Tracking Every Thing in the Wild

Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E. Huang, Fisher Yu

Computer Vision Lab, ETH Zürich
http://vis.xyz/pub/tet

Abstract. Current multi-category Multiple Object Tracking (MOT) metrics use class labels to group tracking results for per-class evaluation. Similarly, MOT methods typically only associate objects with the same class predictions. These two prevalent strategies in MOT implicitly assume that the classification performance is near-perfect. However, this is far from the case in recent large-scale MOT datasets, which contain large numbers of classes with many rare or semantically similar categories. Therefore, the resulting inaccurate classification leads to sub-optimal tracking and inadequate benchmarking of trackers. We address these issues by disentangling classification from tracking. We introduce a new metric, Track Every Thing Accuracy (TETA), breaking tracking measurement into three sub-factors: localization, association, and classification, allowing comprehensive benchmarking of tracking performance even under inaccurate classification. TETA also deals with the challenging incomplete annotation problem in large-scale tracking datasets. We further introduce a Track Every Thing tracker (TETer), that performs association using Class Exemplar Matching (CEM). Our experiments show that TETA evaluates trackers more comprehensively, and TETer achieves significant improvements on the challenging large-scale datasets BDD100K and TAO compared to the state-of-the-art.

Keywords: Large-scale Long-tailed MOT, Class Exemplar Matching, TETA Metric

1 Introduction

Multiple Object Tracking (MOT) aims to estimate the trajectory of objects in a video sequence. While common MOT benchmarks [15,30,10] only consider tracking objects from very few pre-defined categories, *e.g.*, pedestrian and car, the number of categories of interest in the real world is overwhelming. Although the recent extension of MOT to a large number of categories [47,8] may seem trivial, it raises profound questions about the definition and formulation of the problem itself, which are yet to be addressed by the community.

In Fig. 1, we show tracking results from two different trackers on the same video sequence. Tracker A tracks the object perfectly, but with a slightly incorrect classification on a fine-grained level. Tracker B classifies the object perfectly but does not track the object at all. *Which one is the better tracker?* The mMOTA [3] metric gives a 0 score for tracker A and a score of 33 for tracker

Fig. 1: Tracking results from two different trackers (A and B). The same color means the same track. Tracker A gets 0 score in terms of the MOTA [3], IDF1 [40], and HOTA [28] metrics, while the tracker B gets 33 for first two and 44 for HOTA

B. The above example raises an interesting question: ***Is tracking still meaningful if the class prediction is wrong?*** In many cases, the trajectories of wrongly classified or even unknown objects are still valuable. For instance, an autonomous vehicle may occasionally track a van as a bus, but the estimated trajectory can equally well be used for path planning and collision avoidance.

Current MOT models and metrics [2,48,3,40,41,28] are mainly designed for single-category multiple object tracking. When extending MOT to the large-scale multi-category scenarios, they simply adopt the same single-category metrics and models by treating each class independently. The models first detect and classify each object, and then the association is only done between objects of the same class. Similarly, the metrics use class labels to group tracking results and evaluate each class separately. This implicitly assumes that the classification is good enough since it is the prerequisite for conducting association and evaluating tracking performance.

The aforementioned near-perfect classification accuracy is mostly valid on benchmarks consisting of only a handful of common categories, such as humans and cars. However, it does not hold when MOT extends to a large number of categories with many rare or semantically similar classes. The classification itself becomes a very challenging task on imbalanced large-scale datasets such as LVIS [16]. Also, it is difficult to distinguish similar fine-grained classes because of the naturally existing class hierarchy, *e.g.*, the bus and van in Fig. 1. Besides, many objects do not belong to any predefined category in real-world settings. Thus, treating every class independently without accounting for the inaccuracy in classification leads to inadequate benchmarking and non-desired tracking behavior. To expand tracking to a more general scenario, we propose that classification should be disentangled from tracking, in both evaluation and model design, for multi-category MOT. To achieve this, we design a new metric, Track Every Thing Accuracy (TETA), and a new model, Track Every Thing tracker (TETer).

The proposed TETA metric disentangles classification performance from tracking. Instead of using the predicted class labels to group per-class tracking results, we use location with the help of local cluster evaluation. We treat each ground truth bounding box of the target class as the anchor of each cluster and group prediction results inside each cluster to evaluate the localization and association performance. Our local clusters enable us to evaluate tracks even when the class prediction is wrong. Furthermore, the local cluster evaluation makes

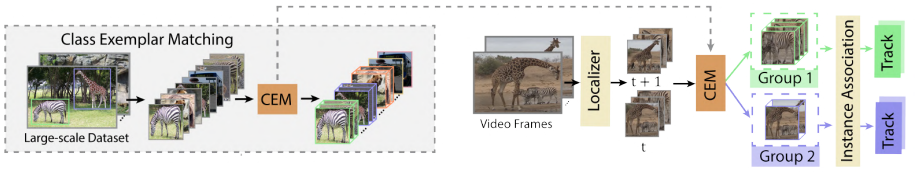


Fig. 2: CEM can be trained with large-scale datasets and directly employed for tracking

TETA competent to deal with incomplete annotations, which are common in datasets with a large number of classes, such as TAO [8].

