# 聚焦视频显著性物体检测

范登平[1]　王文冠[2]　程明明[1] *　沈建冰[2,3]

[1] 南开大学计算机学院　[2] 起源人工智能研究院(IIAI)　[3] 北京理工大学计算机学院

http://dpfan.net/DAVSOD/

## Abstract

过去十年中，人们对视频显著物体检测（*VSOD*）的兴趣日益浓厚。然而，研究界长期缺少一个完善的、具有高质量标注的真实动态场景下的*VSOD*数据集。为此，我们精心构建了一个和人视觉注意力相一致的、稠密标注的*VSOD* 数据集，它有*226*个视频、*2.4*万帧，涵盖了不同的真实场景、对象、实例和动作。借助相应的真实眼动注视点数据，我们得到了精确的用户标注。从而首次明确强调了具有挑战性的**显著性转移**现象，即视频中的显著对象可能会动态改变。

为了进一步给*VSOD*领域提供一个全面的评测，本文在*7*个现有的*VSOD*数据集和我们构建的*DAVSOD*数据集上（总共*4*万帧，当前规模最大）系统地评估了*17*个最具代表性的*VSOD*算法。利用三个经典的指标，我们呈现了全面而深刻的性能分析。此外，我们还提出了一个基准模型。它配备了一个面向显著性转移的长短时记忆卷积网络（*convLSTM*），可通过学习人类注意力转移行为来有效地捕获视频动态显著性。广泛的评测结果为模型开发和比较开辟了光明的前景。所有的显著图、评测工具包、以及数据集见：*https://github.com/DengPingFan/DAVSOD/*。

## 1. 引言

显著性物体检测（SOD）旨在从静止图像或动态视频中提取最吸引注意力的物体。该任务源于认知研究中人类的视觉注意行为，即人类视觉系统(HVS)中的一项惊人能力——能够快速地将注意力转移到视觉场景中最具信息量的区域。早前的研究 [6,40]已经定量地证实了在这种显式的、对象-级的显著性判断（对象显著性）和隐式的视觉注意力分配行为（视觉注意机制）之间存在高度的相关性。

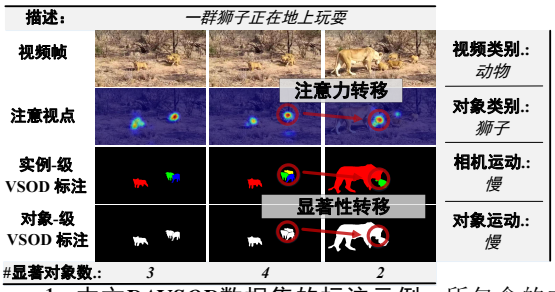


Figure 1: 本文*DAVSOD*数据集的标注示例。所包含的丰富标注如，显著性转移，对象/实例-级VSOD用户标注，显著对象的数目，场景/对象类别以及相机/对象运动模式，这为VSOD任务提供了坚实的基础并使得各种潜在应用受益。

人在观察真实世界的过程中，视觉的动态性无处不在。因此，视频显著物体检测（VSOD）对于理解HVS的潜在机理非常重要且有助于现实中各种应用程序的发展。如，视频分割 [69]，视频字幕 [52]，视频压缩 [22, 24]，自动驾驶 [85]，人机互动 [77]等。除了其学术价值和实际意义外，由于视频数据（各种运动模式、遮挡、模糊、物体形变等）自身的挑战以及人类在动态场景中视觉注意行为（选择性注意分配，注意转移 [5, 32, 55]）固有的复杂性，都使VSOD面临着巨大的困难。因此，这几年引起了广泛的研究兴趣 [7, 20, 26, 31, 33, 34, 56] （见表2）。

然而，具有代表性的VSOD评测的构建仍然严重滞后，这与VSOD建模的蓬勃发展形成了鲜明的对比。尽管有几个针对VSOD任务提出的数据集 [30, 35, 38, 47, 51, 54, 70]，但存在以下缺陷：首先，人在动态浏览的过程中，注意力会随着视频内容的变化而有选择性地动态分配资源到不同的部分。但是，以前的数据集标注时并没有将动态的人眼注视点数据考虑在内，而是

*本文为CVPR2019论文 [15]的中文翻译版。

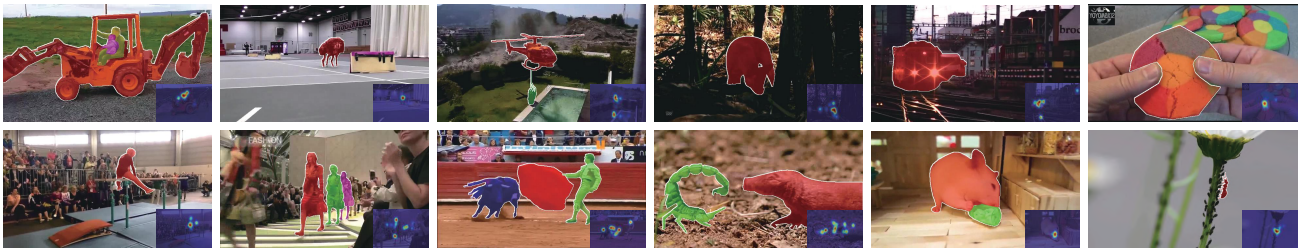

Figure 2: ***DAVSOD*数据集中的视频示例**，其结果由实例-级用户标注的分割结果和注意视点图（右下角）叠加而成。

直接将视频拆分成离散的静态帧来标注，因此不能够揭示人类动态观察期间真实的注意行为。其次，它们的可扩展性，覆盖范围，多样性和难度通常受到限制。从而，现有数据集的这些限制抑制了该分支的进一步发展。

本文的两个贡献如下：首先，我们专门为VSOD任务构建了一个大规模的DAVSOD（稠密标注的视频显著对象检测）数据集。

- 它包含226个视频序列，严格地根据真实的人类注视点记录（见图.2）来标注。更重要的是，选择性注意和注意力转移这两个重要的动态注意特性都被考虑到了。在*DAVSOD*数据集中，显著对象可能会在不同时刻有所改变（见图.1），这更符合实际且需要对视频内容有更完全的理解。这样就构建了一个和视觉注意力相一致性的VSOD数据集。
- 此外，视频序列通过精心筛选得以涵盖多样的场景/对象类别、运动模式并以逐帧逐像素精确地标注（约2.4万帧）。
- DAVSOD的另一个特点是提供了对象和实例-级标注以及简短的文字描述。这有利于促进各种潜在研究方向的发展，如实例-级VSOD，视频显著对象感数，基于显著性的视频字幕等。

其次，利用已建立的*DAVSOD*数据集和之前的7个VSOD数据集[30,35,38,47,51,54,70]，我们对17种最先进的模型 [8, 11, 30, 36, 39, 47, 48, 57, 62, 63, 65, 69–71, 76, 81, 86]进行了全面的评测，使其成为最完整的VSOD评测。此外，我们还提出了一个名为SSAV（面向显著性转移的VSOD）的基础模型。它通过使用显著性转移感知convLSTM模块来学习并预测视频显著性，该模块显式地模拟人类在动态场景中的视觉注意力转移行为。上述评测结果清楚地证明了SSAV模型的有效性。

