UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Towards Open-Vocabulary Video Instance Segmentation

Wang, H.; Yan, C.; Wang, S.; Jiang, X.; Tang, X.; Hu, Y.; Xie, W.; Gavves, E.

DOI
10.48550/arXiv.2304.01715
10.1109/ICCV51070.2023.00375

Publication date
2023

Document Version
Author accepted manuscript

Published in
2023 IEEE/CVF International Conference on Computer Vision

Link to publication

Citation for published version (APA):
Wang, H., Yan, C., Wang, S., Jiang, X., Tang, X., Hu, Y., Xie, W., & Gavves, E. (2023). Towards Open-Vocabulary Video Instance Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision: ICCV 2023 : Paris, France, 2-6 October 2023 : proceedings* (pp. 4034-4043). IEEE Computer Society. https://doi.org/10.48550/arXiv.2304.01715, https://doi.org/10.1109/ICCV51070.2023.00375

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Oct 2026

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Open-Vocabulary Video Instance Segmentation

Haochen Wang*
University of Amsterdam
h.wang3@uva.nl

Cilin Yan*
Beihang University
clyan@buaa.edu.cn

Shuai Wang
University of Amsterdam
s.wang3@uva.nl

Xiaolong Jiang, Xu Tang, Yao Hu
Xiaohongshu Inc.
laige@xiaohongshu.com

Weidi Xie†
Shanghai Jiao Tong University
weidi@robots.ox.ac.uk

Efstratios Gavves
University of Amsterdam
egavves@uva.nl

Abstract

Video Instance Segmentation (VIS) aims at segmenting and categorizing objects in videos from a closed set of training categories, lacking the generalization ability to handle novel categories in real-world videos. To address this limitation, we make the following three contributions. First, we introduce the novel task of Open-Vocabulary Video Instance Segmentation, which aims to simultaneously segment, track, and classify objects in videos from open-set categories, including novel categories unseen during training. Second, to benchmark Open-Vocabulary VIS, we collect a Large-Vocabulary Video Instance Segmentation dataset (LV-VIS), that contains well-annotated objects from 1,196 diverse categories, significantly surpassing the category size of existing datasets by more than one order of magnitude. Third, we propose an efficient Memory-Induced Transformer architecture, OV2Seg, to first achieve Open-Vocabulary VIS in an end-to-end manner with near real-time inference speed. Extensive experiments on LV-VIS and four existing VIS datasets demonstrate the strong zero-shot generalization ability of OV2Seg on novel categories. The dataset and code are released here https://github.com/haochenheheda/LVVIS.

1. Introduction

Despite impressive efforts, Video Instance Segmentation (VIS) [39, 9, 36] is fundamentally constrained to segment and classify objects from a closed set of training categories, thus limiting the capacity to generalize to novel concepts, see Fig. 1 (a). In real-world scenarios, this closed-vocabulary paradigm lacks practical value, as the model usually encounters objects from novel categories unseen in the training stage. In contrast, recent works Open-World Tracking (OWTB)[27], UVO [35], and BURST [1] aim to segment and track all visible objects, see Fig. 1 (b). However, those works still have no capacity to classify the objects from open-set categories, which is significant for video-level tasks, such as video captioning and action recognition. To enhance the practicality of VIS, and make VIS more suited to downstream video-level tasks, *we propose Open-Vocabulary Video Instance Segmentation* to simultaneously classify, track, and segment arbitrary objects from an open set of categories, as illustrated in Fig. 1 (c).

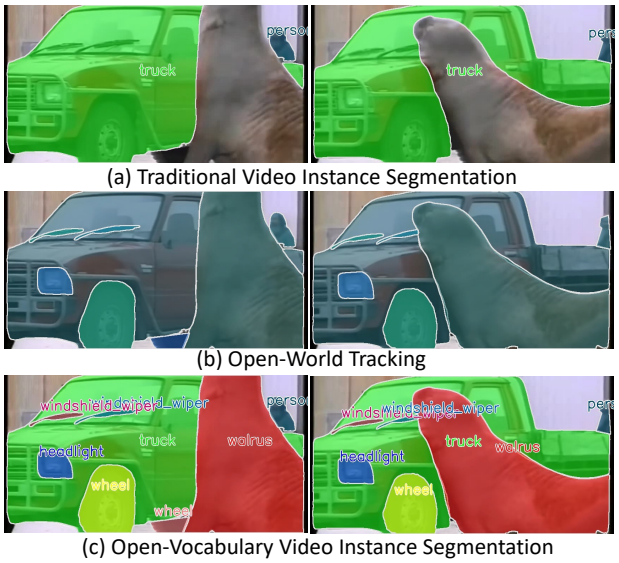

Figure 1. (a) Traditional VIS fails to tackle objects from novel categories (unseen during training, e.g. walrus in the figure); (b) Open-World Tracking aims to segment and track all objects in a class-agnostic manner, while lacking the ability for object categorization in videos; (c) Open-Vocabulary VIS aims to simultaneously segment, track, and classify objects for both training categories, e.g., person, and novel categories, e.g., walrus in the figure. Different colors in the figures indicate different object instances.

To properly benchmark the Open-Vocabulary VIS, a video dataset with large and diverse object categories is

*Equal contribution.
†Corresponding author.

Dataset	UVO [35]	YT19 [39]	YT21 [39]	OVIS [31]	BURST [1]	LV-VIS
Videos	11228	2883	3859	901	2914	4828
Instances	104898	4883	8171	5223	16089	25588
Masks	593k	131k	232k	296k	600k	544k
Mask/Frame	12.3	1.7	2.0	4.7	3.1	4.9
Object/Video	9.3	1.6	2.1	5.8	5.5	5.3
Categories	80*	40	40	25	482	1196

Table 1. Key statistics comparison between our LV-VIS dataset and published video segmentation datasets. $*$ indicates the UVO only provides category labels for objects from 80 common categories defined in MS-COCO. Our LV-VIS dataset contains a significantly larger category set than the existing datasets.

necessary. However, existing datasets Youtube-VIS [39], OVIS [31], and UVO [35] are not sufficient as they contain only a few dozen of categories, see Tab.1. BURST [1] contains a relevant larger category set. Still, 81.7% objects in BURST are from the common categories in MS-COCO, shown in Fig. 2 (b), and thus not diverse enough for the open-vocabulary evaluation of novel categories.