Our TETer follows an *Associate-Every-Thing (AET)* strategy. Instead of associating objects in the same class, we associate every object in neighboring frames. The AET strategy frees association from the challenging classification/detection issue under large-scale long-tailed settings. However, despite wholly disregarding the class information during association, we propose a new way of leveraging it, which is robust to classification errors. We introduce Class Exemplar Matching (CEM), where the learned class exemplars incorporate valuable class information in a soft manner. In this way, we effectively exploit semantic supervision on large-scale detection datasets while not relying on the often incorrect classification output. CEM can be seamlessly incorporated into existing MOT methods and consistently improve performance. Moreover, our tracking strategy enables us to correct the per-frame class predictions using rich temporal information.

We analyze our methods on the newly introduced large-scale multi-category tracking datasets, TAO [8] and BDD100K [47]. Our comprehensive analysis show that our metric evaluate trackers more comprehensively and achieve better cross dataset consistency despite incomplete annotations. Moreover, our tracker achieves state-of-the-art performance on TAO and BDD100K, both when using previously established metrics and the proposed TETA.

2 Related Work

Multi-Object Tracking (MOT) aims to track multiple objects in video sequences. Earlier methods follow a track-first paradigm, which do not rely on classification during tracking [35,36,1]. Some utilize LiDAR data with model-free detection [18,33,11] or point cloud segmentation [44,43]. Others [35,32,34] first segment the scene [12], which enables tracking of generic objects. Recently, the most common paradigm for MOT is tracking-by-detection, focusing on learning better appearance features to strengthen association [21,31,46,27,29,24], modeling the displacement of each tracked object [2,48,38], or using a graph-based approach [42,5]. Previous MOT approaches mostly focus on benchmarks with a few common categories, while recent works [25,9] study the MOT in open-set settings where the goal is to track and segment any objects regardless of their categories. Those methods use a class agnostic trained detector or RPN network to generate object proposals, while classification is essential in many applications,

e.g., video analysis. The close-set settings with large-scale, long-tailed datasets are severely under-explored. We study the MOT in such a scenario, identifying issues and proposing solutions in both model design and evaluation metric.

MOT Metrics often evaluate both detection and association performance. Multi-Object Tracking Accuracy (MOTA) [3] was first introduced to unify the two measures. MOTA performs matching on the detection level and measures association performance by counting the number of identity switches. IDF1 [40] and Track-mAP instead performs matching on the trajectory level. Recently, Higher-Order Tracking Accuracy (HOTA) [28] was proposed to fairly balance both components by computing a separate score for each. Liu et al. [25] proposes a recall-based evaluation to extend MOT into open-world settings. All above metrics do not independently access the classification performance, making them unsuitable for large-scale multi-category MOT. TETA extends HOTA by further breaking down detection into localization and classification, enabling TETA to evaluate association despite classification failures. Furthermore, current metrics have issues when evaluating trackers on non-exhaustively annotated datasets such as TAO, which TETA can handle.

3 Tracking-Every-Thing Metric

Here we introduce the Track Every Thing Accuracy (TETA) metric. We first discuss how classification is handled in current metrics and the incomplete annotation problem in section 3.1. Then, we formulate TETA in section 3.2 to address the existing issues.

3.1 Limitations for Large-scale MOT Evaluation

How to handle classification. How to evaluate classification in MOT is an important but under-explored problem. MOT metrics such as MOTA [3], IDF1 [40], and HOTA [28] are designed for the single category MOT. When extending to multiple classes, they require trackers to predict a class label for each object, then they group tracking results based on the labels and evaluate each class separately. However, the wrong classification happens frequently in long-tailed scenarios, which leads to failures in the grouping based on class labels, and the tracking performance will not be evaluated even if the tracker localizes and tracks the object perfectly as shown in Fig. 1.

One simple solution is to ignore classification and evaluate every object class-agnostically. However, large vocabulary datasets often follow long-tailed distributions where few classes dominate the dataset. Ignoring class information leads to the evaluation being dominated by those classes, resulting in trackers' performance in tracking rare classes being negligible. The class-aware HOTA proposes to use a geometric mean between with classification confidence and HOTA, which requires the trackers to output the class probability distribution, while most only output the final categories. Moreover, it still cannot access classification independently.

Fig. 3: Left: TAO ground truth sample. TAO is partially annotated. **Right**: Corresponding prediction from the best tracker AOA [13] ranked by TAO metric. AOA generates many low confidence bounding boxes, making it difficult to use in practice

Incomplete Annotations. MOT metrics such as MOTA [3], IDF1 [40], and HOTA [28] are designed for datasets with exhaustive annotations of every object. However, it is prohibitively expensive to annotate every object when constructing large-scale datasets with many categories. The TAO dataset contains over 800 categories, but most of them are not exhaustively annotated (see Fig. 3). Incomplete annotations pose a new challenge: *how can we identify and penalize false positive (FP) predictions?* MOTA, IDF1, and HOTA metrics treat every unmatched prediction as FP, but this falsely penalizes correct predictions with no corresponding annotations. On the other hand, TAO metric [8] adopts the same federated evaluation strategy as the LVIS [16] dataset and does not penalize categories if there is no ground-truth information about their presence or absence. This strategy inadvertently rewards a large number of false positives. In Fig. 3, we visualize predictions from the best tracker on TAO. Since TAO does not punish most false positives, trackers are incentivized to generate many low confidence tracks to increase the chances that objects from rare categories get tracked, making their results difficult to be used in practice. Also, this makes TAO a game-able metric. We show a simple copying and pasting trick that can drastically improve the TAO metric's score in section 5.1. A similar issue is also observed in the LVIS mAP metric [7].

