Bridging Images and Videos: 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, Daejeon, South Korea
[2] Adobe Research, San Jose, USA
jolee@adobe.com

Abstract. Scaling object taxonomies is one of the important steps toward a robust real-world deployment of recognition systems. We have faced remarkable progress in images since the introduction of the LVIS benchmark. To continue this success in videos, a new video benchmark, TAO, was recently presented. Given the recent encouraging results from both detection and tracking communities, we are interested in marrying those two advances and building a strong large vocabulary video tracker. However, supervisions in LVIS and TAO are inherently sparse or even missing, posing two new challenges for training the large vocabulary trackers. First, no tracking supervisions are in LVIS, which leads to inconsistent learning of detection (with LVIS and TAO) and tracking (only with TAO). Second, the detection supervisions in TAO are partial, which results in catastrophic forgetting of absent LVIS categories during video fine-tuning. To resolve these challenges, we present a simple but effective learning framework that takes full advantage of all available training data to learn detection and tracking while not losing any LVIS categories to recognize. With this new learning scheme, we show that consistent improvements of various large vocabulary trackers are capable, setting strong baseline results on the challenging TAO benchmarks.

**Keywords:** Large vocabulary · Video object detection and tracking

1 Introduction

A central goal of computer vision is to produce a general-purpose perception system that robustly works in the wild. Towards this ambitious goal, extending the current short category regime is one of the essential key milestones. As an

S. Woo and K. Park—This work was done during an internship at Adobe Research.

initial effort in this direction, the large-scale image benchmark, LVIS [16], was introduced and fostered significant progress in developing solid image domain solutions [24, 58, 70, 71, 76, 104]. Recently, a video benchmark, TAO [12], calls for a shift from image to video, opening the new task of detecting and tracking large vocabulary objects.
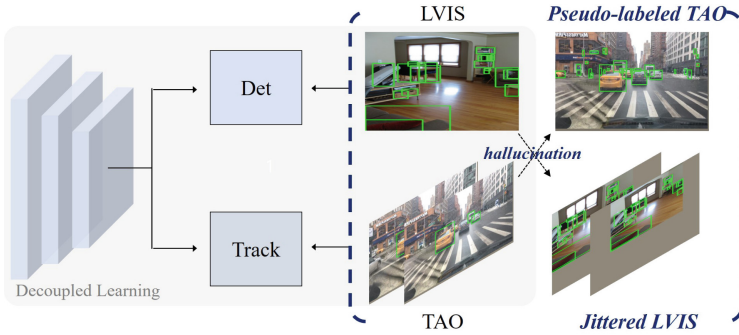
Fig. 1. The Proposed Learning Framework for Large Vocabulary Tracker Training. While the current learning paradigm learns detection and tracking separately from LVIS and TAO (decoupled), our proposal takes all training data to learn detection and tracking jointly (**unified**). This is achieved through missing supervision hallucination.

With these new datasets of images and videos, LVIS and TAO, we are interested in building a strong large vocabulary video tracker. However, as the annotation difficulty between images and videos is even more severe in large vocabulary datasets, the significant gap in dataset scale and label vocabularies naturally exists. Therefore, pre-training the model on images for learning large vocabularies and then fine-tuning on video for seamless video domain adaptation is a standard learning protocol. Given this context, can the current advances of large-vocabulary detection and multi-object tracking be successfully unified and tied into a single model? In particular, we see there are two main challenges for the successful marriage of two streams: **First, no tracking supervisions are in LVIS.** This essentially leads to inconsistent learning of detection (with LVIS and TAO) and tracking (only with TAO), resulting in sub-optimal video feature representations. **Second, detection supervisions in TAO are partial[1].** Thus, catastrophic forgetting [45] is inevitable if one naively fine-tunes the LVIS tracker directly on the TAO.

In this work, we present simple, effective, and generic methods for hallucinating missing supervisions in each dataset. Below, we describe the challenges and our solutions in turn.

---

[1] TAO dataset annotates 482 classes in total, which are the subset of LVIS dataset [16], and only 216 classes are in the training set.

First, how can we simulate the tracking supervisions only with images in LVIS? Given an image, our idea is to apply spatial jittering artifacts to mimic temporal changes in video and form a natural pair for tracking. Here, we present two new spatial jittering methods. The first is strong zoom-in/out augmentation, which has a large scale-jittering effect that can effectively simulate the low sampling rate test-time inputs in large vocabulary tracking. It yields significant performance improvements over the conventional image affine augmentation [20,48,52,66,102,105]. Plus, our findings are in line with recent work that shows a large scale jittering is effective in image detection and segmentation [15], and here we examine this observation on video for robust large vocabulary tracker training. The second is mosaicing augmentation [5,94], which is originally presented for object detection with enriched background. We extend this augmentation for combining foreground objects in different images in a class-balanced manner [16] and simulate test-time hard, dense tracking scenarios suitable for "many object" trackers. We show that both are effective and complementary to each other.

Second, how can we fill the missing detection supervisions in TAO? The TAO training data partially spans the LVIS categories, and thus the direct fine-tuning of LVIS tracker on TAO causes catastrophic forgetting of absent categories. A straightforward way to avoid this issue is to learn only the tracking part of the model with TAO [51]. However, this hinders full model training and abandons all the TAO detection labels, limiting the overall performance. We instead approach this problem by combining the self-training [60,61,63] with a teacher-student framework [21,89]. In practice, the teacher and student are identical copies of the LVIS pre-trained model, and we freeze the weights of the teacher during training. The overall learning pipeline consists of two steps: First, given an input, we predict pseudo labels using the teacher model. The idea behind the pseudo labeling is to leverage the past knowledge acquired from LVIS and fill in the missing annotations in TAO. Second, using the augmented labels, we train the student model with both distillation loss and ordinal detection loss. Unlike the typical teacher-student schemes used in semi-supervised object detection studies [67,91,92], we introduce two new adaptations suitable for large vocabulary learning setup. The first is using *soft* pseudo labels, i.e., distilling class logits directly, to fire all the student's classifier weights, rather than using common one-hot (hard) pseudo labels. This is crucial as standard hard pseudo labels tend to bias distillation only toward the frequent class objects due to inherent classifier calibration issue [11,50]. The second is to use *MSE* loss in order to equally impact all the classifier weights [28,72], rather than using picky KL-divergence loss [21]. We found the type of the loss function is also very important for the successful large vocabulary classifier distillation. Despite the simplicity, we empirically show that the distillation results are greatly improved with these adaptations. We also show that our proposal works well on the common vocabulary setup, e.g., COCO, and can be easily extended to new class learning scenarios, COCO → YTVIS.