In this work, *we collect a Large-Vocabulary Video Instance Segmentation dataset (LV-VIS)*. LV-VIS contains 4,828 videos and over 544k instance masks from 1,196 categories, which is significantly larger than all existing video segmentation datasets. Notably, as illustrated in Fig. 2 (b), a large percentage of annotated objects in LV-VIS are distinct from the categories in the commonly used datasets such as MS-COCO [25] and LVIS [17], making it well suited for evaluating the generalization ability on novel categories of Open-Vocabulary VIS methods, and also a valuable supplement to the existing datasets.

Architecturally, a straightforward approach for Open-Vocabulary VIS is to associate per-frame results of open-vocabulary detectors [45, 43] with open-world trackers [27]. However, this propose-reduce-association approach desires intricate hand-crafted modules such as non-maximum suppression, and neglects video-level features for stable tracking and open-vocabulary classification in videos, leading to sub-optimal performance and inference speed.

In this work, we propose the *first end-to-end Open-Vocabulary Video Instance Segmentation model*, OV2Seg, which simplifies the intricate propose-reduce-association paradigm and attains long-term awareness with a Memory-Induced Transformer. Specifically, it starts by proposing and segmenting all objects with a Universal Object Proposal module, then a set of Memory Queries are introduced to incrementally encode object features through time, enabling long-term awareness for efficiently tracking all objects through time. Lastly, given arbitrary category names as input, a language transformer encoder is adopted to classify the tracked objects based on the Memory Queries. The Memory Queries incrementally aggregate the object features from different frames, thus leading to robust video object classification. To our knowledge, OV2Seg is the first end-to-end model with the capability of segmenting, tracking, and classifying objects in videos from arbitrary open-set categories with near real-time inference speed.

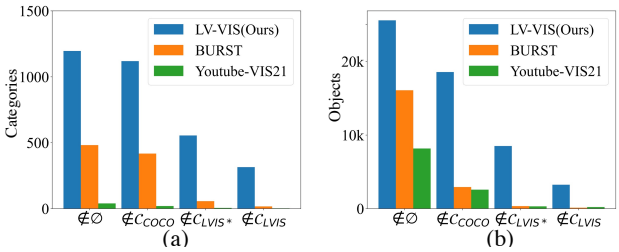

Figure 2. Figure (a) and (b) show the number of categories and object instances disjointed with a certain category set $\mathcal{C}$, respectively. $\mathcal{C}_{\text{COCO}}$, $\mathcal{C}_{\text{LVIS}*}$, and $\mathcal{C}_{\text{LVIS}}$ represent all categories in MS-COCO, frequent/common categories in LVIS, and all categories in LVIS. A larger percentage of categories and objects in LV-VIS are disjointed with the categories in commonly used datasets.

We evaluate OV2Seg on LV-VIS as well as four existing video segmentation datasets: Youtube-VIS2021, Youtube-VIS2019, OVIS, and BURST. *Without finetuning on downstream video datasets*, OV2Seg matches the performance of several fully-supervised competitors [39, 7] that have been trained on target video datasets, while exhibiting strong zero-shot generalization ability for novel categories unseen during the training phase.

To sum up, our main contributions are:

(i) we introduce Open-Vocabulary VIS, which simultaneously segments, tracks, and classifies objects of arbitrary open-set categories in videos, generalizing the traditional VIS for closed-set training categories;

(ii) we collect a large-scale, pixel-level annotated video dataset LV-VIS, with a significantly larger vocabulary set than the existing video datasets, thus being a suitable testbed for Open-Vocabulary VIS;

(iii) we propose the first end-to-end Open-Vocabulary VIS model, OV2Seg, that can segment, track, and classify objects from novel categories with a Memory-Induced Transformer architecture.

2. Related Work

Video Instance Segmentation aims to segment, track and classify object instances from pre-defined training categories, such as people, cars, and dogs. Existing methods could be divided into frame-based methods [4, 26, 24, 41, 19, 40, 18] and clip-based methods [36, 2, 23, 37, 9, 42, 21]. Frame-based methods, such as MaskTrack R-CNN [39], propose to include an additional tracking head to the image instance segmentation model Mask R-CNN [19] for video instance segmentation. The clip-based methods, such as IFS [23], SeqFormer [37] and Mask2Former [9], take the whole video or a clip of video as input and simultaneously output the masks and classification scores for all the input frames. All the aforementioned models necessitate training on a substantial dataset of videos that include densely

annotated objects from the training categories. In contrast, our proposed OV2Seg is trained on image datasets, yet it can be directly applied to videos during testing and has the capability to tackle objects from novel categories.

Open-Vocabulary Object Detection aims to detect objects in images beyond a close vocabulary set, which is proposed in ZSD [3]. Inspired by CLIP [32], ViLD [16] and ZSD-YOLO [38] propose to explicitly distill knowledge from the pre-trained CLIP visual embedding into the backbone of Faster RCNN. Detic [45] and PromptDet [15] strive to expand the vocabulary of the detector by self-training on additional images. Those methods inherit the standard architecture of a two-stage detector, which follows a complex propose-and-reduce pipeline. OV-DETR [43] made the first attempt to model an end-to-end open vocabulary detector with an architecture like DETR [8], while still limited in the inference speed caused by class-dependent object queries.