3.2 Tracking-Every-Thing Accuracy (TETA)

TETA builds upon the HOTA [28] metric, while extending it to better deal with multiple categories and incomplete annotations. TETA consists of three parts: a localization score, an association score, and a classification score, which enable us to evaluate the different aspects of each tracker properly.

Local cluster evaluation. We design the local clusters to deal with incomplete annotations and disentangle classification from large-scale tracking evaluation. The main challenge for evaluation with incomplete annotations is determining false positives. We propose local cluster evaluation to strike a balance between false-penalizing or the non-penalizing phenomenon as discussed in 3.1. We have observed that even though we do not have exhaustive annotations, we can still identify a specific type of false positives with high confidence. Unlike previous metrics, we only consider predictions within local clusters. We view each ground truth bounding box as an anchor point of a cluster and assign each prediction to the closest anchor points within an IoU margin of r. The predictions inside

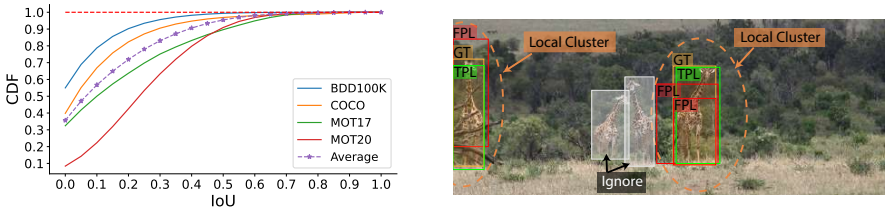


Fig. 4: Left: Inter-object overlap in real datasets. We compute the cumulative probability of ground truth bounding boxes that have different level of IoU overlaps in four different datasets with exhaustive annotations along with their average. Extreme inter-object overlap is very rare in real data. **Right**: Local cluster evaluation. TPL, FPL, and GT are the true positive localization, false positive localization, and ground truth, respectively. We create a cluster for each ground truth bounding box based on the IoU similarities. For evaluation, we only consider predictions inside each cluster. The predictions that do not belong to any cluster will be ignored

the clusters not chosen as matched true positives are considered false positives. Fig. 4 shows the inter-object overlap in popular object detection and tracking datasets, which indicates extreme inter-object overlap is rare in the real world. If we set the r to 0.7 or higher, there is less than 1% chances that we make mistakes even in highly crowded dataset like MOT20.

To avoid false punishments, we ignore the predictions that are not assigned to any clusters during evaluation. This process is illustrated in Fig. 4. The margin r of the clusters can be set according to different scenarios. The bigger the r is, the more conservative the metric is regarding choosing false positives. Also, it means fewer false punishments. If the dataset is super crowded and lacks annotation, we can select a higher r to avoid false punishment. The local cluster design also allows us to disentangle classification. For evaluation of a particular class, we evaluate predictions that are assigned to clusters with ground truth bounding boxes of that class. Since the per class result grouping is done using location instead of classification. Thus, within each local cluster, we are able to evaluate the tracking performance even if the class predictions are wrong.

Localization Score. The localization score measures the localization performance of a tracker. A true positive candidate $b \in \text{TPL}$ is a prediction box (pBox) that has an IoU higher than a localization threshold α with a ground truth box (gBox). We use the Hungarian algorithm [20] to choose the final matched TPL that optimizes both localization and association scores. The chosen assignment represents the best-localized tracks. A false negative localization (FNL) is a gBox that is not matched to any other pBox. A false positive localization (FPL) is defined based on each cluster. If a pBox is in a cluster but is not matched to any ground truth, it is a false positive. The localization score is computed using the Jaccard Index,

$$\text{LocA} = \frac{|\text{TPL}|}{|\text{TPL}| + |\text{FPL}| + |\text{FNL}|}. \tag{1}$$

Association Score. Our association score follows the definition of HOTA but redefines true positive associations (TPA), false negative associations (FNA),

and false positive associations (FPA) to be based on each b. The association score of b is

$$\text{AssocA}(b) = \frac{|\text{TPA}(b)|}{|\text{TPA}(b)| + |\text{FPA}(b)| + |\text{FNA}(b)|}. \tag{2}$$

The final association score is the average over all TPLs,

$$\text{AssocA} = \frac{1}{|\text{TPL}|} \sum_{b \in \text{TPL}} \text{AssocA}(b). \tag{3}$$

