



Occluded Video Instance Segmentation: A Benchmark

Jiyang Qi^{1,2} · Yan Gao² · Yao Hu² · Xinggang Wang¹ · Xiaoyu Liu² · Xiang Bai¹ · Serge Belongie³ · Alan Yuille⁴ · Philip H. S. Torr⁵ · Song Bai^{2,5}

Received: 8 November 2021 / Accepted: 17 May 2022 / Published online: 18 June 2022
© The Author(s) 2022

Abstract

Can our video understanding systems perceive objects when a heavy occlusion exists in a scene? To answer this question, we collect a large-scale dataset called OVIS for occluded video instance segmentation, that is, to simultaneously detect, segment, and track instances in occluded scenes. OVIS consists of 296k high-quality instance masks from 25 semantic categories, where object occlusions usually occur. While our human vision systems can understand those occluded instances by contextual reasoning and association, our experiments suggest that current video understanding systems cannot. On the OVIS dataset, the highest AP achieved by state-of-the-art algorithms is only 16.3, which reveals that we are still at a nascent stage for understanding objects, instances, and videos in a real-world scenario. We also present a simple plug-and-play module that performs temporal feature calibration to complement missing object cues caused by occlusion. Built upon MaskTrack R-CNN and SipMask, we obtain a remarkable AP improvement on the OVIS dataset. The OVIS dataset and project code are available at <http://songbai.site/ovis>.

Keywords Video instance segmentation · Occlusion reasoning · Dataset · Video understanding · Benchmark

Mathematics Subject Classification 68T07 · 68T45

Communicated by Chen Change Loy.

Jiyang Qi and Yan Gao these authors contributed equally to this work.

✉ Song Bai
songbai.site@gmail.com

Jiyang Qi
jiyangqi@hust.edu.cn

Yan Gao
yangao0119@gmail.com

Yao Hu
yaoohu@alibaba-inc.com

Xinggang Wang
xgwang@hust.edu.cn

Xiaoyu Liu
xiaoyuliu1991xyl@gmail.com

Xiang Bai
xbai@hust.edu.cn

Serge Belongie
s.belongie@di.ku.dk

Alan Yuille
alan.l.yuille@gmail.com

Philip H. S. Torr
philip.torr@eng.ox.ac.uk

1 Introduction

In the visual world, objects rarely occur in isolation. The psychophysical and computational studies (Hegdé et al., 2008; Nakayama et al., 1989) have demonstrated that human vision systems can perceive heavily occluded objects with contextual reasoning and association. The question then becomes, *can our video understanding system perceive objects that are severely obscured?*

Our work aims to explore this matter in the context of video instance segmentation, a popular task proposed in Yang et al. (2019) that targets a comprehensive understanding of objects in videos. To this end, we explore a new and challenging scenario called *Occluded Video Instance Segmentation (OVIS)*, which requests a model to simultaneously detect, segment, and track object instances in occluded scenes.

¹ Huazhong University of Science and Technology, Wuhan, China

² Alibaba Group, Beijing, China

³ University of Copenhagen, Copenhagen, Denmark

⁴ Johns Hopkins University, Baltimore, MD, USA

⁵ University of Oxford, Oxford, UK



Fig. 1 Sample video clips from OVIS. Click them to watch the animations (best viewed with Acrobat/Foxit Reader). The hairs and whiskers of animals are all exhaustively annotated

As the major contribution of this work, we collect a large-scale dataset called OVIS, specifically for video instance segmentation in occluded scenes. While being the second video instance segmentation dataset after YouTube-VIS (Yang et al., 2019), OVIS consists of 296k high-quality instance masks out of 25 commonly seen semantic categories. Some example clips are given in Fig. 1. The most distinctive property of OVIS dataset is that most objects are under severe occlusions. The occlusion level of each object is also labeled (as shown in Fig. 2) and we also present an AP (average precision) based metric to measure performance under different occlusion degrees. Therefore, OVIS is a useful testbed to evaluate video instance segmentation models for dealing with heavy object occlusions.

To dissect the OVIS dataset, we conduct a thorough evaluation of 9 state-of-the-art algorithms whose code is publicly

available, including FEELVOS (Voigtlaender et al., 2019a), IoUTacker+ (Yang et al., 2019), MaskTrack R-CNN (Yang et al., 2019), SipMask (Cao et al., 2020), STEm-Seg (Athar et al., 2020), STMask (Li et al., 2021), TraDeS (Wu et al., 2021), CrossVIS (Yang et al., 2021), and QueryVIS (Fang et al., 2021). However, the experimental results suggest that current video understanding systems fall behind the capability of human beings in terms of occlusion perception. The highest AP is only 16.3 achieved by Yang et al. (2021) and the highest AP on the heavily occluded group is only 6.3 achieved by Li et al. (2021). In this sense, we are still far from deploying those techniques into practical applications, especially considering the complexity and diversity of scenes in the real visual world.

To alleviate the occlusion issue, we also present a plug-and-play module called temporal feature calibration. For a given query frame in a video, we resort to a reference frame to complement its missing object cues. Specifically, the proposed module learns a calibration offset for the reference frame with the guidance of the query frame, and then the offset is used to adjust the feature embedding of the reference frame via deformable convolution (Dai et al., 2017). The refined reference embedding is used in turn to assist the object recognition of the query frame. Our module is a highly flexible plug-in. While applied to MaskTrack R-CNN (Yang et al., 2019) and SipMask (Cao et al., 2020) respectively, we obtain an AP of 15.4 and 14.3, significantly outperforming the corresponding baselines by 4.6 and 4.1 in AP respectively.

To summarize, our contributions are three-fold:

- We advance occlusion handling and video instance segmentation by releasing a new benchmark dataset named *OVIS* (short for *Occluded Video Instance Segmentation*). OVIS is designed with the philosophy of perceiving object occlusions in videos, which could reveal the complexity and the diversity of real-world scenes.

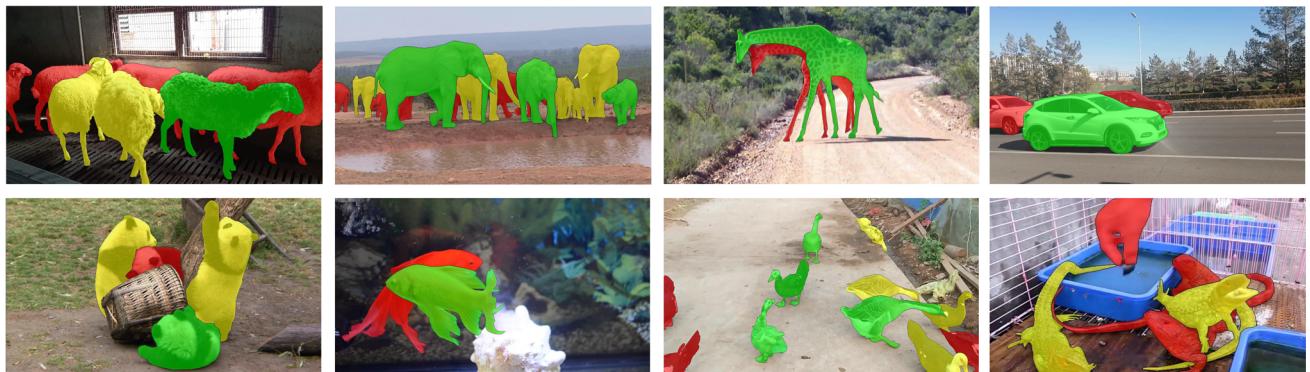


Fig. 2 Different occlusions levels in OVIS. Unoccluded objects are colored green, slightly occluded objects are colored yellow, and severely occluded objects are colored red (Color figure online)

- We streamline the research over the OVIS dataset by conducting a comprehensive evaluation of 9 state-of-the-art video instance segmentation algorithms, which could be a baseline reference for future research on OVIS.
- As a minor contribution, we present a plug-and-play module called Temporal Feature Calibration to alleviate the occlusion issue. Using MaskTrack R-CNN (Yang et al., 2019) and SipMask (Cao et al., 2020) as baselines, the proposed module obtains remarkable improvements on both OVIS and YouTube-VIS. More importantly, its “plug-and-play” nature makes it widely applicable to future endeavors on OVIS.

Compared with our conference version (Qi et al., 2021) that briefly describes the OVIS dataset and challenge held in 2021, the improvements are concluded as follows: (1) more thorough experiments (e.g., oracle experiments, error analysis, per-class result analysis) are conducted to dissect the OVIS dataset and the occlusion problem; (2) we comprehensively evaluate the effect of leveraging temporal context and adjusting the NMS threshold adaptively on occlusion handling; (3) more baseline results (e.g., the results that training with augmented image sequences, the results obtained with larger backbone or larger input resolutions) are provided, which can be a better reference for future work; (4) we further summarize remaining difficulties and future directions that deserve attention in OVIS.

2 Related Work

2.1 Video Instance Segmentation

Our work focuses on Video Instance Segmentation in occluded scenes. The most relevant work to ours is Yang et al. (2019), which formally defines the concept of video instance segmentation and releases the first dataset called YouTube-VIS. Built upon the large-scale video object segmentation dataset YouTube-VOS (Xu et al., 2018), the 2019 version of YouTube-VIS dataset contains a total of 2883 videos, 4883 instances, and 131k masks in 40 categories. Its latest 2021 version contains a total of 3859 videos, 8171 instances, and 232k masks. While YouTube-VIS is not designed to study the occluded video understanding problem, most objects in the OVIS dataset are under severe occlusions. Our experimental results show that OVIS is much more challenging.

Since the release of the YouTube-VIS dataset, video instance segmentation has attracted great attention in the computer vision community, arising a series of algorithms recently. MaskTrack R-CNN (Yang et al., 2019) is the first unified model for video instance segmentation. It fulfills video instance segmentation by adding a tracking branch to the popular image instance segmentation method Mask

R-CNN (He et al., 2017). Lin et al. (2020) propose a modified variational auto-encoder architecture built on the top of Mask R-CNN. MaskProp (Bertasius & Torresani, 2020) is also a video extension of Mask R-CNN which adds a mask propagation branch to track instances by the propagated masks. SipMask (Cao et al., 2020) extends single-stage image instance segmentation to the video level by adding a fully-convolutional branch for tracking instances. STMask (Li et al., 2021) improves feature representation by spatial feature calibration and temporal feature fusion. Different from those top-down methods, STEm-Seg (Athar et al., 2020) proposes a bottom-up method, which performs video instance segmentation by clustering the pixels of the same instance. Built upon Transformers, VisTR (Wang et al., 2020) supervises and segments instances at the sequence level as a whole. IFC (Hwang et al., 2021) further reduces the computations of full space-time transformers by only executing attention between memory tokens. QueryVIS (Fang et al., 2021) follows a multi-stage paradigm and leverages the intrinsic one-to-one correspondence in queries across different stages. Based on FCOS (Tian et al., 2019), SGNet (Liu et al., 2021) dynamically divides instances into sub-regions and performs segmentation on each region. CrossVIS (Yang et al., 2021) uses the instance feature in the current frame to localize the same instance in other frames. Different from the tracking-by-detection paradigm, TraDeS (Wu et al., 2021) integrates tracking cues to assist detection.

