如此,鉴于 SALICON 数据集的较大规模,还是被当前主流的基于深度学习技术的显著性检测模型广泛使用.SALICON 数据集公开了训练集(10 000 张)和验证集(5 000 张)的眼动数据,但保留了测试集(5 000 张)的眼动数据.

(6) DUT-OMRON 数据集

DUT-OMRON 数据集[58]由大连理工大学的 Yang 等人于 2013 年建立.该数据集包含 5 168 张图像,每张图像提供了 5 名观测者的眼动数据.该数据集主要关注显著物体检测,因而在物体之外的视觉注意点在后处理过程中被移除.

我们将上述常用的静态场景的眼动数据集的相关信息进行了总结,见表 1.

Table 1 Information of eye-tracking datasets collected in static scenes

表 1 关于静态场景下眼动数据集的相关信息

数据集	建立时间	出版物	图像数目	观测者数目	图像大小
MIT300[38]	2012	技术报告	300	39	max(宽,高)=1024
MIT1003[25]	2009	ICCV	1 003	15	max(宽,高)=1024
TORONTO[22]	2006	NIPS	120	20	511×681
PASCAL-S[85]	2014	CVPR	850	8	max(宽,高)=500
SALICON[37]	2015	CVPR	20 000	–	640×480
DUT-OMRON[58]	2013	CVPR	5 168	5	max(宽,高)=400

3.2 视频眼动数据集

与图像眼动数据集相比,动态场景下的眼动数据集较少.这主要是由于收集人类在观测动态视频时的眼动数据更为困难,对眼动仪器的要求更高,并且需要的工作量也更多.目前,代表性的动态眼动数据集主要有 4 个:Hollywood-2[88]、UCF-sports[88]、DIEM[89]以及最新提出的 DHF1K[52].

(1) Hollywood-2 数据集

Hollywood-2 眼动数据集[88]由多伦多大学的 Mathe 等人在 2012 年建立,包括了 Hollywood-2 动作识别数据集[90]中的所有 1 770 个视频.这些视频是从 69 个电影中收集的,并按照 12 个动作类别进行了标注,例如吃饭、接吻和跑步等.眼动数据的收集过程共有 19 个观测对象参与完成,这些观测对象被分为 3 组:自由观看组(3 个观测对象)、人类动作标注组(12 个观测对象)和视频内容标注组(4 个观测对象).虽然 Hollywood-2 数据集的视频数量较大,但这些视频的内容仅限于常见的人类动作行为和电影场景,并且该数据集主要关注在任务驱动(动作识别)的观看模式下,由于人类视觉系统的显著性机制,自由观看模式下的人眼关注点数据仅占所有数据的很小一部分比例.Wang 等人的研究[52]指出,当从 Hollywood-2 数据集中随机抽取 1 000 个视频帧后,统计结果显示,84.5%的人眼注视点都位于场景中的人脸位置附近.

(2) UCF-sports 数据集

UCF-sports 眼动数据集[88]也是由 Mathe 等人在 2012 年建立的.该数据集包含了 UCF sports action 数据集[91]中的 150 个视频,这些视频涵盖了 9 种常见的体育运动类别,如潜水、游泳和跑步等.与 Hollywood-2 数据集相类似,该数据集偏向于任务驱动的观看方式,即观测对象在观看过程中被指示“识别在视频序列中发生的动作”,因此,观测对象在观看时具有偏向于动作识别的目的性.Wang 等人的研究[52]指出,当从 UCF sports 数据集上随机选择 1 000 个视频帧进行统计后,结果表明,有 82.3%的人眼注视点位于运动人物的身体区域内.

(3) DIEM 数据集

DIEM 数据集[89]是伦敦大学的 Mital 等人于 2011 年建立的.该数据集包含了从公共网络中收集到的 84 个视频,包括广告、纪录片、体育赛事和电影预告片等.数据集中的每段视频都有人眼关注点的标注,这些标注来自约 50 名观测对象在自由观看模式下的眼动数据.但该数据集包含的场景内容较为有限,并且数据规模较小.

(4) DHF1K 数据集

DHF1K 数据集[52]是由北京理工大学的 Wang 等人于 2018 年建立的,是学术领域迄今为止规模最大的、用于动态场景自由观看模式下的眼动数据集.整个数据集的收集、标定过程耗时近半年.Wang 等人通过 Youtube 搜索引擎搜索了大约 200 个关键字(如狗、行人、汽车等),并忽略了返回结果中包含较大图标、文字或分辨率较低的视频,最终从检索结果中选择了 1 000 个视频序列,这些视频被统一地转换为 30fps 的 Xvid MPEG-4 视频格式,并统一地缩放到 640×360 的分辨率.DHF1K 数据集一共包含了 1 000 个视频序列和 582 605 个视频帧,总持续时间达 19 420s.同时,DHF1K 数据集还提供了更丰富的标定,每个视频都被人工标记了一个场景子类别(共有 150 类),这些子类别进一步被聚类为 7 种主要类别,即动物、景物、人造物以及 4 种人类活动(日常活动、运动、群体行为、艺术表演),这些场景的语义标注帮助人们更深入地理解引导动态注意力机制的高层信息,对将来的研究很有帮助.此外,DHF1K 还提供了运动模式、场景明暗、物体数量等标定.共有 17 位志愿者作为观测

对象参与了眼动数据收集,这些观测对象包括 10 名男性和 7 名女性,年龄范围在 20 岁~28 岁之间,得到共计 51 038 600 组眼动数据.DHF1K 数据集的 1 000 个视频被分为 3 部分,包括:600 个视频作为训练集、100 个视频作为验证集和 300 个视频的测试集.Wang 等人公开发布了训练集和验证集的眼动数据,用于模型的训练和验证,测试集作为对各方法进行统一评估的标准,保留了标注数据.此外,Wang 等人还在 DHF1K、Hollywood-2 和 UCF-sports 这 3 个数据集上对 16 个视觉注意力模型进行了评估,这也是当前动态视觉注意力检测领域规模最大的一次测评.

(5) 其他数据集

除了以上数据集之外,还有 Itti 等人在 2004 年建立的 CRCNS 数据集[92]以及 Hadizadeh 等人在 2012 年建立的 SFU 数据集[93],但是这些数据集的规模和影响力都相对较小.

我们在表 2 中对上述静态场景的眼动数据集的相关信息进行了总结.

Table 2 Information of eye-tracking datasets collected in dynamic scenes

表 2 关于动态场景下眼动数据集的相关信息

数据集	建立时间	出版物	视频数目	分辨率	观测者数目	观看方式
CRCNS[92]	2004	TIP	50	640×480	15	任务驱动
Hollywood-2[88]	2012	ECCV	1 707	720×480	19	任务驱动
UCF sports[88]	2012	ECCV	150	720×480	19	任务驱动
DIEM[89]	2011	Cognitive computation	84	1280×720	~50	自由观看
SFU[93]	2012	TIP	12	352×288	15	自由观看
DHF1K[52]	2018	CVPR	1 000	640×360	17	自由观看

3.3 图像显著物体检测数据集

常用的图像显著物体检测数据集有 MSRA10K[12,53]、ASD[12,13]、ECSSD[55]、PASCAL-S[85]、DUT-OMRON[58]和 HKU-IS[65].

(1) MSRA10K 数据集

2007 年,西安交通大学与微软亚洲研究院的 Liu 等人[12]提出了第 1 篇显著物体检测的论文,同时也提出了第 1 个显著物体检测数据集,但是该数据集只提供了物体边界框这一级别的显著性真值标定.之后,Cheng 等人[53]对该数据集[12]中的 10 000 张数据进行了像素级的标定,这一重标定的数据集被称为 MSRA10K 数据集,是目前显著物体检测领域最常用的数据集之一(主要作为深度显著物体检测模型的训练样本).

(2) ASD 数据集

ASD 数据集是最早使用的显著物体检测数据集之一,由洛桑联邦理工学院的 Achanta 等人[13]在 2009 年建立.该数据集包含了 Liu 等人[12]建立的数据集中的 1 000 张图像,Achanta 等人对这 1 000 张图像进行了像素级的显著物体真值标定,该数据集也常被称为 MSRA1000.

(3) ECSSD 数据集

ECSSD 数据集[55]由香港中文大学的 Yan 等人于 2013 年建立,包含了 1 000 张图像,这些图像由互联网得到.该数据集中的显著物体包含较复杂的结构,且背景具备一定的复杂性.

