arXiv:2304.01715v2 [cs.CV] 6 Aug 2023

迈向开放词汇视频实例分割

Haochen Wang*阿姆斯特丹大学
h.wang3@uva.nl

Cilin Yan*北京航空航天大学
clyan@buaa.edu.cn

Shuai Wang 阿姆斯特丹大学s.wang3@uva.nl

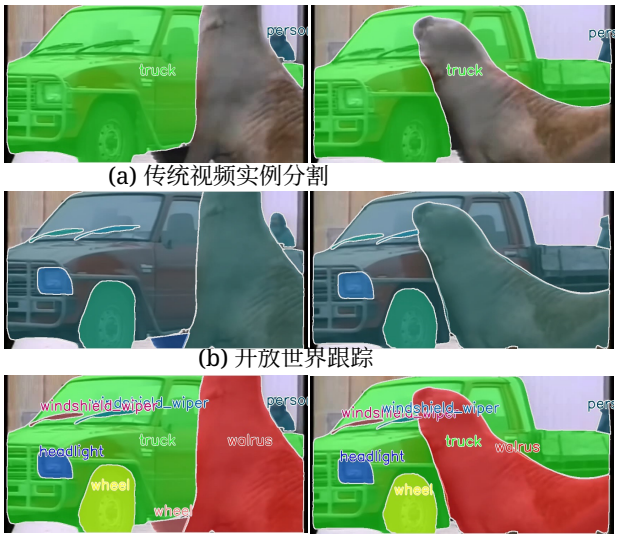
Xiaolong Jiang, Xu Tang, Yao Hu 小红书
laige@xiaohongshu.com

Weidi Xie†上海交通大学
weidi@robots.ox.ac.uk

Efstratios Gavves 阿姆斯特丹大学
egavves@uva.nl

摘要

视频实例分割（*VIS*）旨在对视频中属于一组封闭训练类别的物体进行分割和分类，但缺乏处理真实世界视频中新类别的泛化能力。为解决这一局限，我们作出以下三项贡献。首先，我们提出了开放词汇视频实例分割这一新任务，旨在同时对来自开放集类别的视频物体进行分割、跟踪和分类，其中包括训练期间未见过的新类别。其次，为开放词汇视频实例分割建立基准，我们收集了一个大词汇量视频实例分割数据集（*LV-VIS*），其中包含来自 *1,196* 个不同类别、经过完善标注的物体，其类别规模比现有数据集高出一个数量级以上。第三，我们提出了一种高效的记忆诱导变换器架构 *OV2Seg*，首次以端到端方式实现开放词汇视频实例分割，并达到接近实时的推理速度。在 *LV-VIS* 和四个现有视频实例分割数据集上的大量实验表明，*OV2Seg* 对新类别具有很强的零样本泛化能力。数据集和代码已在此发布 *https://github.com/haochenheheda/LVVIS*。



(a) 传统视频实例分割

(b) 开放世界跟踪

(c) 开放词汇视频实例分割

图 1。（a）传统的视频实例分割无法处理新类别中的物体（训练期间未见过，例如图中的海象）；（b）开放世界跟踪旨在以类别无关的方式分割并跟踪所有物体，但缺乏对视频中的物体进行分类的能力；（c）开放词汇视频实例分割旨在同时分割、跟踪并分类训练类别中的物体（例如人）和新类别中的物体（例如图中的海象）。图中不同的颜色表示不同的目标实例。

1. 引言

尽管视频实例分割（VIS）取得了令人瞩目的进展，但其本质上仍受限于对训练类别这一封闭集合中的物体进行分割和分类，因而限制了模型泛化至新概念的能力，见图 [50, 12, 46]1（a）。在现实场景中，这种封闭词汇范式缺乏实用价值，因为模型通常会遇到训练阶段未见过的新类别物体。相比之下，近期的开放世界跟踪（OWTB）[31], UVO [45], 和 BURST [4] 旨在对所有可见物体进行分割和跟踪，见图1 (b)。然而，这些方法仍无法对开放集类别中的物体进行分类，而这对于视频描述和动作识别等视频级任务至关重要。为了提升视频实例分割的实用性，使其更适用于下游视频级任务，我们提出了开放词汇视频实例分割，旨在同时对开放类别集合中的任意物体进行分类、跟踪和分割，如图1(c)所示。

为了对开放词汇视频实例分割进行恰当的基准评测，需要一个包含大量且多样化物体类别的视频数据集

*共同贡献。
†通讯作者。

数据集	UVO [45]	YT19 [50]	YT21 [50]	OVIS [39]	BURST [4]	LV-VIS
视频	11228	2883	3859	901	2914	4828
实例	104898	4883	8171	5223	16089	25588
掩码	593k	131k	232k	296k	600k	544k
掩码/帧	12.3	1.7	2.0	4.7	3.1	4.9
对象/视频	9.3	1.6	2.1	5.8	5.5	5.3
类别	80*	40	40	25	482	1196

表1. 我们的 LV-VIS 数据集与已发布视频分割数据集的关键统计数据对比。 $*$ 表示 UVO 仅为 MS-COCO 定义的80个常见类别中的物体提供类别标签。我们的 LV-VIS 数据集包含的类别集合显著大于现有数据集。

这是必要的。然而，现有数据集 Youtube-VIS [50], OVIS [39], 和 UVO [45]并不充分，因为它们仅包含几十个类别，见表1。BURST [4]包含的类别集合要大得多。不过，BURST 中 81.7% 的物体属于 MS-COCO 中的常见类别，如图 2 (b) 所示，因此其多样性仍不足以用于新类别的开放词汇评估。

在这项工作中，我们收集并构建了一个大词汇量视频实例分割数据集（*LV-VIS*）。LV-VIS 包含 4,828 个视频，以及来自 1,196 个类别的 54.4 万多个实例掩码，其规模显著超过所有现有视频分割数据集。值得注意的是，如图 2 (b) 所示，LV-VIS 中很大一部分已标注物体不同于 MS-COCO 等常用数据集中的类别 [29]和 LVIS [20], , 因此非常适合评估开放词汇视频实例分割方法对新类别的泛化能力，也可作为现有数据集的宝贵补充。

从架构上看，开放词汇视频实例分割的一种直接方法是将开放词汇检测器 [57, 55] 的逐帧结果与开放世界跟踪器 [31]关联起来。然而，这种提议-精简-关联方法需要非极大值抑制等复杂的手工设计模块，并且忽略了视频级特征在视频中的稳定跟踪和开放词汇分类中的作用，从而导致性能和推理速度欠佳。

在本文中，我们提出了首个端到端开放词汇视频实例分割模型OV2Seg。该模型简化了复杂的提议-精简-关联范式，并通过记忆诱导Transformer实现长期感知。具体而言，该模型首先通过通用物体提议模块提出并分割所有物体，然后引入一组记忆查询

通过时间逐步编码对象特征，从而实现长期感知，高效地跨时间跟踪所有物体。最后，给定任意类别名称作为输入，采用语言变换器编码器，基于记忆查询对跟踪对象进行分类。记忆查询逐步汇集来自不同视频帧的对象特征，从而实现稳健的视频物体分类。据我们所知，OV2Seg 是首个能够对任意开放集类别的视频中的物体进行分割、跟踪——

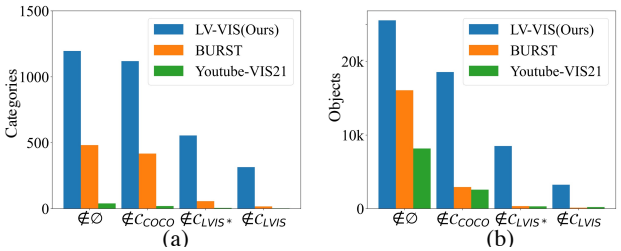


图2。图(a)和图(b)分别展示了与某一类别集合 $\mathcal{C}$互斥的类别数量和物体实例数量。$\mathcal{C}_{\text{COCO}}$、 $\mathcal{C}_{\text{LVIS}*}$和 $\mathcal{C}_{\text{LVIS}}$ 分别表示 MS-COCO中的所有类别、LVIS中的高频/常见类别以及 LVIS中的所有类别。LV-VIS中有更大比例的类别和物体与常用数据集中的类别互斥。

并分类的端到端模型，且推理速度接近实时。

我们在 LV-VIS 以及四个现有视频分割数据集上评估 OV2Seg：Youtube-VIS2021、Youtube- VIS2019、OVIS 和 BURST。无需在下游视频数据集上进行微调，OV2Seg 的性能可与若干经过充分监督、在目标视频数据集上训练的竞争方法相当 [50, 10]，同时对训练阶段未见过的新类别展现出很强的零样本泛化能力。

总而言之，我们的主要贡献如下：

(i) 我们提出了开放词汇视频实例分割，它能够同时对视频中任意开放集类别的物体进行分割、跟踪和分类，并将传统视频实例分割对封闭集训练类别的适用范围加以扩展；

(ii) 我们收集并标注了大规模、像素级标注的视频数据集 LV-VIS，其词汇集远大于现有视频数据集，因此可作为开放词汇视频实例分割的理想测试平台；

(iii) 我们提出了首个端到端开放词汇视频实例分割模型 OV2Seg。该模型采用记忆诱导变换器架构，能够对新类别的物体进行分割、跟踪和分类。

2. 相关工作

视频实例分割 旨在对预定义训练类别中的目标实例进行分割、跟踪和分类，例如人物、汽车和狗。现有方法可分为基于帧的方法 [7, 30, 28,52, 22, 51, 21] 和基于视频片段的方法 [46, 5, 26, 47, 12,53, 24]。基于帧的方法（如 MaskTrack R-CNN [50],）提出在图像实例分割模型 Mask R-CNN [22]中加入一个额外的跟踪头，用于视频实例分割。基于视频片段的方法（如 IFS [26], 、SeqFormer [47] 和 Mask2Former [12],）以整段视频或视频片段为输入，同时输出所有输入视频帧的掩码和分类得分。上述所有模型都需要在包含大量视频的数据集上进行训练，这些视频中包含