本文的两个贡献组成了一个完整的评测平台。作为一个补充性的评测，有了它可以更加深入地了解VSOD 任务并促进更多的研究工作朝着这个方向发展。

| 数据集 | 年份 | #Vi. | #AF. | DL | AS | FP | EF | IL |
|---|---|---|---|---|---|---|---|---|
| *SegV2* [35] | 2013 | 14 | 1, 065 | ✓ | | | | |
| *FBMS* [51] | 2014 | 59 | 720 | | | | | |
| *MCL* [30] | 2015 | 9 | 463 | | | | | |
| *ViSal* [70] | 2015 | 17 | 193 | | | | | |
| *DAVIS* [54] | 2016 | 50 | 3, 455 | ✓ | | | | |
| *UVSD* [47] | 2017 | 18 | 3, 262 | ✓ | | | | |
| *VOS* [38] | 2018 | 200 | 7, 467 | | | ✓ | | |
| ***DAVSOD*** | 2019 | **226** | **23, 938** | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: ***DAVSOD*数据集以及当前VSOD数据集的统计数据。**显然，*DAVSOD*提供了更加丰富的标注。**#Vi.**: 视频数量。**#AF.**: 标注帧的数量。**DL**: 是否是稠密（逐帧）标注。**AS**: 是否考虑了注意力转移。**FP**: 显著物体的标注是否根据人眼注视点。**EF**: 是否为标注的显著对象提供人眼注视点。**IL**: 是否提供了实例-级标注。

## 2. 相关工作

**VSOD数据集.** 这些年，有几个数据集被建立或引入VSOD领域。表1列出了这些数据集的统计数据。其中，*SegV2* [35]和*FBMS* [51] 是两个早期被采纳的数据集。由于它们是为特定目的而设计的，因此不太适合VSOD任务。另一个*MCL* [30]数据集仅有9个简单的视频序列。*ViSal* [70]则是第一个专门设计的VSOD数据集，包含17个带有明显对象的视频序列。最近，Wang等人 [71]把著名的视频分割数据集DAVIS [54] 引入到VSOD中，它由50个具有挑战性的场景组成。尽管上述数据集从不同程度上促进了VSOD的发展，但其规模（仅几十个视频）严重受限。并且这些数据集的构建未考虑动态场景中人类真实的注意视点，仅仅通过几个标注者以手工的方式武断地找出显著物体。标注过程中不考虑复杂场景中的帧间时序特性而是单帧独立标注。最近，一个较大(200个视频)规模的VOS [38]数据集部分地弥补了上述限制。但它的多样性和普遍性非常有限，由于它含有大量简单的室内

| 编号. | 模型 | 年份 | 出版社. | 训练数量 | 训练集 | Basic | 类型 | OF | SP | S-measure | PCT | 代码 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SIVM [57] | 2010 | ECCV | | | CRF, 统计量 | T | | | 0.481∼0.606 | 72.4* | M&C++ |
| 2 | DCSM [31] | 2011 | TCSVT | | | SORM距离 | T | | | | 0.023* | C++ |
| 3 | RDCM [42] | 2013 | TCSVT | | | gabor，区域对比 | T | ✓ | | | 9.8* | N/A |
| 4 | SPVM [48] | 2014 | TCSVT | | | 超像素，直方图 | T | | ✓ | 0.470∼0.724 | 56.1* | M&C++ |
| 5 | CDVM [16] | 2014 | TCSVT | | | 压缩域 | T | | | | 1.73* | M |
| 6 | TIMP [86] | 2014 | CVPR | | | 时间映射 | T | ✓ | | 0.539∼0.667 | 69.2* | M&C++ |
| 7 | STUW [17] | 2014 | TIP | | | 不定加权 | T | ✓ | | | 50.7* | M |
| 8 | EBSG [50] | 2015 | CVPR | | | 格式塔原理 | T | ✓ | | | | N/A |
| 9 | SAGM [69] | 2015 | CVPR | | | 测地距离 | T | ✓ | ✓ | 0.615∼0.749 | 45.4* | M&C++ |
| 10 | ETPM [59] | 2015 | CVPR | | | 眼动追踪先验 | T | ✓ | | | | N/A |
| 11 | RWRV [30] | 2015 | TIP | | | 随机游走 | T | | | 0.330∼0.595 | 18.3* | M |
| 12 | GFVM [70] | 2015 | TIP | | | 梯度流 | T | ✓ | ✓ | 0.613∼0.757 | 53.7* | M&C++ |
| 13 | MB+M [81] | 2015 | ICCV | | | 最小障碍距离 | T | | | 0.552∼0.726 | 0.02* | M&C++ |
| 14 | MSTM [65] | 2016 | CVPR | | | 最小生成树 | T | | | 0.540∼ 0.657 | 0.02* | M&C++ |
| 15 | SGSP [47] | 2017 | TCSVT | | | 图，直方图 | T | ✓ | ✓ | 0.557∼0.706 | 51.7* | M&C++ |
| 16 | SFLR [8] | 2017 | TIP | | | 低秩一致性 | T | ✓ | ✓ | 0.470∼0.724 | 119.4* | M&C++ |
| 17 | STBP [76] | 2017 | TIP | | | 背景先验 | T | | ✓ | 0.533∼0.752 | 49.49* | M&C++ |
| 18 | VSOP [23] | 2017 | TC | | | 对象候选框 | T | ✓ | ✓ | | | M&C++ |
| 19 | DSR3 [33] | 2017 | BMVC | 44 (6+8+30) clips | 10C+S2+DV | RCL [43] | D | | | | | Py&Ca |
| 20 | VQCU [3] | 2018 | TMM | | | 光谱，图结构 | T | | ✓ | | 0.78* | M |
| 21 | CSGM [72] | 2018 | TCSVT | | | 视频联合显著性 | T | ✓ | ✓ | | 3.86* | M&C++ |
| 22 | STUM [2] | 2018 | TIP | | | 局部时空邻域线索 | T | | | | | N.A. |
| 23 | SAVM [73] | 2018 | PAMI | | | 测地距离 | T | ✓ | ✓ | 0.615∼0.749 | 45.4* | M&C++ |
| 24 | bMRF [7] | 2018 | TMM | | | MRF | T | ✓ | ✓ | | 2.63* | N/A |
| 25 | LESR [87] | 2018 | TMM | | | 局部估计，时空估计 | T | ✓ | ✓ | | 5.93* | N/A |
| 26 | TVPI [56] | 2018 | TIP | | | 测地距离, CRF | T | | ✓ | | 2.78* | M&C |
| 27 | SDVM [4] | 2018 | TIP | | | 时空分解 | T | | | | | N/A |
| 28 | SCOM [11] | 2018 | TIP | ∼10K frame pairs | MK | DCL [37] | D | ✓ | ✓ | 0.555∼0.832 | 38.8 | N/A |
| 29 | STCR [34] | 2018 | TIP | 44 (6+8+30) clips | 10C+S2+DV | CRF | D | | ✓ | | | N/A |
| 30 | DLVS [71] | 2018 | TIP | ∼18K frame pairs | MK+DO+S2+FS | FCN [49] | D | ✓ | ✓ | 0.682∼0.881 | 0.47 | Py&Ca |
| 31 | SCNN [63] | 2018 | TCSVT | ∼11K frame pairs | MK+S2+FS | VGGNet [61] | D | ✓ | ✓ | 0.657∼0.847 | 38.5 | N/A |
| 32 | FGRN [36] | 2018 | CVPR | ∼10K frame pairs | S2+FS+DV | 长短时记忆 | D | ✓ | | 0.674∼0.861 | 0.09 | Py&Ca |
| 33 | SCOV [28] | 2018 | ECCV | | | BOW [18], 候选框, FCIS [41] | T | ✓ | ✓ | | 3.44 | N/A |
| 34 | MBNM [39] | 2018 | ECCV | ∼13K frame pairs | Voc12 + Coco [44] + DV | 运动特征, DeepLab [9] | D | ✓ | | 0.637∼0.898 | 2.63 | N/A |
| 35 | PDBM [62] | 2018 | ECCV | ∼18K frame pairs | MK+DO+DV | DC [79] | D | | | 0.698∼0.907 | 0.05 | Py&Ca |
| 36 | UVOS [26] | 2018 | ECCV | | | 标准边缘提取器 | D | ✓ | ✓ | | | N/A |
| 37 | **SSAV (Ours)** | 2019 | CVPR | ∼13K frame pairs | DAVSOD val + DO +DV | SSLSTM, PDC [62] | D | | | 0.724∼0.941 | 0.05 | Py&Ca |