Open World Tracking. Multiple Object Tracking [44, 33, 6] aims to track multiple objects through time. OWTB [27] proposed Open World Tracking to segment and track all the objects in videos. OWTB achieves state-of-the-art performance on several multi-object tracking datasets. Recent work UVO [35] focuses on class-agnostic object segmentation and tracking in videos, neglecting the object classification performance during evaluation. Methods adopted in BURST [1], such as STCN [11] and MaskTrack [39], have no ability to classify objects from novel categories. Therefore BURST mainly follows the evaluation protocols of class-agnostic Multiple Object Tracking, while only measuring category-wise evaluation for the 80 training categories in MS-COCO. On the contrary, we introduce Open-Vocabulary VIS to entail open-vocabulary object categorization along with segmentation and tracking in videos. We establish several baseline models for Open-Vocabulary VIS by integrating open-world tracking methods [27] with open-vocabulary detectors [45, 43] to experimentally demonstrate the effectiveness of our proposed OV2Seg.

3. Setting of Open-Vocabulary VIS

Given a training dataset $\mathcal{D}_{train}$ containing pixel-level annotations for a set of training categories $\mathcal{C}_{\text{train}}$, traditional VIS aims to train a model $f_\theta(\cdot)$ which could be tested on a dataset $\mathcal{D}_{\text{test}} = \{V_i\}_{i=1}^{N}$, where $V_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ refers to a video clip with shape of (H_i, W_i) and T_i frames. $f_\theta(\cdot)$ is supposed to predict a sequence of segmentation mask $\{m_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times H_i \times W_i}$ and a category label $c \in \mathcal{C}_{train}$ for each object in videos from the training categories, while the objects from novel categories $\mathcal{C}_{novel}$ are ignored.

In contrast, Open-Vocabulary VIS aims to train a model $f_\theta(\cdot)$ on $\mathcal{D}_{train}$, and then test on $\mathcal{D}_{test}$ for both training categories $\mathcal{C}_{train}$ and novel categories $\mathcal{C}_{novel}$. Specifically, given a test video clip $V_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ during inference, the trained model $f_\theta(\cdot)$ is supposed to predict the segmentation mask sequence $\{m_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times H_i \times W_i}$ and the category label $c \in (\mathcal{C}_{train} \cup \mathcal{C}_{novel})$ for each object p in V_i:

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

where P is the total number of objects in video V_i, class label c belongs to the union of training and novel categories, and $m_t \in \mathbb{R}^{H_i \times W_i}$ is the segmentation mask for each frame t. In the experiments, we term the training categories as base categories, while the categories disjointed with base categories as novel categories.

Evaluation Metrics. We adopt the mean average precision (mAP) metric for both base categories and novel categories. The mAP in videos is defined in Youtube-VIS [39]. Specifically, we first compute the average precision (AP) of each object category among all the test videos, defined as the area under the precision-recall curve plotting with the category confidence score. The value of AP is measured under 10 intersection-over-union (IoU) thresholds from 50% to 95% at step 5%, following the standard setting in MS-COCO [25]. The IoU in videos is defined in Youtube-VIS [39], which computes the spatial-temporal consistency of predicted and ground truth segmentation masks. Finally, the mean average precision (mAP) is computed by averaging the AP over the base category set and novel category set, termed as mAP_b and mAP_n.

4. LV-VIS Dataset

To properly evaluate the Open-Vocabulary VIS, we establish a new dataset named LV-VIS with a large vocabulary set. This section outlines the data collection and annotation procedure, and provides a comparison of key dataset statistics with existing video-level datasets.

Data Collection and Manual Annotation. We start with collecting 7,569 video clips through the Internet by querying 1,621 well-defined category names. Then we manually remove inappropriate video clips such as videos with sensitive personal information, and videos with too many objects to annotate. Then we keep 4,828 videos in the LV-VIS. We maintain the original resolution of the videos, varying from 480p to 1080p. All the videos are pre-processed to a length between 2 and 10 seconds, and we extract 5 frames per second for all the processed videos. We thoroughly annotate all clearly visible objects in the selected videos. Each annotated object is assigned a category label, instance identity, and mask sequence in polygon format. Then we remove the uncovered categories, merge categories with synonyms, and then retain 1,196 well-defined object categories. The details are shown in Supplemental Materials Sec. A. We provide the visualization of annotated video clips in Supplemental Materials Sec. B.

Dataset Statistics. The LV-VIS dataset comprises 4,828 real-world videos with a total length of 6.2 hours and

