人类的视觉系统具有极强的数据处理能力，能够在复杂场景中快速地选择比较醒目的区域或者是感兴趣的区域，并且只对选择出的区域进行处理而忽略其他区域中的信息。这种视觉信息处理机制被人们称之为“选择性视觉注意机制”或者“视觉注意力机制”，又常常简称“注意机制”。由于选择性视觉注意机制能够快速地锁定视觉场景中的感兴趣区域或者目标区域，从而极大地减少数据的处理量、加快信息处理的速度，这对于计算资源有限以及实时性要求较高的各种机器视觉应用来说具有非常大的吸引力，因而受到学术界的广泛关注。将人类这种选择性视觉注意机制引入到计算机视觉的信息处理中，近些年来已经成为计算机视觉领域的研究热点，并在各种视觉计算任务中得到了快速的发展，如目标检测、目标跟踪和图像理解等。

人工神经网络被认为是一种以简化的方式模仿人类大脑并行计算机制的数学模型，人们同样试图建立一些计算模型来模仿人类视觉系统注意机制，以实现选择性地专注于一些与视觉任务目的相关的事物而忽略其他事物，从而可以利用有限的计算资源来快速完成视觉场景的处理和理解等过程。

早在 1980 年，Treisman 等就从认知心理学与神经科学方面对人类注意力机制进行了研究，提出了显著特征整合理论。Koch 等于 1985 年发表了被认为是视觉注意力领域的奠基文章，并于 1987 年发表了该文的扩展版。到 1998 年，Itti 等发表了第一个基于显著度的视觉注意计算模型，使用多种特征的融合获得最终的显著图，视觉注意力机制模型正式引入到计算机视觉领域。之后，视觉注意模型逐渐受到研究者的关注，并运用于视觉任务计算的方方面面。在今年前不久召开的 IEEE 国际计算机视觉与模式识别会议( CVPＲ) 和全国图象图形学学术会议( NCIG) 中，都各有超过 15 篇论文/报告的主题或内容涉及注意力机制或显著性检测。

对视觉注意力的研究主要有两个分支———关注点检测( Fixation Prediction) 和显著物体检测( Salient Ob- ject Detection) 。关注点检测可以应用在广告和人类行为研究等领域，显著物体检测可广泛应用于各种视觉任务的预处理过程。用于这两个分支研究的数据集都需要人为地标注真值。关注点的真值需要利用专门设备———眼动仪对实验者的注视落点( 如图 1 ( b) 中的白色点) 进行标注，标注通常侧重于物体的某块区域或某些像素值; 而显著物体的真值则是根据人眼主观感知通过手工标注，一般标注整个显著物体的像素区域，并表示为二值图片( 如图 2( b)所示) ，通常值为 1 的白色部分表示显著物体区域，值为 0 的黑色部分表示背景。人眼关注点检测示意图如图 1 所示，显著性物体检测示意图如图 2 所示。在本综述中，主要关注的是显著物体检测。



图 1 人眼关注点检测示意图



图 2 显著性物体检测示意图

本文首先简单叙述视觉注意力研究的发展历程，然后主要介绍显著性物体检测的各种方法，包括传统的方法和基于深度学习的方法，并根据不同的算法特点和网络结构对这两大类的方法进行了进一步的分类和小结，同时从不同角度对现有的显著性物体检测的数据集进行了整理和列表、对评测算法效果的方法和指标作了详细的介绍，另外还探讨了显著性物体检测在不同领域的应用，最后对显著性物体检测研究的发展趋势和方向进行了分析和总结展望。

