Opening up Open World Tracking

Supplementary Material

A. Defining Distractor Classes

We observe that some of the objects in TAO are similar to COCO classes, *i.e.*, visually similar to the set of *known* classes, but have different labels. Thus, they cannot be easily separated into *known* and *unknown*. We treat these categories as *distractor* classes and ignore them during the evaluation (similarly as in existing closed-world multi-object tracking datasets, *e.g.*, [18, 24]).

Different from prior work, finding *distractor* classes from over 800 TAO classes is not an easy task. Therefore, we develop a semi-automatic way to determine the *distractor* classes by looking at the COCO-class prediction frequency of a COCO pre-trained Mask R-CNN. We run the detector on all frames in the TAO validation set and predict proposals with COCO class labels. These COCO class prediction frequencies automatically highlight TAO object classes that are very visually similar to COCO classes. As an example, if a class 'minivan' (in TAO vocabulary) is frequently detected as a 'car' class (in COCO vocabulary), we tag this class for manual verification. This way we mark 42 classes as distractors, see Fig. 8. To ensure distractor classes do not "leak" into the set of *unknown* classes, we assign classes to distractors whenever there is any ambiguity among annotators of whether they are visually or semantically similar to their associated COCO classes.

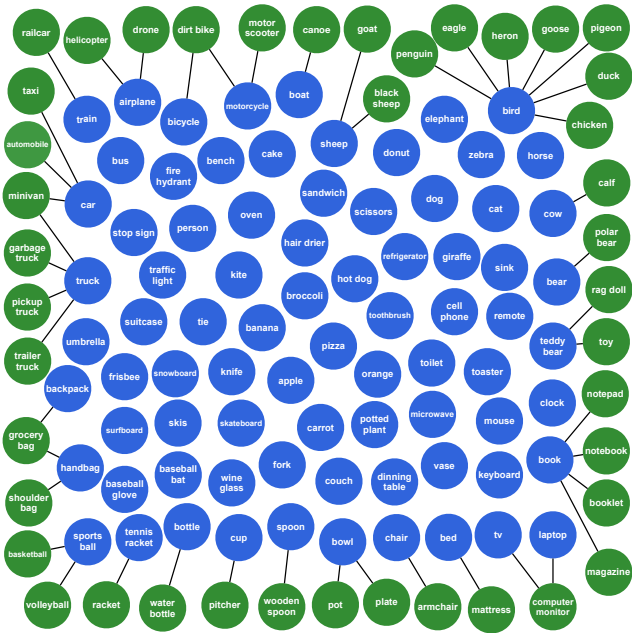

Figure 8. **Known and distractor object classes** The link between the *known* (blue) and the *distractor* (green) indicates visual similarity.

B. Implementation Details

Proposal generation We use a Cascade Mask R-CNN model directly from Detectron 2 [86]. During the inference, we disable the non-maxima-suppression (NMS) of Mask-RCNN and obtain 1000 proposals per frame.

Association similarity In the TAO dataset, the videos are annotated at one frame per second (every 30th frame annotated), and we gather pairs of contiguous annotated frames (*current frame* and *next frame*, 1 second apart). We extract matching ground truth objects in each frame-pair and form the set of paired ground truth objects as our evaluation set. As described in Sec. 5, we apply different similarity measures to see if given a proposal that matches with ground truth in the *current frame*, can be successfully associated with a proposal matching with the paired ground truth in *next frame*. To form a reasonable set of paired ground truths, two factors must be considered: (i) the ground truths with the same ID (in the same track) must be present in both frames; (ii) for each ground truth in *current frame*, there should be at least one proposal that overlaps with it with an IoU > 0.5, and the same for *next frame*. By forcing these two constraints, we ensure that the evaluation score will not be affected by unpaired ground truth or missing proposals, and therefore the evaluation can sorely focus on the association ability of various similarity measures. To compare performance, we use top-1 accuracy:

$$\text{accuracy} = \frac{\text{number of successful associations}}{\text{total number of paired GTs}}$$

For methods using optical flow in Table 3, we use PWC-Net [74] (with fine-tuned weights on MPI Sintel [12]) to calculate optical flow vectors.