来自训练类别且经过密集标注的物体。相比之下，我们提出的 OV2Seg 使用图像数据集进行训练，但在测试时可直接应用于视频，并能够处理来自新类别的物体。

开放词汇目标检测旨在检测图像中超出封闭词汇集范围的物体，这一任务最早由 ZSD 提出 [6]。受 CLIP 启发 [40], ViLD [19] 和 ZSD- YOLO [48]提出将预训练 CLIP视觉嵌入中的知识显式蒸馏到 Faster R-CNN 的骨干网络中。Detic [57] 和 PromptDet [18] 则尝试通过在额外图像上进行自训练来扩展检测器的词汇量。这些方法沿用了两阶段检测器的标准架构，该架构采用复杂的候选生成与筛选流程。OV-DETR [55] 首次尝试采用类似 DETR 的架构构建端到端开放词汇检测器 [11], , 但类别相关的目标查询仍然限制了其推理速度。

开放世界跟踪。多目标跟踪 [56, 42,9] 旨在随时间跟踪多个物体。OWTB [31]提出了开放世界跟踪，以分割并跟踪视频中的所有物体。OWTB 在多个多目标跟踪数据集上取得了最先进的性能。近期工作 UVO [45]专注于视频中的类别无关物体分割与跟踪，却忽视了评估期间的物体分类性能。BURST [4], 采用的方法（如 STCN [14]和 MaskTrack [50], ）无法对新类别中的物体进行分类。因此，BURST 主要遵循类别无关多目标跟踪的评估协议，同时仅针对 MS-COCO 中的 80 个训练类别进行按类别评估。相反，我们引入开放词汇视频实例分割，使视频中的开放词汇物体分类能够与分割和跟踪相结合。我们将开放世界跟踪方法 [31] 与开放词汇检测器 [57, 55]相结合，为开放词汇视频实例分割构建了多个基线模型，并通过实验验证了所提出的 OV2Seg 的有效性。

3. 开放词汇视频实例分割的设定

给定训练数据集 $\mathcal{D}_{train}$ ，其中包含一组训练类别 $\mathcal{C}_{\text{train}}$的像素级标注，传统视频实例分割旨在训练模型 f_θ (·)，并在数据集 $\mathcal{D}_{\text{test}} = \{V_i\}_{i=1}^{N}$上进行测试，其中 $V_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ 指形状为 $(H_i,\ W_i)$、包含 T_i 帧的视频片段。 f_θ(·)需要为训练类别中的视频内每个物体预测一系列分割掩码$\{m_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times H_i \times W_i}$ 及类别标签 $c \in \mathcal{C}_{train}$，而新类别 $\mathcal{C}_{novel}$ 中的物体则会被忽略。

相比之下，开放词汇视频实例分割旨在使用 $\mathcal{D}_{train}$训练模型f_θ(·)，然后在 $\mathcal{D}_{test}$ 上测试，使其能够处理训练类别 $\mathcal{C}_{train}$ 和新类别 $\mathcal{C}_{novel}$。 具体而言，

在推理过程中，给定测试视频片段 $V_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ ，训练好的模型 f_θ(·)应为 V_i中的每个物体 p 预测分割掩码序列 $\{m_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times H_i \times W_i}$ 和类别标签 $c \in (\mathcal{C}_{train} \cup \mathcal{C}_{novel})$:

$$\{\{m_1, m_2, ..., m_{T_i}\}, c\}_p^P = f_\theta(V_i), \tag{1}$$

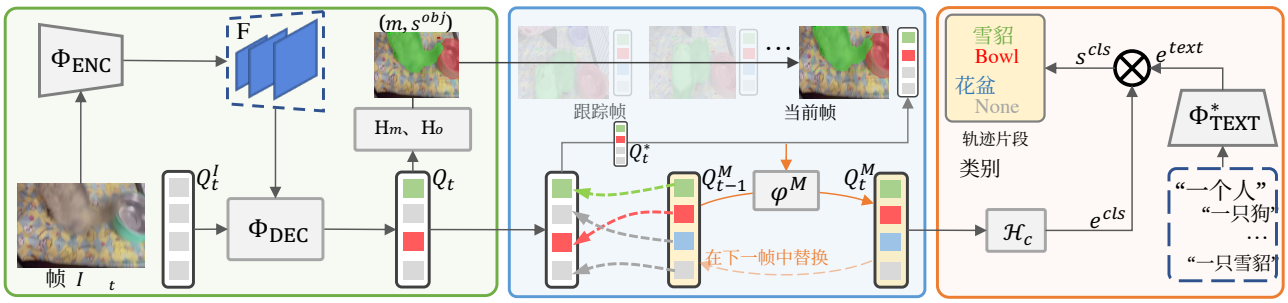
其中， P 是视频 V_i中的物体总数，类别标签 c 属于训练类别与新类别的并集，而 $m_t \in \mathbb{R}^{H_i \times W_i}$ 是每个视频帧t对应的分割掩码。在实验中，我们将训练类别称为基类，而将与基类互斥的类别称为新类别。

评估指标。我们采用平均精度均值（mAP）指标评估基类和新类别。视频中的 mAP 定义见 Youtube-VIS [50]。具体而言，我们首先计算所有测试视频中每个物体类别的平均精度（AP），即根据类别置信度绘制的精确率-召回率曲线下的面积。按照 MS-COCO [29] 中的标准设置，在 50% 至 95% 的 10 个交并比（ IoU）阈值上（步长为 5%）测量 AP 值。视频中的 IoU 定义见 Youtube-VIS [50], , 其计算预测分割掩码与真实分割掩码在时空上的一致性。最后，对基类集合和新类别集合上的 AP 求平均，得到平均精度均值（mAP），分别记为 mAP_b 和 mAP_n。

4. LV-VIS 数据集

为正确评估开放词汇视频实例分割，我们建立了一个名为 LV-VIS 的新数据集，具有较大的词汇集。本节概述数据收集和标注流程，并将该数据集的关键统计数据与现有视频级数据集进行比较。

数据收集与人工标注。 我们首先通过查询 1,621 个定义明确的类别名称，从互联网上收集了 7,569 个视频片段。随后，我们手动剔除不合适的视频片段，例如包含敏感个人信息的视频，以及物体过多、难以标注的视频。最终，我们在 LV-VIS 中保留了 4,828 个视频。我们保留视频的原始分辨率，范围从 480p 到 1080p。所有视频均经过预处理，时长控制在 2 至 10 秒之间，并从每个处理后的视频中按每秒 5 帧提取视频帧。我们对所选视频中所有清晰可见的物体进行全面标注。每个标注物体都被赋予一个类别标签、一个实例身份，以及一个多边形格式的掩码序列。随后，我们剔除未覆盖的类别，合并同义类别，最终保留 1,196 个定义明确的物体类别。详情见补充材料第 A 节。我们在补充材料第 B 节展示了已标注视频片段的可视化结果。



(a) 通用物体提议　(b) 记忆诱导跟踪　(c) 开放词汇分类

图3。OV2Seg 概览。（a）通用物体提议：输入帧 I_t，采用 Transformer 编码器 Φ_{ENC} 提取多尺度特征 $\mathcal{F}$。N 类别无关查询 $Q_t^I \in \mathbb{R}^{N\times d}$ 被输入 Transformer 解码器 Φ_{DEC}，以生成以对象为中心的查询 Q_t。随后，利用 Q_t，通过掩码头 $\mathcal{H}_m$ 和对象评分头 $\mathcal{H}_o$分别生成分割掩码 $m \in \mathbb{R}^{N\times H\times W}$ 和对象分数 $s^{\text{obj}} \in \mathbb{R}^{N\times 1}$。（b）记忆诱导跟踪：提出一组记忆查询 $Q_{t-1}^M \in \mathbb{R}^{N\times d}$，并通过匈牙利算法与 Q_t 进行关联。Q_{t-1}^M 通过函数 φ^M 进行更新，以动态聚合随时间变化的关联以对象为中心的查询 Q_t^*，并获得 Q_t^M，用于下一帧的跟踪。（c）开放词汇目标分类：在 Q_t^M上应用类别头 $\mathcal{H}_c$，生成类别嵌入 $\text{e}^{\text{cls}} \in \mathbb{R}^{N\times d}$。随后，将 e^{cls} 与文本编码器 $\Phi_{\text{TEXT}} \in \mathbb{R}^{|\mathcal{C}|\times d}$ 生成的文本嵌入 e^{text} 进行点积，以获得针对开放类别集合 $\mathcal{C}$的分类分数 $s^{cls} \in \mathbb{R}^{N\times|\mathcal{C}|}$。

数据集统计。 LV-VIS 数据集包含 4,828 个真实世界视频，总时长为 6.2 小时，并包含来自 1,196 个类别的 544,451 个像素级标注分割掩码。LV-VIS 中的类别分为 641 个基类（训练期间见过），这些类别继承自 LVIS 中的高频和常见类别 [20], ，以及 555 个与基类互斥的新类别。如表 1 所示，LV-VIS 的词汇量显著大于 Youtube-VIS（40 个类别）、OVIS（25 个类别）、UVO（80 个类别）和 BURST（482 个类别）。此外，LV-VIS 中有 93.5%/46.4% 的类别分别与 MS-COCO 中的类别以及 LVIS 中的高频/常见类别互斥；相比之下，BURST 和 Youtube-VIS 的对应数值要低得多，分别为 86.7%/11.8% 和 50%/15%。同时，LV-VIS 中有 33.3% 的物体属于新类别，而 BURST 中只有 2.2% 的物体来自新类别，见图 2。这意味着 LV-VIS 包含更多 MS-COCO 和 LVIS 等常用数据集中未出现的新目标，使其成为现有数据集的有价值补充，因此更适合对新类别进行按类别评估。受篇幅限制，更全面的数据集统计见补充材料第 A 节。

