arXiv:2302.01872v1 [cs.CV] 3 Feb 2023

MOSE：复杂场景下视频目标分割的新数据集

Henghui Ding[1] Chang Liu[1] Shuting He[2] XudongJiang[1] Philip H.S.Torr[3] SongBai[4][1]南洋理工大学 [2]浙江大学 [3]牛津大学 [4]字节跳动https://henghuiding.github.io/MOSE



图 1. 来自MOSE 数据集的视频片段示例。这些视频来自复杂视频Object SEgmentation (MOSE) 数据集。选定的目标物体以橙色 ■掩码标示。MOSE 最显著的特点是场景复杂，包括物体消失后重新出现、小型/不显眼目标、严重遮挡、拥挤环境等。例如，第 2 行中的目标球员在第 3 列消失后，于第 4 列和第 5 列重新出现时转过身来，这给重新识别他带来了挑战。MOSE 中的大多数视频都包含拥挤且被遮挡的物体，而目标对象很少是显著的。MOSE 数据集旨在提供一个平台，推动更全面、更稳健的视频目标分割算法的发展。

摘要

视频目标分割（*VOS*）旨在对整个视频片段序列中的特定物体进行分割。最先进的视频目标分割方法在现有数据集上取得了出色的性能（例如，***90+%*** $\mathcal{J}\&\mathcal{F}$）。然而，由于这些现有数据集中的目标物体通常较为显著、突出且孤立，因此复杂场景下的视频目标分割很少受到研究。为了重新审视视频目标分割并使其更适用于现实世界，我们收集了一个新的视频目标分割数据集，称为 *coMplex video ObjectSEgmentation*（*MOSE*），用于研究复杂环境中的物体跟踪与分割。*MOSE* 包含来自 *36* 个类别的 *2,149* 个视频片段和 *5,200* 个物体，并配有 *431,725* 个高质量物体分割掩码。*MOSE* 数据集最显著的特点是包含拥挤且存在遮挡物体的复杂场景。视频中的目标物体通常会被其他物体遮挡，并在某些视频帧中消失。为分析所提出的 *MOSE* 数据集，我们在该数据集上针对 *4* 种不同设置对 *18* 种现有视频目标分割方法进行了基准测试，并开展了全面比较。实验表明，当前的视频目标分割算法无法很好地感知复杂场景中的物体。例如，在半监督视频目标分割设置下，现有最先进的视频目标分割方法所取得的最高 $\mathcal{J}\&\mathcal{F}$ 性能仅为 ***59.4%***，在 *MOSE* 上远低于其 *~90%* $\mathcal{J}\&\mathcal{F}$ 在 *DAVIS* 上的性能。结果表明，尽管现有基准上已经取得了出色的性能，但复杂场景下仍存在尚未解决的挑战，未来还需要投入更多努力来探索这些挑战。所提出的 *MOSE* 数据集已发布于 *https://henghuiding.github.io/MOSE*。

✉henghui.ding@gmail.com

1. 引言

视频目标分割（VOS） [1, 2, 3] 旨在对整个视频序列中的特定物体进行分割，例如，主要目标或用户指定的物体。这是最基本且最具挑战性的计算机视觉任务之一，在许多涉及视频分析与理解的实际应用中发挥着至关重要的作用，例如，自动驾驶车辆、增强现实、视频编辑、等等。视频目标分割有许多不同的设置，例如半监督视频目标分割 [4, 5]，该设置提供目标对象的首帧掩码；无监督视频目标分割 [6, 7]，该设置会自动

寻找主要目标，以及依赖用户与目标对象进行交互的交互式视频目标分割 [8, 9] 。过去，人们使用传统技术或深度学习方法对视频目标分割进行了广泛研究。尤其是，基于深度学习的方法大幅提升了视频目标分割的性能，并以显著优势超越了传统技术。

当前最先进的方法在两个最常用的视频目标分割数据集 DAVIS [1, 2]和 YouTube-VOS [3]上取得了极高的性能。例如，XMem [10] 在 **DAVIS 上的表现达到** $\mathcal{J}\&\mathcal{F}$ 92.0% 2016 [1]，**在 DAVIS 上达到** $\mathcal{J}\&\mathcal{F}$ 87.7% 2017 [2], ，并且在 **YouTube-VOS 上达到** $\mathcal{G}$ 86.1% [3]。凭借如此出色的性能，视频目标分割似乎已经得到了很好的解决。然而，在真实场景中，我们真的能感知到物体吗？为回答这一问题，我们重新审视真实且更加复杂场景下的视频目标分割。现有数据集中的目标物体 [1, 2, 3] 通常显著且占主导地位。在真实世界环境中，孤立且显著的物体很少出现，而复杂和遮挡场景却频繁发生。为了评估当前最先进的视频目标分割方法在更复杂场景下的表现，我们收集了 2,149 个复杂场景视频，并构建了一个新的、大规模且具有挑战性的视频目标分割基准，名为复**杂**场景视频**目**标**分**割（MOSE）。具体而言，MOSE 包含来自 36 个类别的 5,200 个物体，以及 431,725 个高质量分割掩码。如图 1所示，MOSE 最显著的特点是环境复杂，包括物体消失后重新出现、小型或不显眼目标、严重遮挡、拥挤环境等。例如，图 1第一行中的白色轿车被公共汽车遮挡，而第三幅图中的严重遮挡使轿车完全消失。在图1第二行中，人群中的目标运动员并不显眼，并因受到人群遮挡而在第三帧中消失。当目标运动员再次出现时，他转过身来，外观与前两帧不同，因此极难跟踪。复杂场景下物体的严重遮挡和消失给视频目标分割带来了巨大挑战。我们希望推动复杂环境下的视频目标分割研究，并使视频目标分割能够应用于现实世界。

为了分析所提出的 MOSE 数据集，我们在 MOSE 上重新训练并评估了一些现有的视频目标分割方法。具体而言，我们在半监督设定下重新训练了 6 种最先进方法，并使用掩码作为首帧参考；在半监督设定下重新训练了 2 种方法，并使用边界框作为首帧参考；在多物体零样本视频目标分割设定下重新训练了 3 种方法；在交互式设置下重新训练了 7 种方法。实验结果表明，复杂场景中的视频会削弱当前最先进视频目标分割方法的表现，尤其是在跟踪因遮挡而暂时消失的物体方面。例如， $\mathcal{J}\&\mathcal{F}$ XMem 的表现 [10] 在 DAVIS 2016 上为**92.0%** ，但在 MOSE 上降至**57.6%** ； $\mathcal{J}\&\mathcal{F}$ DeAOT 的表现 [11] 在 DAVIS 2016 上为**92.9%** ，但降至**59.4%** 在 MOSE 上，这持续揭示了复杂场景带来的困难。

MOSE 上的表现不佳，不仅是因为静态图像中的遮挡、人群和小物体，还因为物体在时间域中会消失后重新出现并发生闪烁。严重遮挡、人群和小物体给图像中的物体分割带来了挑战，而物体消失后重新出现则使跟踪被遮挡物体变得极其困难，增加了关联的难度。

总而言之，我们的主要贡献如下：

- 我们构建了一个新的视频目标分割基准数据集，名为 MOSE（复杂视频目标分割）。MOSE 致力于理解复杂环境下视频中的物体。
- 我们在 MOSE 数据集上，对最先进的视频目标分割方法进行了全面比较和评估，涵盖 4 种不同设置，包括掩码初始化半监督、边界框初始化半监督、无监督和交互式设置。
- 通过深入考察 MOSE，我们分析了复杂场景下未来视频理解研究所面临的挑战和潜在方向。

2. 相关工作

2.1. 视频目标分割（VOS）

视频目标分割（VOS）旨在对整个视频片段序列中的特定目标进行分割。根据指定特定目标的方式，主要有四种不同的设置：即., 半监督视频目标分割（或半自动视频目标分割 [12]或单次视频目标分割）、无监督视频目标分割（或自动视频目标分割 [12] 或零样本视频目标分割）、交互式视频目标分割和指代表达视频目标分割。