C. Additional Experimental Evaluation Results

In Tab. 4 we outline the results of ablation studies on various long-term tracking and overlap removal strategies on TAO-OW val. These results are discussed in Sec. 5.2 and Sec. 5.3 in the main paper.

In Tab. 5 we outline results we obtain with our open-world tracking baseline (OWTB). We discuss these results in the Sec. 6 of the main paper.

		Known				Unknown			
		OWTA	D.Re	A.Re	A.Pr	OWTA	D.Re	A.Re	A.Pr
NO→T	Hung.	60.6	78.8	56.6	61.3	39.8	49.7	39.7	53.4
	Hung.+KA	60.0	78.8	57.4	57.2	39.7	49.7	40.7	48.9
	Hung.+OffTM	60.5	78.8	58.4	57.4	40.2	49.7	42.0	48.6
T→NO	Hung.	59.8	78.5	55.5	60.7	39.8	49.6	39.5	54.1
	Hung.+KA	59.3	78.4	56.4	56.9	40.0	49.6	41.0	50.1
	Hung.+OffTM	59.7	78.4	57.1	57.0	40.1	49.6	41.5	49.6

Table 4. **Long-term tracking and Overlap removal.** Ablation of various long-term tracking and overlap removal strategies on TAO-OW val. Hung.: Online Hungarian algorithm; KA: Online multi-step keep-alive strategy, OffTM: Offline tracklet merging. NO→T: Non-overlap first, and then track. T→NO: Track first, and then non-overlap.

DAVIS Unsup.	$\mathcal{J}\ \&\ \mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
RVOS [77]	41.2	36.8	45.7
PDB [72]	55.1	53.2	57.0
AGS [81]	57.5	55.5	59.5
ALBA [25]	58.4	56.6	60.2
MATNet [92]	58.6	56.7	60.4
STEm-Seg [2]	64.7	61.5	67.8
UnOVOST [47]	67.9	66.4	69.3
OWTB (Ours)	65.5	63.7	67.4

KITTI-MOTS	HOTA car	HOTA ped.
TrackR-CNN [78]	56.5	41.9
PointTrack [89]	61.9	54.4
OWTB (Ours)	64.0	52.7

Table 5. Results of our OWTB on closed-world benchmarks DAVIS Unsupervised (val) and KITTI-MOTS (test), compared to all previous published methods. *MOTSFusion additionally uses stereo-depth information.

C.1. Long-term tracking

After obtaining object proposals and determining a method for calculating the similarity between proposals over time, we combine all the proposals together into long-term tracks. We investigate various approaches from prior work, excluding certain expensive or complex approaches, such as QBPO optimization [56]. [17] uses Hungarian matching with a keep-alive mechanism to keep tracks alive through occlusions or missing detections. [47] first builds tracklets using Hungarian matching and then merges these tracklets in a second offline step.

In Table 4 we compare both of these approaches to long-term tracking, along with just a simple online Hungarian approach which both build upon. In general, the keep-alive strategy performs slightly better than without it, but the offline tracklet merging approach works the best of all. Note, however, that there is only a small difference between all these approaches. Generally, what one approach gains in association-recall, it loses most of in association precision. Successful long-term tracking is still an open challenge in both open-world and closed-world tracking.

C.2. Overlap removal

Finally, we investigate different approaches for removing overlaps between different tracks. This boils down to assigning a score per proposal such that we remove segments of proposals that overlap with any proposal with a higher score. We investigate two approaches for scoring proposals for overlap removal. The first, inspired by [47] and [17], simply takes the per frame proposal score (which we investigated earlier) and uses this for determining which proposals should be given priority to occlude other proposals, such that occluded proposals are made smaller to not overlap. Since this is done at the proposal level, it can be done before long-term tracking takes place. We call this 'Non-overlap and then track'. The second approach follows [56] and performs tracking first on the set of non-overlapping proposals. Each track as a whole is then scored using the mean score of each proposal in a track. Then overlap removal occurs using these track scores. Table 4 shows that these two approaches generally perform very similarly, though the simpler 'non-overlap then track' approach produces slightly better results.