2.2 Other Related Tasks

Our work is also relevant to several other tasks, including:

Video Object Segmentation Video object segmentation (VOS) is a popular task in video analysis. According to whether to provide the mask for the first frame, VOS can be divided into semi-supervised and unsupervised scenarios. Semi-supervised VOS (Hu et al., 2018; Johnander et al., 2019; Khoreva et al., 2017; Li & Loy, 2018; Li et al., 2020b; Oh et al., 2018, 2019; Voigtlaender and Leibe, 2017; Wang et al., 2021a) aims to track and segment a given object with a mask. Many Semi-supervised VOS methods (Khoreva et al., 2017; Li & Loy, 2018; Voigtlaender and Leibe, 2017) adopt an online learning manner which fine-tunes the network on the mask of the first frame during inference. Recently, some other works (Hu et al., 2018; Johnander et al., 2019; Li et al., 2020b; Oh et al., 2018, 2019; Wang et al., 2021a) aim to avoid online learning for the sake of faster inference speed. Unsupervised VOS methods (Li et al., 2018; Tokmakov et al., 2017; Wang et al., 2019) aim to segment the primary objects in a video without the first frame annotations.

As the first video object segmentation dataset, DAVIS (Caelles et al., 2019; Perazzi et al., 2016) contains 150 videos and 376 densely annotated objects. Xu et al. (2018) further

proposes the larger YouTube-VOS dataset with 4453 video clips and 7755 objects based on the large-scale YouTube-8M (Abu-El-Haija et al., 2016) dataset. Different from video instance segmentation that needs to classify objects, both unsupervised and semi-supervised VOS does not distinguish semantic categories. In addition, only one or several salient objects are annotated in these VOS datasets, while we annotate all the objects belonging to the pre-defined category set.

Video Semantic Segmentation Video semantic segmentation requires semantic segmentation for each frame in a video. The popular video semantic segmentation datasets include Cityscapes (Cordts et al., 2016), CamVid (Brostow et al., 2009), etc. There are 5000 video clips in the Cityscapes (Cordts et al., 2016) dataset. Each clip consists of 30 frames and only the 20th frame is annotated. CamVid (Brostow et al., 2009) dataset contains 4 videos and the authors annotate one frame every 30 frames, obtaining 800 annotated frames finally. LSTM (Fayyaz et al., 2016), GRU (Nilsson & Sminchisescu, 2018), and optical flow (Zhu et al., 2017) are introduced to leverage temporal contextual information for more accurate or faster video semantic segmentation. Video semantic segmentation does not require distinguishing instances and tracking objects across frames.

Video Panoptic Segmentation Kim et al. (2020) define a video extension of panoptic segmentation (Kirillov et al., 2019), which requires generating consistent panoptic segmentation, and in the meantime, associating instances across frames. They further reformatted the VIPER dataset with 124 videos and proposed the Cityscapes-VPS dataset which contains 500 videos.

Open-World Video Object Segmentation Different from the aforementioned tasks, open-world video object segmentation (Wang et al., 2021b) is taxonomy-free and requires segmenting and tracking all the objects class-agnostically. The proposed UVOD dataset (Wang et al., 2021b) contains 1200 videos and all the videos are densely annotated.

Multi-Object Tracking Multi-object tracking (MOT) (Smeulders et al., 2013) aims to detect the bounding boxes of objects and track them in a given video. Some popular datasets focus on the tracking of pedestrians and cars in street scenes, such as MOT16 (Milan et al., 2016) and KITTI (Geiger et al., 2012). Meanwhile, UA-DETRA (Wen et al., 2020) features vehicle tracking only.

Multi-Object Tracking and Segmentation Multi-object tracking and segmentation (MOTS) (Voigtlaender et al., 2019b) extends multi-object tracking (MOT) (Smeulders et al., 2013) from a bounding box level to a pixel level. Voigtlaender et al. (2019b) release the KITTI MOTS and MOTSChallenge datasets, and propose Track R-CNN that extends Mask R-CNN by 3D convolutions to incorporate temporal context and an extra tracking branch for object tracking. Xu et

al. (2020) release the ApolloScape dataset which provides more crowded scenes and proposes a new track-by-points paradigm. The task definition of MOTS is similar to video instance segmentation, which means an algorithm needs to simultaneously detect, segment, and track objects. While MOTS mainly focuses on pedestrians and cars in the streets, VIS targets more diverse scenes and more general objects in our daily life, such as animals.

Video Object Detection Video object detection (VOD) is a direct extension of image-level object detection. Compared with multi-object tracking, the video object detection task does not require tracking an object. Some commonly used datasets include the ImageNet-VID dataset (Russakovsky et al., 2015), which contains 3862 snippets for training, 555 snippets for validation, and 937 snippets for evaluation.

Our work is of course relevant to some image-level recognition tasks, such as semantic segmentation (Chen et al., 2017, 2018; Long et al., 2015), instance segmentation (He et al., 2017; Huang et al., 2019; Kirillov et al., 2020), panoptic segmentation (Kirillov et al., 2019; Li et al., 2020a; Xiong et al., 2019), large vocabulary instance segmentation (Gupta et al., 2019; Wu et al., 2020a), etc.

2.3 Occlusion Understanding

There are also some works focusing on occlusion understanding and handling. BCNet (Ke et al., 2021) adds a new branch to infer the occluders and utilizes the obtained occluder features to enhance the feature of occludees. OCFusion (Lazarow et al., 2020) introduces the occlusion head to indicate the occlusion relation between each pair of mask proposals. Zhan et al. (2020) proposes a self-supervised method that can recover the occlusion ordering and complete the invisible parts of occluded objects. Different from the full-DNN paradigm described above, Some methods (Kortylewski et al., 2020a, 2021, 2020b) integrate compositional models and deep convolutional neural networks into a unified model which is more robust to partial occlusions. As for pedestrian detection in crowded scenes, Wang et al. (2018b) and Zhang et al. (2018) propose new loss functions to enforce predicted boxes to locate compactly to the corresponding ground-truth objects while far from other objects. Zhou and Yuan (2018) regresses two bounding boxes for each object to localize the full body and visible part of a pedestrian respectively. Liu et al. (2019a) introduces adaptive-NMS which adaptively increases the NMS threshold in crowd scenes. Wu et al. (2020b) aggregates the temporal context to enhance the feature representations. Chu et al. (2020) predicts multiple instances in one proposal. In Multi-Object Tracking, Chu et al. (2017) and Zhu et al. (2018) utilize the attention module to attend to the visible parts of objects. Liu et al. (2020) and Xu et al. (2019) exploit

the topology between objects to track the occluded objects. In our experiments, to test the effect of temporal aggregation on occlusion handling, a temporal feature calibration module is presented, in which the calibrated features from neighboring frames are fused with the current frame for reasoning occluded objects and improving the recognition in each frame.

3 OVIS Dataset

Given an input video, video instance segmentation requires detecting, segmenting, and tracking object instances simultaneously from a predefined set of object categories. An algorithm is supposed to output the class label, confidence score, and a sequence of binary masks of each instance.

The focus of this work is on collecting a large-scale benchmark dataset for video instance segmentation with severe object occlusions. In this section, we mainly review the data collection process, the annotation process, and the dataset statistics.

3.1 Video Collection

We begin with selecting 25 semantic categories, including *Person, Bird, Cat, Dog, Horse, Sheep, Cow, Elephant, Bear, Zebra, Giraffe, Poultry, Giant panda, Lizard, Parrot, Monkey, Rabbit, Tiger, Fish, Turtle, Bicycle, Motorcycle, Airplane, Boat, and Vehicle*. The categories are carefully chosen mainly for three motivations: (1) most of them are animals, with which object occlusions extensively happen, (2) they are commonly seen in our life, (3) these categories have a high overlap with popular large-scale image instance segmentation datasets (Gupta et al., 2019; Lin et al., 2014) so that models trained on those datasets are easier to be transferred. The number of instances per category is given in Fig. 3.

As the dataset is to study the capability of our video understanding systems to perceive occlusions, we ask the annotation team to (1) exclude those videos, where only one single object stands in the foreground; (2) exclude those

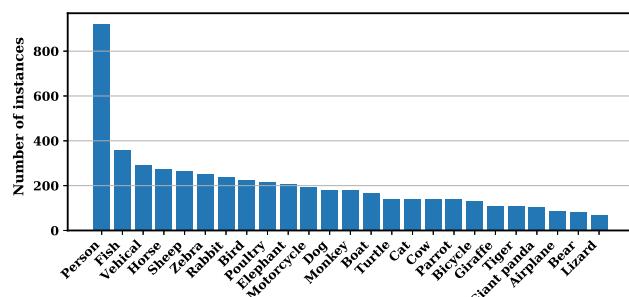


Fig. 3 Number of instances per category in the OVIS dataset

videos with a clean background; (3) exclude those videos, where the complete contour of objects is visible all the time. Some other objective rules include (1) video length is generally between 5 and 60 s, and (2) video resolution is generally 1920×1080 .

After applying the objective rules, the annotation team delivers 8644 video candidates and our research team only accepts 901 challenging videos after a careful re-check. It should be mentioned that due to the stringent standard of video collection, the pass rate is as low as 10%.

3.2 Annotation

Given an accepted video, the annotation team is asked to exhaustively annotate all the objects belonging to the predefined category set. Each object is given an instance identity and a class label. In addition to some common rules (e.g., no ID switch, mask fitness ≤ 1 pixel), the annotation team is trained with several criteria particularly about occlusions: (1) if an existing object disappears because of full occlusions and then re-appears, the instance identity should keep the same; (2) if a new instance appears in an in-between frame, a new instance identity is needed; and (3) the case of “object reappears” and “new instances” should be distinguishable by you after you watch the contextual frames therein. All the videos are annotated every 5 frames and the final annotation granularity of most videos is 5 or 6 fps.