Combining all these proposals together, unified learning of detection and tracking with both LVIS images and TAO videos becomes possible without forgetting any LVIS categories (see Fig. 1). Furthermore, we also introduce a new regularization objective, semantic consistency loss. It aims to prevent the common tracking failure in large vocabulary tracking due to semantic flicker between similar classes. We study the efficacy of our final framework on the TAO benchmark and achieve new state-of-the-art results. Our extensive ablation studies confirm that the proposals are generic and effective.

2   Related Work

Large Vocabulary Recognition. The object categories in natural images follow the Zipfian distribution [44], and thus, the large vocabulary recognition is naturally tied with the long-tailed recognition [17,42,86]. Based on this connection, lots of solid approaches are introduced. The existing methods can be roughly categorized into *data re-sampling* or *loss re-weighting*. The data re-sampling methods more often sample data from rare classes to balance the long-tailed training distribution [8,16]. The loss re-weighting aims at adjusting the loss of each data instance based on their labels or train-time accumulated statistics [22,70,71,76]. Some approaches perform multi-staged training upon these methods, which first pre-train the model in a standard way and then fine-tune using either data re-sampling or loss re-weighting [23,24,39,58,78,79,82,98]. Also, there are new approaches based on data augmentations [15,95,97] or test time calibration [50].

Apart from all these previous efforts on the image, we study a new video extension of the task [12]. We show that our proposal is generic and not sensitive to a specific method, data re-sampling or loss re-weighting, in successfully converting the current large vocabulary detectors to large vocabulary trackers.

Multi-object Tracking. Most modern multi object trackers [35] follow the tracking-by-detection paradigm [56]. An off-the-shelf object detector is first employed to localize all objects in each frame, and then track association is performed between adjacent frames. The main difference among existing methods is in how they estimate the similarity between detected objects and previous tracks for the association. To name a few, Kalman Filter [4,84], optical flow [88], displacement regression [53,105], and appearance similarities [3,25,34,35,43,47,51,62,68,83,93,99,100] are the representatives. On the other side, there are also efforts on joining detection and tracking [13,85,101], and recently by transformer-based architectures [46,69,96]. We note that all these methods only focus on a few object categories such as people or vehicles, ignoring the vast majority of objects in the world.

Our work is an early attempt for extending the current short category regime of modern trackers [12,41,80,107]. In this paper, we build our proposal upon the tracking-by-detection paradigm. We choose the state-of-the-art method, QDTrack [51], which adopts Faster R-CNN [59] and lightweight embedding head for detection and tracking, respectively. The tracking is learned through a dense matching between quasi-dense samples on the pair of images and optimized with

multiple positive contrastive learning. Given the state-of-the-art large vocabulary detection [16,70,76] and multi-object tracking [51] methods, we primarily investigate the new challenges in developing a strong large vocabulary tracker.

Tracking Without Video Annotations. There is a line of recent research on self-supervised learning for tracking, either using unlabeled videos [32,33,38,55, 73,77,81,90] or images [14,20,48,52,66,102,105]. Our work belongs to the latter category. Applying random affine augmentation to the original image provides a spatially jittered version, which mimics the temporal changes in the video. By letting the model find the correspondence between those two images, meaningful tracking supervision can be provided [17,48,105]. Sio *et al.* [66] present that an image and any cropped region of it can generate a similar effect. Zheng *et al.* [102] extend this idea to incorporate only the foreground objects in cropped regions for stable training.

In this work, we explore this general idea under a more specific large vocabulary tracking setting. First, we focus on the fact that conventional motion cues are not applicable for large-vocabulary trackers as the input are temporally distant (1FPS) due to the annotation difficulties and there are severe camera movements in natural videos. This motivate us to train the tracker's vision feature matching more discriminative. To this end, we present a strong zoom-in/out augmentation that can not only simulate low sampling rate input but also includes large scale-jittering effect [15] which is known to be effective in the image domain vision tasks. Second, we recast the image mosaicing augmentation [5], which was initially proposed for robust object detection with enriched backgrounds [94], to simulate test-time dense tracking in the large vocabulary setting. We show that both are complementary in providing discriminative tracking supervisions for this task.

Catastrophic Forgetting. The phenomenon wherein neural networks forget how to solve past tasks because of the exposure to new tasks is known as catastrophic forgetting [45]. It occurs because the model weights that contain important information for the old task are over-written by information relevant to the new one. While the catastrophic forgetting can occur in various scenarios, many existing efforts are focused on the class incremental learning setup, where it incrementally adds new object categories phase-by-phase, in image classification [1,9,29,36,54,57,64,87]. Also, there are some few approaches tackling incremental object detection [30,65,75,103].

We target a different setup, transfer learning from image to video without forgetting. Specifically, we aim to train the model on images covering the entire evaluation categories and then fine-tune it on videos, which partially covers the evaluation categories, without forgetting. While lots of current video models [27, 49,93] are trained in this way for generic feature learning, the label difference issue between images and videos has been rarely studied and explored. We study this issue, as this is a practical setup for training large vocabulary trackers using both images and videos.
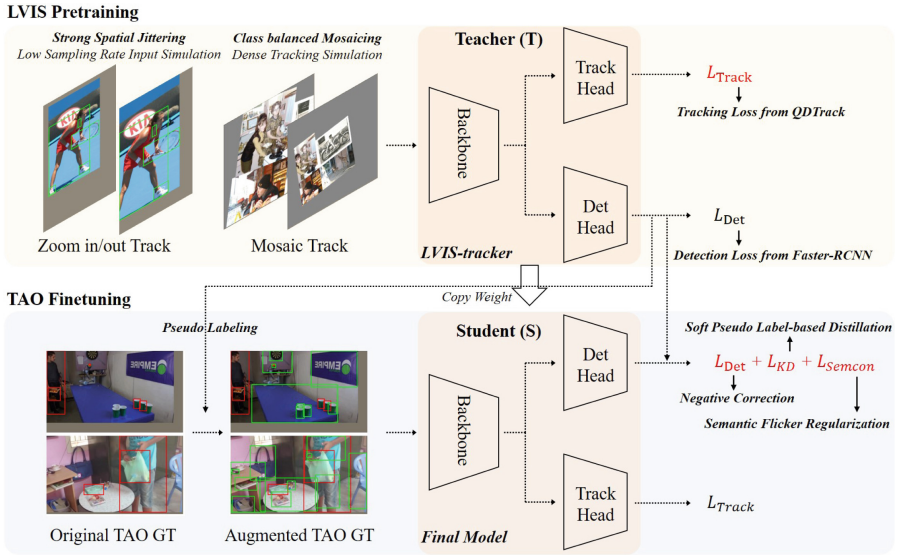
Fig. 2. Overview of the proposed learning framework. The red colored objective functions are generated supervisions with our proposals. (Color figure online)