Table 2: **36个先前具有代表性的VSOD方法和提出的SSAV模型。 训练集:** 10C = *10-Clips* [19]。 S2 = *SegV2* [35]。 DV = *DAVIS* [54]. DO = *DUT-OMRON* [78]. MK = *MSRA10K* [12]。 MB = *MSRA-B* [46]。 FS = *FBMS* [51]。 Voc12= PASCAL VOC2012 [13]。 **Basic:** CRF = 条件随机场。 SP = 超级像素。 SORM = 自序相似度量。 MRF = 马尔可夫随机场。 **类型:** T = 传统方法。 D = 深度学习。 **OF:** 是否使用光流技术。 **SP:** 是否使用超像素过分割技术。 **S-measure [14]:**表4中8个数据集Smeasure的得分范围。 **PCT:**每帧计算时间（秒）。 由于 [3, 7, 11, 28, 39, 42, 63, 87]模型没有公布代码，因此运行时间PCTs来源于原文或者由作者提供。 **代码:** M = Matlab。 Py = Python。 Ca= Caffe。 N/A = 文中无法获得。 "*" 表示CPU运行时间。

场景和相机稳定拍摄的场景。

总的来说，*DAVSOD*与上述数据集有明显的区别：i) 通过深入分析动态场景下人类真实的注意行为，我们发现了视觉注意力转移现象。从而，首次强调动态场景中的显著对象转移，并提供了唯一的、与视觉注意力相一致的标注。ii) 其多样性、大规模稠密标注、完整的对象/实例-级显著对象的标注、视频描述以及丰富的属性标注（例如，显著对象的数量，运动模式以及场景/对象类别等），共同为VSOD 任务打下坚实而独特的基础。

**VSOD模型。** 早期的*VSOD*模型 [8, 21, 23, 30, 47, 48, 57, 58, 69, 70]建立在手工设计的特征（如，颜色、运动等）之上，并在很大程度上依赖于图像显著对象检测领域(例如，中心周边对比 [12]，背景先验 [74]) 中典型的启发式方法和视觉注意认知理论（例如，特征整合理论 [64]，引导式搜索 [75]等）。以外，它们还探索了采用不同的计算机制整合空间和时间显著特征的方法，如梯度流场 [70]，测地距离 [69]，重启随机游走 [30]和光谱图结构 [3]。因此，繁重的特征工程以及表达能力有限的手工特征必然会束缚传统VSOD模型的发展。详见表2。

最近，因深度神经网络在图像显著性检测 [27, 45, 66, 67, 80, 82–84]领域的成功应用，基于深度学习的VSOD模型 [26, 33, 34, 36, 62, 63, 71]由此备受关注。具体而言，Wang等人的研究 [71]是VSOD中最早采用全卷积神经网络的。另一项同期工作 [33]使用3D滤波器将空间和时间信息合并到时空CRF框架中。后来，时空深度特征 [34]，RNN [36]，金字塔扩张convLSTM [62] 陆续被提出来以便更好地捕捉空间和时间显著特征。由于神经网络的强大的学习能力，这些基于深度

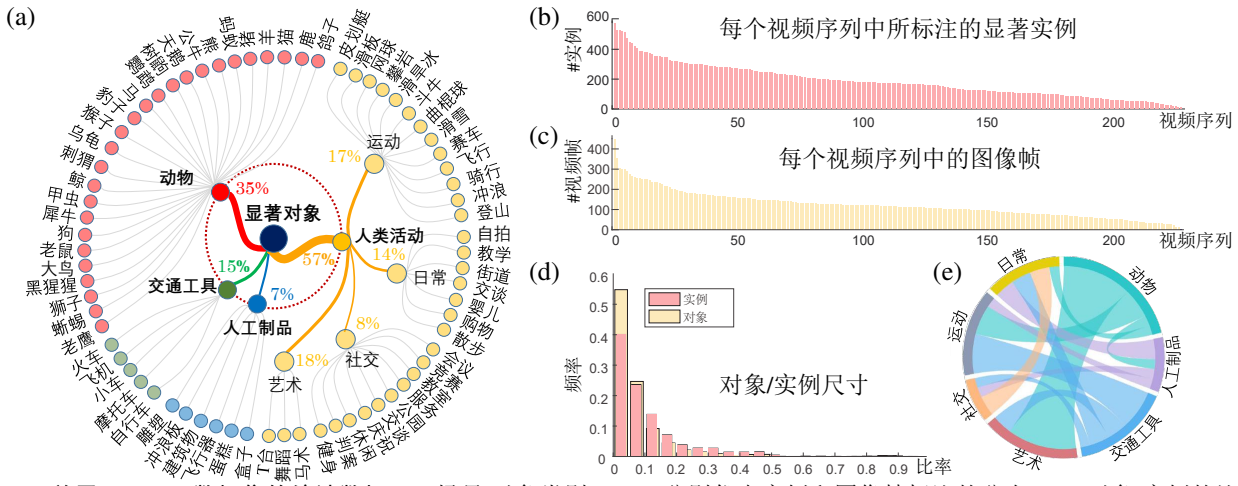


Figure 3: **关于*DAVSOD*数据集的统计数据**: (a) 场景/对象类别。(b, c)分别代表实例和图像帧标注的分布。(d) 对象/实例的比例分布。(e) 表示(a)中场景类别之间的相互依赖关系。

的VSOD模型通常取得了更好的性能。然而，这些模型都忽略了对于理解人类视觉注意机制非常重要的显著性转移现象。相对而言，我们的基础模型（SSAV）明确地利用显著性转移线索，得到了卓有成效的结果。

我们在七个以前的数据集和提出的*DAVSOD*数据集上系统地对17个最先进的VSOD模型 [30, 35, 38, 47, 51, 54, 70] 进行了评测，这项工作代表了迄今为止在VSOD领域中最大的性能评估。我们借助大量的定量结果，为VSOD领域呈现了一系列重要的结论并指出了若干有前景的研究方向。

## 3. 提出的数据集

本节将详细阐述所提的*DAVSOD*数据集，它是专门为VSOD任务设计的。图1和图2展示了带标注的图像帧。细节请读者参阅补充材料。我们将从以下4个关键方面来介绍*DAVSOD*。

### 3.1. 视频采集

*DAVSOD*的视频序列源自*DHF1K* [68]，DHF1K是当前最大规模的动态眼动追踪数据集。使用DHF1K构建*DAVSOD*数据集有以下好处。DHF1K是从Youtube上收集的，涵盖了各种现实场景、多种物体外观和运动模式、丰富的对象类别，以及动态场景中大部分常见的挑战，这为我们构建大规模和具有代表性的评测提供了坚实的基础。更重要的是，DHF1K所提供的视觉注视点使我们能够得到更合理的、生物启发的对象-级显著性标注。我们以手工的方式将视频分为小片段(图.3(c))并删除那些带黑屏过渡的片段。通过这种方式，最终得到了一个大型数据集，它包含226个视频，共计23,938帧，798秒。视频分辨率为640×360像素。

### 3.2. 数据标注

**显著性转移标注.** 在真实的动态场景中 [32,55]，人类的注意力行为更加复杂，即选择性注意力分配和显式的注意力转移（由于突然的攻击，新的动态事件等）都可能发生。通过*DHF1K*的眼动追踪记录，我们观察到数据驱动的注意力转移普遍存在，如图1所示。然而，VSOD领域中之前的研究都没有明确强调这种基本的视觉注意行为。在*DAVSOD*中，我们根据真实的人类注视点来标注显著的对象，并且首次标注了注意力转移所发生的时刻，强调了该领域中显著性转移这一更具挑战的任务。