Classification Score. The classification score reflects the pure performance of the classifier in each tracker. Unlike all other tracking metrics where classification performance is entangled with tracking performance, our metric makes it an independent score. We only consider the well-matched TPL, where α is at least 0.5. The classification score is defined for each class. The true positive classification (TPC) for certain class c is defined as

$$\text{TPC}(c) = \{b | b \in \text{TPL} \ \wedge \ \text{pc}(b) = \text{gc}(b) = c\}, \tag{4}$$

where $\text{pc}(b)$ is the class ID of b and $\text{gc}(b)$ is the class ID of the ground truth that is matched to b. This set includes all TPL that have the same predicted class as the corresponding ground truth. The false negative classification for class c is defined as

$$\text{FNC}(c) = \{b | b \in \text{TPL} \ \wedge \ \text{pc}(b) \neq c \wedge \text{gc}(b) = c\}, \tag{5}$$

which includes all TPL that have incorrect class predictions with ground truth class c. The false positive classification for class c is defined as

$$\text{FPC}(c) = \{b | b \in \text{TPL} \ \wedge \ \text{pc}(b) = c \wedge \text{gc}(b) \neq c\} \tag{6}$$

which includes all TPL with class c but is matched to an incorrect ground truth class. If the dataset is fully annotated, the $b \in \text{P}$, which includes TPL and the predictions outside clusters. Full annotations indicate that the predictions that are far away from gBox wrongly classify background or other classes as c. The final classification score is

$$\text{ClsA} = \frac{|\text{TPC}|}{|\text{TPC}| + |\text{FPC}| + |\text{FNC}|}. \tag{7}$$

Combined Score. HOTA uses geometric mean to balance detection and association. However, the geometric mean becomes zero if any term is zero. If the classification performance of a tracker is close to zero, *e.g.* due to a long-tail class distribution, it will completely dominate the final tracking metric if computed as a geometric mean. Therefore, we use an arithmetic mean to compute the final score:

$$\text{TETA} = \frac{\text{LocA} + \text{AssocA} + \text{ClsA}}{3}. \tag{8}$$

Besides, since different applications focus on different aspects, we encourage users to look at each subfactor based on the needs instead of focusing on a single score.

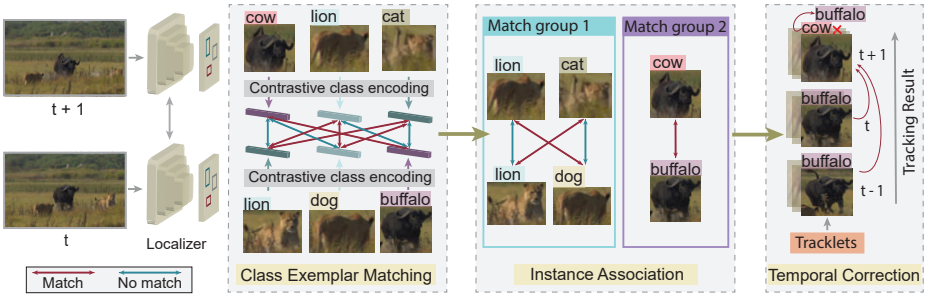


Fig. 5: Association protocol of TETer. For every pair of frames, we first compute and match the class exemplars for each localized object to determine potential matching candidates. Then, we perform instance association to determine the final matches. Finally, we use rich temporal information to correct classification errors in each frame

4 Tracking-Every-Thing Tracker

We here introduce our Tracking Every Thing tracker (TETer). TETer aims to discover every foreground object, associate, and classify them across time. The full pipeline is shown in Fig. 5.

4.1 Class-Agnostic Localization (CAL)

To track every thing, we first need to localize them. However, object detectors struggle on large-scale, long-tailed datasets, especially for rare categories. Interestingly, when decoupling localization and classification in common object detectors, we find that the detector can still localize rare or even novel objects well. Fig. 6 shows a comparison of the object detector's performance with and without considering classification on the TAO validation set. When we do not consider class predictions during evaluation, the performance of the detector is stable across rare, common, and frequent classes. This strongly suggests that the bottleneck in detection performance lies in the classifier. With this in mind, we replace the commonly used intra-class non-maximum suppression (NMS) using class confidence with a class-agnostic counterpart to better localize every object in the wild.

4.2 Associating Every Thing

Association is often done by considering a single or a combination of cues, *e.g.*, location, appearance, and class. Learning motion priors under large-scale, long-tailed settings is challenging since the motion patterns are irregular among classes. Moreover, there are also many objects in the real world that are not in any predefined categories. In contrast, objects in different categories usually have very different appearances. Thus, we adopt appearance similarity as our primary cue.

We propose an alternative approach for utilizing class information as feature clues during association. Instead of trusting the class predictions from object detectors and using them as hard prior, we learn class exemplars by directly

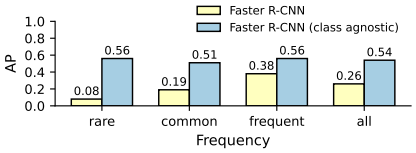


Fig. 6: Faster R-CNN's performance w/ and w/o considering classification

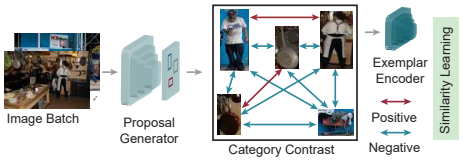


Fig. 7: Training pipeline of CEM

contrasting samples from different categories. This enables us to compute similarity scores for pairs of objects at the category-level, which can better deal with semantically similar classes compared to discrete class labels. During association, we use the class exemplars to determine potential matching candidates for each object. This process can be thought of as using class information as soft prior. Therefore, it can integrate fine-grained cues required for classification (*e.g.* the difference between a big red bus and a red truck), which are difficult to learn for the purely class-agnostic appearance embedding.