数据集划分。 LV-VIS 划分为训练集、验证集和测试集，分别包含 3083、837 和 908 个视频。我们将公开发布训练集和验证集的视频及标注，并搭建在线评估服务器，以基于具有挑战性的测试集对相关研究进行基准评测。

5. OV2Seg 的架构

构建数据集后，我们在此提出端到端的开放词汇视频实例分割架构 OV2Seg，如图所示 33。一般而言，所提出的架构由三个模块组成：(i) 通用对象提议模块，使用一组与类别无关的对象查询来提议并分割物体；(ii) 记忆诱导跟踪模块，使用一组记忆查询跟踪来自连续视频帧的对象提议，这些查询会随时间逐步聚合对象特征，以建立长期感知；(iii) 开放词汇分类模块，根据任意输入的类别名称，使用预训练文本编码器计算得到的文本嵌入对跟踪对象进行分类。我们将在以下各节详细介绍该架构。

5.1. 通用物体提议

我们将每个输入帧 I_t 输入到一个由多个可变形注意力层组成的 Transformer 编码器 Φ_{ENC}，以获取多尺度特征图 [58], : $\mathcal{F}$

$$\mathcal{F} = \{f^1, \ldots, f^l\} = \Phi_{\text{ENC}}(I_t) \tag{2}$$

其中，f^l 表示第 l 层的特征图。为了从所有类别中提出物体，一种先进的图像级开放词汇检测器 OV-DETR [55] 使用由给定类别名称条件化的类别相关对象查询。这样，每个类别都由一组独立的对象查询处理，导致其在真实世界视频中的推理速度难以满足实际需求。

在这项工作中，我们通过实验验证，一组类别无关的对象查询能够高效地从所有类别中提议物体，无需针对每个类别分别处理。因此，我们采用 N 可学习的类别无关对象查询 $Q^I \in \mathbb{R}^{N\times d}$ 来获取类别无关对象提议。具体而言，将对象查询 Q^I 和多尺度特征图 $\mathcal{F}$ 输入 Transformer解码器 Φ_{DEC}，其中 Q^I 迭代地关注每一层

的 $\mathcal{F}$，生成以对象为中心的查询 Q:

$$Q = \Phi_{\text{DEC}}(\mathcal{F}, Q^I) \in \mathbb{R}^{N\times d}. \tag{3}$$

随后，来自 Transformer 解码器的以对象为中心的查询 Q 被传递到掩码生成头 $\mathcal{H}_m$ 和对象评分头 $\mathcal{H}_o$:

$$m, s^{\text{obj}} = \mathcal{H}_m(Q) \circledast f^{-1}, \mathcal{H}_o(Q), \tag{4}$$

其中，f^{-1} 是 $\mathcal{F}$ 的最后一层。$\mathcal{H}_m(\cdot)$ 由 3 个多层感知机（MLP）层组成，并为每个对象查询输出动态卷积核。随后，动态卷积核与 f^{-1} 进行卷积，得到每帧的分割掩码 $m \in \mathbb{R}^{N\times H\times W}$。$\mathcal{H}_o(\cdot)$ 同样由 3 个 MLP 层组成，并输出二元分数 $s^{\text{obj}} \in \mathbb{R}^{N\times 1}$，用于指示某个查询是否代表一个物体。

5.2. 记忆诱导跟踪

连续帧中的以对象为中心的查询 Q 可以通过匈牙利算法进行关联，从而生成视频中的对象轨迹片段；这一方法受 MinVIS 启发 [25]。然而，直接关联连续帧中的物体会忽略视频中的长期依赖关系。因此，这种方法在复杂情况下会失效，例如物体在视频中消失后又重新出现。

为此，我们提出维护一组记忆查询 $Q^M \in \mathbb{R}^{N\times d}$ 以获取长期依赖关系，见图3(b)。具体而言，我们计算最后一帧的记忆查询 Q^M_{t-1} 与以对象为中心的查询 Q_t 之间的内积相似度。根据相似度矩阵，使用匈牙利算法将每个以对象为中心的查询与一个记忆查询关联起来。Q^M_0由第一帧的 Q_0 初始化。随后，通过更新函数$\varphi^M(\cdot,\ \cdot)$随时间进行动量更新，从而逐步编码视频级对象特征，以应对消失或严重遮挡的物体。具体而言，$\varphi^M(\cdot,\ \cdot)$ 定义如下:

$$\begin{aligned} Q^M_t &= \varphi^M(Q^M_{t-1}, Q^*_t) && (5)\\ &= \alpha\cdot s^{\text{obj}}\cdot Q^*_t + (1-\alpha\cdot s^{\text{obj}})\cdot Q^M_{t-1}, && (6)\end{aligned}$$

其中，α 是用于控制更新比例的因子，s^{obj} 是每个查询对应的目标分数。Q^*_t 是经过匈牙利算法匹配后关联的以对象为中心的查询。如果物体在视频中被严重遮挡或消失，相应的目标分数 s^{obj} 往往较低，因此会限制记忆更新过程，使物体记忆特征保持不变。通过维护一小组记忆查询，OV2Seg 能够高效地随时间跟踪所有分割出的物体，并借助增量视频对象特征增强后续的对象分类。

5.3. 开放词汇分类

我们将词汇集输入冻结的预训练 CLIP 文本编码器 Φ_{TEXT}，生成作为分类器的文本嵌入。例如，狗分类器可以表示为:

$$\mathbf{e}^{\text{text}} = \Phi_{\text{TEXT}}(\text{“this is a photo of [dog]”}), \tag{7}$$

推理过程中，词汇表可以替换为任意新类别。

通过$\mathbf{e}^{\text{cls}} \in \mathbb{R}^{N\times d}$ 获得被跟踪物体的类别嵌入，其中$\mathbf{e}^{\text{cls}} = \mathcal{H}_c(Q^M_t)$，$\mathcal{H}_c(\cdot)$表示由三个 MLP 层构成的类别头。随后，通过计算类别嵌入$\mathbf{e}^{\text{cls}} \in \mathbb{R}^{N\times d}$ 与文本嵌入$\mathbf{e}^{\text{text}} \in \mathbb{R}^{|\mathcal{C}|\times d}$之间的余弦相似度，即可获得每条物体轨迹的分类分数 s^{cls}：

$$s^{\text{cls}}_{i,j} = \sigma(\cos(\mathbf{e}^{\text{cls}}_i, \mathbf{e}^{\text{text}}_j)/\epsilon), \tag{8}$$

其中，$i \in [1, N]$ 和 $j \in [1, |\mathcal{C}|]$ 分别是对象查询和文本嵌入的索引。σ 是Sigmoid函数。$\cos(\cdot,\cdot)$表示余弦相似度，ϵ是温度超参数。记忆查询 Q^M_t根据目标分数对所有视频帧中的对象特征进行加权聚合，对低置信度的对象特征加以约束，从而实现鲁棒的视频对象分类。

5.4. 训练与损失

由于记忆诱导跟踪中不存在可学习参数，我们可以排除跟踪模块，并在图像级数据集（例如LVIS）上训练OV2Seg。这降低了大规模视频数据集的训练资源需求，并提高了训练效率。我们采用了与DETR类似的训练损失 [11] 进行集合预测:

$$\begin{aligned}\mathcal{L}_{\text{match}}(\hat{y}, y) =& \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}(\hat{s}^{\text{obj}}, s^{\text{obj}}) + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}(\hat{s}^{\text{cls}}, s^{\text{cls}})\\ &+ \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}(\hat{m}, m).\end{aligned} \tag{9}$$

我们使用二元交叉熵损失作为分类损失 $\mathcal{L}_{\text{cls}}$ 和目标分数损失 $\mathcal{L}_{\text{obj}}$。掩码损失 $\mathcal{L}_{\text{mask}}$ 是Dice损失 [35] 与二元焦点损失之和。真实值$\hat{}y$ 由每个物体的对象标签$\hat{}s^{\text{obj}}$、类别标签$\hat{}s^{\text{cls}}$ 和分割掩码$\hat{}m$ 组成。我们通过最小化总损失函数 $\mathcal{L}_{\text{match}}$，将真实值集合$\hat{}y$ 分配给预测集合 y。

6. 实验

我们首先介绍数据集和评估指标，然后描述方法的实现，最后对实验结果进行全面分析。

6.1. 数据集与评估指标

我们在 LVIS [20]，中常见和频繁类别的并集上训练 OV2Seg，并评估其性能，

评估对象包括多个视频实例分割数据集，其2019 [50]，中包括我们收集的 LV-VIS 数据集，以及 Youtube-2021 [50]，VIS、OVIS [39]，和 BURST [4]。值得注意的是，我们没有使用任何来自目标数据集的视频对 OV2Seg 进行微调。

LVIS 是一个广泛使用的图像开放词汇检测数据集，包含 1,203 个类别。遵循 ViLD [19]，，我们将高频类别和常见类别作为基类，并将与基类互斥的所有类别作为新类别。

Youtube-VIS 是应用最广泛的视频实例分割数据集，YVIS19 和 YVIS21 两个版本均包含 40 个类别。我们遵循 LVIS 中的划分方式，将 YVIS19 和 YVIS21 中的类别划分为基类和新类别。

