Supplementary Materials: A Simple Learning Framework for Large Vocabulary Video Object Detection

Sanghyun Woo[1,†], Kwanyong Park[1,†],
Seoung Wug Oh[2], In So Kweon[1], and Joon-Young Lee[2]

[1] KAIST
[2] Adobe Research

1 Appendix

In this appendix, we provide,

A1. Our view of the proposal from video data scaling perspectives,
A2. Datasets specifics used in the experiments,
A3. Implementation details including COCO → YTVIS transfer learning setup,
A4. Oracle analyses to investigate the disentangled impact of the method on object classification and tracking,

1.1 Bridging Images and Videos

Applying deep learning in the video domain fundamentally suffers from the data-hungry issue, and the situation will become even more severe for more complex and challenging tasks. One promising direction we believe is leveraging already well-curated large-scale image data to complement the insufficient video data. However, jointly using multiple datasets [14], image and video labels, leads to several issues, detailed below.

In this paper, we investigate the new problem of large vocabulary tracking, one of the essential milestones for dynamic world understanding AI agents. The task naturally lacks training labels as the data collection and annotation procedure is extremely expensive. As a remedy, leveraging the large-scale images is an attractive solution [3]. However, in doing so, we face three main issues: 1) lacking video supervision in images, 2) semantic label inconsistency between images and videos, and 3) the domain gap (e.g., explicit data styles or implicit data distributions are different) between images and videos. The current learning paradigm bypasses the first two issues by independently training the detection head and tracking head with images and videos (*decoupled*). Instead, our learning framework explicitly handles the former two issues by hallucinating the supervisions and enables end-to-end video model learning from all training data, leading

[†] This work was done during an internship at Adobe Research.

to better feature representations (*unified*). The last issue is implicitly handled by the two-step training of image pre-training and video fine-tuning. In the preliminary experiments, we observe a slight performance drop when concatenating image and video datasets as a single dataset, possibly due to the weaker feature adaptation toward the video domain.

The abovementioned issues are fundamental and compounded when jointly using the image and video labels for video recognition models. We thus see they should be carefully considered and adequately handled. Our proposal is an initial effort in this direction, and we believe more clever and innovative solutions will be developed and presented in the future.

1.2 Data

LVIS LVIS [4] is a large-scale benchmark for large vocabulary image recognition. It provides precise bounding boxes and masks annotations for various categories with the long-tailed distribution. To be consistent with the prior works [3,9], we use LVIS v0.5 dataset and pre-train the model on 482 (out of 1230) LVIS categories that correspond to TAO categories.

TAO TAO [3] is the first video benchmark for large vocabulary video recognition. TAO dataset annotates 482 classes in total, which are the subset of LVIS dataset. It has 400 videos, 216 classes in the training set, 988 videos, 302 classes in the validation set, and 1419 videos, 369 classes in the test set. The videos are annotated in 1 FPS. We fine-tune the model on TAO-train and evaluate on TAO-val (or TAO-test).

We additionally use COCO [8] and YTVIS [12] to evaluate the proposed teacher-student scheme on COCO $\rightarrow$ YTVIS transfer learning setup.

COCO 2017 COCO contains 118k training images and 5k validation images. We pre-train the model on 20 (out of 80) COCO categories, as the remaining 60 categories cannot be evaluated with YTVIS annotations.

YTVIS YTVIS is largest video benchmark for video instance segmentation. YTVIS annotates 40 classes in total. It has 2238 training, 302 validation, 343 test video clips. The videos are annotated in 5 FPS. We fine-tune the model on 30 (out of 40) YTVIS categories to simulate missing 10 categories and new 20 categories during transfer learning (see Fig. 1). We evaluate the model on YTVIS-val.

1.3 Implementation Details

Training. The proposals are implemented under the MMdetection framework [1]. The COCO-style training schedule of 2$\times$ and 1$\times$ are adopted for LVIS pre-training and TAO fine-tuning, respectively. We set the maximum number of predictions per image to 1000 for not losing correct predictions at frame-level [2], which makes the inter-frame object affinity matrix large and the subsequent tracking

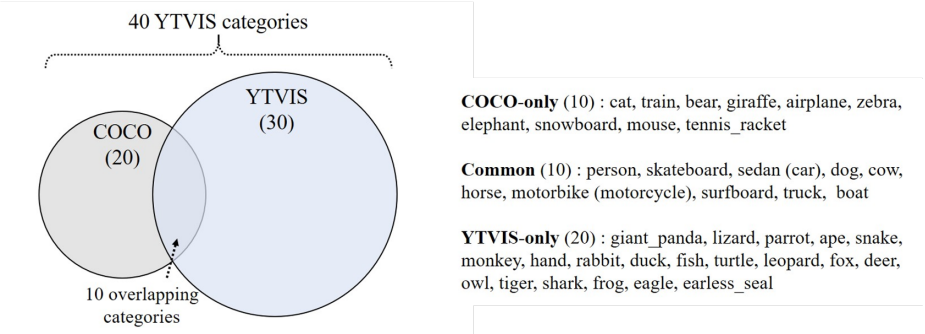

Fig. 1: We provide the detailed category distribution setup of COCO → YTVIS transfer learning used in the experiments.

challenging. However, thanks to the proposed mosaic training, we see our tracker is robust to dense object tracking. Batch size of 16 (2 per GPU) and an initial learning rate of 0.02 are used. We randomly select a scale between 640 to 800 to resize the shorter side of images during training. For the hyper-parameters of the models, we follow the original implementations [4,11,10,13,9]. The standard learning protocol is first pre-training the model in LVIS and then fine-tuning on TAO, *i.e.*, decoupled learning.