一、显著性物体检测方法

显著性物体检测可以分为静态物体检测和动态物体检测，静态物体检测主要是针对一幅图像中的显著物体进行检测，而动态显著物体检测主要是对视频场景中的显著物体进行检测。静态图像和动态视频中物体显著性的区别有: (1)静态图像中的显著性物体主要表现在显著目标与环境中其他目标在特征上和语义上的对比，而视频中运动的物体一般被定义为显著物体，并且人眼会随着视频中物体的运动关注和跟踪该物体，或者随着物体的消失和出现而不断改变显著物体; (2)由于视频播放的过程是一个连续动态的过程，因此在观看视频时人眼的注意力在每一帧的停留时间很短，新的物体出现更能吸引人的注意，另外视频显著性目标的能量相对较为集中，当没有显著物体时，人的关注点往往会集中在视频画面的中央。一般来说，对于视频中的运动物体，其显著性检测依赖于每一帧图像 ( 如图 3 所示) ，当存在物体消失和出现时，其显著性检测也依赖于前后帧图像的对比。因此，图像的显著性物体检测是基础，所以本文对显著性物体检测方法的综述主要集中在静态图像的显著性物体检测上。

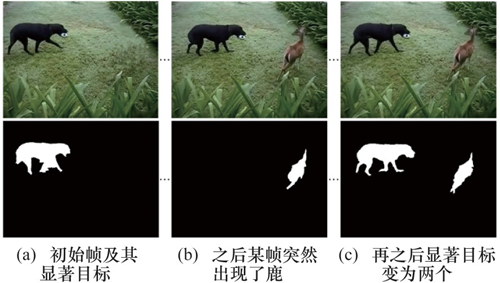


图 3 视频显著性检测示意图( 每列图像的上图为视频中的帧、下图为对应的显著目标真值)

近些年来，一些高性能的显著性物体检测方法相继被提出，总的来说可分为传统的显著性物体检测方法和基于深度学习的显著性物体检测方法。传统方法属于自底向上的检测方法，主要利用低级特征，如颜色、纹理、亮度等信息，通过颜色对比、图像周边和中心的差异等获得显著图。基于深度学习的方法属于自顶向下的检测方法，主要利用任务先验和高级特征，可以被认为是任务驱动的检测方法，如在驾驶场景下司机更关注于前方的信号灯、标识牌、前车尾灯等，通过这些目标的先验特征，进行显著目标的检测。下面分别对这两类方法进行介绍。

**1.1 传统的显著性物体检测方法**

自从 Itti 等将视觉注意力机制模型引入到计算机视觉领域，其后研究的显著度检测或显著图计算主要针对图像中具有各种显著特征的区域而不是具有明确语义的整个物体。对于具备显著性的整个物体的检测研究一般被认为是从 2007 年开始兴起的。

根据在计算过程中的不同特点，将显著性物体检测方法划分为: 基于全局和基于局部对比的计算方法、基于频域的计算方法、基于稀疏理论的计算方法、基于图的计算方法和基于背景及前景先验的计算方法等。

**1.1.1 基于全局和基于局部对比的计算方法**

全局信息和局部信息在显著性物体检测当中分别起到不同的作用。对基于全局对比和基于局部对比的计算方法的划分，主要考虑视觉系统是先加工图形的整体性质，还是加工其局部的性质。HC 方法是通过像素点之间的颜色差异来得到像素点的显著值。ＲC 方法将像素的空间关系和颜色差异结合到一起，首先将图像划分为若干个区域块，然后计算区域块之间的颜色对比度，再将区域之间的对比度加权和作为该区域的显著值。

**1.1.2 基于频域的计算方法**

基于频域的计算方法，主要是根据信息论的观点来对图像进行处理，该观点认为图像由冗余和突出两部分组成，那么找出图像当中的冗余部分，将其除去即为突出部分，也就是对应的显著部分。图像的统计特性具有尺度变换不变性，首先将原图进行傅里叶变换，得到振幅谱和相位谱; 然后根据得到的振幅谱，求其对数振幅谱; 将对数振幅谱减去对数振幅谱均值滤波后的结果，即为剩余谱; 最后利用傅里叶反变换将相位谱和剩余谱求得的自然指数转换到空间域，再进行高斯滤波，即得到最终的显著图。但是该方法得到的结果仅仅是对图像中显著物体部分像素进行高亮处理，没有检测到整个显著物体。在此基础上，复旦大学 Guo 等提出相位谱方法。从图像的频率域角度进行图像中的显著性物体检测，即通过滤波器将图像分为高频域和低频域，然后通过高斯平滑和均值的计算得到了显著图，随后作者在之前的基础上，考虑到显著物体的边缘与物体中心的频率特征不同，又提出了改进的频域方法，从而提升了显著性物体的检测性能。