BURST是近期发布的一个扩展 TAO [15] 的视频数据集。按照 LVIS 中的划分方式，BURST 包含 425 个基类和 57 个新类别。

OVIS 包含 25 个物体类别，其中只有一个类别不属于 LVIS 的基类；因此，我们仅报告 OVIS 全部类别的性能。

指标。 我们报告全部类别的平均精度均值（mAP），并进一步细分为基类的 mAP_b 和新类别的 mAP_n 。对于 LVIS [20], 的评估，我们报告全部类别的 mAP，以及新类别（稀有类别）的 mAP_n 。

6.2. 实现细节

基线模型。我们遵循提议-筛减-关联方案，将现成的两阶段开放词汇检测方法与开放世界跟踪方法相结合，构建了多个基线模型。具体而言，我们使用开放词汇检测器 [57, 17]为每个视频帧生成 300 个对象提议。然后，将这些提议输入 SORT [8] 和 OWTB [31]，以实现物体跟踪。SORT 是一种经典的多物体跟踪方法，而 OWTB 则是先进的开放世界跟踪方法之一，它结合 IOU 矩阵和感兴趣区域特征相似度，估计不同视频帧间提议的相似度。为提高评估效率，我们移除了 OWTB 中的光流。由此，我们构建了四种基线模型，分别为 Detic-SORT、Detic-OWTB、Detpro-SORT 和 Detpro-OWTB，并与我们提出的 OV2Seg 进行比较。此外，我们还设计了一种两阶段方法，使用先进的半监督视频目标分割方法 XMem [13]传播 Detic 的分割掩码。

OV2Seg。我们分别采用 ResNet50 [23] 和 SwinB [32] 作为骨干网络。我们使用 6 层多尺度可变形注意力层 [58] 作为 Transformer 编码器，并使用 6 层掩码注意力层 [12]作为 Transformer 解码器。我们采用从 CLIP ViT-32 模型中提取的预训练文本编码器来生成开放词汇分类器。文本编码器中的参数在训练期间保持冻结。温度超参数 ϵ 设为 50。我们采用 Detpro [17] 中的提示学习，替代文本编码器模块中的手工设计的提示。所有实验的对象查询数量均设为 100。我们将 λ_{obj}=2、 λ_{cls}=2、 λ_{mask}=5 设为各项损失的权重。记忆更新因子 α 设为 0.7。

训练细节。我们在 LVIS 上训练 OV2Seg 30 个 epoch，批量大小为 8。我们采用 AdamW [33] 优化器。基础学习率设为 1e-4，并在训练总步数的 0.9 和 0.95 处分别降低为原来的 1/10。骨干网络参数的学习率乘以 0.1。我们使用大尺度抖动 [16]，随机缩放比例从 0.1 到 2.0 的范围内采样。为提高训练效率，我们每次迭代仅采样 50 个类别来计算二元交叉熵损失。训练在 4 个 A100 GPU 上进行，耗时约 36 小时；所有推理均在单个 A100 GPU 上进行。

6.3. 结果

本节首先报告 OV2Seg 与一组强基线模型在 LV-VIS 数据集上的结果对比。随后，我们进行了广泛的消融研究，以说明 OV2Seg 中主要模块的有效性。最后，我们报告 OV2Seg 在现有视频实例分割数据集上的零样本泛化性能：Youtube-VIS2019、Youtube-VIS2021、BURST 和 OVIS。

6.3.1 LV-VIS 数据集上的结果

我们将 OV2Seg 与采用 propose-reduce 的关联规则基线模型在 LV-VIS 数据集上的表现进行比较，结果见表2。采用 ResNet-50 骨干网络的 OV2Seg 在 LV-VIS 验证集和测试集上的平均精度均值分别达到 11.9 和 8.9$_n$，$_n$ 分别比 DetPro-XMem 高出 6.5 和 5.3。如表3所示，OV2Seg 在图像数据集上的表现与 Detpro [17]和 Detic [57] 相当，而在视频数据集 LV-VIS 上则大幅超越 Detpro-OWTB 和 Detic-OWTB。这一比较展现了 OV2Seg 在视频级任务上的有效性。这主要是因为多阶段基线模型需要复杂的手工关联规则，并通过对逐帧分类得分取平均来分类物体，从而忽略了视频级对应关系。

相比之下，OV2Seg 将对象查询与记忆查询关联起来，后者编码长期对象特征，以便在视频中实现稳健跟踪和开放词汇分类。值得注意的是，在 LV-VIS 验证集的新类别上，OV2Seg 的表现优于 Detic-XMem，相对提升了 120%。这显著高于其在基类上 28.4% 的提升。因此，我们得出结论：OV2Seg 在真实世界视频的新类别上具备泛化能力。如表 2 所示，

方法	骨干网络	Val			Test			fps
		平均精度	平均精度$_b$	平均精度$_n$	平均精度	平均精度$_b$	平均精度$_n$	
DetPro [17]-SORT [8]	R50	6.4	10.3	3.5	5.8	10.8	2.1	3.1
Detic [57]-SORT [8]	R50	6.5	10.7	3.4	5.7	10.6	2.1	6.0
DetPro [17]-OWTB [31]	R50	7.9	12.9	4.2	7.0	12.6	2.9	3.1
Detic [57]-OWTB [31]	R50	7.7	12.6	4.2	7.0	12.8	2.8	5.9
Detic [57]-XMem [13]	R50	8.8	13.4	5.4	7.7	13.3	3.6	16.4
OV2Seg（我们的方法）	R50	**14.2**	**17.2**	**11.9**	**11.4**	**14.9**	**8.9**	**20.1**
Detic [57]-SORT [8]	SwinB	12.8	21.1	6.6	9.4	15.8	4.7	6.7
Detic [57]-OWTB [31]	SwinB	14.5	22.6	8.5	11.8	19.6	6.1	6.8
Detic [57]-XMem [13]	SwinB	16.3	24.1	10.6	13.1	20.5	7.7	13.4
OV2Seg（我们的方法）	SwinB	**21.1**	**27.5**	**16.3**	**16.4**	**23.3**	**11.5**	**16.8**

表2. LV-VIS验证集和测试集上的性能比较。AP、AP$_b$和AP$_n$分别表示所有类别、基地类别和小说类别的平均精度。

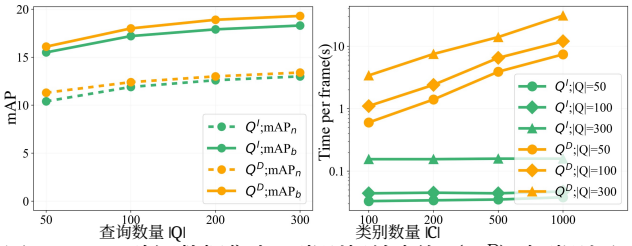


图4. LV-VIS验证数据集上，类别相关查询（Q^D）与类别无关查询（Q^I）的平均精度均值和每帧推理时间（秒）比较。

采用 ResNet50 骨干网络的 OV2Seg 在 LV-VIS 上达到每秒 20.1 帧，比 Detic-OWTB 高出 14.2。我们在图 6 中提供了一些可视化结果。由于篇幅有限，更多可视化结果和失败案例见补充材料第 C 节。

6.3.2 消融研究

我们在 LVIS、LV-VIS 验证集和 Youtube-VIS2019 数据集上进行了大量消融实验，以全面验证 OV2Seg 各主要模块的有效性。

图像级开放词汇检测。 移除记忆诱导跟踪模块后，OV2Seg 也可以进行图像级开放词汇检测。我们在 LVIS 和 LV-VIS(O)（图像预知评估）上，将 OV2Seg 与多种最先进的开放词汇检测方法进行比较，以展示 OV2Seg 在图像级任务上的有效性和效率。如表 3所示，使用 ResNet50 骨干网络时，OV2Seg 在 LVIS 和 LV-VIS(O) 上分别取得了 17.5 和 7.4 的平均精度均值 $_n$ 。这一结果与现有两阶段最先进方法 Detic [57]和 DetPro [17], 相当，同时推理速度更快。具体而言，Detic 和 DetPro 需要单独的区域提议网络和边界框回归模块来生成物体，还需要通过非极大值抑制去除冗余物体，因此推理速度较慢，分别为 6.2 帧/秒和 3.4 帧/秒。作为首个基于Transformer的开放词汇检测器，OV-DETR [55], 取得了 17.4 和 7.2 的平均精度均值$_n$ ，在

方法	骨干网络阶段	LVIS			LV-VIS(O)			fps
		平均精度	平均精度	$_n$	平均精度	平均精度	$_n$	
ViLD [19]	R50	2	25.5	16.6	-	-	-	-
DetPro* [17]	R50	2	25.7	18.7	10.8	7.6	3.4	
Detic [57]	R50	2	30.2	16.4	11.0	6.9	6.2	
OV-DETR [55]	R50	1	26.6	17.4	11.4	7.2	0.10	
OV2Seg（我们的方法）	R50	1	25.4	17.5	11.0	7.4	21.6	
Detic [57]	SwinB	2	38.4	21.9	13.9	7.8	7.6	
OV2Seg（我们的工作）	SwinB	1	31.4	21.8	14.2	8.3	18.2	

表3. 在 LVIS 数据集和 LV-VIS 验证数据集上的图像级性能。LV-VIS(O) 表示预知图像评估，即将 LV-VIS 中的所有视频帧视为图像数据集，并分别在这些图像上进行评估。为公平比较，DetPro* 采用单尺度推理。