(4) PSCAL-S 数据集

PASCAL-S 数据集[85]于 2014 年由乔治亚理工学院的 Li 等人建立.该数据集使用了 PASCAL VOC 2010[86]数据集的验证集的 850 张图像.Li 等人根据该数据集上的眼动数据(参见第 3.1 节),对该数据集中每张图像的显著物体进行了标定.该数据集与其他显著物体检测数据集区别较大,没有非常明显的、较少的显著物体,并主要根据人类的眼动数据集进行标注,因此该数据集的难度较大.

(5) DUT-OMRON 数据集

DUT-OMRON 数据集[58]由大连理工的 Yang 等人于 2013 年建立,包含了 5 168 张图像,每张图像提供了 5 名观测者的眼动数据.该数据集的主要任务是显著物体检测,但也提供了眼动数据集(参见第 3.1 节),同时也包括了物体的标定框.该数据集每张图像由 5 人标注完成.

(6) HKU-IS 数据集

HKU-IS 数据集[65]由香港大学的 Li 等人于 2015 年建立,包含了 4 447 张图像和相应的像素级显著物体真值标定.该数据集的每张图像至少满足以下的 3 个标准之一:(1) 含有多个分散的显著物体;(2) 至少有 1 个显著物体在图像边界;(3) 显著物体与背景表观相似.

我们在表 3 中对上述常用的图像显著物体检测数据集的相关信息进行了总结.

Table 3 Information of image salient object detection datasets

表 3 关于图像显著物体检测数据集的相关信息

数据集	建立时间	出版物	图像数目	分辨率
MSRA10K[12,53]	2015	TPAMI	10 000	300×400
ASD[12,13]	2009	CVPR	1 000	300×400
ECSSD[55]	2013	CVPR	1 000	max(宽,高)=400
PSCAL-S[85]	2014	CVPR	850	max(宽,高)=500
DUT-OMRON[58]	2013	CVPR	5 168	max(宽,高)=400
HKU-IS[65]	2015	CVPR	4 447	max(宽,高)=400

3.4 视频显著物体检测数据集

在视频显著物体检测领域常用的数据集有 ViSal[74]、MCL[77]、UVSD[78]、VOS[82]、SegTrack[94,95]、FBMS[96,97]和 DAVIS[98],其中,ViSal、MCL、UVSD、VOS 是专门用于视频显著物体检测任务的数据集,SegTrack、FBMS 和 DAVIS 则在视频物体分割领域有较多的应用.

(1) ViSal 数据集

ViSal 数据集[74]由北京理工大学的 Wang 等人于 2015 年建立,是第 1 个明确提出用于视频显著物体检测的数据集.该数据集包含了 17 个从 Youtube 上收集的视频序列,包含了多种类别的显著物体,如人类、动物等,视频的分辨率多为 320×240,长度为 30 帧~500 帧.该数据集每间隔 5 帧提供像素级的显著物体真值标定.该数据集涵盖了丰富的场景内容、不同的目标运动模式、较为复杂的背景、快速物体形状变化以及相机移动.

(2) MCL 数据集

2015 年,高丽大学的 Kim 等人建立了 MCL 数据集[77].该数据集包含了 9 个分辨率为 480×270 的视频序列,每个视频序列包含约 100 帧~400 帧视频图像,涉及室内和室外场景,包含了多个快速运动的目标以及相机运动.该数据集每隔 8 帧视频图像给出了视频显著物体的像素级真值标定.

(3) UVSD 数据集

UVSD 数据集[78]由上海大学的 Liu 等人于 2017 年建立.该数据集含有 18 个视频序列,每一帧视频图像均进行了像素级的显著性标注.该数据集的视频分辨率以 320×240 为主,长度为 70 帧~300 帧.该数据集的难度主要在于显著的物体相对较小,且显著物体与背景具有一定的相似性.

(4) VOS 数据集

VOS 数据集[82]由北京航空航天大学的 Li 等人于 2018 年建立,该数据集包含了 200 个室内/室外场景下的视频序列,时长共 64min,包含 116 103 帧视频图像,帧率统一为 30fps.该数据集对 7 650 个关键帧进行了像素级的标定;同时,该数据集还收集了 23 名观测者的眼动数据,以此作为确定显著物体的依据.

(5) SegTrack 数据集

SegTrack 数据集的初始版本(V1)于 2010 年由佐治亚理工学院的 Tsai 等人[94]建立.该数据集建立的初始目标是用于视频跟踪分割,在视频分割领域曾经极为流行.之后,在 2015 年被 Wang 等人[74]引入视频显著物体检测,SegTrack-V1 数据集包含了 6 个视频,共 224 帧,其中,penguin 这一视频中不包含显著的前景物体,故这一视频在无监督的视频物体分割和视频显著物体检测任务上不予以采用.之后,在 2014 年,Li 等人[95]建立了 SegTrack 的扩充版本(V2),增添了 8 个视频,并提供了多个目标的标定,SegTrack-V2 数据集因而共包含了 14 个视频序列以及 1 065 帧像素级的标定.

(6) FBMS 数据集

FBMS 数据集[96]的早期版本由加州大学伯克利分校的 Brox 等人在 2010 年建立,包含了 26 个视频.之后,Ochs 等人[97]在 2014 年对其进行了扩展,最终版本共包含了 59 个视频.该数据集最早是用来进行运动分割(motion segmentation)的.该任务主要是在无监督条件下对视频中的运动物体进行分割.之后,由 Wang 等人[74]引入到视频显著物体检测任务中.该数据集的标定较为稀疏,13 860 帧视频中共有 720 帧的真值标定;并且,该数据集的标定较为简单,且并不完全符合视频显著物体的定义.

(7) DAVIS 数据集[98]

DAVIS 数据集于 2016 年由苏黎世联邦理工学院的 Perazzi 等人建立,主要用于视频物体分割.该数据集经过精心设计,因此一经提出就在视频分割领域获得了极大的影响力.该数据集包含了 50 个高质量的视频序列,含有 480p 和 1 080p 两个版本,视频长度约为 2s~4s,且提供了对每帧视频图像的像素级真值标注.该数据集包含了多种挑战,如遮挡、运动模糊、表观变化等,因而有较高的难度.由于该数据集有明显的前景目标,在标注时主要考虑单一的前景目标或相连的两个明显前景目标,较为符合视频显著物体的定义,Wang 等人于 2018 年[84]将其引入视频显著物体检测任务.

我们对上述常用的视频显著物体检测数据集的相关信息进行了总结,见表 4.

Table 4 Information of video salient object detection datasets

表 4 关于视频显著物体检测数据集的相关信息

数据集	建立时间	出版物	视频数目	最高分辨率	视频帧数	标记数量
ViSal[74]	2015	TIP	17	512×288	963	193
MCL[77]	2015	TIP	9	480×270	3 680	3 680
UVSD[78]	2017	TCSVT	18	320×240	3 262	3 262
VOS[82]	2018	TIP	200	800×800	116 103	7 467
SegTrack[94,95]	2010	BMVC	14	640×360	1 065	1 065
FBMS[96,97]	2010	ECCV	59	960×540	13 860	720
DAVIS[98]	2016	CVPR	50	1920×1080	3 455	3 455

4 视觉显著性检测评估指标

本节主要介绍视觉显著性检测任务中常用的评估指标.

4.1 人眼关注点检测评估指标

在人眼关注点检测任务中,研究者们提出了较多的评估指标,其中较为典型的包括 EMD 距离(earth movers distance)、交叉熵(kullback-leibler divergence)、标准化扫描路径显著性(normalized scanpath saliency,简称 NSS)、相似性测度(similarity metric,简称 SIM)、线性相关系数(linear correlation coefficient,简称 CC)、AUC 指标(the area under the receiver operating characteristic (ROC) curve).

这些指标遵循了不同的设计原则,如,交叉熵指标将显著性预测结果与真实的人眼注意力标定视为概率分布;AUC 指标将显著性预测结果视为二分类结果,并使用信号检测理论,从分析分类器分类性能的角度进行评估;或将显著性预测结果与真实的人眼注意力标定二者都视为随机变量,从而可以采用线性相关系数或标准化扫描路径显著性来度量二者相关性.本质上,这些评估指标为显著性检测结果和真实的人眼注意力分布之间的一致性提供了不同维度上的评估,从实际效果而言,综合采取多种评估方式对模型进行评估的做法更可取.