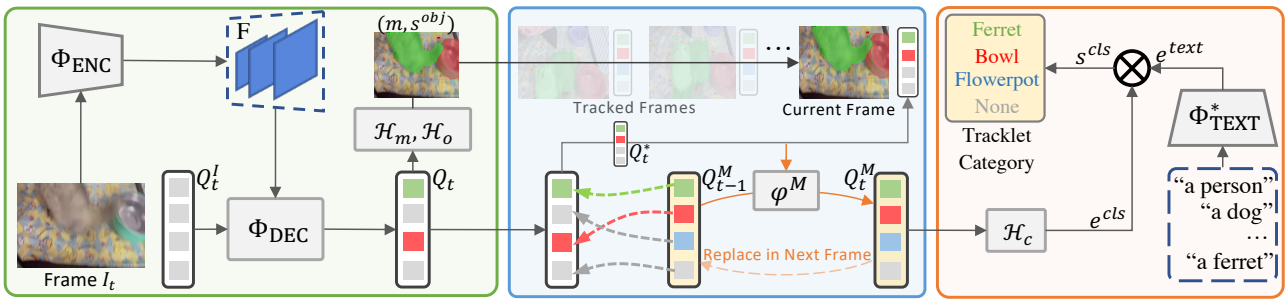

Figure 3. Overview of OV2Seg. (a) Universal Object Proposal: Input with frame I_t, a transformer encoder Φ_{ENC} is adopted to extract multi-scale features $\mathcal{F}$. N class-independent queries $Q_t^I \in \mathbb{R}^{N\times d}$ are fed to a transformer decoder Φ_{DEC} to generate object-centric queries Q_t. Then Q_t is utilized to produce the segmentation mask $m \in \mathbb{R}^{N\times H\times W}$ and object score $s^{\text{obj}} \in \mathbb{R}^{N\times 1}$ with a mask head $\mathcal{H}_m$ and an object score head $\mathcal{H}_o$, respectively. (b) Memory-Induced Tracking: A set of Memory Queries $Q_{t-1}^M \in \mathbb{R}^{N\times d}$ is proposed to associate Q_t with Hungarian Algorithm. Q_{t-1}^M is updated by a function φ^M to dynamically aggregate the associated object-centric queries Q_t^* through time and obtain Q_t^M for the tracking of the next frame. (c) Open-Vocabulary Object Classification: A class head $\mathcal{H}_c$ is applied on Q_t^M to generate the class embedding $\mathbf{e}^{\text{cls}} \in \mathbb{R}^{N\times d}$. $\mathbf{e}^{\text{cls}}$ is then dot-producted with the text embedding $\mathbf{e}^{\text{text}}$ generated by a text encoder $\Phi_{\text{TEXT}} \in \mathbb{R}^{|\mathcal{C}|\times d}$ to obtain the classification score $s^{cls} \in \mathbb{R}^{N\times|\mathcal{C}|}$ for a open set of categories $\mathcal{C}$.

contains 544,451 pixel-level annotated segmentation masks from 1,196 categories. The categories in LV-VIS are split into 641 base categories (seen during training) inherited from frequent and common categories in LVIS [17], and 555 novel categories disjoint with the base categories. As shown in Tab. 1, the vocabulary size of LV-VIS is significantly larger than Youtube-VIS (40 categories), OVIS (25 categories), UVO (80 categories), and BURST (482 categories) respectively. Moreover, 93.5%/46.4% of the categories in LV-VIS are disjoint from categories in MS-COCO and frequent/common categories in LVIS, whereas the corresponding values for BURST and Youtube-VIS are much lower at 86.7%/11.8% and 50%/15%, respectively. In the meantime, 33.3% objects in LV-VIS are from novel categories, while the BURST only contains 2.2% objects from novel categories, see Fig. 2. That means LV-VIS contains a greater number of novel objects not shown in commonly used datasets such as MS-COCO and LVIS, making it a valuable supplement to the existing datasets and thus more suitable for the category-wise evaluation of novel categories. Due to the space limitation, more extensive dataset statistics are shown in Supplementary Material Sec. A.

Dataset Split. LV-VIS is split into a train set, a validation set, and a test set, which contains 3083, 837, and 908 videos respectively. We will publicly release the videos and annotations of the train set and validation set, and host an online evaluation server to benchmark related efforts with the challenging test set.

5. Architecture of OV2Seg

After constructing the dataset, we here present an end-to-end Open-Vocabulary VIS architecture, OV2Seg, as shown in Fig. 3. Generally speaking, the proposed architecture consists of three modules, (i) a Universal Object Proposal module, proposes and segments objects with a set of class-independent object queries; (ii) a Memory-Induced Tracking module, tracks the object proposals from consecutive frames with a set of Memory Queries, which incrementally aggregate object features through time to establish long-term awareness; (iii) an Open-Vocabulary Classification module, classifies the tracked objects with the text embeddings computed from a pre-trained text encoder given arbitrary category names as input. We detail the architecture in the following sections.

5.1. Universal Object Proposal

We feed each input frame I_t into a transformer encoder Φ_{ENC}, consisting of multiple deformable attention layers [46], to get the multi-scale feature maps $\mathcal{F}$:

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

where f^l refers to the feature map at l-th level. To propose objects from all categories, one state-of-the-art image-level open-vocabulary detector OV-DETR [43] uses class-dependent object queries conditioned on the given category names. In this way, each category is processed by a set of separate object queries, leading to an infeasible inference speed on real-world videos.

In this work, we experimentally prove that a set of class-independent object queries could efficiently propose objects from all categories, avoiding separate processes for each category. Therefore, we adopt N class-independent learnable object queries $Q^I \in \mathbb{R}^{N\times d}$ to obtain the class-agnostic object proposals. Specifically, the object queries Q^I and the multi-scale feature maps $\mathcal{F}$ are fed into a transformer decoder Φ_{DEC}, where the Q^I iteratively attend to each layer of $\mathcal{F}$, generating object-centric queries Q:

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

Then the output object-centric queries Q from the transformer decoder are passed into a mask generation head $\mathcal{H}_m$ and an object score head $\mathcal{H}_o$:

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

where f^{-1} is the last layer of $\mathcal{F}$. $\mathcal{H}_m(\cdot)$ consists of 3 multilayer perceptron (MLP) layers and outputs the dynamic convolutional kernel for each object query. The dynamic convolution kernels are then convolved with f^{-1} to obtain the per-frame segmentation masks $m \in \mathbb{R}^{N\times H\times W}$. $\mathcal{H}_o(\cdot)$ also consists of 3 MLP layers and outputs binary scores $s^{\text{obj}} \in \mathbb{R}^{N\times 1}$, indicating if a query represents an object.

5.2. Memory-Induced Tracking

The object-centric queries Q in consecutive frames could be associated with the Hungarian algorithm to generate the object tracklets in videos, inspired by MinVIS [22]. However, directly associating objects of consecutive frames neglects long-term dependencies in videos. As such, this approach fails in complex cases such as objects disappearing and reappearing in the videos.