3 Proposed Method

We introduce a general learning framework that allows joint learning of detection and tracking from all training data, LVIS and TAO, for robust large vocabulary tracking. The overview of our pipeline is shown in Fig. 2. We first present how we can learn tracking from images through zoom-in/out and mosaicing augmentations in Sect. 3.1. We then describe how we avoid catastrophic forgetting when the videos for fine-tuning have fewer label vocabularies than pre-trained images in Sect. 3.2. Finally, we present a new regularization loss term, namely semantic consistency loss, for preventing semantic flicker in Sect. 3.3.

3.1 Learn to Track in LVIS

Our approach is straightforward. An original image and a transformed image with the spatial jittering artifacts can form a natural input pair for tracking. For the jittering artifacts, we present two new augmentations, zoom-in/out and mosaicing (see Fig. 2). Note that tracking annotations come for free as we know the exact transformation relationship between the images. We assign the same unique track-id to the same object in the transformed image.

Strong Zoom-in/Out Track. Due to the annotation difficulty of the large-vocabulary tracking dataset, the train and test time inputs are temporally sparse, i.e., low sampling rate, which naturally results in conventional motion cues not applicable and rather rely on pure vision feature matching. To make the vision

feature matching more discriminative, and to effectively simulate test-time low sampling rate inputs, we present strong zoom-in/out augmentation.

It is mainly composed of scaling and cropping operations, which essentially vary the scale and position of the objects. Specifically, for an image I, we generate a input pair, I_t and $I_{t+\tau}$, by applying the scale_and_crop$(\cdot)$ function to each image. In practice, it scales an image up to 2 times and crops the image to have a minimum IoU of 0.4 or above with original bounding boxes to avoid heavy object truncation and ensure stable tracker training. Prior works either adopt standard random affine transformation [17, 48, 105] or cropping without scaling [66, 102], which generally provide *weak* scale-jittering effect. Instead, we focus on enlarging the scaling effect and show that our proposal significantly outperforms the baselines.

Mosaicing Track. While the zoom-in/out augmentation is already effective in providing tracking supervisions, it is limited in the tracking of a few objects due to the federated annotations of LVIS [16]. To resolve the issue, we present to combine multiple images and perform tracking with the increased foreground objects. We implement our idea by extending the image mosaicing augmentation [5], which stitches four random training images with certain ratios. While it was originally presented for object detection with enriched background [94], we recast it to simulate hard, dense tracking scenarios in large vocabulary tracking. In practice, four random images, $\{I_a, I_b, I_c, I_d\}$, are sampled from RFS (Repeat Factor Sampling)-based dataset [16] to maintain the class-balance. Then, image stitching followed by random affine (with large scale jittering within a range of 0.1 to 2) and crop is applied. We summarize these procedure as mosaic$(\cdot)$. The tracking pair then can be obtained by applying the mosaic$(\cdot)$ function to the sampled images twice. However, we see that unnatural layout pair results in train and test time inconsistency. To this end, we propose to sample tracking input pairs in a mixed way from two different augmentations, zoom-in/out and mosaicing, with equal probability during training. We empirically confirm that this works well in practice.

With our proposal, the model can receive tracking supervisions from all LVIS object categories. The tracking objective function is adopted from the QDtrack [51] (see Fig. 2-top), and we call this model LVIS-Tracker. While the model is only trained on LVIS dataset, it already outperforms the previous state-of-the-art tracker (trained with the standard decoupled learning scheme) significantly (see Table 2a).

3.2   Learn to Unforget in TAO

Due to the fundamental annotation difficulties in videos, the images are in general bigger in dataset scale and larger in taxonomies. Therefore, pre-training the model on images to acquire generic features and fine-tuning on videos for target domain adaptation has become a common protocol for obtaining satisfactory performance in various video tasks [27, 49, 93]. This also applies to training the large vocabulary video trackers, where we first learn a large number of vocabulary from LVIS images and then adapt to the evaluation domain with TAO

videos. However, as TAO partially spans the full LVIS vocabularies, a naive transfer learning scheme results in catastrophic forgetting.

Here, our goal is to keep the ability to detect the previously seen object categories while also adapting to learn from new video labels. We mainly focus on the catastrophic forgetting in the detector, as the tracking head is learned in a category-agnostic manner. We detail the proposal using the standard two-staged Faster-RCNN detector (FPN backbone) [40,59]. Without loss of generality, the proposals can be extended to multi-staged architectures [6,7,10,74], where we apply the proposal for each RCNN head and average them. In fact, the main issue is missing annotations for the seen, known object categories during the image to video transfer learning. Since they are not annotated, we can neither provide detection supervision nor prevent them from being treated as background. This basically perturbs the pre-trained classifier boundaries of both RPN and RCNN, leading to catastrophic forgetting. We remedy this issue by presenting a pseudo-label guided teacher-student framework.

Our key idea is intuitive. The pre-trained model already has sufficient knowledge to detect the seen, known categories. Based on this fact, we first fill in the missing annotations by pseudo-labeling the input. We adopt the basic pseudo-labeling scheme with a threshold of 0.3. The redundant pseudo labels that highly overlap with the current labels are filtered out with NMS. With these augmented labels, we 1) design a teacher-student network to provide (soft) supervisions, i.e., class logit, for preserving the past knowledge, and 2) update the incorrect background samples, i.e., negatives, in RPN and RCNN to prevent seen objects from being background (see Fig. 2-bottom). Using soft class logit is important for the large vocabulary classifier distillation, as the hard pseudo labels bias the operation towards the frequent class objects. Moreover, we use MSE loss instead of Kullback-Leibler (KL) divergence loss [21] for the logit matching. This is because the MSE loss treats all classes equally and thus it allows the rare classes with low probability also to be updated properly [72]. This two new adaptation leads to the successful distillation of the previous knowledge of the large vocabulary classifier (see Table 2b).