当给定显著性预测结果 $P=[0,1]^{W\times H}$ 时,真实的二值人眼注意点记录 $R=\{0,1\}^{W\times H}$ 以及连续的视觉注意力真值分布 $Q=[0,1]^{W\times H}$.这里,连续的视觉注意力真值分布 Q 是通过对二值的人眼注意点分布图使用较小的高斯核卷积得到的,高斯核的参数主要根据不同眼动数据集上人眼大小和眼动设备的情况来进行设定.下面我们详细介绍人眼关注点检测任务中常用的评估指标:

(1) EMD 距离

EMD 距离(earth movers distance)衡量的是显著性预测结果 P 与连续的人眼注意力真值分布 Q 之间的相似性,该度量方式被定义为:从显著性预测结果 P 上的概率分布转移到连续的人眼注意力真值分布 Q 上的最小代

价.因而,EMD 距离越小,表示估计结果越准确.

(2) 交叉熵

交叉熵(kullback-leibler divergence)主要基于信息理论,经常被用于衡量两个概率分布之间的距离.在人眼关注点检测中,该指标被定义为:通过显著性预测结果 P 来近似连续的人眼注意力真值分布 Q 时产生的信息损失,可通过式(3)来计算.

$$KL(P,Q)=\sum_i Q_i \log\left(\varepsilon+\frac{Q_i}{\varepsilon+P_i}\right) \tag{3}$$

其中,ε表示很小的正则化系数,i 表示第 i 个像素.交叉熵指标是非对称的度量指标,交叉熵越小,表示显著性估计结果越准确.交叉熵这一指标对零值非常敏感,会对稀疏的人眼关注点预测产生非常大的惩罚.

(3) 标准化扫描路径显著性

标准化扫描路径显著性(normalized scanpath saliency,简称 NSS)是专门为显著性检测设计的评估指标.该指标被定义为:对在人眼关注点位置归一化的显著性(均值为 0 和归一化标准差)求平均,可通过式(4)来计算.

$$NSS(P,R)=\frac{1}{N}\sum_i \bar{P}_i\times R_i, N=\sum_i R_i, \bar{P}=\frac{P-\mu(P)}{\sigma(P)} \tag{4}$$

其中,N 表示所有的人眼关注点数目;$\mu(\cdot)$表示均值;$\sigma(\cdot)$表示标准差,该指标越小,表示显著估计结果越准确.

(4) 线性相关系数

线性相关系数(linear correlation coefficient,简称 CC)是一种用于衡量两个变量之间相关性的统计指标.在使用该度量时,将显著性预测结果 P 和连续的人眼注意力真值分布 Q 视为随机变量.然后,统计它们之间的线性相关性,如式(5)所示.

$$CC(P,Q)=\frac{\mathrm{cov}(P,Q)}{\sigma(P)\sigma(Q)} \tag{5}$$

其中,cov(·,·)表示表示协方差,该统计指标的取值范围是[−1,+1].当该指标的值接近−1 或+1 时,代表显著性预测结果与真值标定高度相似.

(5) 相似性测度

相似性测度(similarity metric,简称 SIM)指标将显著性预测结果 P 和连续的人眼注意力真值分布 Q 视为概率分布,将二者归一化后,通过计算每一个像素上的最小值,最后加和得到,如式(6)所示.

$$SIM(P,Q)=\sum_i \min(P_i',Q_i'), \sum_i P_i'=1, \sum_i Q_i'=1 \tag{6}$$

当相似性测度为 1 时,表示两个概率分布一致;为 0 时,表示二者完全不同.

(6) AUC 指标

AUC 指标(the area under the receiver operating characteristic curve,简称 ROC 曲线),即受试者工作特性曲线下面积.ROC 曲线是以假阳性概率(false positive rate,简称 FPR)为横轴,以真阳性概率(true positive rate,简称 TPR)为纵轴所画出的曲线,如式(7)所示.

$$\begin{cases} FPR=\dfrac{FP}{FP+TN} \\ TPR=\dfrac{TP}{TP+FN} \end{cases} \tag{7}$$

其中,TN 表示二值显著性图中的背景区域且对应于显著性真值图中的背景区域的像素个数.ROC 曲线越趋近于左上方,说明算法的性能越好.AUC 即为 ROC 曲线下的面积,通过在[0,1]上滑动的阈值,能够将显著性检测结果 P 进行二值化,从而得到 ROC 曲线.当采用较小的阈值时,可以视为计算两个概率分布的整体相似度;当取较大的阈值时,进而计算两个分布在峰值处的相似度,通过 ROC 曲线可以计算 AUC 指标,AUC 数值越大,说明算法性能越好.当接近 1 时,代表着显著性估计与真值标定完全一致.根据 ROC 曲线的定义,AUC 指标主要受高阈值的影响.此外,AUC 指标对人眼关注点的中心偏向较为敏感,根据对 FPR 以及 TPR 定义的不同,AUC 指标也产生

了许多变体,典型的包括:Judd 等人[25]提出的 AUC-Judd,真阳性概率是所有真值关注点上预测准确的像素比率,假阳性概率为非关注点上被预测为显著的像素比率;Borji 等人[99]提出了 AUC-Borji 指标,该指标在计算假阳性时,在非关注点上采用了均一的随机采样,而不是直接选取所有的非关注点,但由于采取了随机采样,AUC-Borji 指标容易出现多次对同一个模型评估但结果不一致的现象;shuffled AUC(简称 sAUC)[100]也是一个常用的 AUC 变体,该指标降低了原 AUC 指标对中心偏移的敏感性,sAUC 指标对非显著性点进行采样时,是从其他多张图像上的关注点分布中进行采样,而不是根据在原来图像上的非显著点上进行随机采样,这一采样方法能够导致符合高斯分布的采样,如果在一个模型的检测结果上人为地加入了中心偏向,那么 sAUC 指标在图像中心位置的密集采样会导致这个模型的评估结果下降.

我们对以上人眼关注点评估指标进行了统计和归类,见表 5.根据这些评估指标对视觉显著性做出的不同假设[101],可以将其分为基于位置的评估指标和基于概率分布的评估指标:基于位置的评估指标将显著性视为随机变量,基于概率分布的评估指标将显著性视为概率分布.根据不同评估指标的度量方式,可以分为相似性度量指标和非相似性度量指标:相似性指标越大,表示模型表现越好;非相似性指标越小,表示模型表现越好.根据不同评估指标采用的真值形式,可以将其分为使用连续显著性真值 Q 的评估指标和使用二值离散人眼注意点真值 R 的评估指标.

Table 5 Information of evaluation metrics used in eye fixation prediction

表 5 关于人眼关注点检测评估指标的相关信息

评估指标	基于位置 (location-based)	基于概率分布 (distribution-based)	相似性度量 (similarity↑)	非相似性度量 (dissimilarity↓)	使用的真值
EMD 距离	–	√	–	√	连续显著性 Q
交叉熵	–	√	–	√	连续显著性 Q
标准化扫描路径显著性	√	–	√	–	二值人眼注意点 R
线性相关系数	–	√	√	–	连续显著性 Q
相似性测度	–	√	√	–	连续显著性 Q
AUC-Judd 指标	√	–	√	–	二值人眼注意点 R
AUC-Borji 指标	√	–	√	–	二值人眼注意点 R
sAUC 指标	√	–	√	–	二值人眼注意点 R

4.2 显著物体检测评估指标

在显著物体检测任务中,查准率-查全率曲线(precision-recall curve)、F 值(F-measure)以及平均绝对误差 MAE 值(mean absolute error)是 3 个最常见的评估指标.

(1) 查准率-查全率曲线

给定显著性估计结果,取值范围在[0,255]之间,通过使用从 0 到 255 依次变化的阈值,能够生成一组二值化的显著性结果图(小于阈值的像素标记为 0,大于阈值的像素标记为 1).将每张二值化显著性结果图与显著性真值标定的结果进行比较,可以得到相应的查准率和查全率,如式(8)所示.

$$\begin{cases} Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN} \end{cases} \tag{8}$$

其中,TP 表示二值显著性结果中显著区域与真值显著性标定中一致的像素个数;FP 表示二值显著性结果中被错误划分为显著的像素的个数;FN 表示二值显著性结果中被错误划分为背景的像素的个数.即查准率是指在算法生成的所有前景像素中被正确标定的像素的比率,查全率是指在实际真值标定的前景像素中被算法正确标定的像素的比率.查准率较高,说明有较多的显著区域被正确地检测到,而这往往意味着被误检为显著的像素也较多,从而查全率可能较低;查全率较高,代表检测到的显著区域中检测正确的概率很高,但这也往往意味着有较多的显著区域没有被正确检出,从而精确度可能较低.通过不断变化的阈值,能够得到一组相应的查准率和查全率结果.以查全率为横轴,查准率为纵轴,可以绘得查准率-查全率曲线(precision-recall curve,简称 PR-curve).曲线越靠近右上方,说明算法性能越好.