**1.1.3 基于稀疏理论的计算方法**

基于稀疏理论的计算方法，主要是将样本转换为稀疏表示的形式，达到简化模型、降低数据维度的目的。根据人眼观察的特性，人们在观察场景中的目标时，往往更关注于稀疏区域的物体，而获得的这部分稀疏区域可以用来作为显著性物体检测的结果。Shen等提出了使用稀疏理论进行显著性物体检测，该方法通过分割和特征变换，将图像表示成一个低秩矩阵和稀疏噪声的组合，再融合了底层信息和高层先验知识实现显著区域的优化。该算法存在的主要问题是计算量大、运行速度慢、实验效果一般。随后，Li 等在稀疏理论的基础上，利用图像的边界作为提取背景线索的模板，在得到背景模板的基础上，通过构造稠密和稀疏重建误差来进行显著性检测，最后利用多尺度重构得到显著性结果。

这类基于稀疏理论的计算方法为显著性物体检测开辟了一个新的解决问题的思路，也获得了较好的实验效果，后续还有学者在此基础上进行了改进和优化。

**1.1.4 基于图模型的计算方法**

自从简单线性迭代聚类算法 ( Simple Linear It- erative Cluster，SLIC ) 在图像分割中得到广泛应用之后，出现了越来越多的基于图模型的显著性物体检测方法。基于图模型的计算方法一般是通过 SLIC 算法将图像的像素转换为超像素( Superpixel) ，以各个超像素作为图模型的节点、超像素之间的关系作为模型的边，构建好图模型后，利用图的相关知识来解析图像得到显著图。Jiang 将图像转换为图模型，利用马尔可夫链( Markov Chain，MC) 的吸收特性来计算显著图，即将图像四周边界的节点作为吸收节点，然后计算转移节点到吸收节点的平均吸收时间，吸收时间短的被认为是背景的可能性大，而吸收时间长的被认为是前景的可能性大，根据这种原理最终得到显著图。 Yang 等采用基于图的流形排序( Manifold Ｒanking， MＲ) 方法来计算显著性，首先构造图模型，然后使用图像四边界的节点作为背景种子点，得到 4 个边界的显著值并进行融合，作为第一阶段的显著图，然后对第一阶段的显著图进行阈值分割，从而选出种子节点，再通过优化计算得到最终的显著结果。Zhu 等利用边界连通率来度量节点与边界之间的关系，并提出一个统一的优化框架来优化结果，得到最终的显著图。

**1.1.5 基于背景和前景先验的计算方法**

先验信息在显著性物体检测中非常重要，如颜色先验、中心先验、形状先验等。在这里，主要介绍基于背景和前景先验的显著物体检测方法。

很多基于背景先验的计算方法都将图像四周的边界作为背景先验。Wei 等提出一种基于背景先验的显著性检测方法，认为背景区域通常比显著的物体更加接近边界，这样就将计算显著图的问题转化为计算到边界的距离长短的问题。其他采纳类似思想的方法有: Jiang 等在马尔可夫链的基础上，将四边界作为吸收节点; Yang 等提出的流形排序的显著性检测算法，将图像四边界作为背景节点。基于前景先验的计算方法，主要通过假设前景先验信息进行显著物体检测。Zhu 等提出了一种利用 Harris 角的凸包来近似定位前景物体的方法，将凸包先验图和凸包中心偏置值相结合为初始显著图，然后对其进行优化得到最终结果。

**1.1.6 小结**

总的来说，传统方法使用空域或频域中大量的特征对比，或者利用先验信息等进行图像显著性检测，其针对性较强，一般来说只适用于主体明确、物体颜色及结构较为单调的场景。由于泛化能力差、无法描述和适应复杂的场景和对象结构，导致此类显著性检测方法的研究陷入瓶颈。但因不需要进行大规模的样本训练，并且具有较快的计算速度，可满足实时性需求，此类方法还是具有一定的吸引力。