Teacher-Student Framework Setup. To effectively retain the previous knowledge, we design a teacher-student framework. We first make identical copies of the image pre-trained model, teacher (T) and student (S). The teacher model (T) is frozen to keep the previous knowledge and guide the student. The student model (S) adapts to the new domain with incoming video labels (via detection loss) and also mimics the teacher model to preserve the past information (via distillation loss). We detail the components in the following.

RPN Knowledge Distillation Loss. The RPN takes multi-level features from the ResNet feature pyramid [40]. In particular, each feature map is embedded through the convolution layer, followed by two separate layers, one for objectness classification and the other for proposal regression. We collect the outputs of both heads from the teacher and student to compute RPN distillation loss, which is defined as $L_{\text{KD}}^{\text{RPN}} = \frac{1}{N_{cls}} \sum_{i=1} L_{cls}(u_i, u_i^*) + \frac{1}{N_{reg}} \sum_{i=1} L_{reg}(v_i, v_i^*)$. Here, i is the index of an anchor. u_i and u_i^* are the mean subtracted objectness logits obtained

from the student and the teacher, respectively. v_i and v_i^* are four parameterized coordinates for the anchor refinement obtained from the student and teacher, respectively. L_{cls} and L_{reg} are MSE loss and smooth L1 loss, respectively. Here, we note that L_{reg} is only computed for the positive anchors that have an IoU larger than 0.7 *with the augmented ground-truth boxes.* $N_{cls}(= 256)$ and N_{reg} are the effective number of anchors for the normalization.

RCNN Knowledge Distillation Loss. We perform RoIAlign [18] on top-scoring proposals from RPN, extracting the region features from each feature pyramid level. Each region feature is embedded through two FC layers, one for classification and the other for bounding box regression. We collect the outputs of both heads from the teacher and student to compute RCNN distillation loss, which is defined as $L_{\text{KD}}^{\text{RCNN}} = \frac{1}{M_{cls}}\sum_{j=1} L_{cls}(p_j, p_j^*) + \frac{1}{M_{reg}}\sum_{j=1} L_{reg}(t_j, t_j^*)$. Here, j is the index of a proposal. p_j and p_j^* are the mean subtracted classification logits obtained from the student and the teacher, respectively. t_j and t_j^* are four parameterized coordinates for the proposal refinement obtained from the student and teacher, respectively. L_{cls} and L_{reg} are MSE loss and smooth L1 loss, respectively. We only impose L_{reg} for the positive proposals that have an IoU larger than 0.5 *with the augmented ground-truth boxes.* $M_{cls}(= 512)$ and M_{reg} are the effective number of proposals for the normalization.

Correcting Negatives in Computing the Detection Loss. We avoid sampling the anchors or proposals that have significant IoU overlaps *with the augmented ground-truth boxes* as a background (>0.7 for RPN and >0.5 for RCNN). We note that positives are only sampled based on the provided original ground truth labels. This is because the detectors, especially the large vocabulary detectors, suffer from predicting the precise labels [11,50] while they are good at recalling the objects. We empirically verify this in the experiment.

Extension to Other Transfer Learning Setup. COCO to YTVIS is another important transfer learning setup (see Fig. 3-(a)). This is more challenging than LVIS to TAO, as the superset-subset relationship does not hold, and new object categories to learn are added. To deal with this new pattern, we take a two-step approach (see Fig. 3-(b)). First, we adapt the RCNN classifier of the pre-trained model, increasing the number of output channels to accommodate newly added classes, and train on the videos, YTVIS − COCO, that contain new object categories. In practice, we freeze the original detector, and thus the past information is intact, and only the newly added weight matrices are updated accordingly. The key idea here is to use the original pretrained weight as an anchor and update the newly added weight to be compatible. Second, after sufficient training of the new weights, we now unlock the original detector and update the whole weights with the remaining videos, YTVIS∩COCO, using the presented teacher-student scheme.

3.3   Regularizing Semantic Flickering

One of the common tracking failures in large vocabulary tracking is due to semantic flicker between similar object categories [12]. To cope with this issue,
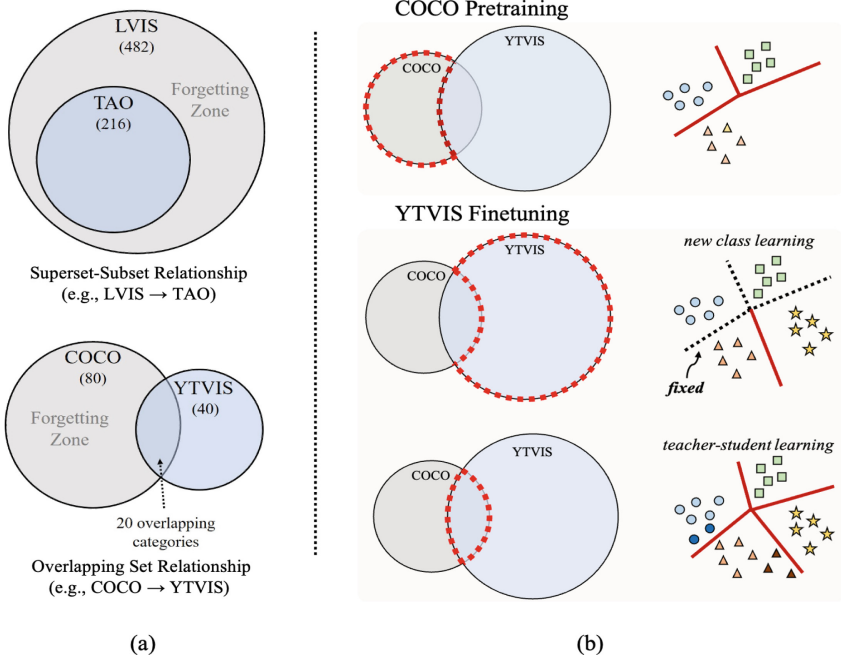
Fig. 3. (a) Two standard image to video transfer learning setups. In typical, a naive transfer learning from images to videos leads to catastrophic forgetting due to the missing annotations in the video. We present a generic teacher-student scheme that works on both scenarios. **(b) Two-step approach for COCO → YTVIS transfer learning setup.** We first learn new object classifier weights with the pre-trained classifier as a fixed anchor and then fine-tune the whole classifier through the proposed teacher-student scheme. The red-dotted line along the circle in the set relationship figure indicates the training data used in each stage. The shape figures (e.g., square, triangle) and the separating line denote class instances and the associated classifier. (Color figure online)