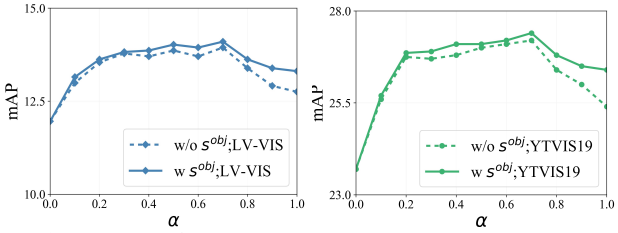


图5. 在不同记忆更新因子 α 下，LV-VIS 和 Youtube-VIS2019 验证数据集上的总体平均精度均值（mAP）。其中，w/o s^{obj} 表示移除记忆更新模块中的目标分数 s^{obj} 。

LVIS 和 LV-VIS(O) 的推理速度为 0.1 帧/秒，受类别相关的目标查询限制。相比之下，我们提出的 OV2Seg 使用类别无关目标查询，避免了针对每个类别分别处理。因此，OV2Seg 达到了 21.6 帧/秒的速度，其帧/秒数比 OV-DETR 高出两个数量级以上。更广泛地说，我们认为，在图像数据集上取得高精度并非本工作的主要目标。相反，我们更倾向于构建一个简单而有效的开放词汇视频实例分割模型。因此，我们的工作没有采用 Detic 和 Detpro 中使用的自训练和信息蒸馏。

类别相关查询与类别无关查询。OV-DETR [55]采用 N 针对每个类别的条件目标查询，我们称之为类别相关查询Q^D。我们将其与使用类别无关查询的 OV2Seg 进行比较 Q^I ，并在图4中展示结果。我们发现，使用类别无关查询的 OV2Seg Q^I 能够达到与使用类别相关查询相当的准确率 Q^D，同时速度快两个数量级。如图 4(b) 所示，随着类别数量增加，使用类别相关查询的 OV2Seg 会显著变慢 $|\mathcal{C}|$ ，因此难以应用于包含来自大型类别集合的物体的开放世界视频。

记忆查询。 我们在不同的记忆更新因子下，使用 LV-VIS 验证集和 Youtube-VIS2019 数据集评估 OV2Seg α 和 s^{obj} 在公式6中的表现。如图 5所示，不进行记忆更新的查询关联模块，即 α=0，在.....上仅达到 23.7 平均精度均值和 11.9 平均精度均值

方法	Open	骨干网络	YTVIS2019			YTVIS2021			BURST			OVIS
			mAP	mAP_b	mAP_n	mAP	mAP_b	mAP_n	mAP	mAP_b	mAP_n	mAP
FEELVOS [50]	✗	R50	26.9	-	-	-	-	-	-	-	-	9.6
MaskTrack [50]	✗	R50	30.3	-	-	28.6	-	-	-	-	-	10.8
SipMask [10]	✗	R50	33.7	-	-	31.7	-	-	-	-	-	10.2
Mask2Former [12]	✗	R50	46.4	-	-	40.6	-	-	-	-	-	17.3
Detic [57]-SORT [8]	✓	R50	14.6	17.0	3.5	12.7	14.4	3.1	1.9	1.8	**2.5**	6.7
Detic [57]-OWTB [31]	✓	R50	17.9	20.7	4.5	16.7	18.6	5.8	2.7	2.8	1.8	9.0
OV2Seg（我们的方法）	✓	R50	**27.2**	**30.1**	**11.1**	**23.6**	**26.5**	**7.3**	**3.7**	**3.9**	2.4	**11.2**
Detic [57]-SORT [8]	✓	SwinB	23.8	27.2	7.9	21.6	23.7	9.8	2.5	2.7	1.0	11.7
Detic [57]-OWTB [31]	✓	SwinB	30.0	34.3	9.7	27.1	29.9	11.4	3.9	4.1	2.4	13.6
OV2Seg（我们的方法）	✓	SwinB	**37.6**	**41.1**	**21.3**	**33.9**	**36.7**	**18.2**	**4.9**	**5.3**	**3.0**	**17.5**

表 4. 四个已发布数据集验证集上的性能比较：Youtube-VIS19、Youtube-VIS21、BURST、OVIS。表中的“开放”表示一种方法是否能够按照开放词汇视频实例分割的设定处理新类别中的物体。双横线以上的方法是在各数据集的目标训练视频上训练的；我们仅报告这些方法的总体平均精度均值（mAP）。双横线以下的方法是在图像数据集 LVIS 上训练，并在不进行微调的情况下于视频实例分割数据集上评估；我们报告这些方法的 mAP、mAP_b 和 mAP_n。

方法	LV-VIS		Youtube-VIS2019	
	平均精度均值。	平均精度均值。	平均精度均值。	mAP_n
平均值	16.9	11.3	29.4	10.8
记忆查询	**17.2**	**11.9**	**30.1**	**11.1**

表 5. 两种获取分类得分的方法比较。表中的“记忆查询”表示从记忆查询中获取分类得分。表中的“平均值”表示直接对逐帧分类得分求平均。

Youtube-VIS2019 和 LV-VIS 验证数据集。当 α设置为 0.3 至 0.8 之间的值时，OV2Seg 的性能均能稳定提升，这表明记忆更新计划对该因素具有鲁棒性 α。将 $\alpha = 0.7$ 和目标分数 s^{obj} 设为最佳值时，在 Youtube-VIS2019 和 LV-VIS 验证数据集上分别达到 27.2 和 14.2 的平均精度均值。记忆查询引入长期线索，以应对遮挡以及物体消失后重新出现的情况，从而有助于实现鲁棒的目标关联。如图6所示，零食在完全被遮挡后重新被跟踪。

在表5中，我们比较了基于记忆查询的分类与对逐帧分类得分取平均值的性能。如表所示，在 LV-VIS 和 Youtube-VIS2019 数据集上，记忆查询的性能均略优于朴素的平均值方法。这是因为记忆查询根据逐帧目标置信度聚合视频对象特征，从而限制低置信度特征的影响，使分类在运动模糊和遮挡等复杂情况下更加鲁棒。

6.3.3 视频实例分割数据集上的零样本泛化。

我们直接在四个视频实例分割数据集上评估 OV2Seg，无需在目标训练视频上进行微调。性能比较见表 4。如表所示，采用 ResNet-50 的 OV2Seg 在 Youtube-VIS2019 和 Youtube-n 上分别达到 11.1 平均精度均值$_n$ 和 7.3 平均精度均值

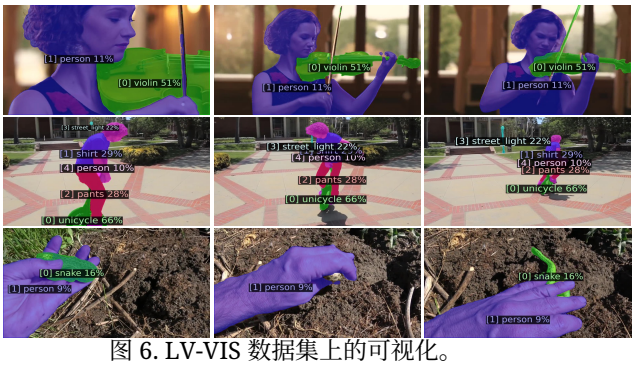


图 6. LV-VIS 数据集上的可视化。

VIS2021 新类别上的表现，分别超过基线模型 Detic-OWTB 6.6 和 1.5。请注意，在未使用任何 Youtube-VIS 训练视频的情况下，OV2Seg 的性能仍显著优于 MaskTrack 等封闭集全监督竞争方法 [50] 和 FEELVOS [43]。在 BURST 上，OV2Seg 达到 3.7 平均精度均值，就整体平均精度均值而言比 Detic-OWTB 高出 1.0。BURST 数据集并未进行穷尽式标注，因此对新类别进行按类别评估在统计上并不可靠。在存在严重物体遮挡的 OVIS 数据集上，OV2Seg 达到 11.2 平均精度均值，与全监督竞争方法 MaskTrack [50],FEELVOS [50], 和 SipMask [10]的表现相当。OVIS 上的比较表明，OV2Seg 在复杂遮挡情况下具有良好的性能。可视化结果见补充材料第 C 节。

7. 结论

我们提出一项新的开放词汇视频实例分割任务，旨在同时分割、分类和跟踪任意类别的物体。为了恰当地评估开放词汇视频实例分割，我们收集了一个大词汇量视频实例分割数据集（LV-VIS），其类别规模和多样性显著超过现有的视频级分割

数据集。此外，我们提出首个端到端开放词汇视频实例分割模型 OV2Seg。该模型采用记忆诱导视觉语言 Transformer，对任意类别的物体进行分割、分类和跟踪。无需在视频上进行微调，我们提出的 OV2Seg 在训练期间未见过的新类别上展现出良好的零样本泛化能力。