Testing. Our method processes video frames recursively, generating detection boxes and matching them with the candidate tracks from the past frames. Apart from the conventional tracking algorithms, we see the motion is highly irregular for tracking the large vocabulary of objects. Thus, the most reliable way is to link detection boxes only based on their appearance features. Here, the main matching strategy is a bi-directional softmax operation that examines the two matched objects being each other's nearest neighbor in the embedding space [9]. For the unmatched tracks, we keep them until it remains for more than 30 frames. We use resized frames of 1080×1080 for testing.

COCO → YTVIS transfer learning setup. We adopt MaskRCNN [5] with ResNet-50 FPN [6,7] backbone for the COCO image pre-training. We use 1× training pipeline. We transfer the pre-trained MaskRCNN model weights to MaskTrack RCNN [12] for the YTVIS video fine-tuning. At this stage, new weights are added to the class head, bounding box head, and mask head to accommodate newly added classes. Also, the track head is appended to the model for object tracking. We follow the original training schedules and hyper-parameters of MaskTrack RCNN [12].

The presented two-step teacher-student scheme is applied to the model during transfer learning (see Fig. 1). The distillation in the class head and the bounding box head follows the methods noted in the main paper. For the mask head distillation, we collect teacher and student mask predictions and minimize their difference through MSE loss.

Method	Oracle Class (pure tracking ability)			Oracle Track (pure classification ability)		
	Track AP50	Track AP75	Track AP	Track AP50	Track AP75	Track AP
SORT [3]	30.2	-	-	31.5	-	-
Decoupled [9]	34.7	12.2	15.1	32.1	12.1	14.8
Unified (Ours)	**38.1**	**17.1**	**18.4**	**43.1**	**16.7**	**19.9**

Table 1: **Oracle analysis.** We analyze the performance of two types of oracles: Oracle Class and Oracle Track. The former provides the pure tracking ability of the model, and the latter allows us to analyze pure classification ability.

As the detailed class-wise evaluation in YTVIS is only possible for the 40 object categories, we simulate the pattern in Fig. 1 by shrinking the original object categories in COCO and YTVIS. In specific, for image pre-training, we trained the MaskRCNN on 20 COCO object categories. For video fine-tuning, we trained the MaskTrack RCNN on YTVIS videos with 30 object categories, which consist of 10 overlapping object categories with the 20 COCO pre-trained categories and 20 new object categories. The 10 overlapping object categories are selected based on the annotation frequency. In this way, we can simulate the co-existence of missing object categories and new object categories during transfer learning.

1.4 Oracle Analysis

To disentangle the impact of methods on object classification and tracking, we use two oracles: class oracle and track oracle on TAO validation set [3].

For ***class oracle***, we first compute the best matching between predicted and ground truth tracks in each video. The predicted tracks that match to a ground truth track with 3D IoU > 0.5 are assigned the category of their matched ground truth track. Tracks that do not match to a ground truth track are treated as false positives. This allows us to analyze the ***pure tracking ability*** of models assuming the classification task is solved.

For ***track oracle***, we compute the best possible assignment of per-frame detection boxes to tracks, by comparing them with ground truth. The class predictions for each detection are held constant. Any detection boxes that are not matched are removed. This allows us to evaluate the ***pure classification ability*** of models given a perfectly linked per-frame detection boxes.

The results are summarized in Table 1. We use the same FasterRCNN-RFS tracker for the Decoupled and Unified methods. We observe that our method outperforms the previous approaches in both oracle types. This shows that the unified learning framework using all training data, LVIS and TAO, essentially improves the model's tracking and classification ability significantly.

References

1. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 2
2. Dave, A., Dollár, P., Ramanan, D., Kirillov, A., Girshick, R.: Evaluating large-vocabulary object detectors: The devil is in the details. arXiv:2102.01066 (2021) 2
3. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: ECCV. pp. 436–454. Springer (2020) 1, 2, 4
4. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR. pp. 5356–5364 (2019) 2, 3
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017) 3
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 3
7. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017) 3
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 2
9. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: CVPR. pp. 164–173 (2021) 2, 3, 4
10. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: CVPR. pp. 1685–1694 (2021) 3
11. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: CVPR. pp. 9695–9704 (2021) 3
12. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV. pp. 5188–5197 (2019) 2, 3
13. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv:2103.07461 (2021) 3
14. Zhou, X., Koltun, V., Krähenbühl, P.: Simple multi-dataset detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7571–7580 (2022) 1