we attempt to regularize the model during training with a new objective function, namely semantic consistency loss. The proposal is motivated by the temporal consistency loss [2, 26, 31, 37], which enforces the outputs of the model for corresponding pixels (or patches) in video frames to be consistent. It is often used in video processing tasks to ensure the output temporal smoothness at a pixel level. The proposal extends this idea from pixels to instances; We enforce the class predictions of the same instances in two different frames to be equivalent. In practice, we forward the ground truth bounding boxes of the same instance in two different frames to the RCNN head. The mean subtracted classification logits, p, are used for the consistency regularization as, $L_{\text{Semcon}} = |p^t - p^{t+\tau}|_2$. ere, p^t and $p^{t+\tau}$ denote the logits of the same instance in two different frames, I_t and $I_{t+\tau}$.

3.4   Unified Learning

Within our proposed learning framework (see Fig. 2), we can train the whole video model, learning detection and tracking jointly, using all available image and video datasets. The final objective function can be summarized as

$$L = \lambda_1 L_{\text{Det}} + \lambda_2 L_{\text{Track}} + \lambda_3 L_{\text{KD}} + \lambda_4 L_{\text{Semcon}}, \tag{1}$$

which consists of four loss terms in total. The detection (L_{Det}) and tracking losses (L_{Track}) are adopted from [59] and [51]. Note that the $L_{\text{KD}}(= L_{\text{KD}}^{\text{RPN}} + L_{\text{KD}}^{\text{RCNN}})$ and L_{Semcon} are used only when fine-tuning on the videos.

4   Experiments

In this section, we conduct extensive experiments to analyze our methods. We investigate the results mainly in two aspects: image-level prediction and cross-frame association, which will be reflected in the BBox AP and Track AP [12], respectively. Considering the task difficulty, we mainly focus on the Track AP of 50, 75 and their average. For the TAO test, we provide Track AP*, a full Track AP average for IoU from 0.5 to 0.95 with a step size of 0.05. We study the impact of unified learning on TAO dataset (Sect. 4.1). We consistently outperformed the current decoupled learning paradigm with healthy margins using various models, and pushed the state-of-the-art performance significantly. Second, to investigate the importance of the major components in our proposals, we provide ablation studies on TAO validation set (Sect. 4.2). Lastly, we evaluate our teacher-student scheme on two representative image-video transfer learning scenarios, LVIS → TAO and COCO → YTVIS[2](Sect. 4.3). In the following, we provide experiment setups, evaluation protocol and results for each section. More details are in supplementary materials.

4.1   Main Results

Upon the state-of-the-art tracking-by-detection framework [51], we instantiate various large vocabulary trackers. In specific, we consider two important detection architecture, two-staged (Faster-RCNN [59]) and multi-staged (CenterNet2 [106]), and three different long-tailed learning methods, Repeat Factor Sampling (RFS) [16], Equalization Loss V2 (EQLv2) [70], and Seesaw Loss [76]. All the models use the same ResNet-101 [19] with feature pyramid [40] backbone following the previous works [12,51]. Based on these baseline models, we compare our learning framework with the current standard learning protocol, decoupled learning. The comparison is in Table 1. We observe that our unified learning scheme consistently outperforms the current decoupled learning paradigm on various models, showing the strong generalizabilty of the proposal. With our method, we push the state-of-the-art performance significantly, achieving 21.6 and 20.1 Track AP50 on TAO-*val* and TAO-*test*, respectively.

---

[2] For the experiment, we contact the authors for the YTVIS-*val* annotations.

Table 1. Our learning framework couples well with different model architectures and learning methods. All the baseline scores are obtained after the decoupled training, i.e., training the detector and tracker on LVIS and TAO, respectively. FasterRCNN-RFS* is a re-implementation of [51] baseline.

Method	Track AP_{50}	Track AP_{75}	Track AP_{avg}
SORT_TAO [12]	13.2	-	-
FasterRCNN-RFS [51]	16.1	5.0	10.6
FasterRCNN-RFS*	13.4	4.9	9.2
w/ SimLearn	19.6	7.3	(+4.4) 13.6
FasterRCNN-EQLv2	14.2	5.5	10.1
w/ SimLearn	19.8	8.8	(+4.1) 14.2
FasterRCNN-Seesaw	15.4	5.7	10.5
w/ SimLearn	20.2	9.4	(+4.3) 14.8
CenterNet2-RFS	18.9	9.1	14.0
w/ SimLearn	21.6	10.4	(+2.1) 16.1

(a) **State-of-the-art results** in TAO-*val*.

Method	Track AP_{50}	Track AP_{75}	Track $AP_{50:95}$
SORT_TAO [12]	10.2	4.4	4.9
FasterRCNN-RFS [51]	12.4	4.5	5.2
FasterRCNN-RFS* + SimLearn.	17.1	6.9	7.8
FasterRCNN-EQLv2 + SimLearn.	16.8	7.2	8.0
FasterRCNN-Seesaw + SimLearn.	17.6	8.0	8.5
CenterNet2-RFS + SimLearn.	20.1	9.6	**10.3**

(b) **State-of-the-art results** in TAO-*test*.

4.2   Ablation Studies

Impact of Image Spatial Jitterings. The results are presented in Table 2b. Compared to the standard affine transformation [17,48,105] or simple cropping without scaling [66,102], the presented strong zoom-in/out and mosaicing provides a large Track AP improvement. This indicates both the low sampling rate input simulation (with large-scale jittering) and dense tracking simulation (with mosaicing) enables more accurate large vocabulary object associations at test-time. To concretely investigate the scaling effects of zoom-in/out, we also provide its variant with small scale-jittering, Z-in/out*, and confirm that the large scale-jittering [15] is indeed important for the performance. We notice that mosaicing augmentation drops the Box AP. We conjecture this happens due to the train and test time inconsistency of input pairs. To this end, we present to form a tracking pair from two different augmentations in equal probability. We found that this mixed sampling strategy provides the best Track AP.

