Unidentified Video Objects: A Benchmark for Dense, Open-World Segmentation
Supplementary material

Weiyao Wang Matt Feiszli Heng Wang Du Tran
Facebook AI Research
{weiyaowang,mdf,hengwang,trandu}@fb.com

Dataset	Attribute	ARlow	ARmid	ARhigh
UVO	IoU	30.5	18.7	2.6
	size change	32.1	16.5	3.9
	velocity	29.3	16.1	7.4
YTVIS	IoU	60.3	41.6	20.9
	size change	61.0	41.2	20.9
	velocity	49.8	44.9	28.4

Table 1: **Object with larger motion is harder.** Performance is measured by AR100. Low, mid and high indicates the magnitude of change with respect to the three attributes. A low inter-frame IoU indicates larger change of the object in time, and we see that performance is consistently lower on UVO and YTVIS. We decouple inter-frame IoU into size change and object velocity, and observe that performance is more distinguishable with respect to size change.

Appendix

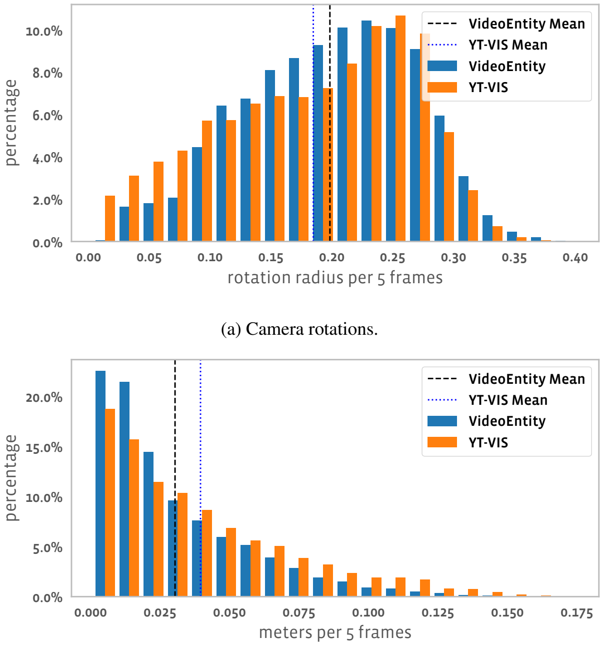
A. Moving objects are harder

In image object segmentation such as COCO, objects are divided into multiple groups based on size, so algorithms are evaluated on how well they perform on large/ medium/ small objects. For video-level evaluation, object motion is another important attribute to look at. We divide the objects by their motion attributes: mask-IoU, object velocity, and object size change. For each attribute, we divide objects into 3 equally-sized groups: high-motion, mid-motion, and low-motion. For example, object with low mask-IoU, high velocity, and large size change is considered as high-motion.

Results are shown in Table 1. Objects with higher motion have worse performance on video models for both UVO and YTVIS. In particular, when motion is quite significant (ARhigh), performance drops significantly. This suggests that current models may fail to handle significant motion in objects.

B. Camera Motion Statistics and Ablations

We use an off-the-shelf camera pose estimator [3] to compute camera motion for UVO and YouTube-VIS (YTVIS). Distributions are shown in Fig. 1. Both datasets offer a wide range of camera translations and rotations: UVO has slightly higher camera rotations and YTVIS has higher camera translations on average.



(a) Camera rotations.

(b) Camera translations.

Figure 1: **Comparing video camera motion statistics between UVO and YouTube-VIS.** For video camera motion, YTVIS and Video-Entity follows similar distributions (a and b).

Open-world segmentation is harder on larger camera motion videos. We also study the impact of two types of camera motions (rotation and translation) on our open-world segmentation task. For each motion type, we divide the videos into two subsets: high motion and low motion. Results are shown in Table 2. ON both YTVIS and UVO, MaskTrack R-CNN performs worse on videos with higher camera motion.

Pre-training ImageNet provides a better initialization. Existing instance segmentation models, such as Mask

Dataset	Attribute	ARlow	ARhigh
UVO	camera rotation	19.0	15.3
	camera translation	18.5	15.9
YTVIS	camera rotation	46.1	36.1
	camera translation	44.6	37.6

Table 2: **Open-world segmentation is harder on larger camera motion videos.** Performance is measured by AR100. Lower translation/ rotation in camera pose has higher performance compared to larger translation/ rotation on both UVO and YTVIS.

Dataset	Pre-train	AP	AR100
COCO	ImageNet	37.0	49.6
	Kinetics	30.0	44.4
UVO-Frame	ImageNet	22.2	41.3
	Kinetics	18.4	37.3
UVO-Video	ImageNet	9.3	23.0
	Kinetics	7.5	14.0

Table 3: **Pre-training on ImageNet is better than on Kinetics**. This indicates the taxonomy of ImageNet provides a better initialization for open-world object segmentation.

R-CNN, are often pretrained on ImageNet [1]. Since our dataset is collected from Kinetics400 [2] videos, it is natural to ask if domain differences matter (from image to video) and pretraining on Kinetics400 is more suitable. We replace the ImageNet pretraining with Kinetics400 pretraining (a frame-based model) for both Mask R-CNN and MaskTrack R-CNN.

Results are summarized in Table 3. When replacing ImageNet with Kinetics for the first stage of pre-training, all results got worse. One possible cause is that Kinetics videos are labeled by human actions (such as shooting soccer ball), and are not usually object-centered. As a result, Kinetics taxonomy and videos may not be proper for pre-training object detectors compared to ImageNet.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2

[2] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 2

[3] Tinghui Zhou, M. Brown, Noah Snavely, and D. Lowe. Unsupervised learning of depth and ego-motion from video. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619, 2017. 1