Class Exemplar Matching (CEM). The training pipeline based on a two-stage detector is shown in Fig. 7. The Region Proposal Network (RPN) computes all the Region of Interest (RoI) proposals from the input images. Then, we use RoI align to extract the feature maps from the multi-scale feature outputs. The feature maps are used as input for the exemplar encoder to learn category similarity. The exemplar encoder generates class exemplars for each RoI. We assign category labels to each RoI with a localization threshold α. If a RoI has an IoU higher than α (in our case, $\alpha = 0.7$) with a ground truth box, then we assign the corresponding category label to the RoI. Positive samples are RoIs from the same category, and negative samples are those from different categories.

We adapt the SupCon loss [19] and propose an unbiased supervised contrastive loss (U-SupCon):

$$\mathcal{L}_{\mathcal{C}} = -\sum_{q \in Q} \frac{1}{|Q^+(q)|} \sum_{q^+ \in Q^+(q)} \log \frac{\exp(\text{sim}(q, q^+)/\tau)}{\text{PosD}(q) + \sum_{q^- \in Q^-} \exp(\text{sim}(q, q^-)/\tau)}, \quad (9)$$

$$\text{PosD}(q) = \frac{1}{|Q^+(q))|} \sum_{q^+ \in Q^+} \exp(\text{sim}(q, q^+)/\tau), \quad (10)$$

where Q is the set of class generated from a random sampled image batch, $Q^+(q)$ is the set of all positive samples to q, $Q^-(q)$ is the set of all negative samples to q, $\text{sim}(\cdot)$ denotes cosine similarity, and τ is a temperature parameter. We set τ to 0.07. We add the PosD(q) to prevent the varying lower bound of the SupCon loss when training with detection proposals, where the number of positive samples is consistently changing.

Association Strategy. For a query object q in a tracklet, we find a group of candidates by comparing their class exemplars. Specifically, assume we have the class exemplar q_c for the query object q in frame t, and a set of detected objects D in frame $t+1$ and their class exemplars $d_c \in D_c$. We compute the similarities between q_c and D_c and select candidates with high similarity. This gives us a candidate list $C = \{d \mid \text{sim}(q_c, d_c) > \delta,\ d \in D\}$. δ is set to 0.5.

To determine the final match from the candidate list, any existing association method can be used. Thus, CEM can be readily used to replace existing hard prior-based matching. For our final model TETer, we further utilize quasi-dense similarity learning [37] to learn instance features for instance-level association. We compute instance-level matching scores with each candidate from C using bidirectional softmax and cosine similarity. We take the candidate with the maximum score and if the score is larger than β, then it is a successful match. We set β to 0.5.

Temporal Class Correction (TCC). The AET strategy allows us to correct the classification using rich temporal information. If we track an object, we assume the class label to be consistent across the entire track. We use a simple majority vote to correct the per-frame class predictions.

5 Experiments

We conduct analyses of different evaluation metrics and investigate the effectiveness of our new tracking method on TAO [8] and BDD100K [47]. TAO provides videos and tracking labels for both common and rare objects with over 800 object classes. Although BDD100K for driving scenes has fewer labeled categories, some, like trains, are much less frequent than common objects such as cars. In this section, we first compare different metrics with TETA. Then we evaluate the proposed TETer on different datasets and plug CEM into existing tracking methods to demonstrate its generalizability.

Implementation Details. For the object detector, we use Faster R-CNN [39] with Feature Pyramid Network (FPN) [23]. We use ResNet-101 as backbone on TAO, same as TAO baseline [8], and use ResNet-50 as backone on BDD100K, same as the QDTrack [37]. On TAO, we train our model on a combination of LVISv0.5 [16] and the COCO dataset using repeat factor sampling. The repeat factor is set to 0.001. We use the SGD optimizer with a learning rate of 0.02 and adopt the step policy for learning rate decay, momentum, and weight decay are set to 0.9 and 0.0001. We train our model for 24 epoch in total and learning rate is decreased at 16 and 22 epoch. For TETer with SwinT backbone [26], we use 3x schedule used by mmdetection [6]. For TETer-HTC, we use the HTC-X101-MS-DCN detector from [22]. On BDD100K, we load the same object detector weights from QDTrack [37] and fine-tune the exemplar encoder with on BDD100K Detection dataset with other weights frozen. For each image, we sample a maximum of 256 object proposals. For more details, please refer to the supplemental materials.

5.1 Metric Analysis

Cross Dataset Consistency. A good metric should correlate with real world tracking performance. Although we face difficulties in incomplete annotations, this principle for metric design should not change. For instance, a tracker designed for tracking objects belonging to hundreds of categories on TAO should also work well on new video sequences that contain a subset of those categories. We evaluate this by using the BDD100K dataset, which has seven out of eight

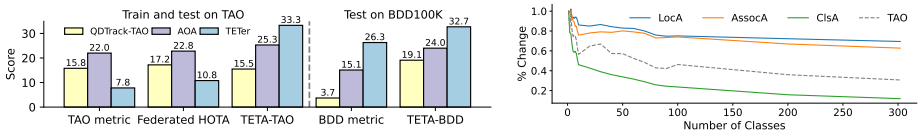


Fig. 8: Left: We pre-train models on TAO (incomplete annotations) and directly test them on BDD (complete annotations) with the default BDD metric (IDF1). We omit MOTA as its value range is $(-\infty, 1]$, which is inconsistent with other metrics. **Right**: Percentage change in score of each metric as number of evaluation classes increases

of its categories overlapped with TAO. We treat BDD100K as the new video sequences in the real world to test two trackers: QDTrack-TAO [37], which is optimized for the TAO metric, and our tracker, which is optimized for TETA. We only evaluate on the overlapped categories, which also contains exhaustive annotations for every object.