**场景和对象类别标注.** 与文献 [68]一致，每个视频都手动标记一个类别（如，*动物*, *交通工具*, *人工制品*, *人类活动*）。*人类活动*有4个子类：*运动*，*日常*-，*社交*-以及*艺术活动*。至于对象类别，和MSCOCO一致，只包含“事物”类别（而不是“东西”）。这样我们就建立了一个大约70个最常出现的场景/对象列表。图.3(a)和(e)中，分别展示了场景/对象类别及其相互依赖性。整个对象标注过程有五个标注者参与。

**实例/对象级显著物体标注.** 我们让20个标注者经过10个视频示例预训练后，从每个待标注的视频帧中选择出最多5个对象并细致地标注它们（用精确的边缘轮廓而不是粗糙的多边形）。标注者还被要求区分出不同的实例并且单独进行标注，从而得到23,938帧对象级显著性标注和39,498 个实例级显著性标注。

### 3.3. 数据集的特点与统计

为深入了解*DAVSOD*数据集，几项重要特征如下。

| *DAVSOD* | 相机运动 | | 对象运动 | | | 实例-级对象个数 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 慢 | 快 | 稳定 | 慢 | 快 | 1 | 2 | 3 | ≥4 |
| 视频数量 | 102 | 124 | 117 | 72 | 37 | 134 | 125 | 46 | 33 |

Table 3: **关于*DAVSOD*数据集中摄像机/对象运动和显著对象实例数量的统计信息。**

**丰富多样的显著对象.** *DAVSOD*中的显著对象涵盖了丰富的类别：动物（如，狮子，大鸟），车辆（汽车，自行车），人工制品（例如，盒子，建筑物）和各种形式的人类活动（例如，舞蹈，骑行），使得全面地理解动态场景下对象-级显著性成为可能。

**显著对象实例的数量.** 现有数据集局限于显著性对象实例的数量（见表1）。之前的研究 [29]表明，人类可以一目了然地精确感知到多达5个物体而无需一个个地计数。因此，如表3所示，*DAVSOD*包含了更多的显著对象（每帧最多5个显著对象实例，平均大约1.65）。图图.3(b)列出了每个视频中标注的实例数量分布。

**显著对象的尺寸.** 对象级显著对象的大小定义为前景对象像素与图像之间的比率。*DAVSOD*数据集中的显著对象尺寸为0.29%～91.3%（平均：11.5%），变化范围更广。（请参阅 图.3(d)）。

**多样化的相机运动模式.** *DAVSOD*包含了各种不同的相机运动模式(见表3)。在这样的数据上进行训练的算法可以更好地处理真实的动态场景，因此更实用。

**不同的对象运动模式.** *DAVSOD*继承了*DHF1K*的优势，囊括了各种各样(见表3) 真实的动态场景(如，对象运动模式从稳定到快速)。这对于避免过度拟合以及客观和准确地进行算法评测至关重要。

**中心偏向.** *DAVSOD*和现有数据集 [30, 35, 38, 47, 51, 54, 70]的中心偏向如图4所示。

### 3.4. 数据集划分

现有数据集没有保留测试集，这样很容易导致模型在数据集上过度拟合。因此，我们按照4：2：4的比例将视频分为训练、验证和测试集合。采用随机筛选的策略，我们得到了一个独特的划分结果，包含了90个训练集视频、46个验证集视频以及90个测试集视频。训练集和测试集的标注结果将会公开而测试集的标注结果将被保留。根据VSOD任务的难度，90个测试集又进一步被分为35个容易子集、30个正常子集和25个困难子集。

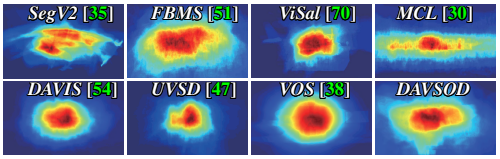


Figure 4: *DAVSOD*和现有VSOD数据集的**中心偏向**.

## 4. 提出的模型

### 4.1. 基于显著性转移的VSOD模型

**模型概述.** 本文所提出的**SSAV**模型由两个基本模块构成：金字塔扩张卷积模块（PDC） [62]和显著性转移感知模块(SSLSTM)。前者用于鲁棒地学习静态显著性特征，后者将传统的长短时记忆卷积网络（convLSTM）[60]与显著性转移感知注意（SSAA）机制相结合。SSAV模型将经由PDC模块得到的静态特征序列作为输入，同时考虑时序变化和显著性转移从而得到相应的VSOD结果。

**金字塔扩张卷积(PDC)模块.** 最新语义分割和VSOD的研究表明 [10, 62]，由于多尺度信息的利用和空间细节的保留，平行叠加一组带有采样率的扩张卷积层可以获得更好的学习性能。因此我们使用PDC模块 [62]作为静态特征提取器。形式上，令$\mathbf{Q}\in\mathbb{R}^{W\times H\times C}$ 表示输入帧$\mathbf{I}\in\mathbb{R}^{w\times h\times 3}$的3D特征张量。扩张率为$d>1$的扩张卷积层$\mathcal{D}_d$可以作用到$\mathbf{Q}$中，从而得到特征$\mathbf{P}\in\mathbb{R}^{W\times H\times C'}$。该输出特征保持了原始空间分辨率，同时获得了更大的感受野（采样步长为$d$）。通过并行排列一组（$K$）不同扩张率$\{d_k\}_{k=1}^{K}$ 的扩张卷积层$\{\mathcal{D}_{d_k}\}_{k=1}^{K}$来组织PDC模块:

$$\mathbf{X}=[\mathbf{Q},\ \mathbf{P}_1,\ldots,\ \mathbf{P}_k,\ldots,\ \mathbf{P}_K], \tag{1}$$

其中，$\mathbf{X}\in\mathbb{R}^{W\times H\times(C+KC')}$，$\mathbf{P}_k=\mathcal{D}_{d_k}(\mathbf{Q})$. $[.,.]$代表连接操作。PDC增强后的特征$\mathbf{X}$是一种更强大的特征（利用多尺度信息）且保留了原始信息$\mathbf{Q}$ （通过残差连接）。

**显著性物体转移感知convLSTM(SSLSTM).** 我们提出了一种显著性转移感知的convLSTM [60]，它使得convLSTM具有显著性转移感知注意机制。它是一个强大的循环模型，不仅可以捕捉时序信息，还可以区分背景中的显著物体以及编码注意力转移信息。更具体地说，通过PDC模块，我们获得了具有$T$帧的输入视频的静态表示$\{\mathbf{X}_t\}_{t=1}^{T}$。在时刻$t$，给定$\mathbf{X}_t$，显著性转移感知的convLSTM输出相应的显著对象掩码$\mathbf{S}_t\in[0,1]^{W\times H}$:

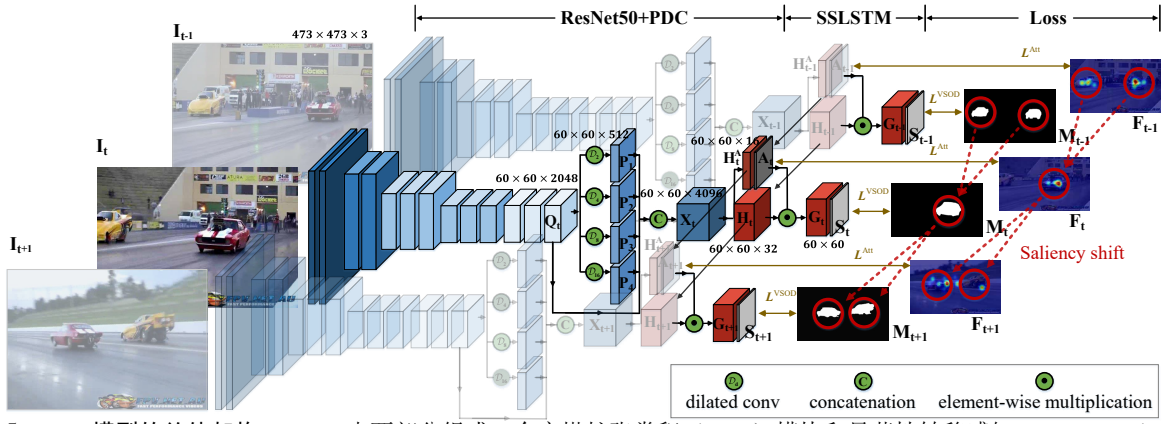


Figure 5: **SSAV模型的总体架构。** SSAV由两部分组成：金字塔扩张卷积（PDC）模块和显著性转移感知convLSTM（SSLSTM）模块。前者用于有效的静态显着性学习，后者用于同时捕获时间动态和显著性转换。有关细节详见第§4节。

$$\begin{aligned} \text{隐藏状态:}\quad & \mathbf{H}_t = \text{convLSTM}(\mathbf{X}_t, \mathbf{H}_{t-1}), \\ \text{显著性转移感知注意:}\quad & \mathbf{A}_t = \mathcal{F}^A(\{\mathbf{X}_1, \cdots, \mathbf{X}_t\}), \\ \text{感知转移:}\quad & \mathbf{G}_{m,t} = \mathbf{A}_t \odot \mathbf{H}_{m,t}, \\ \text{显著性物体预测:}\quad & \mathbf{S}_t = \sigma(\mathbf{w}^S \otimes \mathbf{G}_t), \end{aligned} \tag{2}$$

其中，$\mathbf{H} \in \mathbb{R}^{W\times H\times M}$表示3D张量隐藏状态。注意力图$\mathbf{A} \in [0,1]^{W\times H}$是从显著性转移感知的注意网络$\mathcal{F}^A$计算出来的，它将先前的帧考虑在内。$\mathbf{G} \in \mathbb{R}^{W\times H\times M}$表示感知转移，而$m \in M$表示通道索引下标。$\odot$符号为矩阵元素乘法。$\mathbf{w}^S \in \mathbb{R}^{1\times 1\times M}$是一个$1\times 1$的卷积核，被用作显著对象读取函数。$\otimes$为卷积操作，$\sigma$是*sigmoid*激活函数。

上述模块的关键组成部分是显著性转移感知注意网络$\mathcal{F}^A$。很显然，它充当一个神经元注意机制，因为它被用来对convLSTM输出的特征$\mathbf{H}$进行加权。除此之外，我们还期望它能足以有效地模拟人类注意力转移行为。考虑到这样一个有区别的任务，我们引入了一个小的convLSTM来构建$\mathcal{F}^A$，从而使得convLSTM中嵌套另外的convLSTM结构：

$$\begin{aligned} \text{显著性转移感知注意:}\quad & \mathbf{A}_t = \mathcal{F}^A(\{\mathbf{X}_1, \cdots, \mathbf{X}_t\}), \\ \text{注意力特征提取:}\quad & \mathbf{H}^A_t = \text{convLSTM}^A(\mathbf{X}_t, \mathbf{H}^A_{t-1}), \\ \text{注意力映射:}\quad & \mathbf{A}_t = \sigma(\mathbf{w}^A \otimes \mathbf{H}^A_t), \end{aligned} \tag{3}$$

其中$\mathbf{w}^A \in \mathbb{R}^{1\times 1\times M}$代表一个$1\times 1$的卷积核用来映射注意力特征$\mathbf{H}^A$得到一个重要性矩阵，*sigmoid*函数$\sigma$再把重要性矩阵值归一化到$[0,1]$。然后显著性转移感知注意$\mathbf{A}_t$用于增强公式2中的显著对象分割特征$\mathbf{H}$。由于convLSTM$^A$的应用，我们的注意力模块获得了强大的学习能力，这为学习显式和隐式的注意力转移提供了坚实的基础。假设$\{\mathbf{I}_t \in \mathbb{R}^{w\times h\times 3}\}_{t=1}^T$为包含了$T$帧的一个训练视频，$\{\mathbf{F}_t \in [0,1]^{W\times H}\}_{t=1}^T$为人眼注视标注序列，$\{\mathbf{M}_t \in \{0,1\}^{W\times H}\}_{t=1}^T$为视频显著对象用户标注结果。我们所用的损失函数由注意力模型$\{\mathbf{A}_t \in \{0,1\}^{W\times H}\}_{t=1}^T$的输出和最后视频显著对象预测结果$\{\mathbf{S}_t \in \{0,1\}^{W\times H}\}_{t=1}^T$构成：

$$\mathcal{L} = \sum_{t=1}^{T} \Big( \ell(\mathbf{I}_t) \cdot \mathcal{L}^{\text{Att}}(\mathbf{A}_t, \mathbf{F}_t) + \mathcal{L}^{\text{VSOD}}(\mathbf{S}_t, \mathbf{M}_t) \Big), \tag{4}$$

其中$\mathcal{L}^{\text{Att}}$和$\mathcal{L}^{\text{VSOD}}$都是交叉熵损失函数. $\ell(\cdot) \in \{0,1\}$表示是否存在注意视点标注（因为当前大多数的VSOD数据集缺少人眼注视点数据，见表1)。当缺少相应的注意视点时，误差就不会被回传。更重要的是，当$\ell(\cdot) = 0$时，等式3中的显著性转移感知注意模型$\mathcal{F}^A$就以隐式方式训练。这可以看作是一种典型的神经注意机制。当提供注意视点标注时$(\ell(\cdot) = 1)$，$\mathcal{F}^A$就以显式的方式训练。借助convLSTM结构，$\mathcal{F}^A$就能够准确地将我们的VSOD模型的注意力转移到重要的对象上(见图.6)。

### 4.2. 实验细节

PDC模型的基础CNN网络来自ResNet-50 [25]的卷积层并且最后两层步长设为1。所有输入图像帧都被缩放到$473\times 473$的空间分辨率且$\mathbf{Q} \in \mathbb{R}^{60\times 60\times 2048}$。与文献 [62]一致，我们设置$K = 4$, $C = 512$，$d_k = 2^k$ $(k \in \{1, \cdots, 4\})$。对于等式2中的convLSTM，我们使用一个3×3×32的卷积核。而对于等式3中的convLSTM$^A$则用一个$3\times 3\times 16$的卷积核。训练策略上，我们和 [62]等人保持一致(但是未使用MSRA-10k [12]数据集)。此外，我们进一步利用*DAVSOD*的验证集来显式地训练显著性转移感知注意模块。