•**半监督视频目标分割**。半监督视频目标分割（或单次视频目标分割） [4] 以首帧目标掩码作为输入，并要求在剩余视频帧中分割目标对象。现有的大多数方法可分为基于传播的方法 [3, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]和基于匹配的方法 [19, 26, 27, 28, 29, 30, 31, 32,33, 34]。基于传播的方法利用前一视频帧的掩码来引导当前视频帧的掩码生成，逐帧传播目标对象的线索。基于匹配的方法则记忆目标对象的嵌入，然后进行逐像素分类，以衡量每个像素的特征与目标对象嵌入之间的相似度。

由于难以获取逐像素掩码，一些半监督视频目标分割工作提出使用**边界框**作为首帧线索来指示目标对象 [35, 36,37]。例如，SiamMask [35]在全卷积孪生目标跟踪器上应用掩码预测分支，以生成二值分割掩码。

• **交互式视频目标分割**。交互式视频目标分割是一种特殊形式的半监督视频目标分割 [8, 9, 19, 38, 39, 40, 41, 42, 43], , 旨在根据用户交互（例如点击或涂鸦）分割视频中的目标对象。现有的交互式视频目标分割主要遵循交互传播范式。除了用于提取具有判别性的像素特征的特征编码器外，特征编码器上还设置了另外两个模块，以实现交互式视频目标分割，即基于用户交互修正预测结果的交互式分割模块，以及将用户修正后的掩码传播到其他帧的掩码传播模块。

• **指代表达视频目标分割**。指代表达视频目标分割 [44, 45, 46, 47, 48, 49, 50] 是一种新兴设置，涉及多模态信息。它通过自然语言表达指示目标对象，旨在对视频片段中的目标对象进行全程分割。现有的指代表达视频目标分割方法可分为两类：自底向上方法和自顶向下方法。自底向上方法 [44, 51, 52]直接在第一帧分割目标对象，并将掩码传播到其余帧；或者对每帧独立进行图像分割，然后关联这些掩码。自顶向下方法 [53, 54, 55]穷举提出所有可能的轨迹片段，并选出与语言表达最匹配的轨迹片段作为输出。

• **无监督视频目标分割**。 它也称为自动视频目标分割或零样本视频目标分割 [56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66]。与上述视频目标分割设置不同，无监督视频目标分割不需要任何人工线索来指示物体，而是旨在自动找出视频中的主要物体。无监督视频目标分割只能处理预定义类别中的物体。早期方法通常采用后处理步骤 [57]。随后，端到端训练方法成为无监督视频目标分割的主流，可分为局部内容编码和上下文内容编码。局部内容编码方法 [6,7, 63, 65, 67, 68, 69] 通常采用两个并行网络，分别从光流和RGB图像中提取特征。上下文内容编码方法 [70, 71, 72] 则专注于捕获长期全局信息。

2.2. 相关视频分割任务

还有一些与视频目标分割相关的视频分割任务，例如视频实例分割、视频语义分割和视频全景分割。

• **视频实例分割（VIS）**。Yang等人将图像实例分割扩展为视频实例分割。 [73], 该任务同时对视频中预定义类别的实例进行检测、分割和跟踪。得益于大规模 VIS 数据集 YouTube-VIS [73], , 一系列学习方法相继问世，极大提升了 VIS 的性能 [74,75]。随后， [76] 提出了遮挡视频实例分割，以研究遮挡场景下的视频实例分割。与 [76], 类似，我们研究遮挡等复杂场景下的视频理解；但不同于 [76], , 我们专注于视频目标分割（VOS），且我们提出的 MOSE 数据集包含更多视频和更复杂的场景，尤其是在物体消失后重新出现方面。

• **视频语义分割（VSS）**。受图像语义分割 [77, 78, 79, 80, 81, 82] 以及大规模视频语义分割数据集 [83, 84, 85], 的巨大成功推动，视频语义分割受到广泛关注并取得了显著进展。与图像领域相比，时间一致性和模型效率是视频领域的新努力方向。例如，Sun 等人 [86]提出了由粗到细特征挖掘（CFFM），以捕获静态上下文和运动上下文。

• **视频全景分割（VPS）**。Kim 等人 [87]将全景分割引入视频领域，以同时分割和跟踪前景实例物体与背景填充区域。他们构建了一个包含 124 个视频的视频全景分割数据集。随后，Miao 等人 [88] 构建了一个规模更大的视频全景分割数据集，名为 VIPSeg，包含 3,536 个视频。现有方法 [89, 90] 主要是在图像全景分割模型上添加时间细化或跨帧关联模块，以提升时间一致性和实例跟踪性能。

2.3. 复杂场景理解

复杂场景理解已成为图像理解领域的研究焦点 [91, 92, 93, 94, 95,96, 97, 98, 99, 100]。例如，Ke等人。 [101] 提出了双层卷积网络（BCNet），将重叠物体解耦为遮挡物层和被遮挡物层。Zhang等人。 [94] 提出了一种自监督方法，通过顺序恢复、整体补全和内容补全来实现去遮挡。然而，在视频领域，遮挡理解仍未得到充分探索，目前只有少数多目标跟踪研究涉及这一问题 [102, 103, 104, 105]。例如，Chu 等人。 [102] 提出了时空注意力机制（STAM），以捕捉目标的可见部分，并处理遮挡带来的漂移。Zhu 等人。 [103]提出了双重匹配注意力网络（DMAN），以应对多目标跟踪中的噪声遮挡。Li等人。 [106] 提出在开放世界中跟踪一切

表 1. MOSE 数据集中所有物体类别及其实例数的完整列表。物体类别按出现频率降序排列。

类别	编号		编号				No.
其他人物	519	猴子	181	自行车	130	船	73
Fish	343		175 骑		116		68
马	272	船	166	老虎	108	鸭子	64
绵羊	264		160		108		54
斑马	254	摩托车骑手	153	熊猫	106	骑马者	46
兔子	237		142		90 公		29
Bird	226	猫	139	飞机	89	卡车	15
大象	207		139		86		
摩托车	192	奶牛	139	熊	84	其他家禽	12

通过执行类别无关关联。在这项工作中，我们构建了一个复杂视频目标分割数据集，以支持未来在视频目标分割中开展复杂场景理解的研究。

3. MOSE 数据集

在本节中，我们介绍新建的 MOSE 数据集。首先在第3.1节介绍视频收集和标注流程，然后在第3.2节给出数据集统计和分析。最后，在第3.3节介绍评估指标。

3.1. 视频收集与标注

视频收集。 MOSE 的视频来自两部分，一部分继承自 OVIS [76], ，OVIS 专为视频实例分割设计；另一部分是从现实场景中新拍摄的视频。我们选择 OVIS，是因为其中存在大量严重遮挡，符合复杂场景的要求。对于继承自 OVIS 的视频，由于许多物体首次出现于非首帧，因此无法直接用于需要首帧参考的视频目标分割。一种解决方案是丢弃所有未在首帧出现的物体，但这会浪费大量掩码标注。为了使 OVIS 视频适用于视频目标分割任务并充分利用这些视频，我们根据每个物体在视频中首次出现的视频帧，将每段视频剪切成数段，然后丢弃不符合要求的视频。对于新拍摄的视频，我们首先选取一组在现实世界中常见的语义类别，包括车辆（例如，自行车、公共汽车、飞机、船）、动物（例如，鸟、熊猫、狗、长颈鹿），以及从事不同活动的人类（例如，骑摩托车、驾驶、骑行、跑步），然后收集/拍摄包含这些类别的视频，拍摄地点包括校园、动物园、室内场所、城市街道等。MOSE 数据集中物体类别的完整列表见表 1。这些类别在日常生活中很常见，人群和遮挡现象也频繁出现。此外，它们还包含在 MS-COCO 等流行的大规模图像分割数据集中 [111],，因此很容易使用图像预训练模型。