To deeply analyze the influence of occlusion levels on model performance, OVIS provides the occlusion level annotation of every object in each frame. The occlusion levels are defined as follows: no occlusion, slight occlusion, and severe occlusion. As illustrated in Fig. 2, no occlusion means the object is fully visible, slight occlusion means that more than 50% of the object is visible, and severe occlusion means that more than 50% of the object area is occluded. After the frame-level occlusion degree is annotated, we can quantify the occlusion degree of each instance through the whole video by gathering the occlusion level in all frames of the instance. Specifically, We first map the three occlusion levels mentioned before into numeric scores. The no occlusion, slight occlusion, and server occlusion are mapped into 0, 0.25, and 0.75, respectively. Then, given an instance that appears in multiple frames, we use the averaged occlusion scores of top 50% frames with highest scores to represent the occlusion degree of instances.

Each video is handled by one annotator to get the initial annotation, and the initial annotation is then passed to another annotator to check and correct if necessary. The final annotations will be examined by our research team and sent back for revision if deemed below the required quality.

While being designed for video instance segmentation, it should be noted that OVIS is also suitable for evaluating video object segmentation in either a semi-supervised

Table 1 Comparing OVIS with YouTube-VIS in terms of statistics

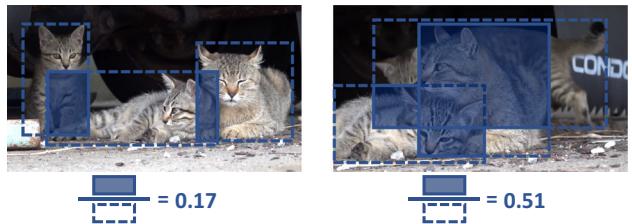
Dataset	YTVIS 19	YTVIS 21	OVIS
Masks	131k	232k	296k
Instances	4883	8171	5223
Categories	40	40	25
Videos	2883	3859	901
Video duration*	4.61 s	5.03 s	12.77 s
Instance duration	4.47 s	4.73 s	10.05 s
mBOR*	0.07	0.06	0.22
Objects/frame*	1.57	1.95	4.72
Instances/video*	1.69	2.10	5.80

See Eq. (1) for the definition of mBOR. *Means the value of YouTube-VIS is estimated from the training set

or unsupervised fashion, and object tracking since the bounding-box annotation is also provided. The relevant experimental settings will be explored as part of our future work.

3.3 Dataset Statistics

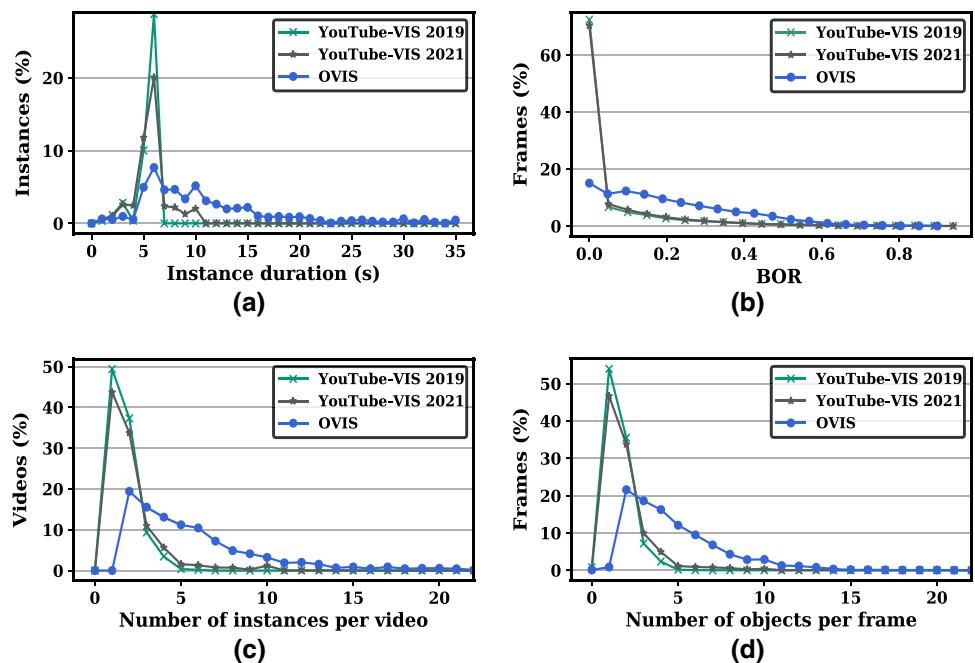
As YouTube-VIS (Yang et al., 2019) is the only dataset that is specifically designed for video instance segmentation nowadays, we analyze the data statistics of OVIS with YouTube-VIS as a reference in Table 1. We compare OVIS with two versions of YouTube-VIS: YouTube-VIS 2019 and YouTube-VIS 2021. Note that some statistics, marked with *, of YouTube-VIS are only calculated from the training set because only the annotation of the training set is publicly

**Fig. 5** Visualization of occlusions with different BOR values

available. Nevertheless, considering the training set occupies 78% of the whole dataset, those statistics could still reflect the properties of YouTube-VIS roughly.

In terms of basic and high-level statistics, OVIS contains 296k masks and 5223 instances. The number of masks in OVIS is larger than YouTube-VIS 2019 and YouTube-VIS 2021 that have 131k and 232k masks, respectively. The number of instances in OVIS is larger than YouTube-VIS 2019 that has 4883 instances, and less than YouTube-VIS 2021 that has 8171 instances. Note that there are fewer categories in OVIS, so the mean instances count per category is larger than YouTube-VIS 2021. Nonetheless, *OVIS has fewer videos than YouTube-VIS as our design philosophy favors long videos and instances so as to preserve enough motion and occlusion scenarios*.

As is shown, the average video duration and the average instance duration of OVIS are 12.77 s and 10.05 s respectively. Fig. 4a presents the distribution of instance duration, which shows that all instances in YouTube-VIS last less than 10 s. Long videos and instances increase the difficulty of tracking and the ability of long-term tracking is required.

Fig. 4 Comparison of OVIS with YouTube-VIS, including the distribution of instance duration (a), BOR (b), the number of instances per video (c), and the number of objects per frame (d)

As for occlusion levels, the proportions of objects with no occlusion, slight occlusion, and severe occlusion in OVIS are 18.2%, 55.5%, and 26.3% respectively. 80.2% of instances are severely occluded in at least one frame, and only 2% of the instances are not occluded in any frame. It supports the focus of our work, that is, to explore the ability of video instance segmentation models in handling occlusion scenes.

In order to compare the occlusion degree with other datasets, we define a metric named Bounding-box Occlusion Rate (BOR) to approximate the degree of occlusion. Given a video frame with N objects denoted by bounding boxes $\{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_N\}$, we compute the BOR for this frame as

$$\text{BOR} = \frac{|\bigcup_{1 \leq i < j \leq N} \mathbf{B}_i \cap \mathbf{B}_j|}{|\bigcup_{1 \leq i \leq N} \mathbf{B}_i|}, \quad (1)$$

where the numerator means the area sum of the intersection between any two or more bounding boxes. The denominator means the area of the union of all the bounding boxes. An illustration is given in Fig. 5, which shows the larger the BOR value is, the heavier the occlusion is.

Then we utilize mBOR, the average value of BORs of all the frames in a dataset (frames that do not contain any objects are ignored), to characterize the dataset in terms of the occlusion. As shown in Table 1, the mBOR of OVIS is 0.22, much higher than that of YouTube-VIS 2019 and YouTube-VIS 2021 (0.07 and 0.06, respectively). The BOR distribution is further compared in Fig. 4b. As can be seen, most frames in YouTube-VIS are located in the region where $\text{BOR} \leq 0.1$. In comparison, the BOR of about half frames in OVIS is no less than 0.2. This supports that there are more severe occlusions in OVIS than YouTube-VIS. However, it should be mentioned here that BOR can only roughly reflect the occlusion between objects. Therefore, *mBOR could serve as an effective indicator for occlusion degrees, but only reflect the occlusion degree in a partial or rough way*.

In addition to long videos&instances and severe occlusions, OVIS features crowded scenes, which is a natural result caused by heavy occlusions. OVIS has 5.80 instances per video and 4.72 objects per frame, while those two values are 2.10 and 1.95 respectively in YouTube-VIS 2021. The comparison of the two distributions is further depicted in Fig. 4c, d.

3.4 Evaluation Metrics

Following previous methods (Yang et al., 2019), we use average precision (AP) at different intersection-over-union (IoU) thresholds and average recall (AR) as the evaluation metrics. The mean value of APs is also employed.

In addition, thanks to the occlusion level annotations in OVIS, we are able to analyze the performance under different occlusion levels. We divide all instances into three groups

called slightly occluded, moderately occluded, and heavily occluded, in which the occlusion scores of instances are in the range of [0, 0.25], [0.25, 0.5], [0.5, 0.75] respectively. The proportions of the three groups are 23%, 44%, and 49% respectively. Then, we can get the AP of each group (denoted by AP_{SO} , AP_{MO} , and AP_{HO} respectively) by ignoring the instances of other groups.

4 Experiments

In this section, we comprehensively study the newly collected OVIS dataset by conducting experiments on 9 existing video instance segmentation algorithms and propose our new baseline method.

4.1 Implementation Details

Datasets On the newly collected OVIS dataset, the whole dataset is divided into 607 training videos, 140 validation videos, and 154 test videos. The split proportions of different categories are approximately the same, and there are at least 4 videos per category in the validation and test set. This split will be fixed as an official split. If not specified, the experiments are conducted on the validation set of OVIS.

A Temporal Feature Calibration Plug-in One of the keys to tackling occlusion is to complement the missing object cues. In a video that has a temporal dimension, a mild assumption is that usually, the missing object cues in the current frame may have appeared in adjacent frames. Hence, it is natural to leverage adjacent frames to alleviate occlusions. However, caused by motions, the features of different frames are not aligned in the spatial dimension. Things get much worse because of the existence of severe occlusions. To solve this issue, following (Bertasius et al., 2018; Dosovitskiy et al., 2015), we present an easy plug-in called temporal feature calibration as illustrated in Fig. 6.