Impact of Teacher-Student Framework. In Table 2b, we study the impact of the key proposals in teacher-student framework. For the baselines, we provide the Naive-ft and Vanilla Teacher-Student schemes. Naive-ft indicates fine-tuning on TAO videos without any proper regulation for forgetting, which results in a significant performance drop. Vanilla Teacher-Student scheme samples the distillation targets only from the original ground truth labels, and no negative correction is performed. While it shows the past knowledge preservation effect to some extent, the performance is still worse than the LVIS-tracker. The vanilla scheme starts to improve over the LVIS-tracker when our proposal is added. This implies that pseudo labeling is essential, and 1) keeping the past knowledge of seen objects (by sampling distillation targets from the augmented labels) and 2) preventing the seen objects from being background (by correcting negatives using the augmented labels) are the key to avoid catastrophic forgetting.

One may wonder if the standard (hard) pseudo-labeling approach can directly preserve the previous knowledge as typical teacher-student scheme do [67,91,92]. However, as can be shown in the results, we instead observe inferior results than the baseline. The large vocabulary classifier fundamentally suffers from the confidence calibration issue [11,50] as it is trained on the long-tailed class-imbalanced data. It results in the classifier bias; predictions are made mainly toward the frequent object categories, missing rare objects in one-hot hard pseudo labels. In contrast, the (soft) pseudo labels essentially affect all classes. Furthermore, we suggest to employ MSE loss rather than standard KL-loss [21] as objective function in distillation. As MSE loss treats all classes equally the impact of the gradient is not attenuated for the rare classes. Recent study also reveals that MSE loss offers better generalization capability due to the direct matching of logits compared to the KL loss [28].

Impact of Semantic Consistency Loss. Finally, we study the impact of semantic consistency loss. It regularizes the model's class logits of the same instance in different frames to be the equivalent. In Table 2b, we observed mean-

Table 2. (a) Zoom-in/out* and Zoom-in/out denote zoom-in/out augmentation with scaling range of [0.8, 1.25] and [0.1, 2.0], respectively. (b) Pseudo Labeled Training denotes the standard (hard) pseudo label-based training.

Method	Box AP	Track AP$_{50}$	Track AP$_{75}$
Decoupled TAO-tracker [51]	18.1	16.1	5.0
Random Affine [20,48,105]	17.4	13.6	5.0
Random Crop [66,102]	15.9	11.5	1.5
Zoom in/out*	18.1	13.2	4.4
Zoom in/out	**19.0**	14.4	**6.0**
Mosaic	16.2	16.5	4.7
Both (LVIS-tracker)	18.5	**17.8**	5.7

(a) **Effect of image spatial jitterings** in LVIS pre-train.

Method	Box AP	Track AP$_{50}$	Track AP$_{75}$
LVIS-tracker	18.5	17.8	5.7
Naive-ft	11.7	11.4	2.7
Vanilla Teacher-Student	18.0	15.9	6.7
+ Pseudo-distill target	18.6	17.8	7.2
+ Pseudo-neg sample	18.8	17.9	7.3
Both (Ours)	19.5	18.5	**7.5**
w. Hard Pseudo label [67,91,92]	17.1	15.9	5.3
w. KL-based distill [21]	17.7	16.8	6.0
w. Semcon. (Final Tracker)	**19.6**	**19.5**	7.3

(b) **Effect of Teacher-Student framework** in TAO fine-tune.

ingful improvement in Track AP. This implies that semantic flicker regularization is indeed effective for the large vocabulary object tracking.

4.3   Image to Video Transfer Learning

Here, we evaluate our teacher-student scheme on two representative image to video transfer learning setups (see Fig. 3). In LVIS → TAO setup, we pre-train FasterRCNN-RFS tracker on LVIS (with 482 categories) and fine-tune on TAO (with 216 categories). We evaluate the model on TAO-*val* with Track AP metric. In COCO → YTVIS setup, we pre-train Mask-RCNN [18] on COCO, transfer the weights to MaskTrack RCNN [93], add new randomly initialized classifier weights to accommodate newly added classes, and fine-tune on YTVIS. More details of the setup are in supplementary materials. We evaluate the model on YTVIS-*val* with Mask AP [93] metric. To quantitatively analyze whether the proposal properly preserves the past knowledge and benefits from the new video labels, we provide the scores of OLD and NEW. Here, OLD indicates the classes that only reveal in the image pre-training stage. NEW denotes the classes that appears in the video fine-tuning stage. For each setup, we provide a baseline of naive fine-tuning, which results in a severe catastrophic forgetting. The results are summarized in Table 3 and Table 4.

Table 3. Teacher-student framework in **LVIS → TAO** transfer learning setup. Evaluated on TAO-*val*.

Method	OLD (LVIS − TAO)			NEW (TAO)			ALL		
	Box AP	Track AP_{50}	Track$_{75}$	Box AP	Track AP_{50}	Track$_{75}$	Box AP	Track AP_{50}	Track$_{75}$
LVIS-tracker	15.7	16.1	5.7	21.1	17.2	9.0	18.5	17.8	5.8
Naive-ft	7.1	7.7	1.3	16.2	14.9	3.7	11.7	11.4	2.6
Track-only	15.7	15.3	5.5	21.1	16.9	7.1	18.5	16.1	6.3
Teacher-Student	**15.7**	**16.3**	**6.5**	**23.1**	**20.6**	**9.0**	**19.5**	**18.5**	**7.5**

Table 4. Teacher-Student framework in **COCO → YTVIS** transfer learning setup. Evaluated on YTVIS-*val*.

Method	OLD (COCO − YTVIS)			NEW (YTVIS)			ALL		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
Naive-ft	5.2	2.6	2.7	38.2	21.0	20.6	30.0	16.4	16.1
Teacher-Student (1-step)	38.9	31.2	26.2	36.2	19.4	20.1	36.9	22.4	21.6
Teacher-Student (2-step)	51.0	44.3	36.1	**43.3**	23.8	**23.8**	**45.2**	28.9	26.9
Full-ft (Oracle)	**54.9**	**46.9**	**38.9**	40.7	**25.1**	23.3	44.3	**30.6**	**27.2**

LVIS → TAO Transfer Learning. Especially in this setup, all necessary vocabularies are already learned at the image pre-training stage. Therefore, we can

avoid catastrophic forgetting by fine-tuning only the tracking part. However, as it only updates the video model partially, it leads to inconsistent video representations, and thus performance rather slightly drops from the baseline LVIS-tracker. Instead, our method preserves the performance of OLD classes (preventing catastrophic forgetting) and significantly improves the NEW class performance (benefiting from labeled learning).