To this end, we propose to maintain a set of Memory Queries $Q^M \in \mathbb{R}^{N\times d}$ to obtain long-term dependency, see Fig. 3 (b). Specifically, we compute the inner-product similarity between Memory Queries Q^M_{t-1} of the last frame and the object-centric queries Q_t. Each object-centric query is associated with one of the Memory Queries by the Hungarian algorithm upon the similarity matrix. The Q^M_0 is initialized by the Q_0 of the first frame. Afterward, it is momentum updated through time with an update function $\varphi^M(\cdot,\cdot)$, thereby gradually encoding the video-level object features to tackle objects with disappear or serious occlusion. Specifically, $\varphi^M(\cdot,\cdot)$ is defined as:

$$Q^M_t = \varphi^M(Q^M_{t-1}, Q^*_t) \tag{5}$$

$$= \alpha \cdot s^{\text{obj}} \cdot Q^*_t + (1 - \alpha \cdot s^{\text{obj}}) \cdot Q^M_{t-1}, \tag{6}$$

where α is a factor to control the update ratio and s^{obj} is the object score for each query. The Q^*_t is the associated object-centric queries after Hungarian algorithm. If an object is seriously occluded or disappears during a video, the corresponding object score s^{obj} tends to be low and therefore constrains the memory update process to keep the object memory features unchanged. By maintaining a small set of Memory Queries, OV2Seg efficiently tracks all the segmented objects through time with long-term awareness, and enhances the following object classification with incremental video object features.

5.3. Open-Vocabulary Classification

We feed the vocabulary set into a frozen pretrained CLIP Text Encoder Φ_{TEXT} to generate the text embedding as the classifiers. For example, a 'dog' classifier can be:

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

in which the vocabularies could be replaced by arbitrary novel categories during inference.

The class embeddings $\mathbf{e}^{\text{cls}} \in \mathbb{R}^{N\times d}$ for the tracked objects are obtained by $\mathbf{e}^{\text{cls}} = \mathcal{H}_c(Q^M_t)$, where $\mathcal{H}_c(\cdot)$ indicates the class head consisting three MLP layers. Then we can obtain the classification score s^{cls} for each object tracklet by computing the cosine similarity between class embedding $\mathbf{e}^{\text{cls}} \in \mathbb{R}^{N\times d}$ and text embedding $\mathbf{e}^{\text{text}} \in \mathbb{R}^{|\mathcal{C}|\times d}$:

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

where $i \in [1, N]$ and $j \in [1, |\mathcal{C}|]$ are indexes of object query and text embedding. σ is the sigmoid function. $\cos(\cdot,\cdot)$ refers to the cosine similarity, and ϵ is the temperature hyper-parameter. The memory queries Q^M_t aggregate the object features from all frames weighted by the object score, the object features with low confidence are constrained thus leading to robust video object classification.

5.4. Training and Loss

As there is no learnable parameter in the Memory-Induced Tracking, we can exclude the tracking module and train OV2Seg on image-level datasets, such as LVIS. This reduces the training resource requirements of large-scale video datasets and improves training efficiency. We adopt a similar training loss in DETR [8] for set prediction:

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

We use the binary cross-entropy loss for classification loss $\mathcal{L}_{\text{cls}}$ and object score loss $\mathcal{L}_{\text{obj}}$. The mask loss $\mathcal{L}_{\text{mask}}$ is the sum of dice loss [30] and binary focal loss. The ground truth $\hat{y}$ consists of an object label $\hat{s}^{\text{obj}}$, a class label $\hat{s}^{\text{cls}}$ and segmentation mask $\hat{m}$ for each object. We assign the ground truth set $\hat{y}$ to the prediction set y by minimizing the total loss function $\mathcal{L}_{\text{match}}$.

6. Experiments

We first introduce the datasets and evaluation metrics. Then we describe the method implementations. Lastly, we present a thorough analysis of the experimental results.

6.1. Datasets and Metrics

We train OV2Seg on the union of common and frequent categories in LVIS [17], and evaluate the performance on multiple video instance segmentation datasets, including our collected LV-VIS dataset, as well as Youtube-VIS2019 [39], Youtube-VIS2021 [39], OVIS [31], and BURST [1]. Notably, we do not finetune OV2Seg on any videos from the target datasets.

LVIS is a widely used image open-vocabulary detection dataset, which contains a large set of 1,203 categories. Following ViLD [16], we take frequent and common categories as the base categories and all categories disjointed with base categories as novel categories.

Youtube-VIS is the most widely used video instance segmentation dataset, which contains 40 categories for both YVIS19 and YVIS21 versions. We divide the categories in YVIS19 and YVIS21 into the base and novel categories following the partitions in LVIS.

BURST is a recently published video dataset extending TAO [12]. BURST contains 425 base categories and 57 novel categories following the partitions in LVIS.

OVIS has 25 object categories, with only one category not in LVIS base categories; therefore, we only report the performance of the overall categories for OVIS.

Metrics. We report the mean Average Precision mAP for overall categories, and further break it down to mAP_b for base categories and mAP_n for novel categories. For the evaluation of LVIS [17], we report the mAP for overall categories and mAP_n for novel (rare) categories.

6.2. Implementation Details

Baseline Models. We build several baseline models by combining off-the-shelf two-stage open-vocabulary detection methods with the open-world tracking methods following the propose-reduce-association scheme. Specifically, we generate 300 object proposals for each frame by the open-vocabulary detectors [45, 14]. Then we feed the proposals to SORT [5] and OWTB [27] to achieve object tracking. The SORT is a classic multiple object tracking method, and the OWTB is one of the state-of-the-art open-world tracking methods, that combines the IOU matrix and ROI feature similarity to estimate proposal similarity across frames. For efficient evaluation, we remove the optical flow in OWTB. In this way, we build four baseline models, namely, Detic-SORT, Detic-OWTB, Detpro-SORT, and Detpro-OWTB, to compare with our proposed OV2Seg. We further design a two-stage method by propagating the segmentation masks of Detic with a state-of-the-art semi-supervised video object segmentation method XMem [10].