我们的重点是复杂场景下的视频目标分割，尤其是包含拥挤目标和被遮挡物体的场景。为确保收集/拍摄的视频中包含拥挤和被遮挡物体，我们制定了以下几条规则：

R1. 每段视频必须包含多个物体，只有单个物体的视频则不予收录。尤其欢迎包含外观相似的拥挤目标的视频。

R2. 视频中应存在遮挡。整段视频的所有帧中均无任何遮挡的视频将被舍弃。我们鼓励收录由其他运动目标造成遮挡的视频。

R3. 应特别关注物体因遮挡或拥挤而消失后又重新出现的场景。

R4. 目标物体应涵盖多种尺度和类型，例如，小尺度或大尺度、显眼或不显眼的物体。

R5. 视频中的物体必须表现出足够的运动。会剔除物体完全静止或几乎没有运动的视频。

除上述规则外，为保证视频质量，我们要求视频分辨率为 1920×1080，视频长度通常为 5 至 60 秒，以确保视频质量。

视频标注。 收集完 MOSE 的所有视频后，我们的研究团队会逐一查看，并为每段视频确定一组感兴趣目标。随后，我们会对视频的开头和结尾进行适当剪辑，以减少视频中运动较少或较简单的视频帧数量。接着，我们标注目标物体的首帧掩码，作为视频目标分割的输入。之后，我们将视频连同首帧掩码一起发送给标注团队，以标注后续视频帧。

标注团队需要以给定的首帧掩码为参考，在其中识别目标物体，然后跟踪并标注首帧之后所有帧中该目标物体的分割掩码。借助我们开发的交互式标注工具，视频标注过程变得更加轻松

表2. MOSE 与现有视频目标分割数据集的规模比较。“标注”表示已标注的目标掩码数量。“时长”表示已标注视频的总时长（分钟）。“平均边界框遮挡率”表示边界框遮挡率的平均值。“消失率”表示至少在一个视频帧中消失的物体所占的频率。新建的 MOSE 视频时长最长，标注数量最多。更重要的是，MOSE 最显著的特点是包含大量人群、遮挡和消失后重现的目标，为视频目标分割提供了复杂得多的场景。

数据集	Year	视频	类别	物体	标注	时长	mBOR	消失率
YouTube-Objects [107]	2012	96	10	96	1,692	9.01	-	-
SegTrack-v2 [108]	2013	14	11	24	1,475	0.59	0.12	8.3%
FBMS [109]	2014	59	16	139	1,465	7.70	0.01	11.2%
JumpCut [110]	2015	22	14	22	6,331	3.52	0	0%
$DAVIS_{16}$[1]	2016	50	-	50	3,440	2.88	-	-
$DAVIS_{17}$[2]	2017	90	-	205	13,543	5.17	0.03	16.1%
YouTube-VOS [3]	2018	4,453	94	7,755	197,272	334.81	0.05	13.0%
MOSE （我们的方法）	**2023**	**2,149**	**36**	**5,200**	**431,725**	**443.62**	**0.23**	**28.8%**

。该标注工具会自动加载视频和所有目标物体。标注人员使用该工具加载并预览视频和首帧掩码，对后续帧中的分割掩码进行标注和可视化，并将其保存。标注工具还内置了交互式物体分割网络 XMem [10]，，以协助标注人员生成高质量掩码。为确保复杂场景下的标注质量，标注人员必须清晰跟踪因严重遮挡或人群遮挡而消失又重新出现的物体。对于目标物体消失或完全被遮挡的帧，标注人员还需要确认这些帧的输出掩码为空。我们要求所有视频至少每五帧进行一次标注。为测试模型的帧率鲁棒性，部分视频会逐帧标注。

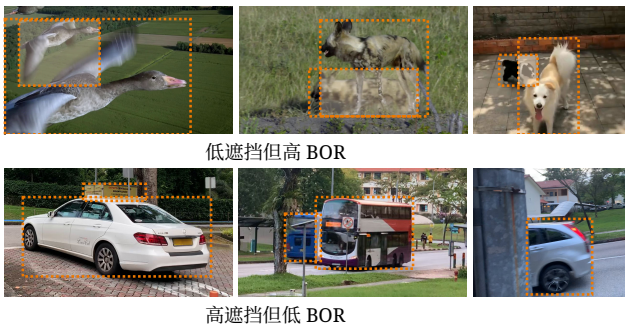


图2. BOR 指标的失败案例。从第一行样本可以看出，它们的 BOR 值较高，但存在的遮挡较少或没有遮挡。第二行的样本 BOR 值很小，但样本中存在严重遮挡。

标注完成后，视频会发送给验证团队，以核验各方面的标注质量，尤其是存在遮挡、人群或目标物体消失后重新出现的视频的标注。

3.2. 数据集统计

在表2中，我们以此前的视频目标分割数据集作为分析参考，对新 MOSE 数据集的数据统计进行了分析，包括 JumpCut [110], SegTrack-v2 [108], YouTube-Objects [107], FBMS [109], DAVIS [1, 2], 和 YouTube-VOS[3]。如表2所示，提出的 MOSE 数据集包含 2,149 个视频、5,200 个物体的 431,725 个掩码标注。与此前规模最大的视频目标分割数据集 YouTube-VOS 相比，MOSE 的掩码标注数量多得多，达到 431k 而197k。在表2中分析的 8 个视频目标分割数据集中，MOSE 的视频时长最长（433.62 分钟），比时长排名第二的 YouTube-VOS（334.81 分钟）长约 100 分钟，也比其他七个数据集长数百倍。与此前的视频目标分割数据集相比，我们在 MOSE 中纳入了更多长视频，以确保 MOSE 数据集中包含多种遮挡场景、运动场景，以及物体消失后重新出现的情形。

消失与遮挡分析。 以往的被遮挡视频数据集 OVIS [76] 定义了边界框遮挡率（BOR），用于反映遮挡程度。BOR 根据视频中边界框的交并比计算得出。表 2 中列出了所有帧的平均 BOR。MOSE 的平均边界框遮挡率最高，表明 MOSE 中发生遮挡的频率高于以往的数据集。然而，我们发现 BOR 只能粗略反映遮挡情况，无法充分揭示 MOSE 数据集中遮挡的程度，而且在这方面可能容易产生误判。如图 2 第一行所示，三张遮挡程度较低甚至没有遮挡的图像中，BOR 却很高。相比之下，第二行的图像中，

存在严重遮挡物体时，BOR 却很低，甚至为 0。

因此，除了统计 BOR 外，我们还计算视频中至少有一帧出现物体消失的物体数量，并进一步计算反映消失频率的消失率，以说明数据集中复杂场景导致物体消失的频繁程度。MOSE 中出现消失情况的物体数量为 1,553，远高于以往的数据集。如表 2所示，MOSE 的消失率（Disapp. Rate）最高，达到 28.8%，这表明在 MOSE 中，物体消失和遮挡现象频繁且严重。

3.3. 评估指标

按照以往的方法 [1, 2], ，我们计算区域相似度 $\mathcal{J}$ 和轮廓准确率 $\mathcal{F}$作为评估指标。给定分割掩码预测 $\hat{M} \in \{0,1\}^{H\times W}$ 和真实掩码 $M \in \{0,1\}^{H\times W}$，区域相似度 $\mathcal{J}$通过计算$\hat{}$与 M 的交并比（IoU） M获得。

$$\mathcal{J} = \frac{\hat{M} \cap M}{\hat{M} \cup M}. \tag{1}$$

随后，计算所有物体的平均区域相似度 $\mathcal{J}_{mean}$，作为最终的区域相似度结果。为简洁起见，我们用$\mathcal{J}$ 表示 $\mathcal{J}_{mean}$。为衡量$\hat{}$ M的轮廓质量，通过二分图匹配 [112] 计算轮廓召回率 R_c 和精确率 P_c。随后，轮廓准确率 $\mathcal{F}$ 是轮廓召回率 R_c 和精确率 P_c的调和平均值，即，

$$\mathcal{F} = \frac{2\mathrm{P}_c\mathrm{R}_c}{\mathrm{P}_c + \mathrm{R}_c}, \tag{2}$$