Denote by $\mathbf{F}_q \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{F}_r \in \mathbb{R}^{H \times W \times C}$ the feature tensor of the query frame (called target or current frame in some literature) and a reference frame, respectively. The feature calibration first computes the spatial correlation (Dosovitskiy et al., 2015) between \mathbf{F}_q and \mathbf{F}_r . Given a location \mathbf{x}_q in \mathbf{F}_q and \mathbf{x}_r in \mathbf{F}_r , we compute

$$\mathbf{c}(\mathbf{x}_q, \mathbf{x}_r) = \sum_{o \in [-k, k] \times [-k, k]} \mathbf{F}_q(\mathbf{x}_q + o) \mathbf{F}_r(\mathbf{x}_r + o)^T. \quad (2)$$

The above operation will transverse the $d \times d$ area centered on \mathbf{x}_q , then outputs a d^2 -dimensional vector.

After enumerating all the positions in \mathbf{F}_q , we obtain $\mathbf{C} \in \mathbb{R}^{H \times W \times d^2}$ and forward it into multiple stacked convolution layers to get the spatial calibration offset $\mathbf{D} \in \mathbb{R}^{H \times W \times 18}$. We

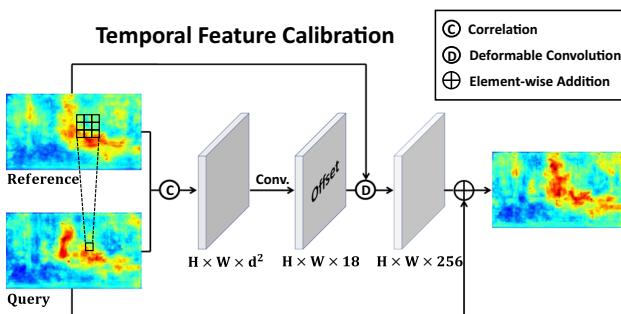


Fig. 6 The pipeline of temporal feature calibration, which can be inserted into different video instance segmentation models by changing the following prediction head

then obtain a calibrated version of \mathbf{F}_r by applying deformable convolutions with \mathbf{D} as the spatial calibration offset, which is denoted as $\bar{\mathbf{F}}_r$. At last, we fuse the calibrated reference feature $\bar{\mathbf{F}}_r$ with the query feature \mathbf{F}_q by element-wise addition for the localization, classification, and segmentation of the current frame afterward.

During training, for each query frame \mathbf{F}_q , we randomly sample a reference frame \mathbf{F}_r from the same video. As compared with the short videos in YouTube-VIS (the longest video in YouTube-VIS contains only 36 frames), the first frame and the last frame of a long video in OVIS (the longest video in OVIS contains 500 frames) may be totally different. In order to ensure that the reference frame has a strong spatial correspondence with the query frame, the sampling is only done locally within $\varepsilon_{\text{train}} = 5$ frames. Since the temporal feature calibration is differentiable, it can be trained end-to-end by the original detection and segmentation loss. When inference, all frames adjacent to the query frame within the range $\varepsilon_{\text{test}} = 5$ are taken as reference frames. We linearly fuse the classification confidences, regression bounding box coordinates, and segmentation masks obtained from each reference frame and output the final results for the query frame.

In the experiments, we denote the new methods as CMaskTrack R-CNN and CSipMask, when Calibrating MaskTrack R-CNN (Yang et al., 2019) models and Calibrating SipMask (Cao et al., 2020) models, respectively.

Experimental Setup For all our experiments, we adopt ResNet-50-FPN (He et al., 2016) as the backbone. The models are initialized by Mask R-CNN which is pre-trained on MS-COCO (Lin et al., 2014). All frames are resized to 640×360 during both training and inference for fair comparisons with previous works (Yang et al., 2019; Cao et al., 2020; Athar et al., 2020). For our new baselines (CMaskTrack R-CNN and CSipMask), we use three convolution layers of kernel size 3×3 in the module for temporal feature calibration. The training epoch is set to 12, and the initial learning rate is set to 0.005 and decays with a factor of 10 at epoch 8 and 11.

4.2 Main Results

On the OVIS dataset, we first produce the performance of several state-of-the-art algorithms whose code is publicly available, including mask propagation methods (e.g., FEELVOS Voigtlaender et al. 2019a), track-by-detect methods (e.g., IoUTracker+ Yang et al. 2019), and recently proposed end-to-end methods (e.g., MaskTrack R-CNN Yang et al. 2019, SipMask Cao et al. 2020, STEm-Seg Athar et al. 2020, STMask Li et al. 2021, TraDeS Wu et al. 2021, CrossVIS Yang et al. 2021, and QueryVIS Fang et al. 2021). The standard deviation of the reported results below is about 0.5.

As presented in Table 2, although most of these methods can obtain more than 30 AP on the YouTube-VIS dataset, all of them encounter a great performance degradation of at least 50% on OVIS compared with that on YouTube-VIS. Especially in the heavily occluded instance group, all methods

Table 2 Overall results of state-of-the-art methods on the OVIS dataset

Methods	OVIS validation set								OVIS test set							
	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
FEELVOS (Voigtlaender et al., 2019a)	9.6	22.0	7.3	7.4	14.8	17.3	11.5	1.7	10.8	23.4	8.7	9.0	16.2	18.9	12.2	2.0
IoUTracker+ (Yang et al., 2019)	7.0	16.9	5.3	5.7	14.3	11.5	7.9	1.8	8.0	18.4	7.5	5.9	15.7	12.8	9.1	2.1
MaskTrack R-CNN (Yang et al., 2019)	10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7	11.8	25.4	10.4	7.9	16.0	22.7	15.0	3.5
SipMask (Cao et al., 2020)	10.2	24.7	7.8	7.9	15.8	19.9	10.5	2.2	11.7	23.7	10.5	8.1	16.6	21.9	13.9	3.2
STEM-Seg (Athar et al., 2020)	13.8	32.1	11.9	9.1	20.0	22.2	16.1	3.9	14.4	30.0	13.0	10.1	20.6	22.5	16.8	4.2
TraDeS (Wu et al., 2021)	11.4	26.5	9.4	7.0	13.8	23.0	12.8	3.0	12.0	26.4	10.8	7.8	14.6	21.6	14.1	3.6
QueryVIS (Fang et al., 2021)	14.7	34.7	11.6	9.0	21.2	27.3	17.2	4.1	16.0	33.7	14.7	9.6	21.7	26.3	17.7	4.5
STMask (Li et al., 2021)	15.4	33.8	12.5	8.9	21.3	24.0	18.7	5.1	15.6	32.5	13.8	9.1	21.8	25.4	17.1	6.3
CrossVIS* (Yang et al., 2021)	14.9	32.7	12.1	10.3	19.8	28.4	16.9	4.1	16.3	31.5	15.4	10.6	21.1	27.3	18.5	5.6

Bold values indicate best performance

AP_{SO}, AP_{MO}, and AP_{HO} respectively denote the AP of “slightly occluded”, “moderately occluded”, and “heavily occluded”.

*Means the baseline model is additionally pre-trained with the YouTube-VIS dataset (Yang et al., 2019)

Table 3 Quantitative comparison between the new methods and their corresponding baselines on the OVIS dataset and the YouTube-VIS dataset

Methods	OVIS validation set								YouTube-VIS 2019 validation set				
	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
SipMask (Cao et al., 2020)	10.2	24.7	7.8	7.9	15.8	19.9	10.5	2.2	32.5	53.0	33.3	33.5	38.9
CSipMask	14.3	29.9	12.5	9.6	19.3	27.1	16.6	3.2	35.1	55.6	38.1	35.8	41.7
MaskTrack R-CNN (Yang et al., 2019)	10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7	30.3	51.1	32.6	31.0	35.5
CMaskTrack R-CNN	15.4	33.9	13.1	9.3	20.0	28.6	18.7	4.1	32.1	52.8	34.9	33.2	37.9

Bold values indicate best performance

suffer from a significant performance drop of more than 80%. For example, SipMask (Cao et al., 2020), which achieves an AP of 32.5 on YouTube-VIS, only obtains an AP of 2.2 in the heavily occluded group of OVIS validation set. It firmly suggests that severe occlusion will greatly improve the difficulty of video instance segmentation, and further attention should be paid to video instance segmentation in the real world where occlusions extensively happen. Benefiting from the feature calibration and temporal fusion, STMask (Li et al., 2021) obtains an AP_{HO} of 5.1 on the validation set and 6.3 on the test set, surpassing all other methods in the heavily occluded group.

It is worth noting that, as the only bottom-up video instance segmentation method, STEm-Seg achieves similar AP_{SO} with MaskTrack R-CNN and TraDeS, but much higher AP_{HO} (3.9 vs. 2.7 vs. 3.0). It demonstrates that the bottom-up paradigm like STEm-Seg may perform better than the general top-down paradigm on occlusion handling. Our interpretation is that the bottom-up architecture avoids the detection process which is difficult in occluded scenes.

In addition, as shown in Table 3, by leveraging the feature calibration module, the performance on OVIS is significantly improved. CMaskTrack R-CNN leads to an AP improvement of 4.6 over MaskTrack R-CNN (10.8 vs. 15.4), and CSipMask leads to an AP improvement of 4.1 over SipMask (10.2 vs. 14.3). Besides, the experiments also show that TFC can boost the performance of all occlusion levels, and the improvement of heavy occlusion and moderate occlusion is more significant (see Fig. 11 for more details). We also evaluate the proposed CMaskTrack R-CNN and CSipMask on the YouTube-VIS dataset. As shown in Table 3, CMaskTrack R-CNN and CSipMask surpass the corresponding baselines by 1.8 and 2.6 in terms of AP, respectively, which demonstrates the flexibility and the generalization power of the proposed feature calibration module.

To present the qualitative evaluation results of methods on OVIS, some evaluation examples of CMaskTrack R-CNN are given in Fig. 7, including 5 successful cases (a)–(e) and 3 failure cases (f) and (h). In (a), the car in the yellow mask first blocks the car in the red mask entirely in the 2nd frame, then is entirely blocked by the car in the purple mask in the 4th frame. It is surprising that even in this extreme case, all the cars are well tracked. In (b), CMaskTrack R-CNN success-

fully tracks the bear in the yellow mask, which is partially occluded by another object, i.e., the bear in the purple mask, and the background, i.e., the tree. In (d), we present a crowded scene where almost all the ducks are correctly detected and tracked. In (f), two persons and two bicycles heavily overlap with each other. CMaskTrack R-CNN fails to track the person and segment the bicycle. In (g), when two cars are intersecting, severe occlusion leads to failure of detection and tracking. In (h), although humans could sense that there are two persons with hats at the bottom, CMaskTrack R-CNN cannot detect and track them because the appeared visual cues are inadequate.