COCO → YTVIS Transfer Learning. This setup is more challenging as the model is required to achieve two goals, new class learning and old class preserving simultaneously. We decompose these goals and approach this setup in two-step as described in Sect. 3.2. As can be shown in the results, the proposed two-step approach performs better than the direct application of teacher student scheme. The final performance is comparable with, and in NEW classes outperforms, the oracle setup that use all the YTVIS training videos. This shows that our proposal is generic and effective for standard image to video transfer learning setups.

5 Conclusion

In this paper, we tackle the challenging problem of learning a large vocabulary video tracker. We present a simple learning framework that uses all LVIS images and TAO videos to jointly learn the detection and tracking. In specific, first, two spatial jittering methods, strong zoom-in/out and mosaicing, which effectively simulate the test-time large vocabulary object tracking are presented to enable tracker training with LVIS. Second, a generic teacher-student scheme is proposed to prevent catastrophic forgetting while fine-tuning the image pre-trained models on videos. We show that two new adaptation of using soft labels with MSE loss is crucial for the large vocabulary classifier distillation. We hope our new learning framework settles as a baseline learning scheme for many follow-up large-vocabulary trackers in the future.

References

1. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. arXiv:1903.08671 (2019)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
3. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV, pp. 941–951 (2019)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP, pp. 3464–3468 (2016)
5. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv:2004.10934 (2020)

6. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: CVPR, pp. 6154–6162 (2018)
7. Cai, Z., Vasconcelos, N.: Cascade R-CNN: high quality object detection and instance segmentation. PAMI **43**, 1483–1498 (2019)
8. Chang, N., Yu, Z., Wang, Y.X., Anandkumar, A., Fidler, S., Alvarez, J.M.: Image-level or object-level? A tale of two resampling strategies for long-tailed detection. arXiv:2104.05702 (2021)
9. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.S.: Riemannian walk for incremental learning: understanding forgetting and intransigence. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 556–572. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_33
10. Chen, K., et al.: Hybrid task cascade for instance segmentation. In: CVPR, pp. 4974–4983 (2019)
11. 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)
12. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: a large-scale benchmark for tracking any object. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 436–454. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_26
13. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: ICCV, pp. 3038–3046 (2017)
14. Fu, Y., Liu, S., Iqbal, U., De Mello, S., Shi, H., Kautz, J.: Learning to track instances without video annotations. In: CVPR, pp. 8680–8689 (2021)
15. Ghiasi, G., et al.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR, pp. 2918–2928 (2021)
16. Gupta, A., Dollar, P., Girshick, R.: LVIS: a dataset for large vocabulary instance segmentation. In: CVPR, pp. 5356–5364 (2019)
17. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
20. Held, D., Thrun, S., Savarese, S.: Learning to track at 100 FPS with deep regression networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 749–765. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_45
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
22. Hsieh, T.I., Robb, E., Chen, H.T., Huang, J.B.: Droploss for long-tail instance segmentation. arXiv:2104.06402 (2021)
23. Hu, X., Jiang, Y., Tang, K., Chen, J., Miao, C., Zhang, H.: Learning to segment the tail. In: CVPR, pp. 14045–14054 (2020)
24. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. arXiv:1910.09217 (2019)
25. Kim, C., Li, F., Ciptadi, A., Rehg, J.M.: Multiple hypothesis tracking revisited. In: ICCV, pp. 4696–4704 (2015)
26. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Deep video inpainting. In: CVPR, pp. 5792–5801 (2019)
27. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Video panoptic segmentation. In: CVPR, pp. 9859–9868 (2020)

28. Kim, T., Oh, J., Kim, N., Cho, S., Yun, S.Y.: Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation. arXiv:2105.08919 (2021)
29. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. **114**(13), 3521–3526 (2017)
30. Kuznetsova, A., Ju Hwang, S., Rosenhahn, B., Sigal, L.: Expanding object detector's horizon: Incremental learning framework for object detection in videos. In: CVPR, pp. 28–36 (2015)
31. Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Yumer, E., Yang, M.-H.: Learning blind video temporal consistency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 179–195. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_11
32. Lai, Z., Lu, E., Xie, W.: MAST: a memory-augmented self-supervised tracker. In: CVPR, pp. 6479–6488 (2020)
33. Lai, Z., Xie, W.: Self-supervised learning for video correspondence flow. arXiv:1905.00875 (2019)
34. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: siamese CNN for robust target association. In: CVPR Workshops, pp. 33–40 (2016)
35. Leal-Taixé, L., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S.: Tracking the trackers: an analysis of the state of the art in multiple object tracking. arXiv:1704.02781 (2017)
36. Lee, S.W., Kim, J.H., Jun, J., Ha, J.W., Zhang, B.T.: Overcoming catastrophic forgetting by incremental moment matching. arXiv:1703.08475 (2017)
37. Lei, C., Xing, Y., Chen, Q.: Blind video temporal consistency via deep video prior. In: Advances in Neural Information Processing Systems 33 (2020)
38. Li, X., Liu, S., De Mello, S., Wang, X., Kautz, J., Yang, M.H.: Joint-task self-supervised learning for temporal correspondence. arXiv:1909.11895 (2019)
39. Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: CVPR, pp. 10991–11000 (2020)
40. 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)
41. Liu, Y., Zulfikar, I.E., et al.: Opening up open-world tracking. arXiv:2104.11221 (2021)
42. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: CVPR, pp. 2537–2546 (2019)
43. Lu, Z., Rathod, V., Votel, R., Huang, J.: RetinaTrack: online single stage joint detection and tracking. In: CVPR, pp. 14668–14678 (2020)
44. Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
45. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: the sequential learning problem. Psychol. Learn. Motiv. **24**, 109–165 (1989)
46. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: TrackFormer: multi-object tracking with transformers. arXiv:2101.02702 (2021)
47. Milan, A., Rezatofighi, S.H., Dick, A., Reid, I., Schindler, K.: Online multi-target tracking using recurrent neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
48. Oh, S.W., Lee, J.Y., Sunkavalli, K., Kim, S.J.: Fast video object segmentation by reference-guided mask propagation. In: CVPR, pp. 7376–7385 (2018)
49. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV, pp. 9226–9235 (2019)