(2) *F*-值

由于查准率和查全率相互制约,且查准率-查全率曲线包含了两个维度的评估指标,不易比较,因而需要就二者进行综合考量.Achanta 等人[13]提出了 *F*-值指标(*F*-measure).该指标同时考虑了查准率和查全率,能够较为全面、直观地反映出算法的性能.其定义如式(9)所示.

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

其中,β^2=0.3,以此强调查全率更高的重要性.*F*-值指标的数值越大,说明算法性能越好.在实际中,有算法使用 *F*-值曲线,而有的则直接给出 *F*-值曲线上的最大值.

(3) MAE 值

MAE 值(mean absolute error,简称 MAE)是指平均每个像素估计的显著性概率与相应的真值显著性标定之间的绝对误差.由于查准率-查全率曲线和 *F*-值这两个评估指标都只考虑了显著像素的划分结果,而没有考虑对背景划分正确的情况(真阴性),因此,MAE 指标经常作为查准率-查全率曲线和 *F*-值这两个评估指标的补充.其定义如式(10)所示.

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

其中,*S* 表示归一化到[0,1]之间的显著性估计结果,*G* 表示显著性真值标定,*W* 和 *H* 对应图像的宽和高.作为相似性度量指标,MAE 值越小,代表算法性能越好.MAE 指标较为直观,对评估显著性检测模型的实际应用能力(如物体分割)十分重要.

5 视觉显著性检测模型性能评估

本节针对静态及动态场景下的人眼关注点检测模型以及显著物体检测模型的性能进行定量评估.

5.1 人眼关注点检测模型在静态场景下的性能评估

本节针对 14 个经典的静态人眼关注点检测模型(DeepFix[30]、SALICON[31]、DVA[35]、Mr-CNN[32]、SalNet[33]、Deep Gaze I[102]、BMS[103]、eDN[29]、CAS[104]、AIM[105]、Judd Model[25]、GBVS[23]、ITTI[9]、SU[106])的性能进行定量测试,使用了 3 个静态人眼关注点检测数据集,分别为 MIT300[38]、MIT1003[25]和 PASCAL-S[85].实验使用了 AUC-Judd、SIM、s-AUC、CC 和 NSS 这 5 种评估指标,相关定量评估结果分别见表 6~表 8.在 MIT300 数据集上的定量评估结果是根据该数据集的公开结果(http://saliency.mit.edu/)得到的,在 MIT1003 及 PASCAL-S 数据集上的定量评估结果是通过运行这些模型的代码或论文中公布的数据得到的.

Table 6 Quantitative evaluation of different static visual fixation prediction models on MIT300 dataset[38]

表 6 对不同的静态人眼关注点检测模型在 MIT300 数据集[38]上性能的定量评估

模型	AUC-Judd↑	SIM↑	AUC-Borji↑	s-AUC↑	CC↑	NSS↑
人类	0.92	1.00	0.88	0.81	1.00	3.29
DeepFix[30]	0.87	0.67	0.80	0.71	0.78	2.26
SALICON[31]	0.87	0.60	0.85	0.74	0.74	2.12
DVA[35]	0.85	0.58	0.78	0.71	0.68	1.98
Mr-CNN[32]	0.77	0.45	0.76	0.69	0.41	1.13
SalNet[33]	0.83	0.51	0.82	0.65	0.55	1.41
Deep Gazel[102]	0.84	0.39	0.83	0.66	0.48	1.22
BMS[103]	0.83	0.51	0.82	0.65	0.55	1.41
eDN[29]	0.82	0.41	0.81	0.62	0.45	1.14
CAS[104]	0.74	0.43	0.73	0.65	0.36	0.95
AIM[105]	0.77	0.40	0.75	0.66	0.31	0.79
Judd model[25]	0.81	0.42	0.80	0.60	0.47	1.18
GBVS[23]	0.81	0.48	0.80	0.63	0.48	1.24
ITTI[9]	0.75	0.44	0.74	0.63	0.37	0.97

Table 7 Quantitative evaluation of different static visual fixation prediction models on MIT1003 dataset[25]

表 7 对不同的静态人眼关注点检测模型在 MIT1003 数据集[25]上性能的定量评估

模型	AUC-Judd↑	SIM↑	AUC-Borji↑	s-AUC↑	CC↑	NSS↑
DVA[35]	0.87	0.50	0.85	0.77	0.64	2.38
SALICON[31]	0.85	0.42	0.83	0.74	0.53	1.86
Mr-CNN[32]	0.80	0.35	0.77	0.73	0.38	1.36
SU[106]	–	–	–	0.71	–	2.08
DeepFix[30]	0.90*	0.54*	0.87*	0.74*	0.72*	2.58*
BMS[103]	0.79	0.33	0.76	0.69	0.36	1.25
eDN[29]	0.85	0.30	0.84	0.66	0.41	1.29
CAS[104]	0.76	0.32	0.74	0.68	0.31	1.07
AIM[105]	0.79	0.27	0.76	0.68	0.26	0.82
Judd model[25]	0.76	0.29	0.74	0.68	0.30	1.02
GBVS[23]	0.83	0.36	0.81	0.66	0.42	1.38
ITTI[9]	0.77	0.32	0.76	0.66	0.33	1.10

Table 8 Quantitative evaluation of different static visual fixation prediction models on PASCAL-S dataset[85]

表 8 对不同的静态人眼关注点检测模型在 PASCAL-S 数据集[85]上性能的定量评估

模型	AUC-Judd↑	SIM↑	AUC-Borji↑	s-AUC↑	CC↑	NSS↑
DVA[35]	0.89	0.52	0.85	0.77	0.66	2.26
SALICON[31]	0.86	0.46	0.83	0.72	0.58	1.88
Mr-CNN[32]	0.79	0.34	–	0.71	0.40	1.35
SU[106]	–	–	–	0.73	–	2.22
BMS[103]	0.79	0.34	0.77	0.67	0.39	1.28
eDN[29]	–	–	–	1.29	–	1.42
CAS[104]	0.78	0.34	0.75	0.67	0.36	1.12
AIM[105]	0.77	0.30	0.75	0.65	0.32	0.97
GBVS[23]	0.84	0.36	0.82	0.65	0.45	1.36
ITTI[9]	0.82	0.36	0.80	0.64	0.42	1.30

5.2 人眼关注点检测模型在动态场景下的性能评估

本节针对 16 个经典的人眼关注点检测模型在动态场景下的性能进行定量测试,其中包括 6 个静态人眼关注点检测模型(ITTI[9]、GBVS[23]、SALICON[31]、Shallow-Net[33]、Deep-Net[33]、DVA[35])以及 10 个动态人眼关注点检测模型(PQFT[107]、SEO[41]、RUDOY[43]、HOU[44]、FANG[108]、OBDL[45]、AWS-D[46]、OM-CMM[51]、Two-stream[50]和 ACLNet[52]),使用了 3 个动态人眼关注点检测数据集,分别为 DHF1K[52]、Hollywood-2[88]和 UCF-sports[88].实验使用了 AUC-Judd、SIM、s-AUC、CC 和 NSS 这 5 种评估指标,相关定量评估结果分别见表 9~表 11.评估结果主要根据 DHF1K 数据集的公开结果(https://github.com/wenguanwang/DHF1K)得到.