| | Metric | 2010-2015 | | | | | | | 2016-2017 | | | | 2018 | | | | | | SSAV† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SIVM [57] | TIMP [86] | SPVM [48] | RWRV [30] | MB+M [81] | SAGM [69] | GFVM [70] | MSTM [65] | STBP [76] | SGSP [47] | SFLR [8] | SCOM [11]† | SCNN [63]† | DLVS [71]† | FGRN [36]† | MBNM [39]† | PDBM [62]† | |
| *ViSal* | max $\mathcal{F}\uparrow$ | .522 | .479 | .700 | .440 | .692 | .688 | .683 | .673 | .622 | .677 | .779 | .831 | .831 | .852 | .848 | .883 | .888 | ***.939*** |
| | $\mathcal{S}\uparrow$ | .606 | .612 | .724 | .595 | .726 | .749 | .757 | .749 | .629 | .706 | .814 | .762 | .847 | .881 | .861 | .898 | .907 | ***.943*** |
| | $\mathcal{M}\downarrow$ | .197 | .170 | .133 | .188 | .129 | .105 | .107 | .095 | .163 | .165 | .062 | .122 | .071 | .048 | .045 | .020 | .032 | ***.020*** |
| *FBMS-T* | max $\mathcal{F}\uparrow$ | .426 | .456 | .330 | .336 | .487 | .564 | .571 | .500 | .595 | .630 | .660 | .797 | .762 | .759 | .767 | .816 | .821 | ***.865*** |
| | $\mathcal{S}\uparrow$ | .545 | .576 | .515 | .521 | .609 | .659 | .651 | .613 | .627 | .661 | .699 | .794 | .794 | .794 | .809 | .857 | .851 | ***.879*** |
| | $\mathcal{M}\downarrow$ | .236 | .192 | .209 | .242 | .206 | .161 | .160 | .177 | .152 | .172 | .117 | .079 | .095 | .091 | .088 | .047 | .064 | ***.040*** |
| *DAVIS-T* | max $\mathcal{F}\uparrow$ | .450 | .488 | .390 | .345 | .470 | .515 | .569 | .429 | .544 | .655 | .727 | .783 | .714 | .708 | .783 | .861 | .855 | ***.861*** |
| | $\mathcal{S}\uparrow$ | .557 | .593 | .592 | .556 | .597 | .676 | .687 | .583 | .677 | .692 | .790 | .832 | .783 | .794 | .838 | .887 | .882 | ***.893*** |
| | $\mathcal{M}\downarrow$ | .212 | .172 | .146 | .199 | .177 | .103 | .103 | .165 | .096 | .138 | .056 | .048 | .064 | .061 | .043 | .031 | .028 | ***.028*** |
| *SegV2* | max $\mathcal{F}\uparrow$ | .581 | .573 | .618 | .438 | .554 | .634 | .592 | .526 | .640 | .673 | .745 | .764 | ** | ** | ** | .716 | *.800* | ***.801*** |
| | $\mathcal{S}\uparrow$ | .605 | .644 | .668 | .583 | .618 | .719 | .699 | .643 | .735 | .681 | .804 | .815 | ** | ** | ** | .809 | ***.864*** | .851 |
| | $\mathcal{M}\downarrow$ | .251 | .116 | .108 | .162 | .146 | .081 | .091 | .114 | .061 | .124 | .037 | .030 | ** | ** | ** | .026 | .024 | ***.023*** |
| *UVSD* | max $\mathcal{F}\uparrow$ | .293 | .338 | .404 | .281 | .339 | .414 | .426 | .336 | .403 | .544 | .562 | .420 | .550 | .564 | .630 | .550 | ***.863*** | .801 |
| | $\mathcal{S}\uparrow$ | .481 | .537 | .581 | .536 | .563 | .629 | .628 | .551 | .614 | .601 | .713 | .555 | .712 | .721 | .745 | .698 | ***.901*** | .861 |
| | $\mathcal{M}\downarrow$ | .260 | .178 | .146 | .180 | .169 | .111 | .106 | .145 | .105 | .165 | .059 | .206 | .075 | .060 | .042 | .079 | ***.018*** | .025 |
| *MCL* | max $\mathcal{F}\uparrow$ | .420 | .598 | .595 | .446 | .261 | .422 | .406 | .313 | .607 | .645 | .669 | .422 | .628 | .551 | .625 | .698 | ***.798*** | .774 |
| | $\mathcal{S}\uparrow$ | .548 | .642 | .665 | .577 | .539 | .615 | .613 | .540 | .700 | .679 | .734 | .569 | .730 | .682 | .709 | .755 | ***.856*** | .819 |
| | $\mathcal{M}\downarrow$ | .185 | .113 | .105 | .167 | .178 | .136 | .132 | .171 | .078 | .100 | .054 | .204 | .054 | .060 | .044 | .119 | ***.021*** | .027 |
| *VOS-T* | max $\mathcal{F}\uparrow$ | .439 | .401 | .351 | .422 | .562 | .482 | .506 | .567 | .526 | .426 | .546 | .690 | .609 | .675 | .669 | .670 | .742 | ***.742*** |
| | $\mathcal{S}\uparrow$ | .558 | .575 | .511 | .552 | .661 | .619 | .615 | .657 | .576 | .557 | .624 | .712 | .704 | .760 | .715 | .742 | .818 | ***.819*** |
| | $\mathcal{M}\downarrow$ | .217 | .215 | .223 | .211 | .158 | .172 | .162 | .144 | .163 | .236 | .145 | .162 | .109 | .099 | .097 | .099 | .078 | ***.073*** |
| *DAVSOD-T* | max $\mathcal{F}\uparrow$ | .298 | .395 | .358 | .283 | .342 | .370 | .334 | .344 | .410 | .426 | .478 | .464 | .532 | .521 | .573 | .520 | .572 | ***.603*** |
| | $\mathcal{S}\uparrow$ | .486 | .563 | .538 | .504 | .538 | .565 | .553 | .532 | .568 | .577 | .624 | .599 | .674 | .657 | .693 | .637 | .698 | ***.724*** |
| | $\mathcal{M}\downarrow$ | .288 | .195 | .202 | .245 | .228 | .184 | .167 | .211 | .160 | .207 | .143 | .220 | .128 | .129 | .098 | .159 | .116 | ***.092*** |

Table 4: **17个最先进的VSOD模型在7个数据集上的评测结果**: *SegV2* [35], *FBMS* [51], *ViSal* [70], *MCL* [30], *DAVIS* [54], *UVSD* [47], *VOS* [38]以及本文的*DAVSOD*测试集中35个简单子集。请注意，TIMP仅在*VOS*上的9个短视频进行测试，因为它无法处理长视频。“**”表示该模型已经在该数据集上进行了训练。“-T”表示结果是在该数据集的测试集上得到的。“†”表示深度学习模型。颜色越深表示性能越好。最佳分数标记为**粗体**。

## 5. 评测结果

### 5.1. 实验设置

**评估指标.** 为定量衡量模型性能，我们用2种流行的指标：平均绝对误差(**MAE**) $\mathcal{M}$ [53], **F-measure** $\mathcal{F}$ [1], 及最新提出的结构性指标**S-measure** $\mathcal{S}$ [14]。评测代码详见：http://dpfan.net/DAVSOD/

**评测的模型.** 我们共测试了17中模型（传统方法11种，深度模型6种）。选择模型的标准为：i）代码已公开，ii）具有代表性。

**评测策略.** 为了提供全面的评测，我们在现有的7个数据集和所提出的*DAVSOD*上评估了17种具有代表性的模型。*VOS* [38]，*FBMS* [51]，*DAVIS* [54]，*DAVSOD*这4个数据集的测试集以及4个完整的数据集*ViSal* [70]，*MCL* [30], *SegV2* [35]，*UVSD* [47] 被用来当作测试集。它们共237个视频，约4万帧。

### 5.2. 性能比较和数据集分析

本节将呈现若干能促进未来研究的有趣结论。

**传统模型的性能.** 基于表4中的不同指标，我们得出的结论为：“*SFLR [8]，SGSP [47]和STBP [76] 是VSOD中非深度学习模型的前3名。*” SFLR和SGSP都显式地考虑了光流策略来提取运动特征。但计算成本通常很高（见表2）。值得注意的是，这3个模型都利用超像素技术在区域级别上整合时空特征。

**深度模型的性能.** 评测中前三名的模型(即SSAV, PDBM [62]，MBNM [39]) 都基于深度学习技术，这表明神经网络具有强大的学习能力。在*ViSal*数据集上（VSOD的第一个专门设计的数据集），它们的平均性能(max $\mathcal{F}$)甚至高于0.9。

