A The OVIS Dataset

The Video frames, annotated labels and masks, dataset format, and evaluation code are all published at the dataset website http://songbai.site/ovis and the challenge website https://competitions.codalab.org/competitions/32377. You can also find OVIS at Papers With Code platform https://paperswithcode.com/dataset/ovis. As the creator of OVIS, we bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

License. The OVIS dataset is published under CC BY-NC-SA license, which means it is for non-commercial research purpose only.

Dataset Maintenance. The OVIS dataset is stored on both Google Drive and Baidu Wangpan for global researchers. Links for data are provided at https://competitions.codalab.org/competitions/32377#participate. OVIS will be maintained for a long time and we will check its accessibility regularly. We also plan to continue hosting the challenges and extend OVIS to more relevant tasks.

Dataset Format. The json format of annotations is COCO-style [5], just like YouTube-VIS [8], which means it's nearly cost-free to adapt the code for YouTube-VIS to running on OVIS. The detailed json format is provided at https://competitions.codalab.org/competitions/32377#participate.

More Details of Data Collection. All videos in the OVIS dataset are collected from our Youku video platform. The terms of the agreement between the video uploaders and us allow us to utilize these submitted videos for research purposes. We also remove the videos with degrading or embarrassing contents and clip the parts containing identity information, *e.g.*, names, watermarks, to protect privacy.

B Results of baselines on YouTube-VIS

Methods	YouTube-VIS validation set				
	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
FEELVOS [6]	26.9	42.0	29.7	29.9	33.4
IoUTracker+ [8]	23.6	39.2	25.5	26.2	30.9
MaskTrack R-CNN [8]	30.3	51.1	32.6	31.0	35.5
SipMask [2]	32.5	53.0	33.3	33.5	38.9
STEm-Seg [1]	30.6	50.7	33.5	31.6	37.1
TraDeS [7]	32.6	52.6	32.8	-	-
QueryInst-VIS [3]	34.6	**55.8**	36.5	**35.4**	**42.4**
STMask [4]	33.5	52.1	36.9	31.1	39.2
CrossVIS [9]	**34.8**	54.6	**37.9**	34.0	39.0

Table 1: Results of nine baseline methods on YouTube-VIS 2019 validation set.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020.

[2] 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 *ECCV*, 2020.

[3] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021.

[4] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *CVPR*, 2021.

[5] Tsung-Yi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[6] 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 *CVPR*, 2019.

[7] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *CVPR*, 2021.

[8] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

[9] 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 *ICCV*, 2021.