Table 9 Evaluation of visual fixation prediction models in dynamic scenes using DHF1K dataset[52]

表 9 DHF1K 数据集[52]上,对不同的人眼关注点检测模型在动态场景下的性能评估

模型		AUC-Judd↑	SIM↑	s-AUC↑	CC↑	NSS↑
中心先验		0.854	0.238	0.503	0.302	0.167
静态人眼关注点检测模型	ITTI[9]	0.774	0.162	0.553	0.233	1.207
	GBVS[23]	0.828	0.186	0.554	0.283	1.474
	SALICON[31]	0.857	0.232	0.590	0.327	1.901
	Shallow-Net[33]	0.833	0.182	0.529	0.295	1.509
	Deep-Net[33]	0.855	0.201	0.592	0.331	1.775
	DVA[35]	0.860	0.262	0.595	0.358	2.013
动态人眼关注点检测模型	PQFT[107]	0.699	0.139	0.562	0.137	0.749
	SEO[41]	0.635	0.142	0.499	0.070	0.334
	RUDOY[43]	0.769	0.214	0.501	0.285	1.498
	HOU[44]	0.726	0.167	0.545	0.150	0.847
	FANG[108]	0.819	0.198	0.537	0.273	1.539
	OBDL[45]	0.638	0.171	0.500	0.117	0.495
	AWS-D[46]	0.703	0.157	0.513	0.174	0.940
	OM-CMM[51]	0.856	0.256	0.583	0.344	1.911
	Two-stream[50]	0.834	0.197	0.581	0.325	1.632
	ACLNet[52]	0.890	0.315	0.601	0.434	2.354

Table 10 Evaluation of visual fixation prediction models in dynamic scenes using Hollywood-2 dataset[88]

表 10 Hollywood-2 数据集[88]上,对不同的人眼关注点检测模型在动态场景下的性能评估

模型		AUC-Judd↑	SIM↑	s-AUC↑	CC↑	NSS↑
中心先验		0.869	0.331	0.615	0.421	1.808
静态人眼关注点检测模型	ITTI[9]	0.788	0.221	0.607	0.257	1.076
	GBVS[23]	0.837	0.257	0.633	0.308	1.336
	SALICON[31]	0.856	0.321	0.711	0.425	2.013
	Shallow-Net[33]	0.851	0.276	0.694	0.423	1.680
	Deep-Net[33]	0.884	0.300	0.736	0.451	2.066
	DVA[35]	0.886	0.372	0.727	0.482	2.459
动态人眼关注点检测模型	PQFT[107]	0.723	0.201	0.621	0.153	0.755
	SEO[41]	0.652	0.155	0.530	0.076	0.346
	RUDOY[43]	0.783	0.315	0.536	0.302	1.570
	HOU[44]	0.731	0.202	0.580	0.146	0.684
	FANG[108]	0.859	0.272	0.659	0.358	1.667
	OBDL[45]	0.640	0.170	0.541	0.106	0.462
	AWS-D[46]	0.694	0.175	0.637	0.146	0.742
	OM-CMM[51]	0.887	0.356	0.693	0.446	2.313
	Two-stream[50]	0.863	0.276	0.710	0.382	1.748
	ACLNet[52]	0.913	0.542	0.757	0.623	3.086

Table 11 Evaluation of visual fixation prediction models in dynamic scenes using UCF-sports dataset[88]

表 11 UCF-sports 数据集[88]上,对不同的人眼关注点检测模型在动态场景下的性能评估

模型		AUC-Judd↑	SIM↑	s-AUC↑	CC↑	NSS↑
中心先验		0.834	0.299	0.566	0.350	1.585
静态人眼关注点检测模型	ITTI[9]	0.847	0.251	0.725	0.356	1.640
	GBVS[23]	0.859	0.274	0.697	0.396	1.818
	SALICON[31]	0.848	0.304	0.738	0.375	1.838
	Shallow-Net[33]	0.846	0.276	0.691	0.382	1.789
	Deep-Net[33]	0.861	0.282	0.719	0.414	1.903
	DVA[35]	0.872	0.339	0.725	0.439	2.311
动态人眼关注点检测模型	PQFT[107]	0.825	0.250	0.722	0.338	1.780
	SEO[41]	0.831	0.308	0.666	0.336	1.690
	RUDOY[43]	0.763	0.271	0.637	0.344	1.619
	HOU[44]	0.819	0.276	0.674	0.292	1.399
	FANG[108]	0.845	0.307	0.674	0.395	1.787
	OBDL[45]	0.759	0.193	0.634	0.234	1.382
	AWS-D[46]	0.823	0.228	0.750	0.306	1.631
	OM-CMM[51]	0.870	0.321	0.691	0.405	2.089
	Two-stream[50]	0.832	0.264	0.685	0.343	1.753
	ACLNet[52]	0.905	0.496	0.767	0.603	3.200

5.3 显著物体检测模型在静态场景下的性能评估

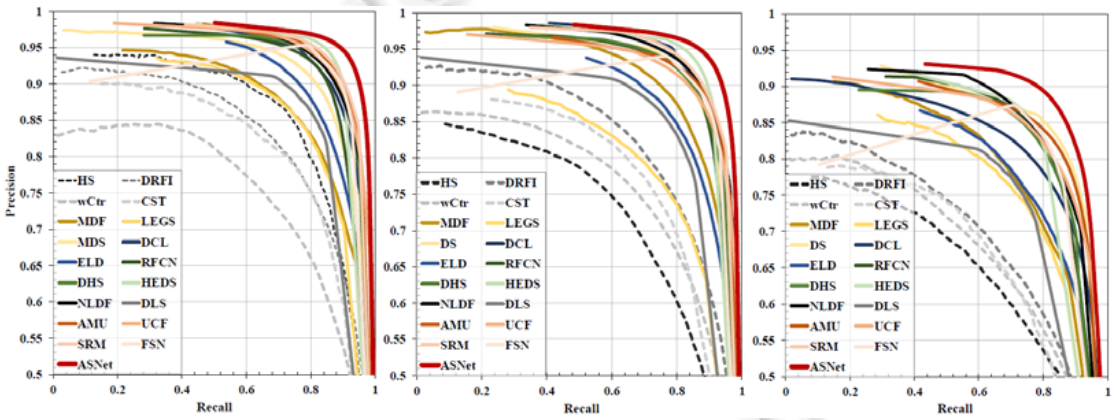
本节针对 20 个经典的静态人眼关注点检测模型,包括 4 个传统的、非深度学习的模型(HS[55]、DRFI[56]、wCtr[57]、CST[109])以及 16 个基于深度学习的模型(MDF[65]、LEG[64]、MDS[110]、DCL[111]、ELD[66]、SU[106]、RFCN[67]、DHS[72]、HEDS[69]、NLDF[112]、DLS[71]、AMU[68]、UCF[113]、SRM[114]、FSN[115]、ASNet[70]),在 ECCSD[55]、HKU-IS[65]和 PASCAL-S[85]这 3 个数据集上的性能进行了定量测试.表 12 总结了使用 *F*-score 和 MAE 作为评估指标的定量结果.

图 1 中以查准率-查全率曲线作为评估指标的定量结果.这些结果是通过运行以上模型的代码或论文中公布的数据得到的,图 1 中没有提供 SU 模型的结果,因为该模型的实现源码和相关查准率-查全率结果都未给出.

Table 12 Quantitative evaluation of different static salient object detection models on ECCSD[55], HKU-IS[65], and PASCAL-S[85] datasets using *F*-score and MAE

表 12 对不同的静态显著物体检测模型在 ECCSD[55]、HKU-IS[65]和 PASCAL-S[85]数据集上性能的定量评估,使用 *F*-score 和 MAE 作为评估指标

模型		ECCSD[55]		HKU-IS[65]		PASCAL-S[85]	
		F-score↑	MAE↓	*F*-score↑	MAE↓	*F*-score↑	MAE↓
非深度学习模型	HS[55]	0.730	0.223	0.710	0.215	0.636	0.259
	DRFI[56]	0.787	0.166	0.783	0.143	0.692	0.196
	wCtr[57]	0.672	0.178	0.694	0.138	0.611	0.193
	CST[109]	0.742	0.147	0.732	0.128	0.598	0.191
深度学习模型	MDF[65]	0.831	0.108	0.860*	0.129*	0.764	0.145
	LEG[64]	0.831	0.119	0.812	0.101	0.749	0.155
	MDS[110]	0.810	0.160	0.848	0.078	0.818	0.170
	DCL[111]	0.898	0.071	0.907	0.048	0.822	0.108
	ELD[66]	0.865	0.080	0.844	0.071	0.767	0.121
	SU[106]	0.88	0.06	–	–	0.77	0.10
	RFCN[67]	0.898	0.097	0.895	0.079	0.827	0.118
	DHS[72]	0.905	0.061	0.892	0.052	0.820	0.091
	HEDS[69]	0.915	0.052	0.913	0.039	0.830	0.080
	NLDF[112]	0.905	0.063	0.902	0.048	0.831	0.099
	DLS[71]	0.825	0.090	0.806	0.072	0.719	0.136
	AMU[68]	0.889	0.058	0.918	0.052	0.834	0.098
	UCF[113]	0.868	0.068	0.905	0.062	0.771	0.116
	SRM[114]	0.910	0.056	0.892	0.046	0.783	0.127
	FSN[115]	0.910	0.053	0.895	0.044	0.827	0.095
	ASNet[70]	0.928	0.043	0.920	0.035	0.857	0.072