该指标表示预测掩码的轮廓与真实掩码的轮廓有多接近。计算所有物体的平均轮廓准确率$\mathcal{F}_{mean}$，作为最终的轮廓准确率结果。为简洁起见，我们用 $\mathcal{F}$ 表示$\mathcal{F}_{mean}$。随后，使用 $\mathcal{J}\&\mathcal{F} = (\mathcal{J} + \mathcal{F})/2$ 衡量整体性能。

4. 实验

在此，我们针对四种不同的视频目标分割（VOS）设置开展实验和基准测试，包括使用掩码初始化的半监督（或单次）VOS、使用边界框初始化的半监督 VOS、无监督（或零次）VOS，以及交互式视频目标分割，以全面分析新构建的 MOSE 数据集。

实现细节。所提出的 MOSE 数据集与 YouTube-VOS [3] 格式一致。我们将以往方法使用的训练数据集从 YouTube-VOS 替换为 MOSE，并严格遵循其在 YouTube-VOS [3]上的训练设置。我们遵循 DAVIS [1, 2] 来评估性能，并报告 $\mathcal{J}_{mean}$、$\mathcal{F}_{mean}$和 $\mathcal{J}\&\mathcal{F}$ 在 MOSE 验证集上的结果。在训练这些视频目标分割方法时，我们遵循其使用图像预训练模型的方式，但不使用任何额外的视频数据集进行预训练。

表 3. 验证集上最先进的半监督方法对比。“$\mathcal{J}$”和“$\mathcal{F}$”分别表示区域相似度的平均值和轮廓准确率的平均值。 $\mathcal{J}\&\mathcal{F}$ 表示 $\mathcal{J}$和 $\mathcal{F}$的平均值。为进行公平比较，训练阶段未加入 BL30K [38]。

方法	Pub.	MOSE（我们的方法）			DAVIS17	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{G}$
AOT [113]	NeurIPS'21	53.1	61.3	57.2	84.9	84.1
STCN [114]	NeurIPS'21	46.6	55.0	50.8	85.4	83.0
RDE [115]	CVPR'22	44.6	52.9	48.8	84.2	-
SWEM [116]	CVPR'22	46.8	54.9	50.9	84.3	82.8
XMem [10]	ECCV 2022	53.3	62.0	57.6	86.2	85.7
DeAOT[11]	NeurIPS 2022	55.1	63.8	59.4	85.2	86.0

设置。 整个 MOSE 数据集包含 2,149 个视频。这些视频分别划分为 1,507 个训练视频、311 个验证视频和 331 个测试视频，用于模型训练、日常评估和竞赛期间评估。在半监督（单次）视频目标分割设置中，每个视频都提供首帧掩码或边界框作为目标对象的参考；在交互式视频目标分割设置中，则提供首帧涂鸦标注。

4.1. 半监督视频目标分割

半监督视频目标分割（又称单次视频目标分割）将目标对象首帧的掩码或边界框作为提示和参考。我们在 MOSE 数据集上训练并评估了最新的 6 种基于掩码初始化的半监督视频目标分割方法和 2 种基于边界框初始化的半监督视频目标分割方法，这些方法基于 ResNet-50 [117]，结果分别如表 3和表 4所示。我们希望这些实验能为未来开发的半监督视频目标分割算法提供基准。

掩码初始化。这是视频目标分割中的经典设置，目前也是最热门的研究方向。过去十年间，针对这一设置涌现了大量优秀的基于深度学习的方法，大幅提升了视频目标分割的性能，使其趋于饱和。例如，最新的最先进方法 DeAOT [11] 在 **DAVIS 上取得了 85.2%** $\mathcal{J}\&\mathcal{F}$ 的成绩，在 YouTube-VOS 上取得了 2017 [2], **86.0%** $\mathcal{G}$ 的成绩， [3], 并且 **92.3%**$\mathcal{J}\&\mathcal{F}$ 的成绩是在 DAVIS 2016 [1], 上取得的。这些表现十分出色，几乎达到了真实标注的水平。为了分析新建的 MOSE 数据集并测试现有方法在复杂场景中的表现，我们对这些方法进行了训练和评估

表 4. 验证集上最先进的边界框初始化半监督方法的比较。“$\mathcal{J}$”和“$\mathcal{F}$”分别表示区域相似度均值和轮廓准确率均值。 $\mathcal{J}\&\mathcal{F}$ 表示 $\mathcal{J}$和 $\mathcal{F}$的均值。

方法	Pub.	MOSE（我们的方法）			DAVIS17	YT-VOS18
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{G}$
SiamMask [35]	CVPR'19	17.3	26.7	22.0	56.4	52.8
FTMU [36]	CVPR'20	19.1	28.5	23.8	70.6	-

，并在 MOSE 上进行测试。我们报告了 6 种近期最先进方法在 MOSE 验证集上的结果，包括 AOT [113], 、STCN [114], 、RDE [115], 、SWEM [116],、XMem [10], 和 DeAOT [11], ，如表 3所示。目前最先进方法在新提出的 MOSE 数据集验证集上的表现仅为 **48.8%** $\mathcal{J}\&\mathcal{F}$ 至 **59.4%** $\mathcal{J}\&\mathcal{F}$，而它们在 DAVIS 2017 [2] 和 YouTube-VOS [3]上的结果通常高于**80%** $\mathcal{J}\&\mathcal{F}$，有些方法甚至 $\mathcal{G}$几乎达到 **90%** $\mathcal{J}\&\mathcal{F}$，或者 $\mathcal{G}$ 在 DAVIS 2017 和 YouTube-VOS 上达到这一水平。结果表明，尽管我们已在以往的基准上取得了出色的视频目标分割性能，但复杂场景下仍存在尚未解决的挑战，未来还需要投入更多精力来探索这些挑战。第 5 节将讨论 MOSE 数据集及一些潜在的未来方向。

边界框初始化。如表 4 所示，使用边界框作为首帧参考的两种半监督（单次）视频目标分割方法在 MOSE 上训练，并在 MOSE 的验证集上进行评估，包括 SimMask [35]和 FTMU [36]。当前最先进方法仅达到**22.0%** 和 **23.8%** $\mathcal{J}\&\mathcal{F}$ 在新提出的 MOSE 数据集验证集上的结果，而它们在 DAVIS 上的结果2017 [2] 已经达到 **70+%** $\mathcal{J}\&\mathcal{F}$。

经过仔细分析，我们认为性能大幅下降的原因之一是，MOSE 数据集的视频中存在许多严重遮挡、小尺度物体和拥挤场景，这使得视频的弱监督分割难度大大增加。由于严重遮挡和人群的影响，即使在目标边界框区域内，目标对象也未必是最显著的对象。物体受到遮挡后可能会被分割成若干彼此不相邻的部分，这大大增加了使用边界框驱动方法分割物体的难度。

4.2. 无监督视频目标分割

无监督视频目标分割，也称为零样本视频目标分割，不需要任何人工线索（例如，掩码或边界框）作为指示物体的参考，而是旨在自动找出视频中的主要物体。零样本 VOS 方法的主流

表 5. 在验证集上对最先进的多目标零样本视频对象分割方法进行比较。“$\mathcal{J}$”和“$\mathcal{F}$”分别表示区域相似度均值和轮廓准确率均值。 $\mathcal{J}\&\mathcal{F}$ 表示 $\mathcal{J}$和 $\mathcal{F}$的均值。

方法	Pub.	MOSE（我们的方法）			DAVIS17
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
RVOS [119]	CVPR'19	24.1	36.9	30.5	43.7
AGNN [71]	ICCV'19	38.6	48.8	43.7	61.1
STEm-Seg [118]	ECCV'20	43.3	50.5	46.9	64.7

表6. 验证集上最先进的交互式视频目标分割方法的比较。 $\mathcal{J}\&\mathcal{F}$ @60s 表示各方法在 60 秒内进行交互所达到的 $\mathcal{J}\&\mathcal{F}$ 性能。