参考文献

[1] Alvin Alpher。《Frobnication》。Journal of Foo，12 778，2002 1 234( ): – 。[2] Alvin Alpher 和 Ferris P. N. Fotheringham-Smythe。《Frob-nication revisited》。Journal of Foo，13(1):234–778，2003。[3] Alvin Alpher、Ferris P. N. Fotheringham-Smythe 和 Gavin Gamow。《Can a machine frobnicate?》。Journal of Foo，14(1):234–778，2004。[4] Ali Athar、Jonathon Luiten、Paul Voigtlaender、Tarasha Khu-rana、Achal Dave、Bastian Leibe 和 Deva Ramanan。《BURST：统一视频中的目标识别、分割和跟踪的基准》。收录于《冬季计算机视觉应用会议论文集》，2023年。1、2、3、6、12、13[5] Ali Athar、Sabarinath Mahadevan、Aljosa Osep、Laura Leal-Taix´e 和 Bastian Leibe。《Stem-seg：用于视频实例分割的时空嵌入》。收录于《欧洲计算机视觉会议论文集》，第158–177页。Springer，2020年。2[6] Ankan Bansal、Karan Sikka、Gaurav Sharma、Rama Chel-lappa 和 Ajay Divakaran。《零样本目标检测》。收录于《欧洲计算机视觉会议论文集》，第384–400页，2018年。3[7] Gedas Bertasius 和 Lorenzo Torresani。《通过掩码传播对视频中的目标实例进行分类、分割和跟踪》。收录于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9739–9748页，2020年。2[8] Alex Bewley、Zongyuan Ge、Lionel Ott、Fabio Ra mos 和 Ben Upcroft。《简单在线实时跟踪》。收录于《IEEE图像处理国际会议论文集》，第3464–3468页。IEEE，2016年。6、7、8[9] Guillem Bras´o 和 Laura Leal-Taix´e。《学习用于多目标跟踪的神经求解器》。收录于《IEEE/CVF计算机视觉与模式识别会议论文集》，第6247–6257页，2020年。3[10] Jiale Cao、Rao Muhammad Anwer、Hisham Cholakkal、Fa-had Shahbaz Khan、Yanwei Pang 和 Ling Shao。《SipMask：用于快速图像和视频实例分割的空间信息保留》。收录于《欧洲计算机视觉会议论文集》，第1–18页。Springer，2020年。2、8[11] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。《使用Transformer进行端到端目标检测》。收录于《欧洲计算机视觉会议论文集》，第213–229页。Springer，2020年。3、5

[12] Bowen Cheng、Anwesa Choudhuri、Ishan Misra、Alexan- der Kirillov、Rohit Girdhar 和 Alexander G Schwing。Mask2Former：用于视频实例分割。arXiv 预印本 arXiv:2112.10764，2021年。1、2、6、8[13] Ho Kei Cheng 和 Alexander G Schwing。XMem：基于阿特金森-希夫林记忆模型的长时视频目标分割。载于《欧洲计算机视觉会议》，第640–658页。Springer，2022。6、7[14] 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的视频目标分割。《神经信息处理系统进展》，34：11781–11794，2021。3、12[15] Achal Dave、Tarasha Khurana、Pavel Tokmakov、Cordelia Schmid 和 Deva Ramanan。TAO：用于跟踪任意物体的大规模基准。载于《欧洲计算机视觉会议论文集》，第436–454页。Springer，2020。6[16] Xianzhi Du、Barret Zoph、Wei-Chih Hung 和 Tsung-Yi Lin。用于目标检测的简易训练策略与模型缩放。arXiv 预印本 arXiv:2107.00057，2021。6[17] Yu Du、Fangyun Wei、Zihe Zhang、Miaojing Shi、Yue Gao 和 Guoqi Li。借助视觉语言模型学习提示，以实现开放词汇目标检测。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第14084–14093页，2022。6、7[18] Chengjian Feng、Yujie Zhong、Zequn Jie、Xiangxiang Chu、Haibing Ren、Xiaolin Wei、Weidi Xie 和 Lin Ma。PromptDet：利用未经筛选的图像扩展检测器词汇表。载于《欧洲计算机视觉会议论文集》，2022。3[19] Xiuye Gu、Tsung-Yi Lin、Weicheng Kuo 和 Yin Cui。通过视觉与语言知识蒸馏实现开放词汇目标检测。载于《国际学习表征会议论文集》，2021。3、6、7[20] Agrim Gupta、Piotr Dollar 和 Ross Girshick。LVIS：大词汇量实例分割数据集。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5356–5364页，2019。2、4、5、6、12[21] Su Ho Han、Sukjun Hwang、Seoung Wug Oh、Yeonchool Park、Hyunwoo Kim、Min-Jung Kim 和 Seon Joo Kim。Visolo：用于高效在线视频实例分割的基于网格的时空聚合。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第2896–2905页，2022。2[22] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Gir- schick。Mask R-CNN。载于《IEEE/CVF国际计算机视觉会议论文集》，第2961 2969页，2017 2 –。[23] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第770–778页，2016。6[24] Miran Heo、Sukjun Hwang、Seoung Wug Oh、Joon-Young Lee 和 Seon Joo Kim。Vita：通过目标标记关联实现的视频实例分割。2022。2

[25]De-An Huang、Zhiding Yu 和 Anima Anandkumar。MinVIS：一种无需基于视频训练的精简视频实例分割框架。arXiv 预印本 arXiv:2208.02245，2022。5[26] Sukjun Hwang、Miran Heo、Seoung Wug Oh 和 Seon Joo Kim。使用帧间通信 Transformer 进行视频实例分割。《神经信息处理系统进展》，34：13352–13363，2021。2[27] Matej Kristan、Ales Leonardis、Jiri Matas、Michael Fels- berg、Roman Pflugfelder、Luka ˇCehovin Zajc、Tomas Vojir、Goutam Bhat、Alan Lukezic、Abdelrahman Eldesokey 等。第六届视觉目标跟踪 VOT2018 挑战赛结果。载于《欧洲计算机视觉会议研讨会论文集》，第0–0页，2018。12、13[28] Huaijia Lin、Ruizheng Wu、Shu Liu、Jiangbo Lu 和 Ji- aya Jia。采用提议-缩减范式的视频实例分割。载于《IEEE/CVF国际计算机视觉会议论文集》，第1739–1748页，2021。2[29] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于《欧洲计算机视觉会议论文集》，第 740 755 2014 2, 3, 14 –页。Springer，。[30] Dongfang Liu、Yiming Cui、Wenbo Tan 和 Yingjie Chen。SG-Net：用于单阶段视频实例分割的空间粒度网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9816–9825页，2021。2[31] Yang Liu、Idil Esen Zulfikar、Jonathon Luiten、Achal Dave、Deva Ramanan、Bastian Leibe、Aljoˇsa Oˇsep 和 Laura Leal- Taix´e。开启开放世界跟踪。载于《 IEEE/CVF 计算机视觉与模式识别会议论文集》，第19045–19055页，2022。1、2、3、6、7、8[32] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：使用移位窗口的层次化视觉 Transformer。载于《IEEE/CVF国际计算机视觉会议论文集》，第 10012 10022, 2021 6 –页。Ilya [33] Loshchilov 和 Frank Hutter。修正 Adam 中的权重衰减正则化。2018。6[34] Jonathon Luiten、Aljosa Osep、Patrick Dendorfer、Philip Torr、Andreas Geiger、Laura Leal-Taix´e 和 Bastian Leibe。HOTA：用于评估多目标跟踪的高阶度量指标。《International Journal of Computer Vision》，129：548–578，2021。[35] Fausto Milletari、Nassir Navab 和 Seyed-Ahmad Ahmadi。V-Net：用于体积医学图像分割的全卷积神经网络。载于2016年第四届国际 *3* 565 571 2016 5 D视觉会议，第–页。IEEE，。[36] 2014 Full Author Name。《可进行 Frobnication 的 foo 过滤器》，。人脸与姿态提交编号324。作为附加材料 fg324.pdf 提供。[37] Full Author Name。Frobnication 教程，2014。作为附加材料 tr.pdf 提供。[38] Jordi Pont-Tuset、Federico Perazzi、Sergi Caelles、Pablo Arbel´aez、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。12、13

[39] Jiyang Qi、Yan Gao、Yao Hu、Xinggang Wang、Xiaoyu Liu、Xiang Bai、Serge Belongie、Alan Yuille、Philip Torr 和 Song Bai。遮挡视频实例分割：一个基准。International Journal of Computer Vision，2022年。2，6，12，13[40]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于 International Conference on Machine Learning，第 8748–8763 页。PMLR，2021。3[41] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：借助区域建议网络实现实时目标检测。神经信息处理系统进展，28，2015年。[42] Siyu Tang、Mykhaylo Andriluka、Bjoern Andres 和 Bernt Schiele。利用提升式多割和人物重识别进行多人跟踪。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 3539–3548 页，2017。3[43] 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 IEEE/CVF计算机视觉与模式识别会议论文集，第 9481 9490 页，2019年 8 – 。[44] Paul Voigtlaender、Michael Krause、Aljosa Osep、Jonathon Luiten、Berin Balachandar Gnana Sekar、Andreas Geiger 和 Bastian Leibe。MOTS：多目标跟踪与分割。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 7942–7951 页，2019 年。12，13[45] Weiyao Wang 、Matt Feiszli、Heng Wang 和 Du Tran。未识别视频目标：密集开放世界分割基准。载于 IEEE/CVF国际计算机视觉会议论文集，第 10776–10785 页，2021。1，2，3，12，13[46] Yuqing Wang、Zhaoliang Xu、Xinlong Wang、Chunhua Shen、Baoshan Cheng、Hao Shen 和 Huaxia Xia。利用Transformer进行端到端视频实例分割。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 8741–8750 页，2021。1，2[47] Junfeng Wu、Yi Jiang、Wenqing Zhang、Xiang Bai 和 Song Bai。SeqFormer：一种简单得令人沮丧的视频实例分割模型。arXiv预印本 arXiv:2112 08275，2021 2 。 。[48] Johnathan Xie 和 Shuai Zheng。ZSD-YOLO：利用视觉语言知识蒸馏进行零样本YOLO检测。arXiv预印本 arXiv:2109.12066，2021。3[49] Ning Xu、Linjie Yang、Yuchen Fan、Dingcheng Yue、Yuchen Liang、Jianchao Yang 和 Thomas Huang。YouTube-VOS：大规模视频目标分割基准。arXiv预印本 arXiv:1809.03327，2018年。12，13[50] Linjie Yang、Yuchen Fan 和 Ning Xu。视频实例分割。载于 IEEE/CVF国际计算机视觉会议论文集，第 5188–5197 页，2019。1，2，3，6，8，12，13