**传统与深度VSOD模型的比较.** 从表4可见几乎所有深度模型都优于传统算法，这归功于深度网络更强大的显著性特征提取能力。另一有趣的发现是

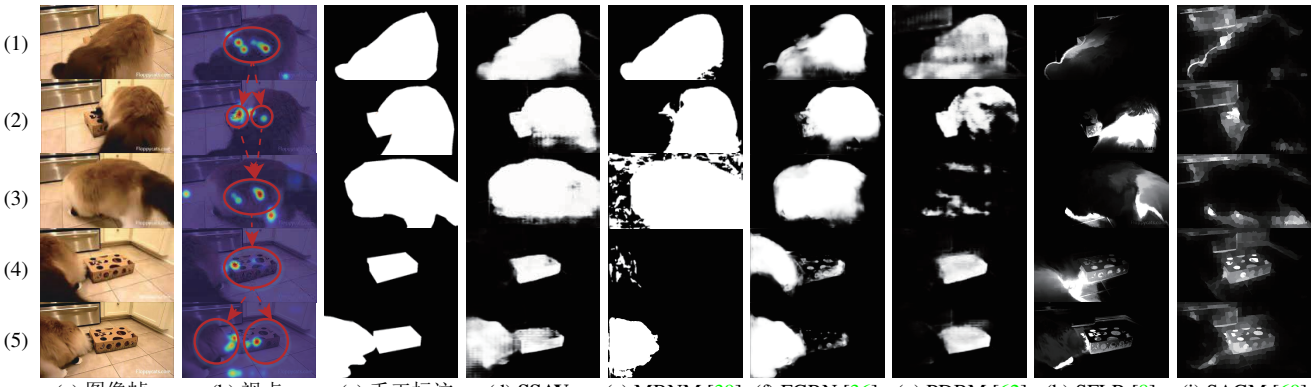


Figure 6: **DAVSOD数据集上深度模型前3名(MBNM [39]，FGRN [36]，PDBM [62])与传统模型前2名(SFLR [8]，SAGM [69])的视觉结果比较。** 我们的SSAV模型成功捕捉了显著性转移现象。

经典方法中最好的模型(SFLR [8])在*MCL*、*UVSD*、*ViSal*及*DAVSOD*数据集上的性能比某些深度模型，如SCOM [11] 的性能更好。说明在深度学习架构中研究如何有效利用人的先验知识是很有前景的方向。

**数据集分析.** 在表4中，我们用灰色标记分数，较暗的颜色意味着特定指标（如，max $\mathcal{F}$，$\mathcal{S}$以及$\mathcal{M}$）具有更好性能。我们发现*ViSal*和*UVSD*数据集相对容易，因为排名前2的模型：SSAV和PDBM [62]获得了非常高的性能（$\mathcal{S}>0.9$）。但是，对于像*DAVSOD*这样更具挑战性的数据集，VOSD模型的性能会急剧下降($\mathcal{S}<0.73$)。这揭示了VOSD模型的整体和单独性能在未来的研究中都还有很大的提升空间。

**运行时间分析.** 表2列出了当前VSOD方法和本文提出的SSAV方法的运行时间（PCF那一列）。对已经公布代码的模型，其测试时间是在相同的硬件平台：Intel Xeon(R) E5-2676v3 @2.4GHz×24 、GTX TITAN X上测试的。其余模型的测试时间则是从原论文中摘录的。注意到，本文提出的模型没有应用任何前/后处理（例如，CRF）算法，因此处理速度仅需约0.05秒。

### 5.3. 分离实验

**隐式和显式显著性转移感知注意机制.** 为了研究所提出的SSAA模块的不同训练策略的影响，我们导出2个基线：显式和隐式，对应于所提出的SSAV模型以显式和隐式方式进行训练。隐式基线训练时，我们采用的是VSOD中对象级标注而没有使用*DAVSOD*数据集中的注意视点标注。由表5可知，SSAV模型采用显式训练方式优于隐式训练。这表明利用眼动数据有助于SSAV模型更好地捕获显著性转移现象，从而进一步提高最终的VSOD性能。

| Type | Baseline | $\mathcal{S}\uparrow$ | max $\mathcal{F}\uparrow$ | $\mathcal{M}\downarrow$ |
|---|---|---|---|---|
| SSAA | *explicit* | **0.724** | **0.603** | **0.092** |
| | *implicit* | 0.684 | 0.593 | 0.103 |
| SSLSTM | *w/o SSLSTM* | 0.667 | 0.541 | 0.132 |

Table 5: **SSAV模型在*DAVSOD*数据集上的分离实验。**

**显著性转移感知convLSTM的有效性.** 为了研究SSLSTM(§ 4)的有效性，我们提供了另一个基线：*w/o SSLSTM*，即从SSAV模型中去掉SSLSTM模块。从表5中发现，基线的性能有所下降($\mathcal{S}$ : 0.724 → 0.667)。这证实了所提出的SSLSTM模块能从具有挑战性的数据中有效地学习到选择性注意力分配和注意力转移。

**与最先进的模型比较.** 表4列出了所提出的SSAV模型与当前最先进的17种VSOD算法的性能。本文的基线SSAV模型性能在大多数数据集上表现比其他模型更好。具体而言，我们的模型在*ViSal*和*FBMS*数据集上的性能得到了显著提高。而在*VOS*，*SegV2*和*DAVIS*数据集上获得了相当的性能。至于更具有挑战性的*DAVSOD*数据集，SSAV模型也获得了最佳性能。我们将这些表现出色的性能归功于SSLSTM的引入，它有效地学习了动态场景中的显著性分配机制，并指导模型准确地处理那些视觉上重要的区域。

图.6表明，与其他最先进的算法相比，本文的SSAV方法获得的视觉效果更为理想。SSAV模型成功捕获了显著性转移现象(从第1到第5帧：猫→[猫,盒子]→猫→盒子→[猫,盒子])。然而，其它高性能的VSOD模型要么无法突显整个显著对象（例如，SFLR，SAGM），要么仅捕获到移动的猫（例如，

MBNM）。我们希望本文提出的基础模型能为模型的开发开辟光明的前景。

## 6. 结论

通过构建新的、与视觉注意力相一致性*DAVSOD*的数据集，建立最大规模的评测，并提出SSAV基础模型，本文呈现了VSOD领域最全面的调研。相比传统或深度学习模型，所提出的SSAV模型获得了卓越的性能并且得到了视觉上更友好的结果。大量实验证明，即使考虑到性能最佳的模型，VSOD问题似乎还远未解决。上述努力以及深入的分析将有利于该领域的发展，并激发更广泛的潜在研究，例如，基于显著性感知的视频字幕，视频显著对象感数和实例级VSOD等。

## References


[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009. 7

[2] Tariq Alshawi, Zhiling Long, and Ghassan AlRegib. Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection. *IEEE TIP*, pages 2818–2827, 2018. 3

[3] Çağlar Aytekin, Horst Possegger, Thomas Mauthner, Serkan Kiranyaz, Horst Bischof, and Moncef Gabbouj. Spatiotemporal saliency estimation by spectral foreground detection. *IEEE TMM*, 20(1):82–95, 2018. 3

[4] Saumik Bhattacharya, K Subramanian Venkatesh, and Sumana Gupta. Visual saliency detection using spatiotemporal decomposition. *IEEE TIP*, 27(4):1665–1675, 2018. 3

[5] Marc D. Binder, Nobutaka Hirokawa, and Uwe Windhorst, editors. *Gaze Shift*, pages 1676–1676. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 1

[6] Ali Borji, Dicky N Sihite, and Laurent Itti. What stands out in a scene? A study of human explicit saliency judgment. *Vision Research*, 91:62–77, 2013. 1

[7] Chenglizhao Chen, Shuai Li, Hong Qin, Zhenkuan Pan, and Guowei Yang. Bi-level feature learning for video saliency detection. *IEEE TMM*, 2018. 1, 3

[8] Chenglizhao Chen, Shuai Li, Yongguang Wang, Hong Qin, and Aimin Hao. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE TIP*, 26(7):3156–3170, 2017. 2, 3, 7, 8