4.3 Discussions

Oracle Results We conduct the image oracle and identity oracle experiments to explore the impact of image-level prediction and cross-frame association on the performance of the OVIS dataset. In order to compare with the YouTube-VIS dataset (Yang et al., 2019), we use MaskTrack-RCNN for experiments. Following (Yang et al., 2019), in the image oracle experiments, we use ground-truth bounding boxes, masks, and category labels to replace the predictions by MaskTrack R-CNN, and then track those ground-truth bounding boxes by the tracking branch. In the identity oracle experiment, we first assign each per-frame prediction to the closest ground-truth bounding box, and then aggregate the bounding boxes with the same identity through the video.

The results are shown in Table 4. On the OVIS dataset, the image oracle experiments and identity oracle experiments obtain 58.4 and 25.5 AP, respectively. This demonstrates that the image level prediction is more critical for the performance of occluded video instance segmentation, which is mainly associated to object segmentation and classification in frames. It can be expected that more advanced image-based techniques could be explored further so as to approach this upper limit. Interestingly, both oracle experiments achieve lower performance on the OVIS dataset than that on the YouTube-VIS dataset, which shows that whether for image-level prediction or cross-frame association, the OVIS dataset is more challenging than the YouTube-VIS dataset. Moreover, in identity experiments, the AP on YouTube-VIS achieves almost no gain (only 4% improvement over the

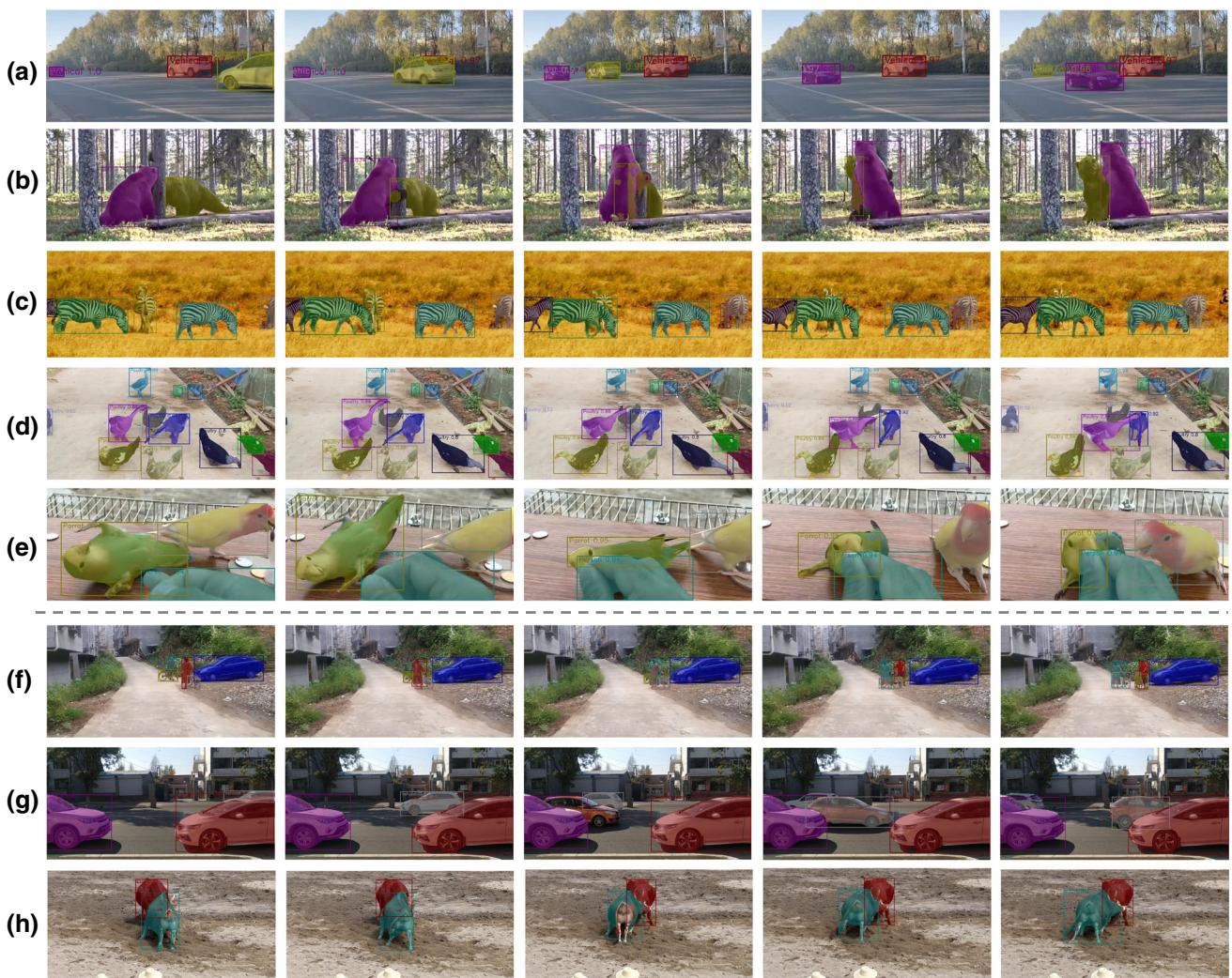


Fig. 7 Evaluation examples on OVIS. Each row presents the results of 5 frames in a video sequence. **a–e** are successful cases and **f, h** are failure cases

Table 4 Oracle results on OVIS and YouTube-VIS

Dataset	OVIS	YouTube-VIS (Yang et al., 2019)
Image Oracle	AP	58.4 ($\uparrow 441\%$)
	AR ₁₀	66.1 ($\uparrow 460\%$)
Identity Oracle	AP	23.9 ($\uparrow 121\%$)
	AR ₁₀	28.2 ($\uparrow 139\%$)

The number in the brackets means the performance improvement ratio over the corresponding baseline

result of MaskTrack R-CNN baseline), while the AP on OVIS is greatly improved (121% improvement over MaskTrack R-CNN), which demonstrates that the tracking task on OVIS is much more difficult than that on YouTube-VIS.

Effect of Leveraging Image Datasets Caused by the high cost of exhaustively annotating high-quality video segmentation masks, video inadequacy is a common problem among existing video segmentation datasets. The lack of diversity in video scenes may affect the generalization capability of

models trained on those datasets. To this end, we further train several models with both the video data in OVIS and additional augmented image sequences/pairs synthesized from other large-scale image instance segmentation datasets. In our experiments, the proportions of video data and augmented image data are 65% and 35% respectively. These pseudo image sequences are generated from the COCO (Lin et al., 2014) dataset by on-the-fly random perspective and affine transformation. The evaluation results are shown in

Table 5. We can see that by leveraging the augmented image sequences, all these three baseline methods can achieve remarkable AP improvements, which can serve as a reference for future research.

Analysis of NMS Threshold Non-Maximum Suppression (NMS) is a necessary post-processing for most detection methods.

To test the impact of NMS threshold on occlusion handling, inspired by Liu et al. (2019a), we design the adaptive NMS oracle experiment. Specifically, for each ground-truth bounding box, we calculate the maximum IoU d between it and all other ground-truth boxes. Then, the NMS threshold of all the predicted boxes that correspond to this ground-truth box will be assigned as $\max(d, 0.5)$. In this way, a larger NMS threshold will be applied to the predictions in dense scenes, which can prevent NMS from removing the true positives that are close to other ground-truth boxes.

As presented in Table 6, based on MaskTrack R-CNN, the adaptive NMS oracle experiment improves AP_{MO} and AP_{HO} by 0.2, which proves that using a higher NMS threshold adaptively improves the performance in occluded scenes. However, even though we exactly know the real density (Liu et al., 2019a) of boxes in the adaptive NMS oracle experiment, AP_{SO} decreases from 23.0 to 22.8. The overall AP only improves from 10.8 to 11.2, which shows that the NMS threshold adjusting is not a bottleneck on OVIS.

One interpretation is that adjusting the NMS threshold is more important for tasks that require detecting the amodal bounding boxes (additionally containing the occluded invisible parts), such as the full-body bounding boxes in crowded pedestrian detection datasets (Shao et al., 2018; Zhang et al., 2017). For two occluded objects, the IoU of amodal bound-

ing boxes will be much higher than the IoU of the bounding boxes of only the visible parts (like the boxes in OVIS). In addition, some learnable NMS methods (Hosang et al., 2017; Liu et al., 2019a) have also been proposed, and many new methods (Carion et al., 2020; Fang et al., 2021) based on set prediction even do not need NMS post-processing. These new methods require further exploration in OVIS.

Error Analysis To explore the detailed influence of occlusion levels on video instance segmentation, in this subsection, we analyze the frame-level error rates of classification, segmentation, and tracking under different occlusion levels. A segmentation error refers to that the IoU between the predicted mask of an object and its ground-truth less than 0.5 and the tracking error is reflected by ID switch rate.

Formally, we denote the predicted masks and labels in all frames as $M = \{m_1, m_2, \dots, m_n\}$ and $Y = \{y_1, y_2, \dots, y_n\}$, respectively, where n is the number of predictions. The corresponding matched ground-truth masks and labels as $M^* = \{m_1^*, m_2^*, \dots, m_n^*\}$ and $Y^* = \{y_1^*, y_2^*, \dots, y_n^*\}$, respectively.