方法	Pub.	MOSE（我们的方法）	DAVIS17
		$\mathcal{J}\&\mathcal{F}$@60s	$\mathcal{J}\&\mathcal{F}$@60s
IPNet [9]	CVPR'19	41.2	78.7
STM [120]	ICCV'19	45.3	84.8
MANet [40]	CVPR'20	43.6	79.5
ATNet [121]	ECCV'20	44.5	82.7
GIS [39]	CVPR'21	50.2	86.6
MiVOS [38]	CVPR'21	51.4	88.5
STCN [114]	NeurIPS'21	56.8	88.8

主要致力于解决单目标视频目标分割问题。然而，MOSE 是一个类似于 DAVIS 的多目标视频对象分割数据集 2017 [2],，因此我们仅在 MOSE 上评测多目标零样本视频目标分割方法，例如STEm-Seg [118], AGNN [71], 和 RVOS [119]。结果如表5所示。在此设置下，仅使用具有完整首帧标注的视频。现有方法依赖现成的、在图像上训练的实例分割方法；得益于在图像领域中对复杂场景的学习，这些方法能够很好地检测/分割静态图像中的物体。性能下降主要是由时序挑战导致的。

4.3. 交互式视频目标分割

遵循 DAVIS 2019年挑战赛的交互赛道， [122], 我们为特定视频序列中的目标对象提供初始涂鸦标注，作为首次交互。交互式视频目标分割方法需要基于首次交互预测整个视频的分割掩码。随后，通过比较视频中预测掩码与真实掩码，在效果最差的视频帧上进一步为方法提供修正涂鸦标注，以细化视频分割预测。上述步骤最多可重复 8 次，每个对象的时间限制为 30 秒。我们报告 $\mathcal{J}\&\mathcal{F}$@60s 指标，以鼓励方法在速度和准确性之间取得良好平衡。如表 6所示，我们在 MOSE 的验证集上评估了七种近期交互式视频目标分割方法。

5. 讨论与未来方向

本文讨论了所提出的 MOSE 数据集因复杂场景带来的挑战，并基于现有方法在 MOSE 上的实验结果分析，提出了一些潜在的未来方向。

• **通过更强的关联跟踪重新出现的物体**。 有必要为视频目标分割方法开发更强大的关联/重识别算法，以便能够跟踪消失后又重新出现的物体。尤其值得关注的是，我们注意到，许多消失后又重新出现的物体，其重新出现时的外观与消失时不同，即，外观变化物体。例如，图1中的目标球员消失前背对镜头，重新出现时则正面朝向镜头。面对这种外观变化，跟踪难度很大。

•**被遮挡物体的视频目标分割**。 MOSE 视频中的频繁遮挡为遮挡视频理解研究提供了数据支持。由于现有数据集中的目标物体通常显著且占据主导地位，因此针对严重遮挡情况下物体的跟踪与分割鲜少得到研究。在实际场景中，孤立物体很少出现，而遮挡场景则十分常见。得益于上下文关联和推理能力，人类能够很好地捕捉那些被遮挡物体。将对遮挡的理解引入视频目标分割过程后，视频目标分割方法在实际应用中将变得更加实用。尤其是，遮挡使边界框初始化半监督视频目标分割设置更具挑战性，因为需要根据边界框分割被遮挡物体。

• **关注小型且不显眼的物体**。尽管小目标的检测与分割是图像领域的热门课题，但在视频目标分割领域，对小型且不显眼的物体进行跟踪与分割仍有待发展。事实上，现有的大多数视频目标分割方法主要关注跟踪大型显著物体。对小尺度物体关注不足，使得视频目标分割方法在实际应用中的表现不够突出，而这些应用可能涉及尺寸和类型各异的目标物体。所提出的 MOSE 数据集中有许多物体尺寸较小，在视频中也不显眼，为研究如何在真实场景中跟踪和分割小型且不显眼的物体提供了更多机会。

• **在人群中跟踪物体**。 MOSE 最显著的特点之一是包含拥挤场景，这类场景在实际应用中十分常见。MOSE 数据集中有许多视频包含群体物体，例如一起移动的羊群、竞速的骑行者群体，在拥挤街道上行走的行人，以及等等。在这类场景中，要在人群中分割和跟踪某个物体颇具挑战，因为人群中的其他物体与目标对象具有相似的外观和运动。在图像/帧领域，视频目标分割算法需要提升识别能力，以区分外观相似的不同物体。此外，在时间域中，当人群由外观相似的物体组成时，在两帧之间进行物体关联也颇具挑战。

• **长时视频分割**：对于电影剪辑和监控等实际应用而言，长时视频理解比短时视频理解实用得多。例如，YouTube 上视频的平均时长约为十二分钟，远长于现有视频目标分割数据集中的视频平均时长。尽管现有视频目标分割方法表现出色，但通常需要大量计算资源，例如 GPU内存，用于存储此前视频帧的目标特征。由于计算成本高昂，现有的大多数视频目标分割方法都无法很好地处理长视频。MOSE 数据集中的视频平均时长长于现有视频目标分割数据集，这为处理长时视频带来了更多挑战与机遇。设计能够以较低计算成本处理长视频，同时实现高质量分割结果的视频目标分割算法，是一个值得探索且具有实用价值的方向。

6. 结论

我们构建了一个名为 MOSE 的大规模视频目标分割数据集，以重新审视并支持复杂场景下的视频目标分割研究。MOSE 包含 2,149 段高分辨率视频，其中针对 36 个类别的 5,200 个物体标注了 431,725 个高质量目标掩码。MOSE 中的视频通常足够长，能够确保包含丰富且充分的遮挡、运动以及消失后重新出现等情形。基于提出的 MOSE 数据集，我们对现有视频目标分割方法进行了基准测试和全面比较。我们在以掩码作为首帧参考的半监督（单次）设置下训练并评估了六种方法，在以边界框作为首帧参考的设置下评估了两种方法，在无监督（零次）设置下评估了三种方法，并在交互式设置下评估了七种方法。在 MOSE 数据集上评估现有视频目标分割方法并全面分析结果后，我们总结了未来视频目标分割研究面临的一些挑战和潜在方向。我们发现，在复杂场景中分割和跟踪物体的研究仍处于起步阶段；这类场景中，人群、物体消失、遮挡以及不显眼或小目标频繁出现。