OV2Seg. We experiment with both ResNet50 [20] and SwinB [28] as backbones. We adopt 6 multi-scale deformable attention layers [46] as the transformer encoder and 6 masked-attention layers [9] as the transformer decoder. We use the pretrained text encoder extracted from the CLIP ViT-32 model to generate the open-vocabulary classifiers. Parameters in the text encoder are frozen during the training. The temperature hyper-parameter ϵ is set to 50. We adopt the prompt learning in Detpro [14] to replace the hand-crafted prompt in the text encoder module. The number of object queries is set to 100 for all experiments. We set λ_{obj}=2, λ_{cls}=2, λ_{mask}=5 as the weight of each loss. The memory update factor α is set to 0.7.

Training Details. We train OV2Seg on LVIS for 30 epochs with a batch size of 8. We adopt AdamW [29] optimizer. The base learning rate is set to 1e-4 and then reduced by a factor of 10 at 0.9 and 0.95 fractions of the total number of training steps. The learning rate of backbone parameters is factorized by 0.1. We use the large-scale jittering [13] with a random scale sampled from the range 0.1 to 2.0. For efficient training, we only sample 50 categories to compute the binary cross-entropy loss in one iteration. The training is conducted on 4 A100 GPUS for around 36 hours and all the inferences are conducted on a single A100 GPU.

6.3. Results

In this section, we first report the result comparison on the LV-VIS dataset between OV2Seg and a set of strong baseline models. Then we conduct extensive ablation studies to illustrate the effectiveness of the main modules in OV2Seg. Finally, we report the zero-shot generalization performance of OV2Seg on existing video instance segmentation datasets: Youtube-VIS2019, Youtube-VIS2021, BURST, and OVIS.

6.3.1 Results on LV-VIS dataset

We compare our OV2Seg with the propose-reduce-association baseline models on the LV-VIS dataset in Tab. 2. OV2Seg with ResNet-50 backbone achieves 11.9 mAP_n and 8.9 mAP_n on LV-VIS validation set and test set, outperforming the DetPro-XMem by 6.5 and 5.3. As shown in Tab. 3, OV2Seg matches the performance of Detpro [14] and Detic [45] on image datasets, while outperforming Detpro-OWTB and Detic-OWTB by a large margin on video dataset LV-VIS. The comparison outlines the effectiveness of OV2Seg on video-level tasks. This is mainly because the several-stage baseline models require complex hand-craft association rules and classify the objects by averaging the per-frame classification scores, ignoring the video-level correspondence.

In contrast, OV2Seg associates the object queries with the Memory Queries, which encodes the long-term object features for robust tracking and open-vocabulary classification in videos. Notably, OV2Seg outperforms Detic-XMem by a relative 120% improvement on novel categories of the LV-VIS validation set. This is significantly higher than the 28.4% improvement on base categories. As such, we conclude that OV2Seg has generalization abilities on novel categories in real-world videos. As shown in Tab. 2, the OV2Seg with Resnet50 backbone achieves 20.1 frame-per-second on LV-VIS, outperforming the Detic-OWTB by 14.2. We provide some visualizations in Fig. 6. Due to space limitations, more visualizations and failure cases are depicted in the Supplementary Material Sec. C.

Method	Backbone	Val			Test			fps
		AP	AP_b	AP_n	AP	AP_b	AP_n	
DetPro [14]-SORT [5]	R50	6.4	10.3	3.5	5.8	10.8	2.1	3.1
Detic [45]-SORT [5]	R50	6.5	10.7	3.4	5.7	10.6	2.1	6.0
DetPro [14]-OWTB [27]	R50	7.9	12.9	4.2	7.0	12.6	2.9	3.1
Detic [45]-OWTB [27]	R50	7.7	12.6	4.2	7.0	12.8	2.8	5.9
Detic [45]-XMem [10]	R50	8.8	13.4	5.4	7.7	13.3	3.6	16.4
OV2Seg(Ours)	R50	**14.2**	**17.2**	**11.9**	**11.4**	**14.9**	**8.9**	**20.1**
Detic [45]-SORT [5]	SwinB	12.8	21.1	6.6	9.4	15.8	4.7	6.7
Detic [45]-OWTB [27]	SwinB	14.5	22.6	8.5	11.8	19.6	6.1	6.8
Detic [45]-XMem [10]	SwinB	16.3	24.1	10.6	13.1	20.5	7.7	13.4
OV2Seg(Ours)	SwinB	**21.1**	**27.5**	**16.3**	**16.4**	**23.3**	**11.5**	**16.8**

Table 2. The performance comparison on LV-VIS validation and test set. The AP, AP_b, and AP_n mean the average precision of overall categories, base categories, and novel categories.

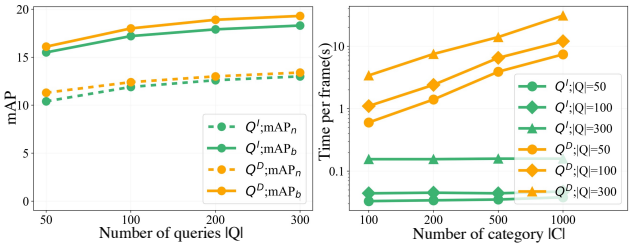

Figure 4. mAP and inference time per frame (s) comparison on LV-VIS validation dataset between class-dependent queries (Q^D) and class-independent queries (Q^I).

6.3.2 Ablation Study

We conduct extensive ablation studies on the LVIS, LV-VIS validation, and Youtube-VIS2019 datasets to thoroughly validate the effectiveness of the main modules in OV2Seg.