Regarding classification error rates, we consider the predicted object whose IoU with its matched ground-truth is greater than 0.5, then count the proportion of classification errors among them, as

$$E_{cls} = \frac{|\{m_j | \text{IoU}(m_j, m_j^*) > 0.5 \wedge y_j \neq y_j^*\}|}{|\{m_i | \text{IoU}(m_i, m_i^*) > 0.5\}|}. \quad (3)$$

For segmentation error rates, following Bolya et al. (2020), we consider masks whose IoU with its matched ground-truth is greater than 0.1. A mask m_i will be counted as a segmentation error if its IoU with the corresponding ground-truth m_i^* is less than 0.5. Then the segmentation error rate is calculated

Table 5 The results of training with and without augmented image sequences

Methods	w/image data	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
MaskTrack R-CNN (Yang et al., 2019)		10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7
	✓	12.0	28.5	8.9	7.8	16.3	24.7	14.2	3.2
SipMask (Cao et al., 2020)		10.2	24.7	7.8	7.9	15.8	19.9	10.5	2.2
	✓	11.8	27.5	9.0	8.3	17.7	23.2	13.4	2.3
STEm-Seg (Athar et al., 2020)		13.8	32.1	11.9	9.1	20.0	22.2	16.1	3.9
	✓	15.2	34.7	12.5	10.7	23.7	25.5	17.4	4.1

Bold values indicate best performance

“w/image data” means training with both video data and the synthesized clips

Table 6 Adaptive NMS oracle results of MaskTrack R-CNN on OVIS

Methods	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
MaskTrack R-CNN	10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7
+ Adaptive NMS (Liu et al., 2019a) Oracle	11.2	26.5	8.6	8.3	15.6	22.8	13.0	2.9

Bold values indicate best performance

Table 7 Error analysis under different occlusion levels

Error type	Methods	No occlusion (%)	Slight occlusion (%)	Severe occlusion (%)	All (%)
Cls. error rate	MaskTrack R-CNN	41.3	41.0	50.1	42.5
	MaskTrack R-CNN + LSS + DCN	30.2 (−11.1)	32.9 (−8.1)	45.8 (−4.3)	34.5 (−8.0)
	CMaskTrack R-CNN (ours)	27.5 (−2.7)	29.0 (−3.9)	39.2 (−6.6)	30.2 (−4.3)
Seg. error rate	MaskTrack R-CNN	12.1	25.6	34.1	28.3
	MaskTrack R-CNN + LSS + DCN	13.2 (+0.9)	25.1 (−0.5)	33.2 (−0.9)	28.0 (−0.3)
	CMaskTrack R-CNN (ours)	13.0 (−0.2)	22.2 (−2.9)	29.5 (−3.7)	25.3 (−2.7)
ID switch rate	MaskTrack R-CNN	18.6	22.5	32.6	22.9
	MaskTrack R-CNN + LSS + DCN	12.5 (−6.1)	16.1 (−6.4)	26.1 (−6.5)	16.8 (−6.1)
	CMaskTrack R-CNN (ours)	11.2 (−1.3)	14.0 (−2.1)	21.6 (−4.5)	14.4 (−2.4)

LSS denotes the local sampling strategy and DCN means applying a deformable convolutional layer on the query frame itself. For a certain row, the number in the brackets means the decrease of error rates over the row above
Bold values indicate best performance

as

$$E_{seg} = \frac{|\{m_j | 0.1 < \text{IoU}(m_j, m_j^*) < 0.5\}|}{|\{m_i | \text{IoU}(m_i, m_i^*) > 0.1\}|}. \quad (4)$$

The ID switch rate refers to the ratio of ID switches in the tracking sequence of all instances. Following Voigtlaender et al. (2019b), the predicted ID of a ground-truth instance in a frame is defined as the tracking ID of the closest predicted mask. If the ID of a ground-truth instance is not equal to that of its latest tracked predecessor, it will be considered as an ID switch.

Based on the error rates defined above, we further evaluate MaskTrack and CMaskTrack. In addition, we define a baseline named “MaskTrack R-CNN + LSS + DCN” by applying the local sampling strategy and applying one deformable convolution layer to the query frame. As a result, by comparing “MaskTrack R-CNN + LSS + DCN” and our method, we could obtain the performance gain purely brought by temporal feature calibration.

As shown in Table 7, the three types of error rates all significantly increase when the occlusion level increases. Among them, the segmentation error rate increases the most, from 12.1 to 34.1% for MaskTrack R-CNN, which demonstrates that severe occlusion will greatly increase the difficulty of the segmentation task. In this sense, accurately localizing the object is helpful for mitigating the impact of occlusions. Meanwhile, among the three error types, the error rate of classification is much higher than that of segmentation and tracking. So a better classification result is important to improving the overall performance.

One could also observe that (1) no matter in terms of classification error rate, segmentation error rate, or ID switch rate, the gain of our method over “MaskTrack R-CNN + LSS + DCN” increases when the occlusion level increases (e.g., CMaskTrack R-CNN decreases the classification error rate by 2.7%, 3.9%, and 6.6% respectively);

(2) in terms of segmentation error rate and ID switch rate, the gain of “MaskTrack R-CNN + LSS + DCN” over the baseline “MaskTrack R-CNN” does not change too much when the occlusion level increases (e.g., “MaskTrack R-CNN + LSS + DCN” decreases the segmentation error rate by 6.1%, 6.4%, and 6.5% respectively); (3) in terms of classification error rate, the gain of “MaskTrack R-CNN + LSS + DCN” over the baseline “MaskTrack R-CNN” even decreases when the occlusion level increases (No occlusion: 11.1%, Slight occlusion: 8.1%, and Severe occlusion: 4.3%).

By comparing Observation (1), (2), and (3), one could conclude that the TFC module improves more in occluded scenes compared with using other training strategies (e.g., the local sampling strategy) and model structure (e.g., applying deformable convolution to the query frame). The same conclusion is also drawn if we compare the relative error decreasing rate.

Effect of Better Feature Representations To test the effect of better feature representations on occlusion, we further try Swin-T (Liu et al., 2019b) and ResNeXt-101 (Xie et al., 2017) backbone on MaskTrack R-CNN and QueryVIS. As can be seen in Table 8, both Swin-T and ResNeXt-101 achieve great improvement (about 4 AP) on OVIS. And these larger backbones can also achieve obvious AP improvement at all occlusion levels.

Effect of Larger Input Resolutions We try to replace the 640 × 360 input resolution with 1280 × 720 which is similar to the commonly used input resolution for COCO (Lin et al., 2014). As shown in Table 9, when the input resolution increases, the performance improves slightly (0.5 AP for MaskTrack R-CNN Yang et al. 2019 and 0.3 AP for SipMask Cao et al. 2020).

Methods Specifically Designed for Occlusion We also migrate three image-level detection methods to the CMaskTrack R-CNN model, including (1) the repulsion loss (Wang

Table 8 Effect of larger backbones

Methods	Backbone	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
MaskTrack R-CNN (Yang et al., 2019)	ResNet-50	10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7
	Swin-T	14.0	30.5	11.5	9.2	19.2	26.9	16.0	3.7
	ResNeXt-101	14.6	33.5	12.0	9.6	19.4	27.1	16.7	3.8
QueryVIS (Fang et al., 2021)	ResNet-50	12.8	28.8	11.0	8.6	19.2	25.2	15.1	2.6
	Swin-T	16.5	36.2	14.5	10.2	22.6	31.2	19.4	4.2
	ResNeXt-101	16.9	36.5	14.7	10.6	23.5	31.8	19.2	4.6

For a fair comparison, the results shown here are all trained only 12 epochs for both pre-training on COCO and training on OVIS
Bold values indicate best performance

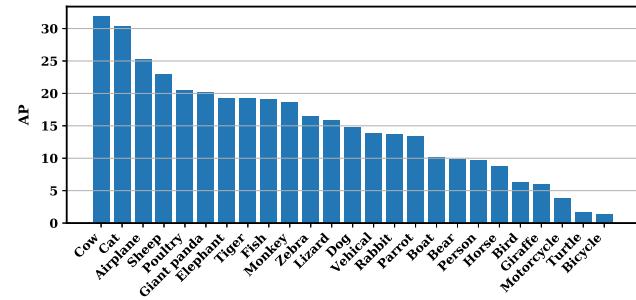
Table 9 Effect of larger input resolutions

Methods	Input size	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
MaskTrack R-CNN (Yang et al., 2019)	640 × 360	10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7
	1280 × 720	11.3	25.8	9.3	7.9	15.7	23.9	12.9	2.6
SipMask (Cao et al., 2020)	640 × 360	10.2	24.7	7.8	7.9	15.8	19.9	10.5	2.2
	1280 × 720	10.5	24.3	8.4	7.1	15.7	19.9	11.8	2.0

Bold values indicate best performance

et al., 2018b) which requires the predicted boxes to keep away from other ground-truth boxes; (2) the compact loss (Zhang et al., 2018) which enforces proposals to be close and locate compactly to the corresponding ground-truth; (3) the occluder branch (Ke et al., 2021) (without any extra designs like the Non-local (Wang et al., 2018a) operation and boundary prediction) which additionally learns the feature of occluders with a new branch and then fuses the feature of occluders and occludees. In particular, the repulsion loss and compact loss are specifically designed for crowded pedestrian detection, and the occluder branch is designed for the occlusion problem of common objects.

As shown in Table 10, the compact loss and occluder branch improve AP_{HO} by 0.4 and 0.3 respectively, while their overall AP improvements are marginal. We believe more gains can be achieved by developing more delicate occlusion handling algorithms and leveraging occluded data (see Sect. 5 for future work discussion).

**Fig. 8** Per-class AP of CMaskTrack R-CNN on OVIS

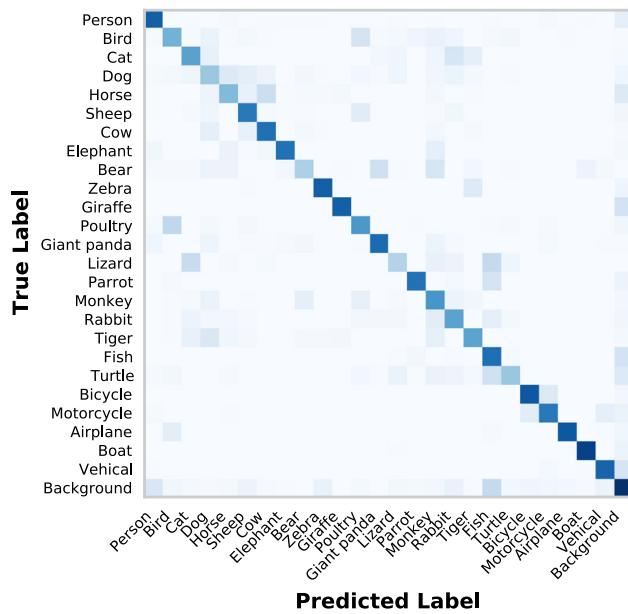
Per-class Results The per-class AP scores of CMaskTrack R-CNN are shown in Fig. 8. It shows that the Top-5 challenging categories are Bicycle, Turtle, Motorcycle, Giraffe, and Bird. The confusion matrix is also given in Fig. 9. As it shows, most categories can be correctly classified except for some visually similar category pairs (e.g., Poultry and Bird, Bicycle and Motorcycle).