[51] Linjie Yang、Yanran Wang、Xuehan Xiong、Jianchao Yang 和 Aggelos K Katsaggelos。通过网络调制实现高效视频目标分割。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 6499–6507 页，2018年。2[52] Shusheng Yang、Yuxin Fang、Xinggang Wang、Yu Li、Chen Fang、Ying Shan、Bin Feng 和 Wenyu Liu。用于快速在线视频实例分割的交叉学习。载于 IEEE/CVF国际计算机视觉会议论文集，第 8043–8052 页，2021。2[53] Shusheng Yang、Xinggang Wang、Yu Li、Yuxin Fang、Jiemin Fang、Wenyu Liu、Xun Zhao 和 Ying Shan。用于视频实例分割的时间高效视觉Transformer。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 2885–2895 页，2022。2[54] Fisher Yu、Haofeng Chen、Xin Wang、Wenqi Xian、Yingying Chen、Fangchen Liu、Vashisht Madhavan 和 Trevor Darrell。BDD100K：用于异构多任务学习的多样化驾驶数据集。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 2636–2645 页，2020年。12，13[55] Yuhang Zang、Wei Li、Kaiyang Zhou、Chen Huang 和 Chen Change Loy。采用条件匹配的开放词汇DETR。arXiv预印本 arXiv:2203.11876，2022。2，3，4，7[56] Li Zhang、Yuan Li 和 Ramakant Nevatia。利用网络流进行多目标跟踪的全局数据关联。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 1–8 页。IEEE，2008年。3[57] Xingyi Zhou、Rohit Girdhar、Armand Joulin、Philipp Kr¨ahenb¨uhl 和 Ishan Misra。利用图像级监督检测两万个类别。载于欧洲计算机视觉会议论文集，2022。2，3，6，7，8[58] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。可变形DETR：用于端到端目标检测的可变形Transformer。载于 International Conference on Learning Representations 论文集，2020年。4，6

附录

A. 数据集统计

统计对比。如表6所示，我们将所提出的 LV-VIS 数据集与现有数据集进行了详细统计对比。

基准 KITTI-MOTS [44]，BDD [54]，和 MOTS-challenge [44]聚焦于自动驾驶场景，其中包含由行驶中的车辆或步行的行人拍摄的街景长视频。基准 DAVIS2017 [38]和 YouTube-VOS [49] 主要聚焦于半监督视频目标分割 [38]。半监督视频目标分割旨在给定第一帧的掩码后，对物体进行跟踪和分割，这与 VOT 中的单目标跟踪类似 [27]UVO [45] 旨在穷尽地分割和跟踪人们在视频中会认为是“物体”的所有内容。UVO 不为特定物体分配类别标签，而是将所有前景物体视为一个类别（在 UVO v1 中，标注了属于 MS-COCO 的物体的类别标签，其余物体则未标注，如图8所示）。在上述数据集中，评估时不提供或考虑类别标签。

Youtube-VIS2019 [50], OVIS [39],和 BURST [4] 都会为每个已标注物体分配一个类别标签。因此，可以进行按类别评估。然而，Youtube-VIS2019 和 OVIS 仅包含 40 个和 25 个类别，类别数量不足以支持开放词汇评估。BURST 包含相对较大的词汇集，共有 482 个类别，但 BURST 中 81% 的目标实例来自 MS-COCO 类别，因此不适合评估新类别，如图 8所示。因此，在按类别评估中，BURST 仅考虑 MS-COCO 中的 80 个常见类别，并主要遵循类别无关的多目标跟踪评估协议。

相比之下，我们的 LV-VIS 数据集不仅包含由 1,196 个类别构成的大型词汇集，而且类别和目标实例的分布也更加多样，如图8所示。具体而言，LV-VIS 中 94 % 的类别与 MS-COCO 中的类别互斥，其中 46% 与 LVIS 中的高频/常见类别互斥。此外，OV-VIS 包含 4,828 个视频和 544,451 个用于评估的标注掩码，规模远大于现有数据集中的大多数验证集/测试集。LV-VIS 的数据集划分详情见表 7。如图所示，我们将 LV-VIS 划分为训练集、验证集和测试集，其中测试集比验证集复杂得多。

类别划分。 在本节中，我们首先说明如何选取这 1,196 个物体类别。我们首先纳入 LVIS 中的全部 1,203 个类别。然后，我们手动选取新类别 ImageNet21K 中的类别，以涵盖动物、植物、交通工具、工具、服装、食物等多种类型的类别。 通过这种方式，我们得到 1,612 个候选类别。对于具有多义性的类别，我们手动添加额外描述，例如 LVIS 数据集中的枣（水果）、三角铁（乐器）。每对类别之间存在两种类型的关系：互斥和非互斥。非互斥类别对可能存在部分重叠、父子关系或等价关系，这意味着单个物体可能具有多个有效的类别标签。因此，我们首先手动合并具有相同语义的物体类别，确保不存在相互等价的类别对。通过这种方式，我们收集了 1,196 个定义明确且彼此不同的物体类别。随后，我们在收集到的类别之间构建父子关系树。例如，“赛车”是“汽车”的子类别。最后，我们根据所定义的父子关系树修订标注，确保视频中的每个物体都被完整标注为所有相应的物体类别。例如，“赛车”也会被赋予相应的父类别标签“汽车”。通过上述流程，我们解决了等价类别、父子类别和部分重叠类别的标注问题。LV-VIS 数据集的类别划分以及每个类别的实例数量如表7所示，其中 1,196 个类别被划分为 641 个基类和 555 个新类别。LV-VIS 中的所有基类均继承自 LVIS 中频繁出现的常见类别 [20]。在进行评估时，Youtube-VIS2019 中的类别被划分为 33 个基类和 7 个新类别。Youtube-VIS2021 中的类别被划分为 34 个基类和 6 个新类别。我们在表2021 [50]中展示了 Youtube-VIS2019 和 Youtube-VIS8的类别划分。

标注详情。 我们基于 Labelme 开发了一个视频分割标注平台，并在此处发布 https://github.com/haochenheheda/segment-anything-annotator。我们首先使用多边形手动标注第一帧中的所有物体掩码，并借助 STCN 将物体掩码传播到第二帧 [14]。之后，我们修正传播得到的掩码，为新出现的物体添加掩码，然后重复传播至下一帧。我们手动识别并为每个已标注的掩码序列指定类别名称。最后，我们进行交叉修订，以确保标注质量。

B. 标注帧的可视化

图 9 展示了 LV-VIS 中标注视频的示例。9。

	基本信息			训练				验证集/测试集			
数据集	类别	时长（小时）	掩码/帧	视频实例		标注帧	Mask	视频实例		标注帧	Mask
VOT [27]	-	10.7	1	0	0	0	0	62	62	19,903	19,903
KITTI-MOTS [44]	2	39	5.4	21	748	8,008	38,197	28	961	11,095	61,904
MOTS-Chal. [44]	1	34.4	10	4	228	2,864	26,894	4	328	3,044	32,369
BDD [54]	7	40	11.4	154	17,838	30,745	347,442	32	4,873	6,475	77,389
DAVIS17 [38]	-	2.9	2.6	60	144	4,219	10,238	90	242	6,240	16,841
YT-VOS19 [49]	-	4.5	1.6	3741	6,459	94,400	12,918	1,048	2,115	28,825	4,310
UVO [45]	80*	3	12.3	5,641	76,627	39,174	416,001	5,587	28,271	18,966	177,153
YT-VIS19 [50]	40	4.5	1.7	2,238	3,774	61,845	103,424	645	1,092	17,415	29,431
OVIS [39]	25	3.2	4.7	607	3,579	42,149	206,092	297	1,644	20,492	89,841
BURST [4]	482	28.9	3.1	500	2,645	107,144	318,200	2,414	13,444	88,569	281,957
LV-VIS（我们的方法）	1196	6.2	4.9	3,083	16,060	70,242	339,533	1,745	9,526	41,253	204,918

表6. 我们的 LV-VIS 与其他视频级数据集的详细统计对比。**类别**：各数据集中的类别总数。类别列中的 - 表示该数据集未提供类别标签，或在评估时未将类别纳入考量。**长度**：视频的总时长。**掩码/帧**：每帧的平均标注掩码数。YT-VIS2019、YT-VOS2019 和 DAVIS2017 的验证集/测试集统计数据根据训练集估算，可能并不准确。

划分	时长（小时）	掩码/帧	视频	实例	标注帧	Mask
训练	3.9	4.8	3,083	16,060	70,242	339,533
Val	1.1	4.0	838	3,646	19,176	76,916
Test	1.3	5.7	908	5749	22,096	124,834
总计	6.2	4.9	4,828	25,588	111,495	544,451