50. Pan, T.Y., et al.: On model calibration for long-tailed object detection and instance segmentation. arXiv:2107.02170 (2021)
51. Pang, J., et al.: Quasi-dense similarity learning for multiple object tracking. In: CVPR, pp. 164–173 (2021)
52. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Per-clip video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1352–1361 (2022)
53. Peng, J., et al.: Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 145–161. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_9
54. Prabhu, A., Torr, P.H.S., Dokania, P.K.: GDumb: a simple approach that questions our progress in continual learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 524–540. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_31
55. Purushwalkam, S., Ye, T., Gupta, S., Gupta, A.: Aligning videos in space and time. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12371, pp. 262–278. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58574-7_16
56. Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: CVPR, vol. 2, pp. II–II. IEEE (2003)
57. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: incremental classifier and representation learning. In: CVPR, pp. 2001–2010 (2017)
58. Ren, J., et al.: Balanced meta-softmax for long-tailed visual recognition. arXiv:2007.10740 (2020)
59. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, vol. 28, pp. 91–99 (2015)
60. Riloff, E.: Automatically generating extraction patterns from untagged text. In: Proceedings of the National Conference on Artificial Intelligence, pp. 1044–1049 (1996)
61. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 105–112 (2003)
62. Sadeghian, A., Alahi, A., Savarese, S.: Tracking the untrackable: learning to track multiple cues with long-term dependencies. In: ICCV, pp. 300–311 (2017)
63. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Trans. Inf. Theory **11**(3), 363–371 (1965)
64. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. arXiv:1705.08690 (2017)
65. Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: ICCV, pp. 3400–3409 (2017)
66. Sio, C.H., Ma, Y.J., Shuai, H.H., Chen, J.C., Cheng, W.H.: S2SiamFC: self-supervised fully convolutional siamese network for visual tracking. In: Proceedings of ACM International Conference on Multimedia, pp. 1948–1957 (2020)
67. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
68. Son, J., Baek, M., Cho, M., Han, B.: Multi-object tracking with quadruplet convolutional neural networks. In: CVPR, pp. 5620–5629 (2017)
69. Sun, P., et al.: Transtrack: multiple-object tracking with transformer. arXiv:2012.15460 (2020)

70. 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)
71. Tan, J., et al.: Equalization loss for long-tailed object recognition. In: CVPR, pp. 11662–11671 (2020)
72. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3132–3141 (2021)
73. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 402–419. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_24
74. Vu, T., Jang, H., Pham, T.X., Yoo, C.D.: Cascade RPN: delving into high-quality region proposal network with adaptive convolution. arXiv:1909.06720 (2019)
75. Wang, J., Wang, X., Shang-Guan, Y., Gupta, A.: Wanderlust: online continual object detection in the real world. In: ICCV, pp. 10829–10838 (2021)
76. Wang, J., et al.: Seesaw loss for long-tailed instance segmentation. In: CVPR, pp. 9695–9704 (2021)
77. Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., Li, H.: Unsupervised deep tracking. In: CVPR, pp. 1308–1317 (2019)
78. Wang, T., et al.: The devil is in classification: a simple framework for long-tail instance segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 728–744. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_43
79. Wang, T., Zhu, Y., Zhao, C., Zeng, W., Wang, J., Tang, M.: Adaptive class suppression loss for long-tail object detection. In: CVPR, pp. 3103–3112 (2021)
80. Wang, W., Feiszli, M., Wang, H., Tran, D.: Unidentified video objects: a benchmark for dense, open-world segmentation. arXiv:2104.04691 (2021)
81. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: CVPR, pp. 2566–2576 (2019)
82. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv:2003.06957 (2020)
83. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 107–122. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_7
84. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP, pp. 3645–3649. IEEE (2017)
85. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: an online multi-object tracker. In: CVPR, pp. 12352–12361 (2021)
86. Wu, T., Huang, Q., Liu, Z., Wang, Yu., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 162–178. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_10
87. Wu, Y., et al.: Large scale incremental learning. In: CVPR, pp. 374–382 (2019)
88. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 472–487. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_29
89. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. In: CVPR, pp. 10687–10698 (2020)

90. Xu, J., Wang, X.: Rethinking self-supervised correspondence learning: a video frame-level similarity perspective. arXiv:2103.17263 (2021)
91. Xu, M., et al.: End-to-end semi-supervised object detection with soft teacher. In: ICCV, pp. 3060–3069 (2021)
92. Xu, M., et al.: Bootstrap your object detector via mixed training 34 (2021)
93. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV, pp. 5188–5197 (2019)
94. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: regularization strategy to train strong classifiers with localizable features. In: ICCV, pp. 6023–6032 (2019)
95. Zang, Y., Huang, C., Loy, C.C.: FASA: feature augmentation and sampling adaptation for long-tailed instance segmentation. arXiv:2102.12867 (2021)
96. Zeng, F., Dong, B., Wang, T., Zhang, X., Wei, Y.: MOTR: end-to-end multiple-object tracking with transformer. arXiv:2105.03247 (2021)
97. Zhang, C., et al.: Mosaicos: A simple and effective use of object-centric images for long-tailed object detection. arXiv:2102.08884 (2021)
98. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: a unified framework for long-tail visual recognition. In: CVPR, pp. 2361–2370 (2021)
99. Zhang, Y., et al.: ByteTrack: multi-object tracking by associating every detection box. arXiv:2110.06864 (2021)
100. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: FairMOT: on the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis., 1–19 (2021)
101. Zhang, Z., Cheng, D., Zhu, X., Lin, S., Dai, J.: Integrated object detection and tracking with tracklet-conditioned detection. arXiv:1811.11167 (2018)
102. Zheng, J., Ma, C., Peng, H., Yang, X.: Learning to track objects from unlabeled videos. In: ICCV, pp. 13546–13555 (2021)
103. Zhou, W., Chang, S., Sosa, N., Hamann, H., Cox, D.: Lifelong object detection. arXiv:2009.01129 (2020)
104. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. arXiv preprint arXiv:2201.02605 (2022)
105. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 474–490. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_28
106. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. arXiv:2103.07461 (2021)
107. Zhou, X., Yin, T., Koltun, V., Krähenbühl, P.: Global tracking transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8771–8780 (2022)