Table 10 Effect of three existing occlusion handling methods that are specifically designed for image-level detection tasks

Methods	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
CMaskTrack R-CNN	15.4	33.9	13.1	9.3	20.0	28.6	18.7	4.1
+ Repulsion loss (Wang et al., 2018b)	14.7	32.0	13.8	9.2	19.3	26.9	17.7	4.0
+ Compact loss (Zhang et al., 2018)	15.4	34.1	12.7	9.4	19.4	27.9	18.3	4.5
+ Occluder branch (Ke et al., 2021) (w/o extra designs)	15.6	34.3	13.5	9.7	20.1	28.3	17.9	4.4

“w/o extra designs” means that we remove the Non-local operation (Wang et al., 2018a) and boundary prediction in the occluder branch for a fair comparison with other methods

Bold values indicate best performance

**Fig. 9** Confusion matrix for classification

Ablation Study of the TFC Module. To verify the rationality of the TFC module, we firstly test the effect of the local sampling strategy of reference frames during training. As shown in Table 11, by only sampling the reference frames locally within $\varepsilon_{\text{train}} = 5$ frames instead of sampling in the

whole video, MaskTrack R-CNN, SipMask, and QueryVIS all obtain significant AP improvements of 2.7, 2.6, and 1.7 respectively, which demonstrates that the local sampling strategy of reference frames during training is necessary and beneficial to learn how to track objects in the long videos of OVIS.

We further study the temporal feature calibration module with a few alternatives. The first option is a naive combination, which sums up the feature of the query frame and the reference frame without any feature alignment. The second option is to replace the correlation operation in our module by calculating the element-wise difference between feature maps, which is similar to the operation used in Bertasius and Torresani (2020). We denote the three options as “+ Uncalibrated Addition” and “+ Calibration_{diff}” respectively and our module as “+ Calibration_{corr}” in Table 12.

As we can see, with the enhanced MaskTrack R-CNN (with local sampling strategy of reference frames during training) as the base model, the naive “+ Uncalibrated Addition” combination even degrades the final AP. This is because the direct addition of the uncalibrated features from other frames may bring noises to the object localization process. In contrast, after applying feature calibration, the performance is improved. “+ Calibration_{corr}” achieves an AP of 15.4, an improvement of 1.9 over the baseline method without feature fusion and 1.0 over “+ Calibration_{diff}”. We argue that

Table 11 Effect of the local sampling strategy on the OVIS validation set

Methods	Local sampling	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
MaskTrack R-CNN (Yang et al., 2019)		10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7
	✓	13.5	29.9	11.3	8.5	18.7	25.4	16.7	3.3
SipMask (Cao et al., 2020)		10.2	24.7	7.8	7.9	15.8	19.9	10.5	2.2
	✓	12.8	29.8	9.6	8.7	17.9	25.5	14.7	2.5
QueryVIS (Fang et al., 2021)		14.7	34.7	11.6	9.0	21.2	27.3	17.2	4.1
	✓	16.4	37.8	12.4	9.9	22.9	31.2	19.0	4.3

Bold values indicate best performance

Table 12 Effect of the local sampling strategy and the comparison of different feature fusion methods

Methods	Local sampling	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀	AP _{SO}	AP _{MO}	AP _{HO}
MaskTrack R-CNN (Yang et al., 2019)		10.8	25.3	8.5	7.9	14.9	23.0	12.8	2.7
MaskTrack R-CNN (Yang et al., 2019)	✓	13.5	29.9	11.3	8.5	18.7	25.4	16.7	3.3
Yang et al. (2019) + DCN	✓	14.0	31.2	11.2	8.8	18.5	26.2	16.2	3.2
Yang et al. (2019) + Uncalibrated Addition	✓	12.9	29.4	11.5	8.2	16.6	25.6	15.3	3.1
Yang et al. (2019) + Calibration _{diff}	✓	14.4	32.6	12.3	8.6	18.9	25.3	17.6	3.8
Yang et al. (2019) + Calibration _{corr}	✓	15.4	33.9	13.1	9.3	20.0	28.6	18.7	4.1

“Local sampling” means only sample the reference frames locally within $\varepsilon_{\text{train}} = 5$ frames during training. “+ DCN” means applying a deformable convolutional layer on the query frame itself. “+ Uncalibrated Addition” means adding feature maps directly without calibration. “+ Calibration_{diff}” means generating the calibration offset based on the element-wise difference between feature maps, similar to Bertasius and Torresani (2020) did. “+ Calibration_{corr}” is the presented method in Sect. 4.1.

Bold values indicate best performance

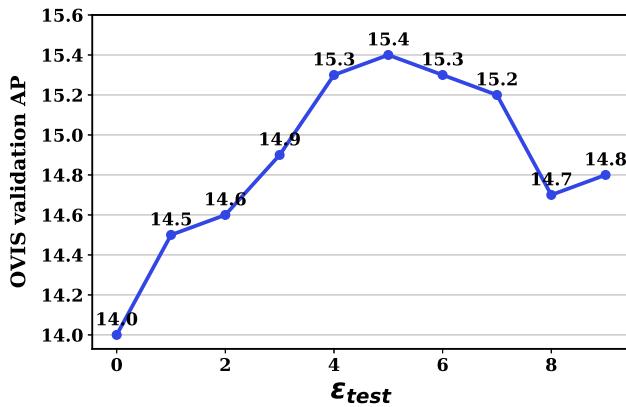


Fig. 10 Results of different reference frame range ε_{test} on the OVIS validation set. Notably, $\varepsilon_{test} = 0$ indicates applying the deformable convolutional layer to the query frame itself, without leveraging adjacent frames

the correlation operation is able to provide a richer context for feature calibration because it calculates the similarity between the query position and its neighboring positions. Testing on P-100 GPU, the speed of CMaskTrack R-CNN when using Calibration_{diff} and Calibration_{corr} are 16 and 7 fps respectively.

We also conduct experiments to analyze the influence of the reference frames range ε_{test} . $\varepsilon_{test} = 0$ means applying the deformable convolutional layer to the query frame itself. As can be seen in Fig. 10, the AP increases when ε_{test} increases, and reaching the highest value at $\varepsilon_{test} = 5$. Even if $\varepsilon_{test} = 1$, the performance exceeds the setting of $\varepsilon_{test} = 0$, which demonstrates that calibrating features from adjacent frames is beneficial to video instance segmentation.

To further compare the improvement of TFC on different occlusion levels, we evaluate the relative gain of AP on different occlusion levels for a fair comparison. As shown in Fig. 11, we report the relative gain by varying ε_{test} . The larger the ε_{test} is, the more temporal context will be aggregated. As can be seen, the relative gain of AP_{HO} is much higher than that of AP_{MO}. The relative gain of AP_{SO} is smallest once the temporal context is considered ($\varepsilon_{test} > 0$). The result demonstrates the effectiveness of temporal feature aggregation on occlusion handling.

5 Future Directions

In the future, there are still many interesting issues that can be studied and many remaining difficulties to be addressed with OVIS, such as:

Occlusion-aware Models Effectively handling occlusions is one of the most straightforward ways to improve the performance in OVIS. In terms of occlusion-aware models, there are a few directions that can be exploited in our future work. For example, compositional models (Kortylewski et

al., 2020a,b, 2021) might be a good choice as they are robust to partial occlusions. It is also interesting to test if completing the invisible parts of occluded objects (*a.k.a.* de-occlusion Zhan et al. 2020) is useful in this scenario.

Occluded Data Generation Due to the high cost of annotation, the scale of video instance segmentation datasets is relatively smaller than image datasets. Some works (DeVries & Taylor, 2017; Yun et al., 2019; Dwibedi et al., 2017; Ghiasi et al., 2021) have proposed augmenting the common datasets (e.g., COCO Lin et al. 2014) with partial occlusions, and some works (Nikolenko, 2019; Kar et al., 2019; Devarajan et al., 2020) synthesize structured amodal data in occluded scenes using simulators. It can be anticipated that utilizing those data with proper training paradigms will improve the performance in VIS.

Learning from Occlusion Annotations In OVIS, a coarse annotation of occlusion levels (no occlusion, slight occlusion, and severer occlusion) is given per object. As a prior knowledge that can be accessed during training, learning paradigms that can abstract such information deserve special attention.

Large Scale Model Pre-Training According to our experiments, it improves the performance to conduct joint training with image datasets. With the development of self-supervised learning (He et al., 2021), exploiting the unlimited amounts of unlabeled data for model pre-training, then transferring the pre-trained model into OVIS will largely enhance the discriminative power of frame embeddings.

Dataset Versatility At last, we are also interested in formalizing the experimental track of OVIS for video object segmentation, either in an unsupervised, semi-supervised, or interactive setting. It is also of paramount importance to extend OVIS to video panoptic segmentation (Kim et al.,

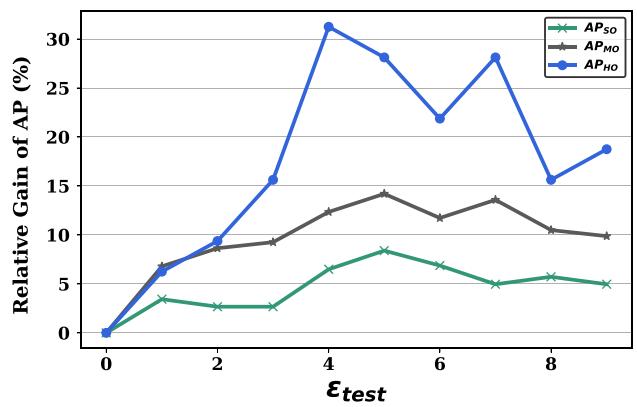


Fig. 11 Relative gain of different occlusion levels with increasing reference frame range ε_{test}

2020). We believe the OVIS dataset will trigger more research in understanding videos in complex and diverse scenes.

6 Conclusions

In this work, we target video instance segmentation in occluded scenes and accordingly contribute a large-scale dataset called OVIS. OVIS consists of 296k high-quality instance masks of 5223 heavily occluded instances. While being the second benchmark dataset after YouTube-VIS, OVIS is designed to examine the ability of current video understanding systems in terms of handling object occlusions. A general conclusion is that the baseline performance on OVIS is far below that on YouTube-VIS, which suggests that more effort should be devoted in the future to tackling object occlusions or de-occluding objects (Zhan et al., 2020). We also explore ways about leveraging temporal context cues to alleviate the occlusion matter and conduct a comprehensive analysis of occlusion handling on OVIS.