表 7. LV-VIS 的数据集划分。

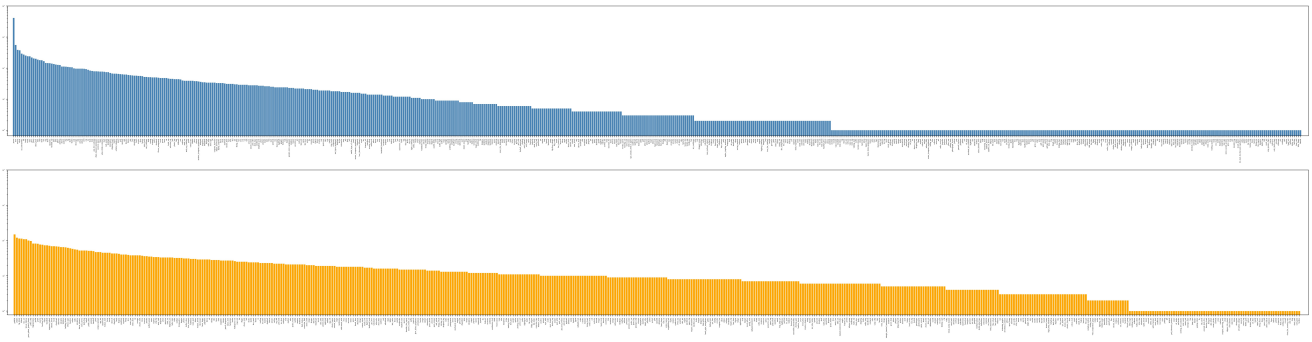


图 7. LV-VIS 数据集中每个类别的实例数。蓝色柱表示基类（LVIS 中的高频/常见类别），橙色柱表示新类别（与基类互斥）。

C. 可视化与失败案例分析

图 11 展示了 OV2Seg 在我们提出的 LV-VIS、Youtube-VIS [50] 和 OVIS 上的结果 [39]，见图11。OV2Seg 在这些视频实例分割数据集上展现出很强的泛化能力，即使在一些困难情况下也不例外，具体包括：(b)视角变化较大、(c)视频模糊、(f)长视频、(a、g)遮挡，以及(g)物体数量较多。

图中展示了失败案例 11 (h)-(i)。主要的失败案例是类别冲突，即新类别物体的分类通常会被视觉上相似的基类主导。具体来说，如图11(i)所示，由于外观相似，该图被识别为“狗”。训练集中展示了“狗”的物体，而没有展示“狼”的物体，因此模型学会了将对象嵌入与“狗”的词嵌入更好地对齐，从而倾向于将物体识别为“狗”而非“狼”。我们认为，类别冲突是所有开放词汇任务面临的一项根本挑战。通过在训练期间纳入大型词汇集，或采用从图像域迁移信息的训练方案（例如知识蒸馏或自训练），有望改善这一问题。另一个失败案例是对一些常见类别的分割遗漏，例如图11(b)、(h)、(i)中的“人”。这是因为 LVIS 不是密集标

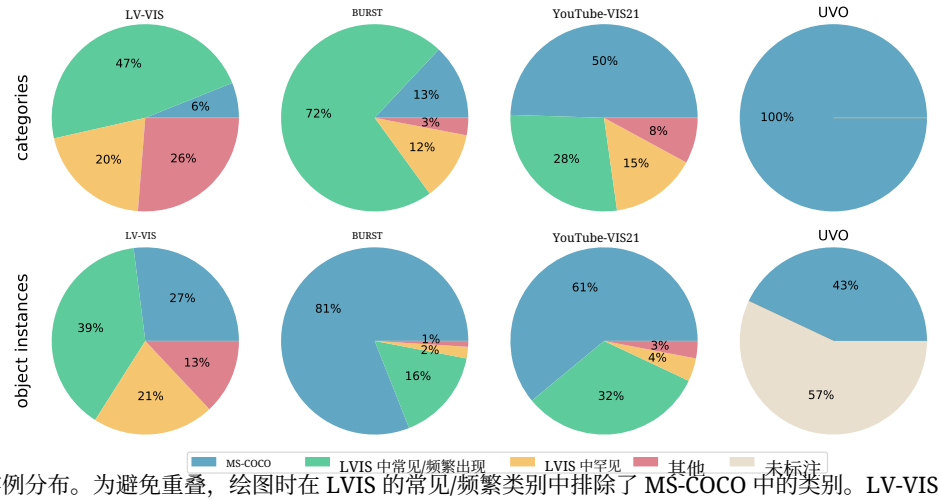


图 8. 类别与对象实例分布。为避免重叠，绘图时在 LVIS 的常见/频繁类别中排除了 MS-COCO 中的类别。LV-VIS 中有相当比例的类别和物体不同于常用数据集中的类别和物体。



图 9. LV-VIS 中的示例视频。

注数据集。数据集中只标注了部分物体，尤其是“人”等最常见物体。具体而言，LVIS 总共只标注了 13,439 个“人”，而 MS-COCO 中标注了 262,465 个 [29]。因此，训练集中大多数“人”物体都被视为背景，导致人类别的召回率较低。采用一些半监督训练方法，或将 LVIS 与 MS-COCO [29] 数据集结合，以补全 LVIS 中常见类别的遗漏标注，有望缓解这一问题。我们希望通过分析这些失败案例，启发未来的研究工作。

数据集	Type	类别					
Youtube-VIS2019	Base	飞机	bear	boat	cat	cow	deer	dog
		duck	鹰	大象	fish	frog	大熊猫	长颈鹿
		马	蜥蜴	猴子	摩托车	老鼠	owl	鹦鹉
		人	兔子	鲨鱼	滑板	单板滑雪板	冲浪板	网球拍
		老虎	训练	卡车	乌龟	斑马		
	小说	无耳海豹	fox	豹	蛇	ape	hand	轿车
Youtube-VIS2021	Base	飞机	bear	boat	car	cat	cow	deer
		dog	duck	鹰	大象	fish	frog	大熊猫
		长颈鹿	房子	蜥蜴	猴子	摩托车	老鼠	owl
		鹦鹉	人	兔子	鲨鱼	滑板	单板滑雪板	冲浪板
		网球	老虎	训练	卡车	乌龟	斑马	
	小说	无耳海豹	fox	豹	蛇	飞盘	鲸鱼	

表8. Youtube-VIS2019 和 Youtube-VIS2021 数据集中的基地类别和新类别。

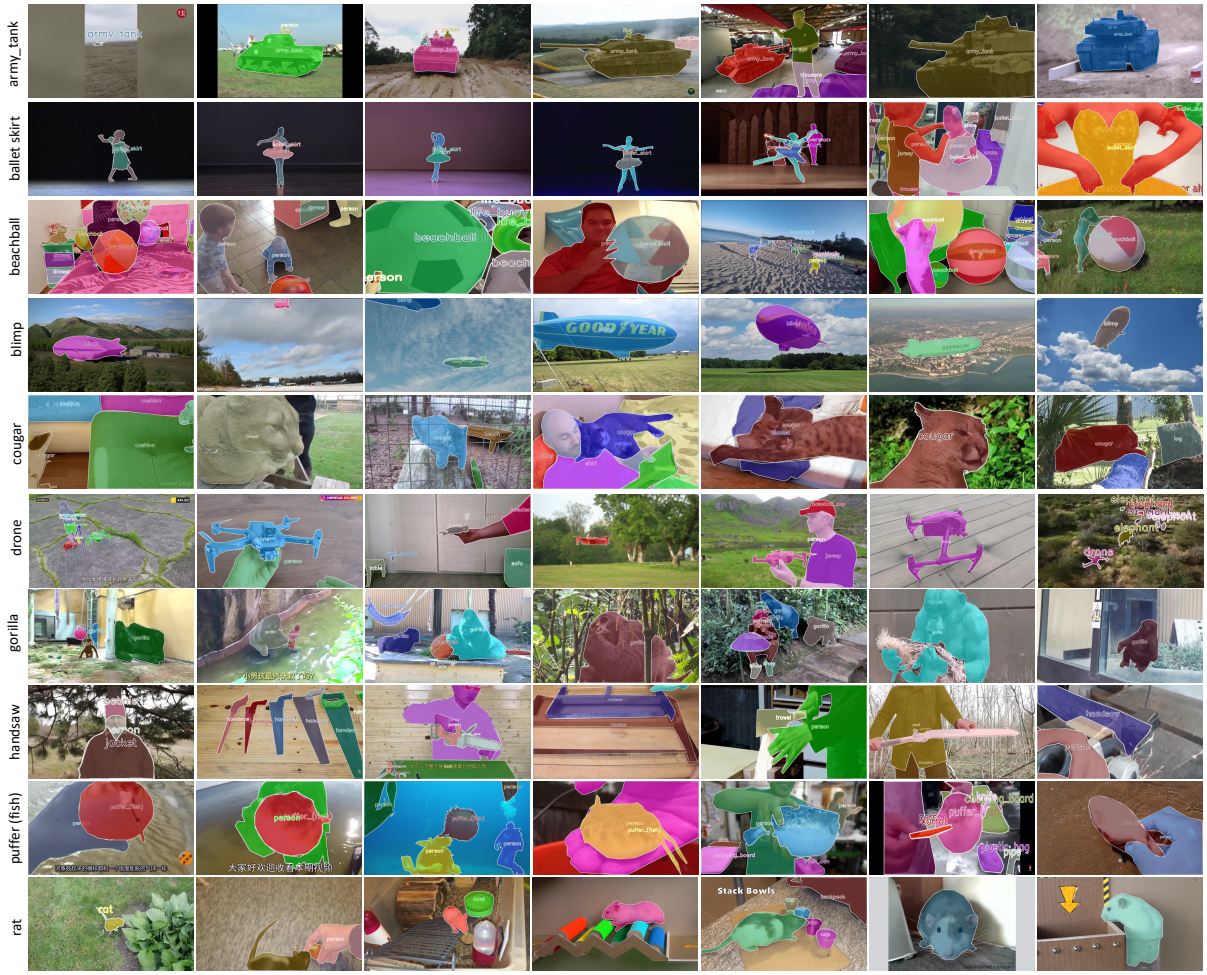


图10. 标注帧的截图。LV-VIS 中标注帧的截图。完整视频和标注将在发表后发布 .



图11。OV2Seg 在视频实例分割数据集上的预测结果。图（a）、（b）和（i）来自 LV-VIS 数据集。图（c）、（d）、（e）和（h）来自 Youtube-VIS2019/2021数据集。图（f）和（g）来自 OVIS数据集。图（h）和（i）是失败案例。