(a) ECCSD (b) HKU-IS (c) PASCAL-S

Fig.1 Quantitative evaluation of different static salient object detection models on ECCSD[55], HKU-IS[65], and PASCAL-S[85] datasets using precision-recall curve

图 1 对不同的静态显著物体检测模型在 ECCSD[55],HKU-IS[65]和 PASCAL-S[85]数据集上性能的定量评估,使用查准率-查全率曲线作为评估指标

5.4 显著物体检测模型在动态场景下的性能评估

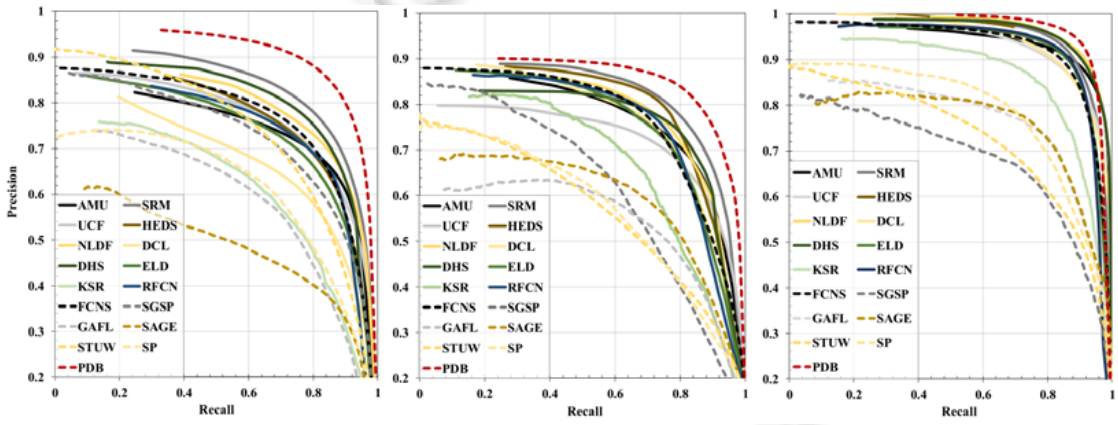
本节针对 17 个经典的显著物体检测模型在动态场景下的性能进行定量测试,其中包括 10 个静态显著物体检测模型(AMU[68]、SRM[114]、UCF[113]、HEDS[69]、NLDF[112]、DCL[111]、DHS[72]、ELD[66]、KSR[116]和 RFCN[67])以及 7 个动态显著物体检测模型(FCNS[84]、SGSP[78]、GAFL[74]、SAGE[75]、STUW[42]、SP[73]和 PDB[117]),使用了 3 个动态视频显著物体检测数据集,分别为 DAVIS[98]、FBMS[97]和 ViSal[74].SGSP、GAFL、SAGE、STUW 和 SP 为非深度学习模型,其余算法均为深度学习模型.表 13 总结了使用 *F*-score 和 MAE 作为评估指标的定量

结果,图 2 中为查准率-查全率曲线作为评估指标的定量结果.这些结果是通过运行以上模型的代码或论文中公布的数据得到的.

Table 13 Evaluation of salient object detection models in dynamic scenes using DAVIS[98], FBMS[97], and ViSal[74] datasets, measured by *F*-score and MAE

表 13 在 DAVIS[98]、FBMS[97]以及 ViSal[74]数据集上,对不同的显著物体检测模型在动态场景下的性能评估,使用 *F*-score 和 MAE 作为评估指标

	模型	DAVIS[98]		FBMS[97]		ViSal[74]	
		F-score↑	MAE↓	*F*-score↑	MAE↓	*F*-score↑	MAE↓
图像显著物体检测模型	AMU[68]	0.699	0.082	0.754	0.115	0.894	0.032
	SRM[114]	0.779	0.039	0.784	0.082	0.890	0.029
	UCF[113]	0.716	0.107	0.686	0.164	0.870	0.068
	HEDS[69]	0.717	0.062	0.776	0.084	0.906	0.028
	NLDF[112]	0.723	0.056	0.749	0.092	0.916	0.022
	DCL[111]	0.631	0.070	0.718	0.098	0.869	0.035
	DHS[72]	0.758	0.039	0.735	0.097	0.911	0.025
	ELD[66]	0.688	0.070	0.739	0.108	0.890	0.038
	KSR[116]	0.601	0.077	0.700	0.101	0.826	0.063
	RFCN[67]	0.710	0.065	0.746	0.105	0.888	0.043
视频显著物体检测模型	FCNS[84]	0.729	0.053	0.753	0.098	0.876	0.041
	SGSP[78]	0.677	0.128	0.589	0.171	0.648	0.172
	GAFL[74]	0.578	0.092	0.563	0.150	0.726	0.099
	SAGE[75]	0.479	0.105	0.597	0.142	0.734	0.096
	STUW[42]	0.691	0.098	0.550	0.143	0.671	0.132
	SP[73]	0.601	0.130	0.566	0.161	0.731	0.126
	PDB[117]	0.849	0.030	0.815	0.069	0.917	0.022



(a) DAVIS (b) FBMS (c) ViSal

Fig.2 Evaluation of salient object detection models in dynamic scenes using DAVIS[98], FBMS[97], and ViSal[74] datasets, measured by precision-recall curve

图 2 在 DAVIS[98]、FBMS[97]以及 ViSal[74]数据集上,对不同的显著物体检测模型在动态场景下的性能评估,使用查准率-查全率曲线作为评估指标

6 总结与展望

随着深度学习技术在计算机视觉领域取得广泛的成功,深度神经网络成为当前视觉注意力机制计算和建模的首选工具,基于深度学习的视觉显著性模型在人眼关注点检测和显著物体检测领域都取得了极佳的效果.我们认为,视觉注意力检测领域未来可能的研究工作主要包括以下几个方面.

(1) 在人眼关注点检测方向,进一步将经典的认知理论与深度学习技术相融合.

传统的认知领域通过对人类和其他灵长类动的观测和研究,积累了很多经典的关于视觉注意力机制的理

论和模型,这些理论更符合生物学原理,如 1985 年 Koch 等人[8]提出的 WTA 理论、1991 年 Leventhal 提出的中央周边差[118]、Treisman 的特征整合(FIT)理论[10]、Wolfe 等人提出的指向搜索模型[11]等.但是,现在计算机视觉领域中,对视觉注意力机制的计算建模主要基于深度学习技术,很少与之前经典的认知理论相结合,虽然基于深度学习技术的计算模型具有较好的性能,但是对研究界理解视觉注意力机制背后更深层次的机理,难以提供更多更有价值的实验支持.因而,有必要将基于深度学习技术的计算模型与经典认知理论相结合,进一步发展新理论和新模型.此外,经典的视觉注意力机制理论指出,人眼注意力的分配是由自底向上与自顶向下两个过程协同完成的,但是当前的基于深度学习技术的显著性检测模型主要通过融合不同网络层抽取的不同层次的特征来得到显著性检测结果,而缺乏有效地、显式地融合自底向上和自顶向下信息的过程,这与之前针对视觉注意力机制的研究成果不符,因此,有必要进一步发掘现有的深度学习技术,将自底向上和自顶向下的显著性检测过程融合到深度神经网络结构中,并能够通过端到端的方式进行学习.

(2) 从注意力机制角度,研究深度神经网络的可解释性.

目前,深度神经网络受到注意力机制的启发,通过特殊的网络结构,能够"迫使"神经网络关注文本或图像中与任务最相关的部分.这种神经网络的注意力模块,可以被视为一种自顶向下、任务相关的注意力机制.这类带有注意力机制的深度神经网络在许多任务上表现出了较好的性能,但是这种通过隐式学习的、与任务相关的注意力是否真的与人类的注意力相一致?这一问题对通过注意力机制来研究深度神经网络的事后解释性(post-hoc explanation)非常重要,但却很少有工作关注这一点.我们有必要在现有的公共数据集上收集人类在执行有关任务时的眼动数据,据此和深度神经网络的注意力机制进行比较;同时,利用人类实际的注意力机制来显式引导神经网络,即观察当神经网络利用有监督的注意力机制时其性能的变化,从而能够从注意力机制的角度,对深度神经网络的可解释性进行更深入的研究.

(3) 通过借鉴认知科学的理论研究,进一步拓宽计算机视觉领域对视觉注意力研究的内涵和外延.