参考文献

[1] Federico Perazzi、Jordi Pont-Tuset、Brian McWilliams、Luc Van Gool、Markus Gross 和 Alexander Sorkine-Hornung。《用于视频目标分割的基准数据集和评估方法》。发表于 CVPR，2016年。1、2、5、6[2] Jordi Pont-Tuset、Federico Perazzi、Sergi Caelles、Pablo Arbeláez、Alex Sorkine-Hornung 和 Luc Van Gool。《2017年 DAVIS 视频目标分割挑战赛》。arXiv 预印本 arXiv:1704.00675，2017年。1、2、5、6、7[3] Ning Xu、Linjie Yang、Yuchen Fan、Jianchao Yang、Dingcheng Yue、Yuchen Liang、Brian Price、Scott Cohen 和 Thomas Huang。《YouTube-VOS：序列到序列视频目标分割》。发表于 ECCV，2018年。1、2、5、6、7[4] Sergi Caelles、Kevis-Kokitsi Maninis、Jordi Pont-Tuset、Laura Leal-T aixé、Daniel Cremers 和 Luc Van Gool。《单次视频目标分割》。发表于 CVPR，2017年。1、2[5] Hyojin Park、Jayeon Yoo、Seohyeong Jeong、Ganesh Venkatesh 和 Nojun Kwak。《在半监督视频目标分割中利用复用门控函数学习动态网络》。发表于 CVPR，2021年。1[6] Suyog Dutt Jain、Bo Xiong 和 Kristen Grauman。《FusionSeg：学习结合运动与外观，实现视频中通用物体的全自动分割》。发表于 CVPR，2017年。1、3[7] Jingchun Cheng、Yi-Hsuan Tsai、Shengjin Wang 和 Ming-Hsuan Yang。《SegFlow：视频目标分割与光流的联合学习》。发表于 ICCV，2017年。1、3[8] Yuhua Chen、Jordi Pont-Tuset、Alberto Montes 和 Luc Van Gool。《利用逐像素度量学习实现极速视频目标分割》。发表于 CVPR，2018年。2、3[9] Seoung Wug Oh、Joon-Young Lee、Ning Xu 和 Seon Joo Kim。《通过交互与传播网络实现快速的用户引导视频目标分割》。发表于 CVPR，2019年。2、3、7[10] Ho Kei Cheng 和 Alexander G Schwing。《XMem：采用 Atkinson-Shiffrin 记忆模型的长期视频目标分割》。发表于 ECCV。施普林格，2022年。2、5、6、7[11] Zongxin Yang 和 Yi Yang。《视频目标分割层次传播中的特征解耦》。发表于 NeurIPS，2022年。2、6、7[12] Wenguan Wang、Tianfei Zhou、Fatih Porikli、David Crandall 和 Luc Van Gool。《深度学习技术在视频分割中的应用综述》。arXiv 预印本 arXiv:2107.01153，2021年。2[13] Federico Perazzi、Anna Khoreva、Rodrigo Benenson、Bernt Schiele 和 Alexander Sorkine-Hornung。《从静态图像中学习视频目标分割》。发表于 CVPR，2017年。2[14] Won-Dong Jang 和 Chang-Su Kim。《通过卷积三叉戟网络实现在线视频目标分割》。发表于 CVPR，2017年。2[15] Varun Jampani、Raghudeep Gadde 和 Peter V. Gehler。《视频传播网络》。发表于 CVPR，2017年。2[16] Huaxin Xiao、Jiashi Feng、Guosheng Lin、Yu Liu 和 Maojun Zhang。《Monet：用于视频目标分割的深度运动利用》。发表于 CVPR，2018年。2[17] Ping Hu、Gang Wang、Xiangfei Kong、Jason Kuen 和 Yap-Peng Tan。《用于视频目标分割的运动引导级联细化网络》。发表于 CVPR，2018年。2[18] Junwei Han、Le Yang、Dingwen Zhang、Xiaojun Chang 和 Xiaodan Liang。《用于视频目标分割的强化学习切割代理》。发表于 CVPR，2018年。2[19] Jingchun Cheng、Yi-Hsuan Tsai、Wei-Chih Hung、Shengjin Wang 和 Ming-Hsuan Yang。《通过部件跟踪实现快速、准确的在线视频目标分割》。发表于 CVPR，2018年。2、3[20] Shuangjie Xu、Daizong Liu、Linchao Bao、Wei Liu 和 Pan Zhou。《MHP-VOS：用于视频目标分割的多假设传播》。发表于 CVPR，2019年。2[21] Xi Chen、Zuoxin Li、Ye Yuan、Gang Yu、Jianxin Shen 和 Donglian Qi. 用于实时视频目标分割的状态感知跟踪器。载于 CVPR，2020年。 2[22] Xuhua Huang、Jiarui Xu、Yu-Wing Tai 和 Chi-Keung Tang。采用时序聚合网络和动态模板匹配的快速视频目标分割。载于 CVPR，2020年。 2[23] Seoung Wug Oh、Joon-Young Lee、Kalyan Sunkavalli 和 Seon Joo Kim。通过参考引导的掩码传播实现快速视频目标分割。载于 CVPR，2018年。 2[24] Huaijia Lin、Xiaojuan Qi 和 Jiaya Jia。Agss-vos：注意力引导的单次视频目标分割。载于 CVPR，2019年。2[25] Lu Zhang 、 Zhe Lin 、 Jianming Zhang 、 Huchuan Lu 和 You He。通过动态目标定位网络实现快速视频目标分割。载于 ICCV，2019年。 2[26] Jae Shin Yoon、François Rameau、Jun-Sik Kim、Seokju Lee、Seunghak Shin 和 In So Kweon。使用卷积神经网络进行视频目标分割的像素级匹配。载于 ICCV，2017年。 2[27] Paul Voigtlaender、Yuning Chai、Florian Schroff、Hartwig Adam、Bastian Leibe 和 Liang-Chieh Chen。Feelvos：用于视频目标分割的快速端到端嵌入学习。载于 CVPR，2019年。 2[28] Ziqin Wang、Jun Xu、Li Liu、Fan Zhu 和 Ling Shao。Ranet：用于快速视频目标分割的排序注意力网络。载于 ICCV，2019年。 2[29] Kevin Duarte、Yogesh S. Rawat 和 Mubarak Shah。Capsulevos：使用胶囊路由进行半监督视频目标分割。载于 ICCV，2019年。 2[30] Seoung Wug Oh、Joon-Young Lee、Ning Xu 和 Seon Joo Kim。使用时空记忆网络进行视频目标分割。载于 ICCV，2019年。 2[31] Yizhuo Zhang、Zhirong Wu、Houwen Peng 和 Stephen Lin。一种用于视频目标分割的传导式方法。载于 CVPR，2020年。 2[32] Zongxin Yang、Yunchao Wei 和 Yi Yang。通过前景-背景整合协同进行视频目标分割。载于 ECCV，2020年。 2[33] Li Hu、Peng Zhang、Bang Zhang、Pan Pan、Yinghui Xu 和 Rong Jin。为基于记忆的视频目标分割学习位置与目标一致性。载于 CVPR，2021年。 2[34] Brendan Duke、Abdalla Ahmed、Christian Wolf、Parham Aarabi 和 Graham W Taylor。Sstvos：用于视频目标分割的稀疏时空变换器。载于 CVPR，2021年。 2[35] Qiang Wang、Li Zhang、Luca Bertinetto、Weiming Hu 和 Philip HS Torr。快速在线目标跟踪与分割：一种统一方法。载于 CVPR，2019年。 3, 7[36] Mingjie Sun、Jimin Xiao、Eng Gee Lim、Bingfeng Zhang 和 Yao Zhao。用于视频目标跟踪与分割的快速模板匹配与更新。载于 CVPR，2020年。 3, 7[37] Fanchao Lin、Hongtao Xie、Yan Li 和 Yongdong Zhang。用于弱监督视频目标分割的查询-记忆再聚合。载于人工智能促进协会年会，2021年。3[38] Ho Kei Cheng、Yu-Wing Tai 和 Chi-Keung Tang。模块化交互式视频目标分割：交互到掩码、传播与差异感知融合。载于 CVPR，2021年。 3, 6, 7[39] Yuk Heo、Yeong Jun Koh 和 Chang-Su Kim。使用基于可靠性的注意力图进行引导式交互视频目标分割。载于 CVPR，2021年。 3, 7[40] Jiaxu Miao、Yunchao Wei 和 Yi Yang。用于高效交互式视频目标分割的记忆聚合网络。载于 CVPR，2020年。 3, 7[41] Henghui Ding、Hui Zhang、Chang Liu 和 Xudong Jiang。通过特征传播实现深度交互式图像抠图。IEEE TIP，31，2022年 3。 [42] Bowen Chen、Huan Ling、Xiaohui Zeng、Gao Jun、Ziyue Xu 和 Sanja Fidler。Scribblebox：交互式视频目标分割标注框架