OWTB vs. previous closed-world trackers. Does a tracker designed for performing well in open-world tracking also work in the traditional closed-world scenario? To test this, we evaluate our OWTB on two previous tracking benchmarks, DAVIS unsupervised [13] and KITTI-MOTS [78]. We choose DAVIS for video object segmentation, and KITTI-MOTS as it is the most commonly used MOTS benchmark. For DAVIS we use our own proposal-generation method. For KITTI-MOTS we use detections supplied by the benchmarks. Table 5 compares our method with prior work.

Despite not being tuned for these datasets, OWTB is competitive on both closed-world benchmarks. Note all other methods are specifically tuned for these benchmarks and the particular classes in the benchmark, reinforcing the strong generalization capability of our method.

D. Limitations and Societal Impacts

Limitations of presented baselines. With our benchmark, for the first time we are able to evaluate the difficulty of detecting and tracking unknown objects, using our proposed OWTB (Open World Tracking Baseline). From this we extract the following insights on the limitations of this approach, and believe that this could motivate future research directions:

- *Unknown object detection is significantly harder* (and thus less accurate) than known object detection: Unknown object detections are often incorrectly grouped into a union of different unknown objects; often only parts of objects are detected instead of the whole; many objects are also simply never recalled at all. Unknown object detection could be improved by taking into account temporal context (multi-frame) for detection, effectively combining elements of tracking and detection.
- *Low quality detections make tracking much harder*. When

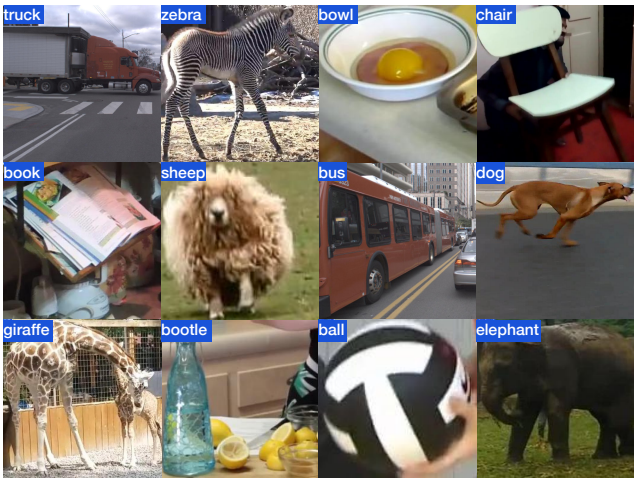

Figure 9. Additional examples of *known* object categories.

detections can be trusted (as with common classes), tracking reduces to identity assignment. When they cannot, the trackers must be robust to missing or partial detections.

- *Fast moving objects, with large deformation, are very challenging*.
- *Unknown objects which completely disappear and reappear again are almost never correctly tracked.* Building robust long-term appearance models of previously unseen objects is a key future research direction.
- *Relying on labeled data alone limits tracking methods.* Our current approach only transfers knowledge from labeled known classes to unknowns. Using unlabeled training data could result in potentially large improvements..

Societal Impacts. Open-world tracking allows operating in a world populated by never-before-seen possibly dynamic obstacles, and learning about semantic concepts with minimal supervision. Unfortunately, object tracking also has privacy and surveillance repercussions, as it can be used for person tracking. Our work focuses more on the class-agnostic setting, rather than the well-established pedestrian tracking, but could be used for this purpose as well.

E. Additional Qualitative Results

We provide the additional examples of *known*, *distractor* and *unknown* object categories in Figures 9, 10 and 11. We also show tracking results of our Open-World Tracking Baseline (OWTB) for *knowns*, *unknowns* and *unknown unknowns* in Figures 12, 13 and 14.