**1.2 基于深度学习的显著物体检测方法**

随着计算机计算能力的提升、深度学习相关模型在众多视觉任务上的成功应用，基于深度学习的显著性物体检测也逐渐成为当前该领域的主流方法。

基于深度学习的显著性物体检测方法，需要对模型进行训练，因此需要大量的样本图片及其对应的标签。虽然在人工标注时耗费大量的时间，以及在网络训练和调参时也花费一定的时间，但是一旦将网络模型训练好，可以适用于较为复杂场景下的显著性物体检测，其性能明显优于传统的显著性物体检测方法。目前，大多数的基于深度学习的显著性物体检测方法都是采取全监督的方式，即需要大量的样本进行训练;按照其网络结构，一般可以将这些方法分为基于传统卷积神经网络的方法和基于完全卷积神经网络的方法。

**1.2.1 基于传统卷积神经网络的方法**

由于传统的卷积神经网络应用于目标检测和图像分类等视觉任务时展现了卓越的性能，它也被用来进行显著性物体检测。在这类方法的网络训练中，一般将图像块对照标签，其包含的显著区域阈值大于某个指定值的被认为是正样本，反之为负样本。Li 等提出了一种多尺度的卷积神经网络检测模型，将图片分为 3 个尺度，利用 3 个卷积网络模型进行特征提取，然后对特征进行融合，最后进行显著区域的检测，并达到了超过传统显著性物体检测的性能。Zhao 等则是利用上下文信息，通过上采样和下采样，构建两个卷积神经网络模型，然后将通过网络得到的显著值进行展平融合获得最终的显著图。He 等提出 SuperC- NN 来进行显著性物体检测，将输入的图片划分为 3个超像素块，然后分别生成颜色类别和颜色分布的特征向量，再进行卷积神经网络的训练，最后将得到的多尺度显著图进行融合获得最终的结果。

通过以上介绍可以看出，传统的卷积神经网络模型的显著性物体检测要将图像块的区域大小调整到固定的尺寸，或进行多尺度的处理，然后进行卷积神经网络的特征提取，最后进行融合，获得每个区域的显著值。从模型结构可以看出，卷积神经网络最后的全连接层丢失了空间信息，从而影响了分割显著物体结果的性能。

**1.2.2 基于完全卷积神经网络的方法**

自从基于完全卷积神经网络( Fully Convolutional Network，FCN) 的语义分割模型被提出，因其可以克服传统卷积神经网络模型的缺陷，具有空间保存的功能，被广泛应用于显著性物体检测当中。FCN 模型是端到端的模型，可以输入任意尺度的图像，并且是像素级别的检测，很好地保留了空间特性。自 FCN 模型被提出，出现了很多基于 FCN 模型的显著性物体检测方法，而这些检测方法中大多使用了 VGGNet或 Ｒes- Net这两种骨干网络架构。下面主要基于这两种骨干网络架构，进行基于完全卷积神经网络显著性检测方法的介绍。

VGGNet 网络都使用了 3 × 3 的卷积核和 2 × 2 的池化层，训练时间相对较少，文献中有各种各样具体的 VGGNet 架构，主要体现在有不同 的网络深度，如 VGG16，VGG19 等。Liu 等提出了一个端到端的显著性物体检测网络 DHSNet，首先从全局结构信息得到一个粗略的显著图，然后再用一个层次结构的循环神经网络 HＲCNN，通过整合局部上下文信息，分层逐步地细化得到最终的显著图。

虽然基于 VGGNet 的显著性物体检测方法大多取得了不错的性能，并且随着其模型深度的增加，往往能提高模型的检测精度，但造成了梯度消失的现象，导致了模型训练误差的提高。相比较于 VGGNet，ＲesNet架构具有更深的卷积层，如 ＲesNet50， ＲesNet101，因 ＲesNet 使用了短连接( Shortcut Connection) ，不仅加快了模型的训练速度、提高了训练效果，而且当模型的深度加深时，这个简单的结构能够很好地解决退化问题。因此 ＲesNet 的提出在图像检测上具有重要的意义，近些年来也被广泛应用在显著性物体检测当中。