As shown in Fig. 8 (Left), the tracker selected by the TAO metric overly optimizes for the incomplete TAO dataset setting, which does not generalize well to BDD100K. In comparison, the tracker selected by TETA generalizes well to BDD100K. Our metric gives the same ranking on the complete annotations setting with the default BDD100K IDF1 metric despite facing the difficulties of ranking trackers under incomplete annotations.

Comprehensively Analyze Trackers. Correctly understanding different aspects of trackers is crucial for designing trackers for various scenarios. For example, it is important for an autonomous vehicle to detect and understand the trajectory of every object to avoid collision, but slightly wrong classification may not be as critical. In this experiment, we evaluate the effect of the number of classes on metric scores on the TAO validation set. We use the same tracking predictions but merge the class predictions by sorting the classes in descending order based on the number of instances and combining the last n classes. For example for $n = 2$, we merge all classes besides humans (the most frequent class) into a single class and only evaluate on two classes. We sample several n between 1 (single class) and 302 (all classes) and evaluate on each set of classes.

The result is shown in Fig. 8 (Right). Although the trajectory predictions are the same, the score produced by the TAO metric drops significantly as the number of classes increases. As the TAO metric entangles classification and tracking into a single metric, this makes it hard to determine which part of the tracker went wrong. On the other hand, with TETA, we can separately investigate different aspects. While the classification performance follows the same trend as the TAO metric, the localization and association scores are still stable. This allows us to instantly understand that the degradation is due to classification.

Cheating TAO track mAP metric. Fig. 9 shows a copy & paste trick that can boost the TAO track mAP. We simply copy and paste existing trajectories with low confidence class predictions from the object detector without additional training. As shown in Table 1, TAO track mAP and Federated HOTA metric increase drastically from 0 to 62.9 and 4.2 to 68.7. In comparison, TETA drops

(a) Original (b) After copy & paste

TAO track mAP	AP↑	AP_{50} ↑	AP_{75} ↑	AR ↑
Before copy	0	0	0	0
After copy	62.9	75.2	50.5	62.9
Federated HOTA	HOTA ↑	DetA ↑	AssA ↑	-
Before copy	4.2	3.0	5.9	-
After copy	68.7	75.7	62.8	-
TETA (ours)	TETA ↑	LocA ↑	AssocA ↑	ClsA ↑
Before copy	47.6	80.1	59.6	3.2
After copy	13.8	3.3	10.2	27.9

Fig. 9 & Table 1: Copy & paste strategy to cheat the TAO track mAP. (a) tracking result from our tracker, which incorrectly classifies the deer as a goat. (b) copying and pasting existing tracks with low confidence class predictions from the object detector. The table shows the comparison between TAO track mAP and TETA with a simple copy & paste trick for the sequence in Fig. 9

Table 2: Results on TAO

Method	TETA	LocA	AssocA	ClsA
SORT [4]	24.845	48.13	14.32	12.08
Tracktor [2]	24.15	47.41	12.96	12.08
DeepSORT [45]	25.98	48.35	17.52	12.09
AOA [14]	25.27	23.40	30.56	**21.86**
Tracktor++ [2]	27.97	49.04	22.81	12.05
QDTrack [37]	30.00	50.53	27.36	12.11
TETer	**33.25**	**51.58**	**35.02**	13.16
QDTrack-SwinT	31.22	51.32	27.27	15.06
TETer-SwinT	**34.61**	**52.10**	**36.71**	15.03
QDTrack-HTC	32.79	56.21	27.43	14.73
TETer-HTC	**36.85**	**57.53**	**37.53**	15.70

Table 3: Results on BDD100K

Method	Split	mMOTA	mIDF1	TETA	LocA	AssocA	ClsA
DeepSORT [45]	val	35.2	49.3	48.03	46.36	46.69	51.04
QDTrack [37]	val	36.6	51.6	47.84	45.86	48.47	49.20
TETer	val	**39.1**	**53.3**	**50.83**	**47.16**	**52.89**	**52.44**
DeepSORT [45]	test	34.0	50.2	46.75	45.26	47.04	47.93
QDTrack [37]	test	35.7	52.3	49.17	**47.19**	50.93	49.38
TETer	test	**37.4**	**53.3**	**50.42**	46.99	**53.56**	**50.71**

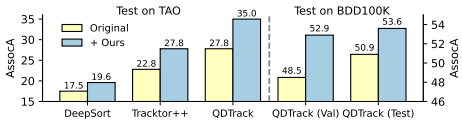
from 47.6 to 13.8, which suggests the trick does not work on TETA. Moreover, we can clearly see consequences brought by copy & paste. The localization score drops sharply as copy & paste generates a lot of false positive localizations. On the other hand, the trick only improves the classification performance.

5.2 TAO Tracking Results

We provide a thorough comparison of TETer against competing methods on the TAO validation set in Table 2 using our TETA metric. We set the margin r of local clusters to 0.5 since we observe the TAO dataset is not crowded, and this choice gives a proper balance between non-penalizing and over-penalizing FPs. We also include results of other margins in the supplemental material. For this experiment, we only use the predefined 302 categories without considering the unknown classes. We allow each tracker maximum outputs 50 predictions per image. We use the same FasterRCNN detectors with class-agnostic NMS for all methods except AOA [13]. Despite the increased difficulty introduced by the large number of categories, TETer outperforms all other methods, providing consistent improvements in TETA, LocA, and AssocA. In particular, TETer improves TETA of QDTrack [37] by over 3 points and AssocA by over 7 points.