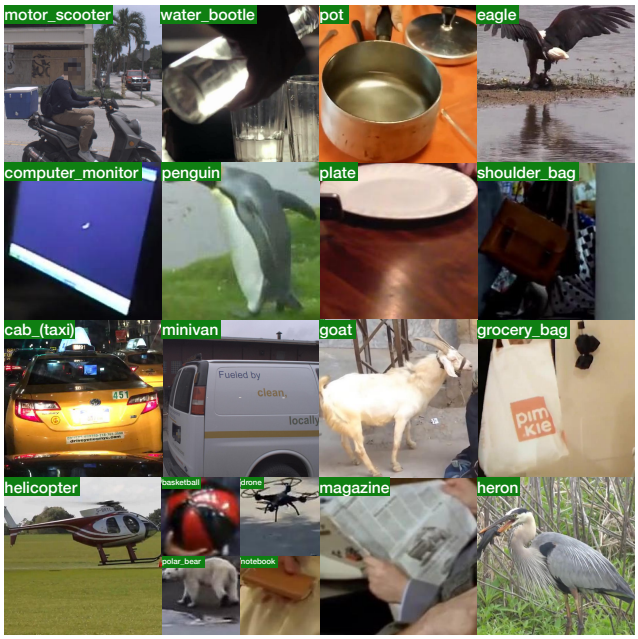

Figure 10. Examples of *distractor* object categories.

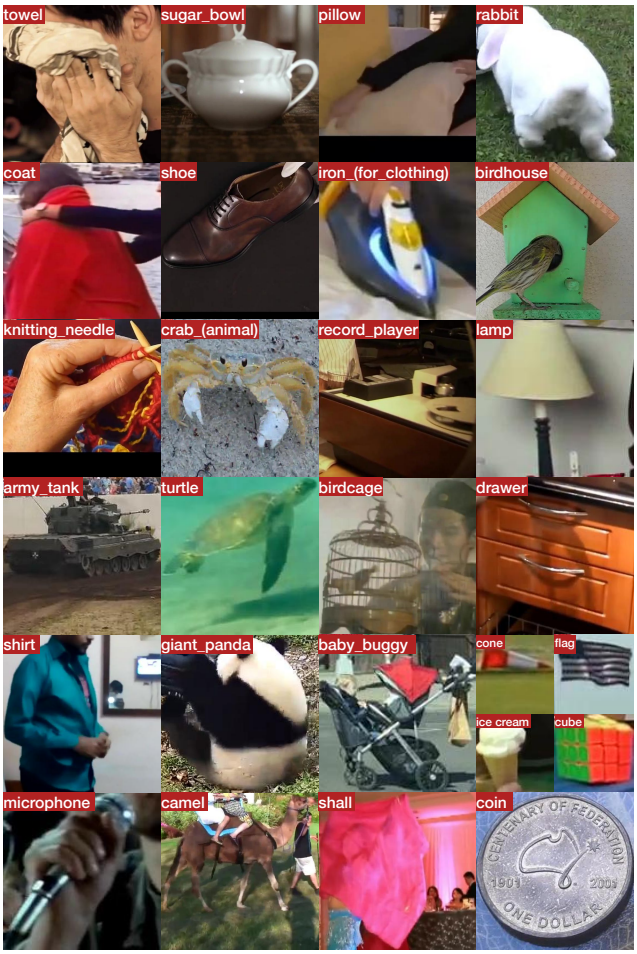

Figure 11. Additional examples of *unknown* object categories.

Figure 12. **Tracking results for *known*.** Examples of *known* objects tracked by OWTB. OWTB is capable of tracking objects in cluttered scenes (*second row*), and making robust associations despite of motion blur (*fifth row*).