认知科学中,对人类的注意力机制进行了更深入、更广泛的研究,如:群体的注意力机制(group attention)、人类在社交场景中的注意力机制(co-attention in social scenes).因此,计算机视觉领域针对视觉注意力计算模型的研究,有必要充分吸收借鉴认知科学领域中对于人类注意力机制的理论成果,进一步研究挖掘人类视觉注意力机制以及更高层次的感知理解,如:研究第一人称视角下的人类注意力机制、研究人类在社交场景下的认知机制、研究人类的多轮注意力分配和转移机制,并基于人类的行为、动作、注意力进一步研究人类的行为意图(intention).

References:

[1] Koch K, McLean J, Segev R, Freed MA, Berry II MJ, Balasubramanian V, Sterling P. How much the eye tells the brain. Current Biology, 2006,16(14):1428−1434.

[2] Borji A, Itti L. State-of-the-art in visual attention modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(1):185−207.

[3] Carrasco M. Visual attention: The past 25 years. Vision Research, 2011,51(13):1484−1525.

[4] Connor C, Egeth H, Yantis S. Visual attention: Bottom-up versus top-down. Current Biology, 2004,14(19):850−852.

[5] Rutishauser U, Walther D, Koch C, Perona P. Is bottom-up attention useful for object recognition? In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2004.

[6] Borji A, Itti L. Scene classification with a sparse set of salient regions. In: Proc. of the IEEE Int'l Conf. on Robotics and Automation. 2011. 1902−1908.

[7] Zhang DW, Han JW, Jiang L, Ye SM, Chang XJ. Revealing event saliency in unconstrained video collection. IEEE Trans. on Image Processing, 2017,26(4):1746−1758.

[8] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, 1985,4(4): 219−227.

[9] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(11):1254−1259.

[10] Treisman AM, Gelade G. A feature-integration theory of attention. Cognitive Psychology, 1980,12(1):97−136.

[11] Wolfe JM, Cave KR, Franzel SL. Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance, 1989,15(3):419−433.

[12] Liu T, Sun J, Zheng NN, Tang XO, Shum HY. Learning to detect a salient object. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2007. 1−8.

[13] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 1597−1604.

[14] Ren ZX, Gao SH, Chia LT, Tsang IWH. Region-based saliency detection and its application in object recognition. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(5):769−779.

[15] Alexe B, Deselaers T, Ferrari V. Measuring theobjectness ofimage windows. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012,34(11):2189−2202.

[16] Lee YJ, Ghosh J, Grauman K. Discovering important people and objects for egocentric video summarization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 1346−1353.

[17] Wang WG, Shen JB. Deep cropping via attention box prediction and aesthetics assessment. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2186−2194.

[18] Wang WG, Shen JB, Yu Y, Ma KL. Stereoscopic thumbnail creation via efficient stereo saliency detection. IEEE Trans. on Visualization and Computer Graphics, 2017,23(8):2014−2027.

[19] Frintrop S, Kessel M. Most salient region tracking. In: Proc. of the IEEE Conf. on Robotics and Automation. 2009. 1869−1874.

[20] Zhang LY, Tong MH, Marks TK, Shan HH, Cottrell GW. SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision, 2008,8(7):Article No.32.

[21] Gao DS, Vasconcelos N. Discriminant saliency for visual recognition from cluttered scenes. In: Proc. of the Advances in Neural Information Processing Systems. 2005. 481−488.

[22] Bruce N, Tsotsos J. Saliency based on information maximization. In: Proc. of the Advances in Neural Information Processing Systems. 2006. 155−162.

[23] Harel J, Koch C, Perona P. Graph-based visual saliency. In: Proc. of the Advances in Neural Information Processing Systems. 2007. 545−552.

[24] Hou XD, Zhang LQ. Saliency detection: A spectral residual approach. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2007. 1−8.

[25] Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2009. 2106−2113.

[26] Gao DS, Han S, Vasconcelos N. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(6):989−1005.

[27] Kanan C, Tong MH, Zhang LY, Cottrell GW. SUN: Top-down saliency using natural statistics. Visual Cognition, 2009,17(6-7): 979−1003.

[28] Borji A, Sihite DN, Itti L. Probabilistic learning of task-specific visual attention. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 470−477.

[29] Vig E, Dorr M, Cox D. Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 2798−2805.

[30] Kruthiventi SS, Ayush K, Babu RV. Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Trans. on Image Processing, 2017,26(9):4446−4456.

[31] Huang X, Shen CY, Boix X, Zhao Q. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 262−270.

[32] Liu N, Han JW, Liu TM, Li XL. Learning to predict eye fixations via multiresolution convolutional neural networks. IEEE Trans. on Neural Networks and Learning Systems, 2016,29(2):392−404.

[33] Pan JT, Sayrol E, GiroiNieto X, McGuinness K, O'Connor NE. Shallow and deep convolutional networks for saliency prediction. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 598−606.

[34] Cornia M, Baraldi L, Serra G, Cucchiara R. Predicting human eye fixations via an LSTM-based saliency attentive model. arXiv preprint arXiv:1611.09571, 2016.

[35] Wang WG, Shen JB. Deep visual attention prediction. IEEE Trans. on Image Processing, 2018,27(5):2368−2378.

[36] Jetley S, Murray N, Vig E. End-to-end saliency mapping via probability distribution prediction. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5753−5761.

[37] Jiang M, Huang SS, Duan JY, Zhao Q. SALICON: Saliency in context. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1072−1080.

[38] Judd T, Durand F, Torralba A. A benchmark of computational models of saliency to predict human fixations. Technical Report, MIT, 2012.

[39] Gao DS, Vijay M, Nuno V. The discriminant center-surround hypothesis for bottom-up saliency. In: Proc. of the Advances in Neural Information Processing Systems. 2008. 497−504.

[40] Mahadevan V, Vasconcelos N. Spatiotemporal saliency in dynamic scenes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(1):171−177.

[41] Seo HJ, Milanfar P. Static and space-time visual saliency detection by self-resemblance. Journal of Vision, 2009,9(12):Article No.15.

[42] Fang YM, Wang Z, Lin WS, Fang ZJ. Video saliency incorporating spatiotemporal cues and uncertainty weighting. IEEE Trans. on Image Processing, 2014,23(9):3910−3921.

[43] Rudoy D, Goldman DB, Shechtman E, Zelnik-Manor L. Learning video saliency from human Gaze using candidate selection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 1147−1154.

[44] Hou XD, Zhang LQ. Dynamic visual attention: Searching for coding length increments. In: Proc. of the Advances in Neural Information Processing Systems. 2008. 681−688.

[45] Hossein KS, Vasconcelos N, Bajic IV, Shan Y. How many bits does it take for a stimulus to be salient? In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 5501−5510.

[46] Leboran V, Garcia-Diaz A, Fdez-Vidal XR, Pardo XM. Dynamic whitening saliency. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,39(5):893−907.

[47] Gao DS, Vasconcelos N. Bottom-up saliency is a discriminant process. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2007. 1−6.

[48] Guo CL, Ma Q, Zhang LM. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2008. 1−8.

[49] Rahtu E, Kannala J, Salo M, Heikkilä J. Segmenting salient objects from images and videos. In: Proc. of the European Conf. on Computer Vision. 2010. 336−379.

[50] Bak C, Kocak A, Erdem E, Erdem A. Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Trans. on Multimedia, 2018,20(7):1688−1698.

[51] Jiang L, Xu M, Wang ZL. Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM. arXiv preprint arXiv:1709.06316, 2017.

[52] Wang WG, Shen JB, Guo F, Cheng MM, Borji A. Revisiting video saliency: A large-scale benchmark and a new model. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4894−4903.

[53] Cheng MM, Mitra NJ, Huang XL, Torr PH, Hu SM. Global contrast based salient region detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(3):569−582.

[54] Wei YC, Wen F, Zhu WJ, Sun J. Geodesic saliency using background priors. In: Proc. of the European Conf. on Computer Vision. 2012. 29−42.

[55] Yan Q, Xu L, Shi JP, Jia JY. Hierarchical saliency detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 1155−1162.

[56] Jiang HX, Wang JD, Yuan ZJ, Wu Y, Zheng NN, Li SP. Salient object detection: A discriminative regional feature integration approach. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 2083−2090.

[57] Zhu WJ, Liang S, Wei YC, Sun J. Saliency optimization from robust background detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 2814−2821.

[58] Yang C, Zhang LH, Lu HC, Ruan X, Yang MH. Saliency detection via graph-based manifold ranking. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 3166−3173.