We also compare our method to AOA [14], the winner of the ECCV 2020 TAO challenge, using the publicly available predictions[1]. AOA combines multiple state-of-the-art few-shot detection and object ReID models that are trained using additional external datasets, which enables it to obtain very strong classification performance. However, as it is optimized using the TAO metric, it makes

[1] https://github.com/feiaxyt/Winner_ECCV20_TAO



Components	Known				Unknown		
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA
QDTrack [37]	28.89	46.60	27.76	12.37	19.89	32.25	27.42
+ CAL	30.00	50.53	27.36	12.11	20.85	36.06	26.50
+ CEM	32.86	51.55	35.00	12.06	24.00	37.38	34.61
+ TCC	**33.25**	**51.55**	**35.00**	**13.16**	**24.00**	**37.38**	**34.61**

Fig. 10 & Table 4: Fig. 10 shows comparison of DeepSORT [17], Tracktor++ [2], and QDTrack [37] with w/ and w/o our CEM module on TAO and BDD100K datasets. CEM consistently improves association performance of all methods. Table 4 shows comparison with different components of TETer on the TAO open set using our TETA metrics

excessive false positive predictions, which are punished by TETA. Additionally, TETer achieves better association performance without using external datasets.

5.3 BDD100K Tracking Results

We provide evaluation results on both the BDD100K validation and test sets in Table 3. We first evaluate each tracker using the established CLEAR MOT metrics, including MOTA, MOTP, and IDF1, each averaged over every class. Without bells and whistles, TETer can obtain performance gains across all three metrics compared to QDTrack [37] on both sets, achieving state-of-the-art performances. In particular, TETer improves the mMOTA and mIDF1 of QDTrack by 2.5 and 1.7 points on the validation set and 1.7 and 1 points on the test set. We also show evaluation results using TETA. TETer again obtains consistent performance gains across all metrics. On both validation and test sets, TETer can improve AssocA of QDTrack by over 2.5 points.

5.4 Generalizability of CEM

To demonstrate the generalizability of our CEM module, we further apply CEM to other MOT methods to replace existing hard prior-based matching. We compare three methods, DeepSORT [45], Tracktor++ [2], and QDTrack [37], with and without CEM across both TAO and BDD100K. The results are shown in Fig. 10. On the TAO validation set, adding CEM results in at least 2 points of improvement in AssocA across all methods. In particular, CEM can improve AssocA of QDTrack by 7 points. On both BDD100K validation and test sets, CEM can obtain over 2.5 points of improvement in AssocA of QDTrack. This shows our CEM module can be applied to various popular MOT methods and achieve consistent improvements in association by better exploiting class information.

5.5 Ablation Study

We conduct ablation studies on TAO and BDD100K. We investigate the importance of our proposed modules on both predefined and unknown categories. For TAO, we use their split for known and unknown (free-form) classes. For unknown split, we only report the LocA and AssocA.

Tracking Components. We evaluate the contributions of each component of TETer on the TAO open set using TETA in Table 4. When we replace the class-dependent intra-class NMS from the object detector with a class-agnostic

Table 5: Comparing different ways of using class information. AET Baseline associates every objects without using any class information. Softmax indicates association happens only within the same class. BERT indicates using BERT word embeddings to group candidates for association

	BDD100K val			TAO val
Class	mMOTA	mIDF1	AssocA	AssocA
AET Baseline	37.3	52.6	52.0	33.5
Softmax	36.6 (-0.7)	51.6 (-0.2)	48.9 (-0.2)	27.4 (-6.1)
BERT	37.2 (-0.1)	52.6	52.0	27.8 (-5.7)
CEM	**39.1** (+1.8)	**53.3** (+0.7)	**52.9** (+0.9)	**35.0** (+1.5)

NMS, we can improve LocA by over 3 points on both known objects and unknown objects. Adding CEM drastically improves its AssocA by over 7 points on known objects and 8 points on unknown objects. Further, using temporal class correction can improve ClsA by over 1 point.

Comparison of using class information. We compare different ways of utilizing class information during association on the validation set of BDD100K and TAO in Table 5. The baseline protocol follows the AET strategy and performs class-agnostic association with pure instance appearance features described in section 4.2. We then add different class prior on top of the AET baseline to study their effectiveness. Softmax use class labels as hard prior and associate objects within the same class. This strategy leads to a severe downgrade in the tracking performance, especially for the TAO dataset.

Alternatively, we use the out-of-the-shelf word embeddings to incorporate class information. The semantically similar classes should be closer in the word embedding space. This way transfers the hard class labels to soft ones. We utilize the BERT model to embed the class names to replace our CEM. While the performance is slightly better than using softmax predictions, it is inferior to the CEM. Our CEM is the only method capable of effectively utilizing semantic information to improve the association by outperforming the AET baseline on large-scale long-tailed datasets.