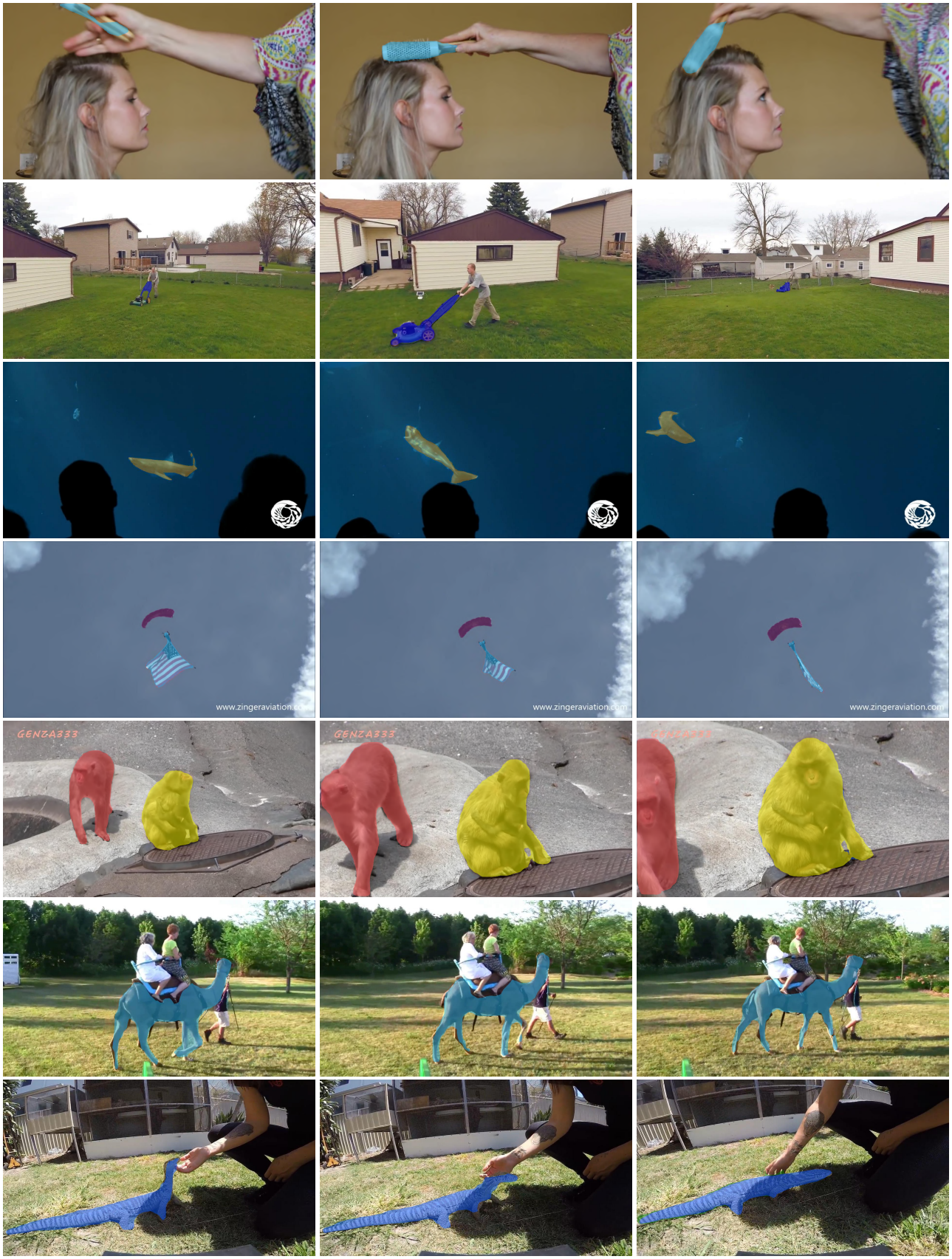

Figure 13. **Tracking results for *unknown*.** Examples of *unknown* objects tracked by OWTB. OWTB is robust for motion blur (*first row, third image*) and large motions (*second row*).