[59] Jiang BW, Zhang LH, Lu HC, Yang C, Yang MH. Saliency detection via absorbing markov chain. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1165−1672.

[60] Qin Y, Lu HC, Xu YQ, Wang H. Saliency detection via cellular automata. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 110−119.

[61] Zhang JM, Sclaroff S, Lin Z, Shen XH, Price B, Mech R. Minimum barrier salient object detection at 80fps. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1404−1412.

[62] Tu WC, He SF, Yang QX, Chien SY. Real-time salient object detection with a minimum spanning tree. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2334−2342.

[63] Zhao R, Ouyang WL, Li HS, Wang XG. Saliency detection by multi-context deep learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1265−1274.

[64] Wang LJ, Lu HC, Ruan X, Yang MH. Deep networks for saliency detection via local estimation and global search. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3183−3192.

[65] Li GB, Yu YZ. Visual saliency based on multiscale deep features. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 5455−5463.

[66] Lee G, Tai YW, Kim J. Deep saliency with encoded low level distance map and high level features. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 660−668.

[67] Wang LZ, Wang LJ, Lu HC, Zhang PP, Ruan X. Saliency detection with recurrent fully convolutional networks. In: Proc. of the European Conf. on Computer Vision. 2016. 825−841.

[68] Zhang PP, Wang D, Lu HC, Wang HY, Ruan X. Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 202−211.

[69] Hou QB, Cheng MM, Hu XW, Borji A, Tu ZW, Torr P. Deeply supervised salient object detection with short connections. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3203−3212.

[70] Wang WG, Shen JB, Dong XP, Borji A. Salient object driven by fixation prediction. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1711−1720.

[71] Hu P, Shuai B, Liu J, Wang G. Deep level sets for salient object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2300−2309.

[72] Liu N, Han JW. DHSNet: Deep hierarchical saliency network for salient object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 678−686.

[73] Liu Z, Zhang X, Luo SH, Meur OL. Superpixel-based spatiotemporal saliency detection. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(9):1522−1540.

[74] Wang WG, Shen JB, Shao L. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans. on Image Processing, 2015,24(11):4185−4196.

[75] Wang WG, Shen JB, Porikli F. Saliency-aware geodesic video object segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3395−3402.

[76] Wang WG, Shen JB, Yang RG, Porikli F. Saliency-aware video object segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,40(1):20−33.

[77] Kim H, Kim Y, Sim JY, Kim CS. Spatiotemporal saliency detection for video sequences based on random walk with restart. IEEE Trans. on Image Processing, 2015,24(8):2552−2564.

[78] Liu Z, Li JH, Ye LW, Sun GL, Shen LQ. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Trans. on Circuits and Systems for Video Technology, 2017,27(12):2527−2542.

[79] Guo F, Wang WG, Shen JB, Shao L, Yang J, Tao DC, Tang YY. Video saliency detection using object proposals. IEEE Trans. on Cybernetcis, 2018,48(11):3159−3170.

[80] Chen CLZ, Li S, Wang YG, Qin H, Hao AM. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans. on Image Processing, 2017,26(7):3156−3170.

[81] Chen YH, Zou WB, Tang Yi, Li X, Xu C, Komodakis N. SCOM: Spatiotemporal constrained optimization for salient object detection. IEEE Trans. on Image Processing, 2018,27(7):3345−3357.

[82] Li J, Xia CQ, Chen XW. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. IEEE Trans. on Image Processing, 2018,27(1):349−364.

[83] Alshawi T, Long ZL, Alregib G. Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection. IEEE Trans. on Image Processing, 2018,27(6):2818−2827.

[84] Wang WG, Shen JB, Shao L. Video salient object detection via fully convolutional networks. IEEE Trans. on Image Processing, 2018,27(1):38−49.

[85] Li Y, Hou XD, Koch C, Rehg M, Yuille AL. The secrets of salient object segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 280−287.

[86] Everingham M, Gool LV, Williams CK, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010,88(2):303−338.

[87] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollar P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the European Conf. on Computer Vision. 2014. 740−755.

[88] Mathe S, Sminchisescu C. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(7):1408−1424.

[89] Mital PK, Smith TJ, Hill RL, Henderson JM. Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive Computation, 2011,3(1):5−24.

[90] Marszalek M, Laptev I, Schmid C. Actions in context. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 2929−2936.

[91] Rodriguez MD, Ahmed J, Shah M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2008. 1−8.

[92] Itti L. Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans. on Image Processing, 2004,13(10):1304−1318.

[93] Hadizadeh H, Enriquez MJ, Bajic IV. Eye-tracking database for a set of standard video sequences. IEEE Trans. on Image Processing, 2012,21(2):898−903.

[94] Tsai D, Flagg M, Rehg JM. Motion coherent tracking with multi-label MRF optimization. In: Proc. of the British Machine Vision Conf. 2010. 1−11.

[95] Li FX, Kim T, Humayun A, Tsai D, Rehg JM. Video segmentation by tracking many figure-ground segments. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 2192−2199.

[96] Brox T, Malik J. Object segmentation by long term analysis of point trajectories. In: Proc. of the European Conf. on Computer Vision. 2010. 282−295.

[97] Ochs P, Malik J, Brox T. Segmentation of moving objects by long term video analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014,36(6):1187−1200.

[98] Perazzi F, Pont-Tuset J, McWilliams B, Gool LV, Gross M, Sorkine-Hornung A. A benchmark dataset and evaluation methodology for video object segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 724−732.

[99] Borji A, Sihite DN, Itti L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Trans. on Image Processing, 2012,22(1):55−69.

[100] Borji A, Tavakoli HR, Sihite DN, Itti L. Analysis of scores, datasets, and models in visual saliency prediction. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 921−928.

[101] Riche N, Duvinage M, Mancas M, Gosselin B, Dutoit T. Saliency and human fixations: State-of-the-art and study of comparison metrics. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1153−1160.

[102] Kummerer M, Theis L, Bethge M. Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet. In: Proc. of the Int'l Conf. on Learning Representations Workshop. 2015. 1−12.

[103] Zhang JM, Sclaroff S. Saliency detection: A Boolean map approach. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 153−160.

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

[105] Bruce ND, Tsotsos JK. Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, 2009,9(3): Article No.5.

[106] Kruthiventi SS, Gudisa V, Dholakiya JH, Venkatesh Babu R. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5781−5790.

[107] Guo CL, Zhang LM. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. on Image Processing, 2010,19(1):185−198.

[108] Fang YM, Lin WS, Chen ZZ, Tsai CM, Lin CW. A video saliency detection model in compressed domain. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(1):27−38.

[109] Wang WG, Shen JB, Shao L, Porikli F. Correspondence driven saliency transfer. IEEE Trans. on Image Processing, 2016,25(11): 5025−5034.

[110] Li X, Zhao LM, Wei LN, Yang MH, Wu F, Zhuang YT, Ling H, Wang JD. Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE Trans. on Image Processing, 2016,25(8):3919−3930.

[111] Li GB, Yu YZ. Deep contrast learning for salient object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 478−487.

[112] Luo ZM, Mishra A, Achkar A, Eichel J, Li SZ, Jodoin PM. Non-local deep features for salient object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6593−6601.

[113] Zhang PP, Wang D, Lu HC, Wang HY, Yin BC. Learning uncertain convolutional features for accurate saliency detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 212−221.

[114] Wang TT, Borji A, Zhang LH, Zhang PP, Lu HC. A stagewise refinement model for detecting salient objects in images. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 4039−4048.

[115] Chen XW, Zheng AL, Li J, Lu F. Look, perceive and segment: Finding the salient objects in images via two-stream fixationsemantic CNNs. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1050−1058.

[116] Wang TT, Zhang LH, Lu HC, Sun C, Qi JQ. Kernelized subspace ranking for saliency detection. In: Proc. of the European Conf. on Computer Vision. 2016. 450−466.

[117] Song HM, Wang WG, Shen JB, Zhao SY, Lam KM. Pyramid dilated deeper convLSTM for video salient object detection. In: Proc. of the European Conf. on Computer Vision. 2018. 715−731.

[118] Leventhal AG. Vision and Visual Dysfunction: The Neural Basis of Visual Function. 1991.

王文冠(1990－),男,河北磁县人,博士,CCF 学生会员,主要研究领域为计算机视觉,人工智能.

贾云得(1962－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为计算机视觉,人工智能.

沈建冰(1978－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为计算机视觉,人工智能.