Image-level Open-Vocabulary Detection. By removing the Memory-Induced Tracking module, OV2Seg can also conduct image-level Open-Vocabulary Detection. We compare OV2Seg with several state-of-the-art open-vocabulary detection methods on LVIS and LV-VIS(O) (image oracle evaluation) to demonstrate the effectiveness and efficiency of OV2Seg on image-level tasks. As shown in Tab. 3, OV2Seg achieves 17.5 and 7.4 mAP_n on LVIS and LV-VIS(O) respectively with Resnet50 backbone. This matches existing two-stage state-of-the-art approaches Detic [45] and Detpro [14], while allowing for faster inference speeds. Specifically, Detic and Detpro require separate region proposal networks and box regression modules to generate the objects, as well as non-maximum suppression to reduce the redundant objects, resulting in a slow inference speed of 6.2 fps and 3.4 fps. OV-DETR [43], the first transformer-based open-vocabulary detector, achieves 17.4 and 7.2 mAP_n on LVIS and LV-VIS(O) with inference speed of 0.1 fps, limited by the class-dependent object queries. In contrast, our proposed OV2Seg uses class-independent object queries, avoiding separate processes for each category. As a result, OV2Seg achieves 21.6 fps, outperforming the fps of OV-DETR for more than two orders of magnitude. More generally, we claim that high accuracy on image datasets is not the main goal of this work. Instead, we prefer a simple and effective model for Open-Vocabulary VIS. Therefore, we do not apply self-training and information distillation in our work, which are adopted in Detic and Detpro.

Method	Backbone	Stage	LVIS		LV-VIS(O)		fps
			AP	AP_n	AP	AP_n	
ViLD [16]	R50	2	25.5	16.6	-	-	-
DetPro* [14]	R50	2	25.7	18.7	10.8	7.6	3.4
Detic [45]	R50	2	30.2	16.4	11.0	6.9	6.2
OV-DETR [43]	R50	1	26.6	17.4	11.4	7.2	0.10
OV2Seg(Ours)	R50	1	25.4	17.5	11.0	7.4	21.6
Detic [45]	SwinB	2	38.4	21.9	13.9	7.8	7.6
OV2Seg(Ours)	SwinB	1	31.4	21.8	14.2	8.3	18.2

Table 3. Image-level performance on LVIS dataset and LV-VIS validation dataset. The LV-VIS(O) means the oracle image evaluation, where we regard all the frames in LV-VIS as an image dataset and evaluate on the images separately. DetPro* is conducted by single-scale inference for fair comparison.

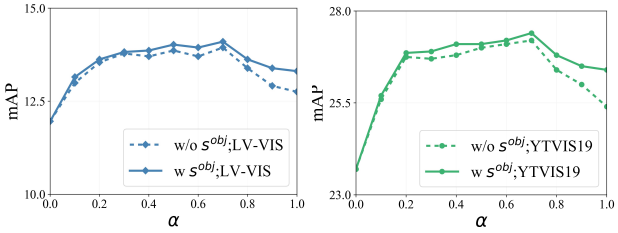

Figure 5. Overall mean Average Precision (mAP) under different memory update factor α on LV-VIS and Youtube-VIS2019 validation datasets. The w/o s^{obj} means removing the object score s^{obj} in the memory update module.

Class-Dependent vs. Class-Independent Queries. OV-DETR [43] utilizes N conditional object queries for each category, which we refer to as class-dependent queries Q^D. We compare it with OV2Seg using class-independent queries Q^I and show the results in Fig. 4. We see that OV2Seg with class-independent queries Q^I achieves comparable accuracy as with class-dependent queries Q^D, while being two orders of magnitude faster. As shown in Fig. 4 (b), OV2Seg with class-dependent queries becomes significantly slower as the number of categories $|\mathcal{C}|$ increases, and thus is intractable in open-world videos containing objects from large category set.

Memory Queries. We evaluate OV2Seg on LV-VIS validation and Youtube-VIS2019 datasets under different memory update factors α and s^{obj} in Equation 6. As shown in Fig. 5, the query association module without memory update, i.e., α=0, only achieves 23.7 mAP and 11.9 mAP on Youtube-VIS2019 and LV-VIS validation datasets. When α is set to a value between 0.3 and 0.8, OV2Seg achieves stable performance improvements, which shows that the memory update schedule is robust to the factor α. The best performance is obtained by setting $\alpha = 0.7$ and the object score s^{obj} is adopted, achieving 27.2 mAP and 14.2 mAP on Youtube-VIS2019 and LV-VIS validation dataset. The

Method	Open	Backbone	YTVIS2019			YTVIS2021			BURST			OVIS
			mAP	mAP_b	mAP_n	mAP	mAP_b	mAP_n	mAP	mAP_b	mAP_n	mAP
FEELVOS [39]	✗	R50	26.9	-	-	-	-	-	-	-	-	9.6
MaskTrack [39]	✗	R50	30.3	-	-	28.6	-	-	-	-	-	10.8
SipMask [7]	✗	R50	33.7	-	-	31.7	-	-	-	-	-	10.2
Mask2Former [9]	✗	R50	46.4	-	-	40.6	-	-	-	-	-	17.3
Detic [45]-SORT [5]	✓	R50	14.6	17.0	3.5	12.7	14.4	3.1	1.9	1.8	**2.5**	6.7
Detic [45]-OWTB [27]	✓	R50	17.9	20.7	4.5	16.7	18.6	5.8	2.7	2.8	1.8	9.0
OV2Seg(Ours)	✓	R50	**27.2**	**30.1**	**11.1**	**23.6**	**26.5**	**7.3**	**3.7**	**3.9**	2.4	**11.2**
Detic [45]-SORT [5]	✓	SwinB	23.8	27.2	7.9	21.6	23.7	9.8	2.5	2.7	1.0	11.7
Detic [45]-OWTB [27]	✓	SwinB	30.0	34.3	9.7	27.1	29.9	11.4	3.9	4.1	2.4	13.6
OV2Seg(Ours)	✓	SwinB	**37.6**	**41.1**	**21.3**	**33.9**	**36.7**	**18.2**	**4.9**	**5.3**	**3.0**	**17.5**

Table 4. Performance comparison on the validation sets of four published datasets: Youtube-VIS19, Youtube-VIS21, BURST, OVIS. The Open in the table indicates whether a method is capable of tackling objects from novel categories following the setting of Open-Vocabulary VIS. The methods above the double horizontal lines are trained on target training videos of each dataset; we only report the overall mean average precision mAP of those methods. The methods below the double horizontal lines are trained on image dataset LVIS and evaluated on the video instance segmentation datasets without fine-tuning; we report both the mAP, mAP_b, and mAP_n of those methods.