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 5

[11] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. Scom: Spatiotemporal constrained optimization for salient object detection. *IEEE TIP*, 27(7):3345–3357, 2018. 2, 3, 7, 8

[12] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 3, 6

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 3

[14] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *IEEE ICCV*, pages 4548–4557, 2017. http://dpfan.net/smeasure/. 3, 7

[15] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE CVPR*, pages 8554–8564, 2019. 1

[16] Yuming Fang, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin. A video saliency detection model in compressed domain. *IEEE TCSVT*, 24(1):27–38, 2014. 3

[17] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE TIP*, 23(9):3910–3921, 2014. 3

[18] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, pages 524–531, 2005. 3

[19] Ken Fukuchi, Kouji Miyazato, Akisato Kimura, Shigeru Takagi, and Junji Yamato. Saliency-based video segmentation with graph cuts and sequentially updated priors. In *ICME*, pages 638–641, 2009. 3

[20] Siavash Gorji and James J Clark. Going From Image to Video Saliency: Augmenting Image Salience With Dynamic Attentional Push. In *IEEE CVPR*, pages 7501–7511, 2018. 1

[21] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE CVPR*, pages 1–8, 2008. 3

[22] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE TIP*, 19(1):185–198, 2010. 1

[23] Fang Guo, Wenguan Wang, Jianbing Shen, Ling Shao, Jian Yang, Dacheng Tao, and Yuan Yan Tang. Video saliency detection using object proposals. *IEEE Transactions on Cybernetics*, 2017. 3

[24] Hadi Hadizadeh and Ivan V Bajic. Saliency-aware video compression. *IEEE TIP*, 23(1):19–33, 2014. 1

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 6

[26] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G. Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*. Springer, 2018. 1, 3

[27] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *IEEE CVPR*, pages 7142–7150, 2018. 3

[28] Yeong Jun Koh, Young-Yoon Lee, and Chang-Su Kim. Sequential clique optimization for video object segmentation. In *ECCV*. Springer, 2018. 3

[29] Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949. 5

[30] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE TIP*, 24(8):2552–2564, 2015. 1, 2, 3, 4, 5, 7

[31] Wonjun Kim, Chanho Jung, and Changick Kim. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE TCSVT*, 21(4):446–456, 2011. 1, 3

[32] Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. 1987. 1, 4

[33] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, 2017. 1, 3

[34] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE TIP*, 27(10):5002–5015, 2018. 1, 3

[35] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE ICCV*, pages 2192–2199, 2013. 1, 2, 3, 4, 5, 7

[36] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *IEEE CVPR*, pages 3243–3252, 2018. 2, 3, 7, 8

[37] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *IEEE CVPR*, pages 478–487, 2016. 3

[38] Jia Li, Changqun Xia, and Xiaowu Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE TIP*, 27(1):349–364, 2018. 1, 2, 4, 5, 7

[39] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C.-C. Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*. Springer, 2018. 2, 3, 7, 8

[40] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *IEEE CVPR*, pages 280–287, 2014. 1

[41] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE CVPR*, pages 2359–2367, 2017. 3

[42] Yong Li, Bin Sheng, Lizhuang Ma, Wen Wu, and Zhifeng Xie. Temporally coherent video saliency using regional dynamic contrast. *IEEE TCSVT*, 23(12):2067–2076, 2013. 3

[43] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *IEEE CVPR*, pages 3367–3375, 2015. 3

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 3

[45] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *IEEE CVPR*, pages 3089–3098, 2018. 3

[46] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to Detect A Salient Object. In *IEEE CVPR*, pages 1–8, 2007. 3

[47] Zhi Liu, Junhao Li, Linwei Ye, Guangling Sun, and Liquan Shen. Saliency detection for unconstrained videos us-

ing superpixel-level graph and spatiotemporal propagation. *IEEE TCSVT*, 27(12):2527–2542, 2017. 1, 2, 3, 4, 5, 7

[48] Zhi Liu, Xiang Zhang, Shuhua Luo, and Olivier Le Meur. Superpixel-based spatiotemporal saliency detection. *IEEE TCSVT*, 24(9):1522–1540, 2014. 2, 3, 7

[49] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, pages 3431–3440, 2015. 3

[50] Thomas Mauthner, Horst Possegger, Georg Waltner, and Horst Bischof. Encoding based saliency detection for videos and images. In *IEEE CVPR*, pages 2494–2502, 2015. 3

[51] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2014. 1, 2, 3, 4, 5, 7

[52] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, pages 6504–6512, 2017. 1

[53] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 7

[54] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE CVPR*, pages 724–732, 2016. 1, 2, 3, 4, 5, 7

[55] Matthew S Peterson, Arthur F Kramer, and David E Irwin. Covert shifts of attention precede involuntary eye movements. *Perception & Psychophysics*, 66(3):398–405, 2004. 1, 4

[56] Wenliang Qiu, Xinbo Gao, and Bing Han. Eye fixation assisted video saliency detection via total variation-based pairwise interaction. *IEEE TIP*, pages 4724–4739, 2018. 1, 3

[57] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting salient objects from images and videos. In *ECCV*, pages 366–379. Springer, 2010. 2, 3, 7

[58] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, 2009. 3

[59] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *CVPR*, pages 3241–3250, 2015. 3

[60] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. 5

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[62] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Sheng, and Kin-Man Lam. Pyramid dilated deeper convLSTM for video salient object detection. In *ECCV*. Springer, 2018. 2, 3, 5, 6, 7, 8

[63] Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen, Yang Hua, and Xia Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE TCSVT*, 2018. 2, 3, 7

[64] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. 3

[65] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *IEEE CVPR*, pages 2334–2342, 2016. 2, 3, 7

[66] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *IEEE CVPR*, pages 3127–3135, 2018. 3

[67] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *IEEE CVPR*, pages 1711–1720, 2018. 3

[68] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *IEEE CVPR*, pages 4894–4903, 2018. 4

[69] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *IEEE CVPR*, pages 3395–3402, 2015. 1, 2, 3, 7, 8

[70] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE TIP*, 24(11):4185–4196, 2015. 1, 2, 3, 4, 5, 7

[71] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2018. 2, 3, 7

[72] Wenguan Wang, Jianbing Shen, Hanqiu Sun, and Ling Shao. Video co-saliency guided co-segmentation. *IEEE TCSVT*, 28(8):1727–1736, 2018. 3

[73] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, pages 20–33, 2018. 3

[74] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42. Springer, 2012. 3

[75] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419, 1989. 3

[76] Tao Xi, Wei Zhao, Han Wang, and Weisi Lin. Salient object detection with spatiotemporal background priors for video. *IEEE TIP*, 26(7):3425–3436, 2017. 2, 3, 7

[77] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *IEEE CVPR*, pages 373–381, 2016. 1

[78] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE CVPR*, pages 3166–3173, 2013. 3

[79] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

[80] Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, and Ali Borji. Learning to promote saliency detectors. In *IEEE CVPR*, pages 1644–1653, 2018. 3

[81] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *IEEE ICCV*, pages 1404–1412, 2015. 2, 3, 7

[82] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *IEEE CVPR*, pages 9029–9038, 2018. 3

[83] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *IEEE CVPR*, pages 1741–1750, 2018. 3

[84] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *IEEE CVPR*, pages 714–722, 2018. 3

[85] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *IEEE CVPR*, pages 669–677, 2016. 1

[86] Feng Zhou, Sing Bing Kang, and Michael F Cohen. Time-mapping using space-time saliency. In *IEEE CVPR*, pages 3358–3365, 2014. 2, 3, 7

[87] Xiaofei Zhou, Zhi Liu, Chen Gong, and Wei Liu. Improving video saliency detection via localized estimation and spatiotemporal refinement. *IEEE TMM*, pages 2993–3007, 2018. 3