视频目标分割。载于 ECCV，2020年。 3[43] Zhaoyuan Yin、Jia Zheng、Weixin Luo、Shenhan Qian、Hanling Zhang 和 Shenghua Gao。学习为野外交互式视频目标分割推荐视频帧。载于 CVPR，2021年。 3[44] Seonguk Seo、Joon-Young Lee 和 Bohyung Han。Urvos：具有大规模基准的统一指代表达视频目标分割网络。载于 ECCV，2020年。 3[45] Chang Liu、Xudong Jiang 和 Henghui Ding。用于指代分割的实例特定特征传播。IEEE多媒体汇刊，2022年。 3[46] Hao Wang、Cheng Deng、Junchi Yan 和 Dacheng Tao。用于根据自然语言查询进行人物与动作视频分割的非对称交叉引导注意力网络。载于 ICCV，2019年。3[47] Zihan Ding 、 Tianrui Hui 、 Junshi Huang 、 Xiao ming Wei、Jizhong Han 和 Si Liu。用于指代表达视频目标分割的语言桥接时空交互。载于 CVPR，2022年。 3[48] Henghui Ding、Scott Cohen、Brian Price 和 Xudong Jiang。Phraseclick：通过短语与点击实现灵活交互式分割。载于 ECCV。施普林格，2020年。 3[49] Linwei Ye、Mrigank Rochan、Zhi Liu、Xiaoqin Zhang 和 Yang Wang。通过跨模态自注意力网络对图像和视频进行指代分割。IEEE TPAMI，2021年。 3[50] Henghui Ding、Chang Liu、Suchen Wang 和 Xudong Jiang。用于指代分割的视觉语言变换器与查询生成。载于 ICCV，2021年。3[51] Henghui Ding 、 Chang Liu 、 Suchen Wang 和 Xudong Jiang。Vlt：用于指代分割的视觉语言变换器与查询生成。IEEE TPAMI，2023年。 3[52] Si Liu、Tianrui Hui、Shaofei Huang、Yunchao Wei、Bo Li 和 Guanbin Li。用于指代分割的跨模态渐进式理解。IEEE TPAMI，2021年。 3[53] Chen Liang、Yu Wu、Tianfei Zhou、Wenguan Wang、Zongxin Yang、Yunchao Wei 和 Yi Yang。从自顶向下视角重新思考指代表达视频目标分割中的跨模态交互。arXiv 预印本 arXiv:2106.01061，2021年。 3[54] Adam Botach、Evgenii Zheltonozhskii 和 Chaim Baskin。使用多模态变换器进行端到端指代表达视频目标分割。载于 CVPR，2022年。 3[55] Dongming Wu、Xingping Dong、Ling Shao 和 Jianbing Shen。通过语义对齐进行多层级表征学习，用于指代表达视频目标分割。载于 CVPR，2022年。 3[56] Hui Zhang 和 Henghui Ding。用于零样本语义分割的原型匹配与开放集拒识。载于 ICCV，2021年。 3[57] Katerina Fragkiadaki、Pablo Arbelaez、Panna Felsen 和 Jitendra Malik。学习分割视频中的运动目标。载于 CVPR，2015年。 3[58] Yanchao Yang、Brian Lai 和 Stefano Soatto。Dystab：通过动静态自举进行无监督物体分割。载于 CVPR，2021年。 3[59] Sucheng Ren、Wenxi Liu、Yongtuo Liu、Haoxin Chen、Guoqiang Han 和 Shengfeng He。用于无监督视频目标分割的互反变换。载于 CVPR，2021年。 3[60] Daizong Liu、Dongdong Yu、Cha nghu Wang 和 Pan Zhou。F2net：学习聚焦前景以实现无监督视频目标分割。载于人工智能促进协会年会，2021年。 3[61] Xinkai Lu、Wenguan Wang、Martin Danelljan、Tianfei Zhou、Jianbing Shen 和 Luc Van Gool。使用情景图记忆网络进行视频目标分割。载于 ECCV，2020年。 3[62] Xiankai Lu、Wenguan Wang、Jianbing Shen、Yu-Wing Tai、David J Crandall 和 Steven CH Hoi。从未标注视频中学习视频目标分割。载于 CVPR，2020年。 3[63] Pavel Tokmakov、Cordelia Schmid 和 Karteek Alahari。学习分割运动目标。IJCV，127(3)，2019年。 3[64] Zhao Yang、Qiang Wang、Luca Bertinetto、Weiming Hu、Song Bai 和 Philip HS Torr。用于无监督视频目标分割的锚点扩散。发表于 ICCV，2019 年。3[65] Haofeng Li、Guanqi Chen、Guanbin Li 和 Yizhou Yu。用于视频显著目标检测的运动引导注意力。发表于 ICCV，2019 年。3[66] Wenguan Wang、Hongmei Song、Shuyang Zhao、Jianbing Shen、Sanyuan Zhao、Steven CH Hoi 和 Haibin Ling。通过视觉注意力学习无监督视频目标分割。发表于 CVPR，2019 年。3[67] Guanbin Li、Yuan Xie、Tianhao Wei、Keze Wang 和 Liang Lin。用于视频显著目标检测的流引导循环神经编码器。发表于 CVPR，2018 年。3[68] Pavel Tokmakov、Karteek Alahari 和 Cordelia Schmid。利用视觉记忆学习视频目标分割。发表于 ICCV，2017 年。3[69] Tianfei Zhou、Shunzhou Wang、Yi Zhou、Yazhou Yao、Jianwu Li 和 Ling Shao。面向零样本视频目标分割的运动注意转换。发表于人工智能促进协会年会，2020 年。3[70] Xiankai Lu、Wenguan Wang、Chao Ma、Jianbing Shen、Ling Shao 和 Fatih Porikli。看得更多，知得更多：利用协同注意力孪生网络进行无监督视频目标分割。发表于 CVPR，2019 年。3[71] Wenguan Wang、Xiankai Lu、Jianbing Shen、David J Crandall 和 Ling Shao。通过注意力图神经网络实现零样本视频目标分割。发表于 ICCV，2019 年。3，7[72] Xiankai Lu、Wenguan Wang、Jianbing Shen、David Crandall 和 Luc Van Gool。从关系视觉数据中分割物体。IEEE TPAMI，2021 年。3[73] Linjie Yang、Yuchen Fan 和 Ning Xu。视频实例分割。发表于 ICCV，2019 年。3[74] Xiangtai Li、Hao He、Yibo Yang、Henghui Ding、Kuiyuan Yang、Guangliang Cheng、Yunhai Tong 和 Dacheng Tao。通过时间金字塔路由改进视频实例分割。IEEE TPAMI，2022 年。3[75] Lei Ke、Henghui Ding、Martin Danelljan、Yu-Wing Tai、Chi-Keung Tang 和 Fisher Yu。用于高质量视频实例分割的视频掩码细化器。发表于 ECCV。施普林格，2022 年。3[76] Jiyang Qi、Yan Gao、Yao Hu、Xinggang Wang、Xiaoyu Liu、Xiang Bai、Serge Belongie、Alan Yuille、Philip HS Torr 和 Song Bai。遮挡视频实例分割：一个基准。IJCV，130(8)，2022 年。3，4，5[77] Henghui Ding、Xudong Jiang、Bing Shuai、Ai Qun Liu 和 Gang Wang。用于场景分割的上下文对比特征与门控多尺度聚合。发表于 CVPR，2018 年。3[78] Henghui Ding、Hui Zhang、Jun Liu、Jiaxin Li、Zijian Feng 和 Xudong Jiang。通过语义区域亲和性的双向图实现交互，用于场景解析。发表于 ICCV，2021 年。3[79] Jonathan Long、Evan Shelhamer 和 Trevor Darrell。用于语义分割的全卷积网络。发表于 CVPR，2015 年。3[80] Liang-Chieh Chen、George Papandreou、Iasonas Kokkinos、Kevin Murphy 和 Alan L Yuille。Deeplab：利用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割。IEEE TPAMI，40(4)，2017 年。3[81] Henghui Ding、Xudong Jiang、Ai Qun Liu、Nadia Magnenat Thalmann 和 Gang Wang。用于场景分割的边界感知特征传播。发表于 ICCV，2019 年。3[82] Henghui Ding、Xudong Jiang、Bing Shuai、Ai Qun Liu 和 Gang Wang。通过语义相关性促进形状变化上下文，用于分割。发表于 CVPR，2019 年。3[83] Gabriel J Brostow、Julien Fauqueur 和 Roberto Cipolla。视频中的语义目标类别：高清真值数据库。Pattern Recognition Letters，30(2)，2009 年。3[84] Marius Cordts、Mohamed Omran、Sebastian Ramos、Timo Rehfeld、Markus Enzweiler、Rodrigo Benenson、Uwe Franke、Stefan Roth、