6 Conclusion

We present a new metric TETA and a new model TETer for tracking every thing in the wild. TETA and TETer disentangle classification from evaluation and model design for the long-tailed MOT. TETA can evaluate trackers more comprehensively and better deal with incomplete annotation issues in large-scale tracking datasets. TETer disentangles the unreliable classifier from both detection and association, resulting in a better tracker which outperforms existing state-of-the-art trackers on large-scale MOT datasets, TAO and BDD100K. The core component of TETer, CEM, can be used as a drop-in module for existing tracking methods and boost their performance.

7 Acknowledgement

Special thanks go to Ruolan Xiang for her help in editing the paper.

References

1. Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., Leibe, B.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: European Conference on Computer Vision. pp. 158–177. Springer (2020) 3
2. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 941–951 (2019) 2, 3, 12, 13
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008) 1, 2, 4, 5
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016) 12
5. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6247–6257 (2020) 3
6. 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) 10
7. Dave, A., Dollár, P., Ramanan, D., Kirillov, A., Girshick, R.: Evaluating large-vocabulary object detectors: The devil is in the details. arXiv preprint arXiv:2102.01066 (2021) 5
8. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: European conference on computer vision. pp. 436–454. Springer (2020) 1, 3, 5, 10
9. Dave, A., Tokmakov, P., Ramanan, D.: Towards segmenting anything that moves. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 3
10. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020) 1
11. Dewan, A., Caselitz, T., Tipaldi, G.D., Burgard, W.: Motion-based detection and tracking in 3d lidar scans. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 4508–4513 (2016). https://doi.org/10.1109/ICRA.2016.7487649 3
12. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2393–2402 (2018) 3
13. Du, F., Xu, B., Tang, J., Zhang, Y., Wang, F., Li, H.: 1st place solution to eccv-tao-2020: Detect and represent any object for tracking. arXiv preprint arXiv:2101.08040 (2021) 5, 12
14. Du, F., Xu, B., Tang, J., Zhang, Y., Wang, F., Li, H.: 1st place solution to eccv-tao-2020: Detect and represent any object for tracking. arXiv preprint arXiv:2101.08040 (2021) 12
15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 1
16. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5356–5364 (2019) 2, 5, 10

17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 13
18. Kaestner, R., Maye, J., Pilat, Y., Siegwart, R.: Generative object detection and tracking in 3d range data. In: 2012 IEEE International Conference on Robotics and Automation. pp. 3075–3081. IEEE (2012) 3
19. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020) 9
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955). https://doi.org/https://doi.org/10.1002/nav.3800020109, https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109 6
21. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: Siamese cnn for robust target association. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 33–40 (2016) 3
22. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10991–11000 (2020) 10
23. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2017) 10
24. Liu, Q., Chu, Q., Liu, B., Yu, N.: Gsm: Graph similarity model for multi-object tracking. In: IJCAI. pp. 530–536 (2020) 3
25. Liu, Y., Zulfikar, I.E., Luiten, J., Dave, A., Ošep, A., Ramanan, D., Leibe, B., Leal-Taixé, L.: Opening up open-world tracking. arXiv preprint arXiv:2104.11221 (2021) 3, 4
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) 10
27. Lu, Z., Rathod, V., Votel, R., Huang, J.: Retinatrack: Online single stage joint detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14668–14678 (2020) 3
28. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**(2), 548–578 (2021) 2, 4, 5
29. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. arXiv preprint arXiv:2101.02702 (2021) 3
30. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 1
31. 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) 3
32. Mitzel, D., Leibe, B.: Taking mobile multi-object tracking to the next level: People, unknown objects, and carried items. In: European Conference on Computer Vision. pp. 566–579. Springer (2012) 3
33. Moosmann, F., Stiller, C.: Joint self-localization and tracking of generic objects in 3d range data. In: 2013 IEEE International Conference on Robotics and Automation. pp. 1146–1152 (2013). https://doi.org/10.1109/ICRA.2013.6630716 3

34. Ošep, A., Hermans, A., Engelmann, F., Klostermann, D., Mathias, M., Leibe, B.: Multi-scale object candidates for generic object tracking in street scenes. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 3180–3187. IEEE (2016) 3
35. Ošep, A., Mehner, W., Voigtlaender, P., Leibe, B.: Track, then decide: Category-agnostic vision-based multi-object tracking. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3494–3501. IEEE (2018) 3
36. Ošep, A., Voigtlaender, P., Luiten, J., Breuers, S., Leibe, B.: Large-scale object mining for object discovery from unlabeled video. pp. 5502–5508 (05 2019). https://doi.org/10.1109/ICRA.2019.8793683 3
37. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (June 2021) 10, 11, 12, 13
38. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: European Conference on Computer Vision. pp. 145–161. Springer (2020) 3
39. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015) 10
40. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016) 2, 4, 5
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 2
42. Schulter, S., Vernaza, P., Choi, W., Chandraker, M.: Deep network flow for multi-object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6951–6960 (2017) 3
43. Teichman, A., Levinson, J., Thrun, S.: Towards 3d object recognition via classification of arbitrary object tracks. In: 2011 IEEE International Conference on Robotics and Automation. pp. 4034–4041. IEEE (2011) 3
44. Teichman, A., Thrun, S.: Tracking-based semi-supervised learning. The International Journal of Robotics Research **31**(7), 804–818 (2012) 3
45. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017) 12, 13
46. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5188–5197 (2019) 3
47. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 1, 3, 10
48. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020) 2, 3