Methods	LV-VIS		Youtube-VIS2019	
	mAP_b	mAP_n	mAP_b	mAP_n
Average	16.9	11.3	29.4	10.8
Memory Queries	**17.2**	**11.9**	**30.1**	**11.1**

Table 5. Comparison between two ways to obtain the classification scores. The Memory Queries in the table indicate obtaining classification scores from the memory queries. The Average in the table means directly averaging the per-frame classification scores.

memory queries introduce long-term clues to handle occlusion and disappear-reappear cases, which helps to achieve robust object association. As shown in Fig 6, the snack is re-tracked after entire occlusion.

In Tab. 5, we compare the performance between the classification by Memory Queries and by Averaging per-frame classification scores. As shown, the performance of Memory Queries slightly outperforms the naive Average on both LV-VIS and Youtube-VIS2019 datasets. This is because Memory Queries aggregate video object features based on per-frame object confidence, which constrains the low confidence features and leads to more robust classification with respect to complex cases such as motion blur and occlusion.

6.3.3 Zero-shot Generalization on VIS Datasets.

We directly evaluate OV2Seg on four video instance segmentation datasets *without finetuning* on the target training videos. The performance comparison is shown in Tab. 4. As shown, OV2Seg with ResNet-50 achieves 11.1 mAP_n and 7.3 mAP_n on Youtube-VIS2019 and Youtube-VIS2021 novel categories respectively, surpassing the baseline model Detic-OWTB by 6.6 and 1.5. Note that without using any training videos of Youtube-VIS, OV2Seg achieves considerable performance compared with close-set fully-supervised competitors such as MaskTrack [39] and FEELVOS [34]. In BURST, OV2Seg achieves 3.7 mAP, outperforming Detic-OWTB by 1.0 mAP in terms of overall mAP. The BURST dataset is not exhaustively annotated, which makes category-wise evaluation of novel categories statistically unreliable. On the OVIS dataset with serious object occlusions, OV2Seg achieves 11.2 mAP, comparable with fully-supervised competitors MaskTrack [39], FEELVOS [39], and SipMask [7]. The comparison on OVIS demonstrates that OV2Seg has promising performance in complex occlusion circumstances. Visualizations are shown in Supplementary Material Sec. C.

Figure 6. Visualizations on LV-VIS dataset.

7. Conclusion

We propose a new task of Open-Vocabulary VIS with the goal of segmenting, classifying, and tracking objects of arbitrary categories simultaneously. To properly evaluate the Open-Vocabulary VIS, we collect a Large-Vocabulary Video Instance Segmentation dataset (LV-VIS), significantly surpassing the existing video-level segmentation datasets in terms of category sizes and diversity. Moreover, we propose the first end-to-end Open-Vocabulary VIS model, OV2Seg, that segments, classifies, and tracks objects from arbitrary categories with a Memory-Induced Vision Language Transformer. Without finetuning on videos, our proposed OV2Seg shows promising zero-shot generalization ability on novel categories unseen during training.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 2023. 1, 2, 3, 5

[2] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Proceedings of the European Conference on Computer Vision*, pages 158–177. Springer, 2020. 2

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision*, pages 384–400, 2018. 3

[4] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2020. 2

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *Proceedings of the IEEE international conference on image processing*, pages 3464–3468. IEEE, 2016. 6, 7, 8

[6] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020. 3

[7] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 1–18. Springer, 2020. 2, 8

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3, 5

[9] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1, 2, 6, 8

[10] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 6, 7

[11] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 3

[12] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *Proceedings of the European conference on computer vision*, pages 436–454. Springer, 2020. 6

[13] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. *arXiv preprint arXiv:2107.00057*, 2021. 6

[14] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 6, 7

[15] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Expand your detector vocabulary with uncurated images. In *Proceedings of the European Conference on Computer Vision*, 2022. 3

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *Proceedings of the International Conference on Learning Representations*, 2021. 3, 6, 7

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 2, 4, 5, 6

[18] Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, and Seon Joo Kim. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2896–2905, 2022. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[21] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. 2022. 2

[22] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *arXiv preprint arXiv:2208.02245*, 2022. 5

[23] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 2

[24] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1739–1748, 2021. 2

[25] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 3

[26] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video in-

stance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9816–9825, 2021. 2

[27] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19045–19055, 2022. 1, 2, 3, 6, 7, 8

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[29] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision*, pages 565–571. IEEE, 2016. 5

[31] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 2022. 2, 5

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[33] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017. 3

[34] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 8

[35] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 1, 2, 3

[36] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 1, 2

[37] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021. 2

[38] Johnathan Xie and Shuai Zheng. Zsd-yolo: Zero-shot yolo detection using vision-language knowledgedistillation. *arXiv preprint arXiv:2109.12066*, 2021. 3

[39] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 1, 2, 3, 5, 8

[40] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018. 2

[41] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8043–8052, 2021. 2

[42] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2885–2895, 2022. 2

[43] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. *arXiv preprint arXiv:2203.11876*, 2022. 2, 3, 4, 7

[44] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3

[45] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision*, 2022. 2, 3, 6, 7, 8

[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations*, 2020. 4, 6