和 Bernt Schiele。用于语义城市场景理解的 Cityscapes 数据集。发表于 CVPR，2016 年。3[85] Jiaxu Miao、Yunchao Wei、Yu Wu、Chen Liang、Guangrui Li 和 Yi Yang。VSPW：野外视频场景解析的大规模数据集。发表于 CVPR，2021 年。3[86] Guolei Sun、Yun Liu、Henghui Ding、Thomas Probst 和 Luc Van Gool。用于视频语义分割的由粗到细的特征挖掘。发表于 CVPR，2022 年。3[87] Dahun Kim、Sanghyun Woo、Joon-Young Lee 和 In So Kweon。视频全景分割。发表于 CVPR，2020 年。3[88] Jiaxu Miao、Xiaohan Wang、Yu Wu、Wei Li、Xu Zhang、Yunchao Wei 和 Yi Yang。野外大规模视频全景分割：一个基准。发表于 CVPR，2022 年。3[89] Sanghyun Woo、Dahun Kim、Joon-Young Lee 和 In So Kweon。学习关联视频全景分割中的每个片段。发表于 CVPR，2021 年。3[90] Siyuan Qiao、Yukun Zhu、Hartwig Adam、Alan Yuille 和 Liang-Chieh Chen。VIP-DeepLab：利用深度感知的视频全景分割学习视觉感知。发表于 CVPR，2021 年。3[91] Bing Shuai、Henghui Ding、Ting Liu、Gang Wang 和 Xudong Jiang。迈向鲁棒的低层和高层场景解析。IEEE TIP，28(3)，2018 年。3[92] Henghui Ding、Xudong Jiang、Bing Shuai、Ai Qun Liu 和 Gang Wang。结合上下文编码与多路径解码的语义分割。IEEE TIP，29，2020 年。3[93] Justin Lazarow、Kwonjoon Lee、Kunyu Shi 和 Zhuowen Tu。学习全景分割中的实例遮挡。发表于 CVPR，2020 年。3[94] Xiaohang Zhan、Xingang Pan、Bo Dai、Ziwei Liu、Dahua Lin 和 Chen Change Loy。自监督场景去遮挡。发表于 CVPR，2020 年。3[95] Adam Kortylewski、Qing Liu、Angtian Wang、Yihong Sun 和 Alan Yuille。组合式卷积神经网络：一种用于遮挡下目标识别的鲁棒且可解释模型。IJCV，129(3)，2021 年。3[96] Xinlong Wang、Tete Xiao、Yuning Jiang、Shuai Shao、Jian Sun 和 Chunhua Shen。排斥损失：在人群中检测行人。发表于 CVPR，2018 年。3[97] Jun Liu、Henghui Ding、Amir Shahroudy、Ling-Yu Duan、Xudong Jiang、Gang Wang 和 Alex C Kot。用于三维姿态估计的特征增强网络。IEEE TPAMI，42(2)：494–501，2019 年。3[98] Shifeng Zhang、Longyin Wen、Xiao Bian、Zhen Lei 和 Stan Z Li。遮挡感知 R-CNN：在人群中检测行人。发表于 ECCV，2018 年。3[99] Xiaohong Wang、Xudong Jiang、Henghui Ding 和 Jun Liu。用于皮损分割的双向皮肤镜特征学习与多尺度一致性决策融合。IEEE TIP，29，2019 年。3[100] Meng-Jiun Chiou、Henghui Ding、Hanshu Yan、Changhu Wang、Roger Zimmermann 和 Jiashi Feng。从有偏场景图中恢复无偏场景图。发表于 ACM MM，2021 年。3[101] Lei Ke、Yu-Wing Tai 和 Chi-Keung Tang。利用重叠双层进行深度遮挡感知实例分割。发表于 CVPR，2021 年。3[102] Qi Chu、Wanli Ouyang、Hongsheng Li、Xiaogang Wang、Bin Liu 和 Nenghai Yu。使用带有时空注意力机制的基于 CNN 的单目标跟踪器进行在线多目标跟踪。发表于 ICCV，2017 年。3[103] Ji Zhu、Hua Yang、Nian Liu、Minyoung Kim、Wenjun Zhang 和 Ming-Hsuan Yang。通过双重匹配实现在线多目标跟踪注意力网络。在ECCV，2018年。3[104] Jiarui Xu、Yue Cao、Zheng Zhang和Han Hu。用于多目标跟踪的时空关系网络。发表于ICCV，2019年。3[105] Qiankun Liu、Qi Chu、Bin Liu和Nenghai Yu。GSM：用于多目标跟踪的图相似度模型。发表于国际人工智能联合会议，2020年。3[106] Siyuan Li、Martin Danelljan、Henghui Ding、Thomas E Huang和Fisher Yu。野外万物跟踪。发表于ECCV。施普林格，2022年。3[107] Alessandro Prest、Christian Leistner、Javier Civera、Cordelia Schmid和Vittorio Ferrari。从弱标注视频中学习目标类别检测器。发表于CVPR，2012年。5[108] Fuxin Li、Taeyoung Kim、Ahmad Humayun、David Tsai和James M Rehg。通过跟踪多个图形-背景片段进行视频分割。发表于ICCV，2013年。5[109] Peter Ochs、Jitendra Malik和Thomas Brox。通过长期视频分析分割运动目标。IEEE TPAMI，36(6)，2014年。5[110] Qingnan Fan、Fan Zhong、Dani Lischinski、Daniel Cohen-Or和Baoquan Chen。JumpCut：用于视频抠图的非连续掩码迁移与插值。ACM图形学汇刊，34(6)，2015年。5[111] Tsung-Yi Lin、Michael Maire、Serge J Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár和C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。发表于ECCV，2014年。4[112] David R Martin、Charless C Fowlkes和Jitendra Malik。利用局部亮度、颜色和纹理线索学习检测自然图像边界。IEEE TPAMI，26(5)，2004年。6[113] Zongxin Yang、Yunchao Wei和Yi Yang。将变换器与物体关联，用于视频目标分割。发表于M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang和J. Wortman Vaughan编，NeurIPS，第34卷。Curran Associates, Inc.，2021年。6, 7[114] Ho Kei Cheng、Yu-Wing Tai和Chi-Keung Tang。重新思考时空网络：通过提升记忆覆盖率实现高效视频目标分割。发表于NeurIPS，2021年。6, 7[115] Mingxing Li、Li Hu、Zhiwei Xiong、Bang Zhang、Pan Pan和Dong Liu。用于视频目标分割的循环动态嵌入。发表于CVPR，2022年。6, 7[116] Zhihui Lin、Tianyu Yang、Maomao Li、Ziyu Wang、Chun Yuan、Wenhao Jiang和Wei Liu。SWEM：通过序列加权期望最大化实现实时视频目标分割。发表于CVPR，2022年。6, 7[117] Kaiming He、Xiangyu Zhang、Shaoqing Ren和Jian Sun。用于图像识别的深度残差学习。发表于CVPR，2016年。6[118] Ali Athar、Sabarinath Mahadevan、Aljoˇsa Oˇsep、Laura Leal-Taixé和Bastian Leibe。STEm-Seg：用于视频实例分割的时空嵌入。发表于ECCV，2020年。7[119] Carles Ventura、Miriam Bellver、Andreu Girbau、Amaia Salvador、Ferran Marques和Xavier Giro-i Nieto。RVOS：用于视频目标分割的端到端循环网络。发表于CVPR，2019年。7[120] Seoung Wug Oh、Joon Young Lee、Ning Xu和Seon Joo Kim。使用时空记忆网络进行视频目标分割。发表于ICCV，2019年。7[121] Yuk Heo、Yeong Jun Koh和Chang-Su Kim。使用全局与局部迁移模块进行交互式视频目标分割。发表于ECCV，2020年。7[122] Sergi Caelles、Jordi Pont-Tuset、Federico Perazzi、Alberto Montes、Kevis-Kokitsi Maninis和Luc Van Gool。2019年DAVIS挑战赛：无监督多目标分割。arXiv:1905.00737，2019年。7