Acknowledgements This work is supported by Turing AI Fellowship EP/W002981/1.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2006). Youtube-8m: A large-scale video classification benchmark. arXiv preprint [arXiv:1609.08675](https://arxiv.org/abs/1609.08675)
- Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., & Leibe, B. (2020). Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*.
- Bertasius, G., & Torresani, L. (2020). Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*
- Bertasius, G., Torresani, L., & Shi, J. (2018). Object detection in video with spatiotemporal sampling networks. In *ECCV* (pp. 331–346).
- Bolya, D., Foley, S., Hays, J., & Hoffman, J. (2020). Tide: A general toolbox for identifying object detection errors. In *ECCV*
- Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2), 88–97.
- Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K. K., & Van Gool, L. (2019). The 2019 davis challenge on vos: Unsupervised multi-object segmentation. arXiv
- Cao, J., Anwer, R. M., Cholakkal, H., Khan, F. S., Pang, Y., & Shao, L. (2020). Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In: *ECCV* (pp. 213–229). Springer.
- Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4), 834–848.
- Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.
- Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., & Yu, N. (2017). Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *ICCV* (pp. 4836–4845).
- Chu, X., Zheng, A., Zhang, X., & Sun, J. (2020). Detection in crowded scenes: One proposal, multiple predictions. In *CVPR*.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *ICCV*.
- Devarajan, J., Kar, A., & Fidler, S. (2020). Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In *ECCV* (pp. 715–733). Springer.
- DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint [arXiv:1708.04552](https://arxiv.org/abs/1708.04552)
- Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *CVPR*.
- Dwibedi, D., Misra, I., & Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV* (pp. 1301–1310)
- Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., & Liu, W. (2021). Instances as queries. In *ICCV*.
- Fayyaz, M., Saffar, M.H., Sabokrou, M., Fathy, M., Klette, R., & Huang, F. (2016). STFCN: spatio-temporal fcn for semantic video segmentation. In *ACCV*.
- Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.
- Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T. Y., Cubuk, E. D., Le, Q. V., & Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR* (pp. 2918–2928).
- Gupta, A., Dollar, P., & Girshick, R. (2019). LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. arXiv preprint [arXiv:2111.06377](https://arxiv.org/abs/2111.06377)
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *CVPR*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
- Hegdé, J., Fang, F., Murray, S. O., & Kersten, D. (2008). Preferential responses to occluded objects in the human visual cortex. *JOV*, 8(4), 16–16.
- Hosang, J., Benenson, R., & Schiele, B. (2017). Learning non-maximum suppression. In *CVPR* (pp. 4507–4515).
- Hu, Y. T., Huang, J. B., & Schwing, A. G. (2018). Videomatch: Matching based video object segmentation. In *ECCV*.

- Huang, Z., Huang, L., Gong, Y., Huang, C., & Wang, X. (2019). Mask scoring R-CNN. In *CVPR*.
- Hwang, S., Heo, M., Oh, S. W., & Kim, S. J. (2021). Video instance segmentation using inter-frame communication transformers. arXiv preprint [arXiv:2106.03299](https://arxiv.org/abs/2106.03299)
- Johnander, J., Danelljan, M., Brissman, E., Khan, F. S., & Felsberg, M. (2019). A generative appearance model for end-to-end video object segmentation. In *CVPR*.
- Kar, A., Prakash, A., Liu, M. Y., Cameracci, E., Yuan, J., Rusiniak, M., Acuna, D., Torralba, A., & Fidler, S. (2019). Meta-sim: Learning to generate synthetic datasets. In *ICCV* (pp. 4551–4560).
- Ke, L., Tai, Y. W., & Tang, C. K. (2021). Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*.
- Khoreva, A., Perazzi, F., Benenson, R., Schiele, B., & Sorkine-Hornung, A. (2017). Learning video object segmentation from static images. In *CVPR*.
- Kim, D., Woo, S., Lee, J. Y., & Kweon, I. S. (2020). Video panoptic segmentation. In *CVPR*.
- Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *CVPR*.
- Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). Pointrend: Image segmentation as rendering. In *CVPR*.
- Kortylewski, A., He, J., Liu, Q., & Yuille, A. L. (2020a). Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *CVPR* (pp. 8940–8949).
- Kortylewski, A., Liu, Q., Wang, A., Sun, Y., & Yuille, A. (2021). Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *IJCV*, 129(3), 736–760.
- Kortylewski, A., Liu, Q., Wang, H., Zhang, Z., & Yuille, A. (2020b). Combining compositional models and deep networks for robust object classification under occlusion. In *WACV* (pp. 1333–1341).
- Lazarow, J., Lee, K., Shi, K., & Tu, Z. (2020). Learning instance occlusion for panoptic segmentation. In *CVPR* (pp. 10720–10729).
- Li, M., Li, S., Li, L., & Zhang, L. (2021). Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *CVPR*.
- Li, Q., Qi, X., & Torr, P. H. (2020). Unifying training and inference for panoptic segmentation. In *CVPR*.
- Li, S., Seybold, B., Vorobyov, A., Fathi, A., & Kuo, C. C. J. (2018). Instance embedding transfer to unsupervised video object segmentation. In *CVPR*.
- Li, X., & Loy, C. C. (2018). Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*.
- Li, Y., Xu, N., Peng, J., See, J., & Lin, W. (2020). Delving into the cyclic mechanism in semi-supervised video object segmentation. *NeurIPS*, 33.
- Lin, C. C., Hung, Y., Feris, R., & He, L. (2020). Video instance segmentation tracking with a modified vae architecture. In *CVPR*.
- Lin, T. Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.
- Liu, D., Cui, Y., Tan, W., & Chen, Y. (2021). SG-Net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*.
- Liu, Q., Chu, Q., Liu, B., & Yu, N. (2020). GSM: Graph similarity model for multi-object tracking. In *IJCAI* (pp. 530–536).
- Liu, S., Huang, D., & Wang, Y. (2019). Adaptive NMS: Refining pedestrian detection in a crowd. In *CVPR*.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV* (pp. 10012–10022).
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.
- Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. arXiv preprint [arXiv:1603.00831](https://arxiv.org/abs/1603.00831)
- Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception*, 18(1), 55–68.
- Nikolenko, S. I. (2019). Synthetic data for deep learning. arXiv
- Nilsson, D., & Sminchisescu, C. (2018). Semantic video segmentation by gated recurrent flow propagation. In *CVPR*.
- Oh, S. W., Lee, J. Y., Sunkavalli, K., & Kim, S. J. (2018). Fast video object segmentation by reference-guided mask propagation. In *CVPR*.
- Oh, S. W., Lee, J. Y., Xu, N., & Kim, S. J. (2019). Video object segmentation using space-time memory networks. In *ICCV*.
- Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*.
- Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P., & Bai, S. (2021). Occluded video instance segmentation: Dataset and ICCV 2021 challenge. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *IJCV*, 115(3), 211–252.
- Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., & Sun, J. (2018). Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint [arXiv:1805.00123](https://arxiv.org/abs/1805.00123)
- Smeulders, A. W., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., & Shah, M. (2013). Visual tracking: An experimental survey. *IEEE TPAMI*, 36(7), 1442–1468.
- Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. In *ICCV* (pp. 9627–9636).
- Tokmakov, P., Alahari, K., & Schmid, C. (2017). Learning motion patterns in videos. In *CVPR*.
- Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., & Chen, L. C. (2019). FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *CVPR*.
- Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., & Leibe, B. (2019). MOTS: Multi-object tracking and segmentation. In *CVPR*.
- Voigtlaender, P., & Leibe, B. (2017). Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*.
- Wang, H., Jiang, X., Ren, H., Hu, Y., & Bai, S. (2021). Swiftnet: Real-time video object segmentation. In *CVPR*.
- Wang, W., Feiszli, M., Wang, H., & Tran, D. (2021). Unidentified video objects: A benchmark for dense, open-world segmentation. arXiv preprint [arXiv:2104.04691](https://arxiv.org/abs/2104.04691)
- Wang, W., Song, H., Zhao, S., Shen, J., & Ling, H. (2019). Learning unsupervised video object segmentation through visual attention. In *CVPR*
- Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *CVPR* (pp. 7794–7803).
- Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., & Shen, C. (2018). Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*.
- Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., & Xia, H. (2020). End-to-end video instance segmentation with transformers. arXiv preprint [arXiv:2011.14503](https://arxiv.org/abs/2011.14503)
- Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M. C., Qi, H., Lim, J., Yang, M. H., & Lyu, S. (2020). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193, 102907.
- Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., & Yuan, J. (2021). Track to detect and segment: An online multi-object tracker. In *CVPR*.
- Wu, J., Song, L., Wang, T., Zhang, Q., & Yuan, J. (2020). Forest R-CNN: Large-vocabulary long-tailed object detection and instance segmentation. In *ACM Multimedia*.

- Wu, J., Zhou, C., Yang, M., Zhang, Q., Li, Y., & Yuan, J. (2020). Temporal-context enhanced detection of heavily occluded pedestrians. In *CVPR*.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *CVPR*.
- Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., & Urtasun, R. (2019). Upsnet: A unified panoptic segmentation network. In *CVPR*.
- Xu, J., Cao, Y., Zhang, Z., & Hu, H. (2019). Spatial-temporal relation networks for multi-object tracking. In *ICCV* (pp. 3988–3998).
- Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., & Huang, T. (2018). Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*.
- Xu, Z., Zhang, W., Tan, X., Yang, W., Huang, H., Wen, S., Ding, E., & Huang, L. (2020). Segment as points for efficient online multi-object tracking and segmentation. In *ECCV*.
- Yang, L., Fan, Y., & Xu, N. (2019). Video instance segmentation. In *ICCV*.
- Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., & Liu, W. (2021). Crossover learning for fast online video instance segmentation. In *ICCV*.
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV* (pp. 6023–6032).
- Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., & Loy, C. C. (2020). Self-supervised scene de-occlusion. In *CVPR*.
- Zhang, S., Benenson, R., & Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *CVPR* (pp. 3213–3221).
- Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In *ECCV*.
- Zhou, C., & Yuan, J. (2018). Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV* (pp. 135–151).
- Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., & Yang, M. H. (2018). Online multi-object tracking with dual matching attention networks. In *ECCV* (pp. 366–382).
- Zhu, X., Xiong, Y., Dai, J., Yuan, L., & Wei, Y. (2017). Deep feature flow for video recognition. In *CVPR*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.