**1.2.3 小结**

总的说来，虽然基于完全卷积神经网络的显著性检测方法相比传统卷积神经网络的显著性检测方法能更好地保存空间信息，但经过卷积、下采样和上采样后，最终的特征图丧失了显著物体的一些细节，在一定程度上影响了检测的精度。因此，在此基础上，基于复杂的特征融合网络结构被提出，如将底层特征和高层特征进行连接，获得更加丰富的语义信息。

二、 显著性物体检测数据集

显著性物体检测近年来得到快速发展的重要原因是其具有很多公开数据集，有系统的评价标准。参考了其他综述文献中的形式，对这些数据集的信息进行了搜集整理，详细列表。主要列举了数据集的名称、容量，数据集中物体的属性、特点，以及相关文献和发表时间等信息。



下面对被广泛应用于深度显著性物体检测模型训练的6 个常用数据集进行介绍。

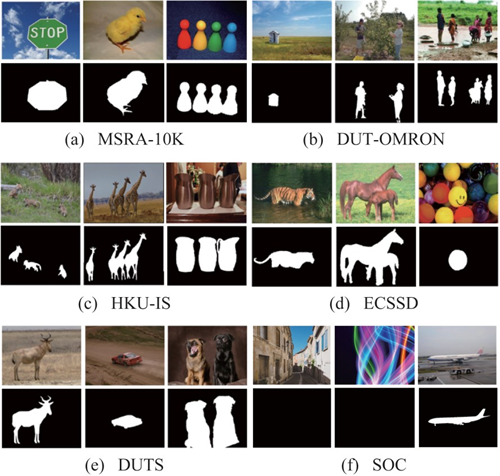


图 4 6 种常见的数据集中的图片和真值示例图

Liu 等公开了第一个用于显著物体检测的数据集 MSＲA-B，，Cheng 等［在其基础上，将原来的只有 5000 张标注物体边界框的图像的数据集，扩展成为有 10K 张标注显著物体像素级别标注的数据集 MSＲA-10K。如图 4( a) 所示，该数据集中的图片目标清晰，多为大目标，并且具有一个目标物体的图片居多，被研究者广泛使用。

DUT-OMＲON 是由 Yang 等在 2013 年建立的，该数据集包含了 5168 张图像，每张图像中有一个或者多个目标( 见图 4( b) ) ，且目标背景复杂，具有一定的相似性。该数据集不仅包含显著物体检测的像素级别的真值，还提供了人眼关注点的真值和物体标注框的真值。

HKU-IS 是Li 等于2015 年建立的，该数据集包含了 4447 张图像，具有多个目标，并且每个目标具有分散性，如图 4( c) 所示，背景和目标具有相似性，给检测带来了一定的难度，具有很大的挑战性。

ECSSD 是 Shi 等于 2015 年在 CSSD 数据集的基础上建立的，将原来 200 张的数据集扩展到 1000张，该数据集中的图片具有复杂的背景结构( 见图 4 ( d) ) ，具有一个或多个目标，具有一定的检测难度，也被广泛使用于各种方法的研究中。

DUTS 是 Wang 等于 2017 年建立的，该数据集包含 10553 张训练数据和 5019 张测试数据，总共 15572 张图片，适用于基于深度学习的检测方法，需要大量标签的数据。该数据集的图片中具有多个目标，并且目标与背景相似度高( 见图 4 ( e) ) ，被广泛应用于基于深度学习的方法当中。

SOC 是 Fan 等于 2018 年建立的，该数据集有 6000 张图片，其中包含 3000 张具有显著目标的图片和 3000 张没有显著目标的图片。SOC 全称为 Salient Objects in Clutter，译为杂乱环境下的显着对象，该数据集具有复杂背景和非显著目标( 如图 4( f) 前两张图片所示，因其没有显著目标，其真值图片是一张全黑色的图片) 。SOC 建立的目的是面向深度学习的方法，作者将该数据集分成了 3600 张训练子集、1200 张测试子集和 1200 张验证子集。