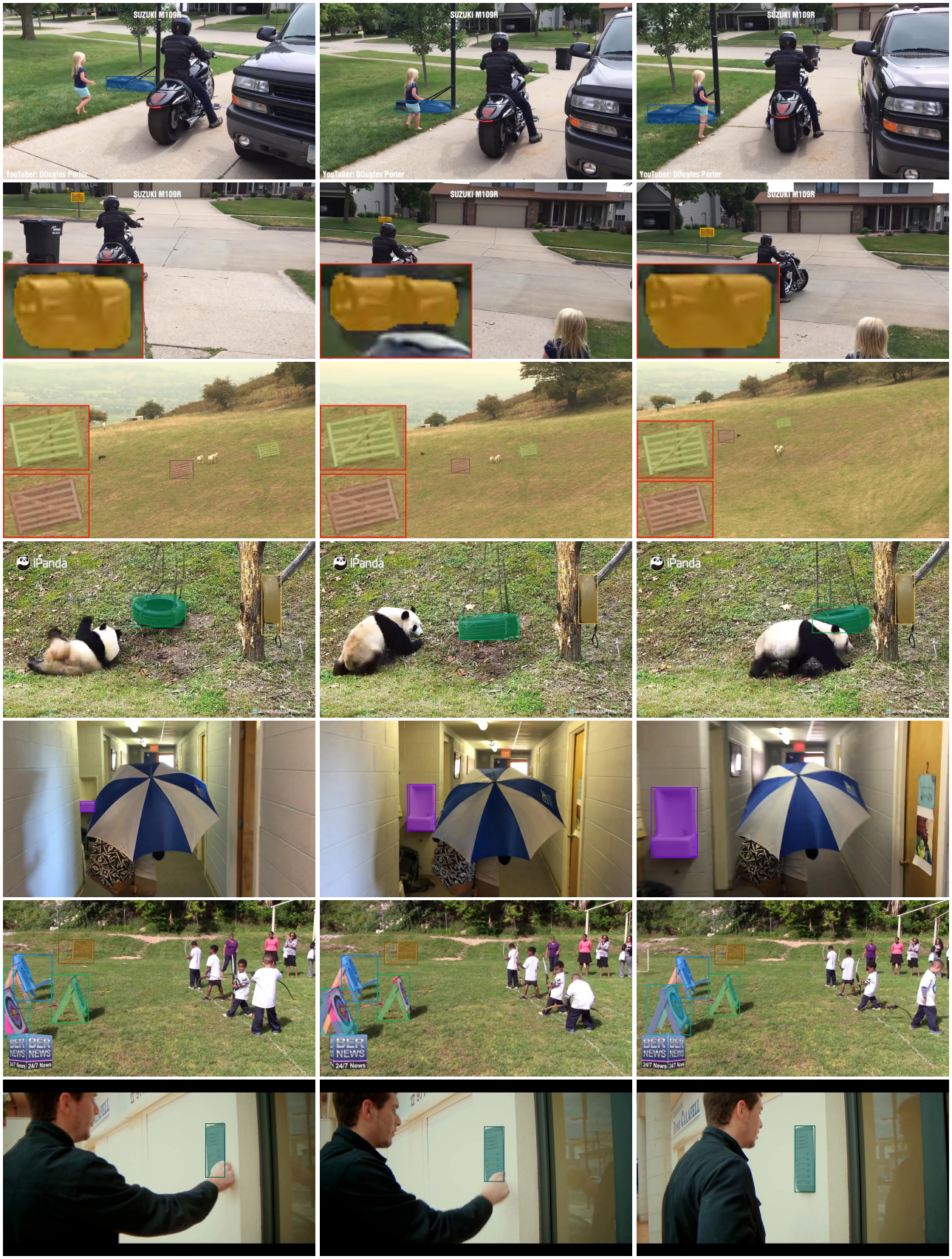

Figure 14. **Tracking results for *unknown unknowns*.** Examples of unlabeled objects outside of the TAO [16] vocabulary which are correctly tracked by OWTB. OWTB performs well even for small objects (*second* and *third row*).