Supplementary Material for Tracking Every Thing in the Wild

Anonymous ECCV submission

Paper ID 4933

This document gives more details on the ablation study regarding our Tracking-Every-Thing Accuracy (TETA) metric and our Tracking-Every-Thing tracker (TETer) and provides additional evaluation training and implementation details.

A BDD100K Per-class Evaluation Results

We provide per-class evaluation results using CLEARMOT [1] and TETA metrics on the BDD100K [11] validation set in Table 1. Data distribution in BDD100k is long-tailed. The Car category consists of most of the tracks in the dataset. The rest of the categories are rare compared to the dominant ones. Thus, we characterized them as rare classes. TETer can achieve significant improvements across all rare classes on both established MOTA, IDF1, and our TETA metrics compared to the previous state-of-the-art QDtrack. In particular, TETer boosts MOTA of buses by over 7 points on the validation set and TETA by over 6 points. We also compare our TETer results with CEM with its class agnostic counterpart, where the model only uses the AET strategy without CEM. The result shows that our model gains significant improvements over rare classes where the class agnostic instance association cannot be well trained due to lacking annotations. For instance, we gain +3.8 MOTA on buses and +4.7 MOTA on riders. Further, we can observe improvements on the TETA score, where we gain +2 on train and +1.5 on motorcycle. This demonstrates that TETer can better handle tracking rare classes. With CEM, we exploit the semantic annotations offered by large-scale object detection datasets. 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 effectively with class-agnostic appearance training on the long-tailed datasets.

B TAO Per Frequency Group Evaluation Results

We provide evaluation results per frequency group on the TAO validation set using TETA in Table 3. We first observe that TETA can effectively evaluate methods across different frequency groups, despite difficulties introduced by classification errors. Although ClsA drops significantly for both QDTrack [8] and TETer as categories become more rare, LocA and AssocA are relatively stable. This enables us to compare different methods even in large-scale, long-tailed settings where classification is the bottleneck.

Compared to QDTrack, TETer can obtain consistent improvements in TETA, LocA, and AssocA across all frequency groups, at the cost of a small degradation

Table 1: Per-class evaluation results on the BDD100K validation set using CLEAR-MOT and TETA metrics. Rare classes are highlighted in gray

Method	Category	MOTA	MOTP	IDF1	TETA	LocA	AssocA	ClsA
QDTrack [8]	Pedestrian	49.3	78.4	59.9	52.1	50.9	46.8	58.7
	Rider	35.0	77.5	51.5	45.1	47.3	39.6	48.5
	Car	69.8	84.6	75.0	69.1	62.2	65.6	79.5
	Truck	39.2	85.4	58.2	55.5	57.0	55.6	53.9
	Bus	40.8	86.2	62.3	57.9	58.1	57.5	58.1
	Train	0.0	-	0.0	12.1	15.6	20.7	0.0
	Motorcycle	28.8	76.9	56.0	46.4	41.6	54.0	43.5
	Bicycle	30.0	76.2	50.1	44.6	34.2	48.0	51.4
	Average	36.6	70.7	51.6	47.8	45.9	48.5	49.2
AET (Class-agnostic)	Pedestrian	47.8	79.1	59.3	53.0	52.7	47.0	59.4
	Rider	35.9	76.4	53.2	48.2	51.6	48.1	45.0
	Car	69.6	85.4	75.0	70.5	64.1	66.5	80.8
	Truck	41.7	85.2	59.8	59.5	59.0	63.2	56.1
	Bus	44.4	86.1	66.2	61.8	61.1	65.4	58.8
	Train	-2.6	-	0.0	12.4	15.7	21.4	0.0
	Motorcycle	31.6	76.8	57.8	47.0	42.8	55.3	42.8
	Bicycle	29.8	76.9	49.7	45.5	35.2	48.7	52.4
	Average	37.3	70.7	52.6	49.7	47.8	52.0	49.4
TETer (CEM)	Pedestrian	49.7	79.1	59.9	54.1	52.3	47.9	62.2
	Rider	40.5	76.5	56.6	49.9	50.4	49.3	50.1
	Car	69.7	85.4	74.2	70.5	63.8	65.6	82.1
	Truck	43.3	85.5	59.7	59.1	58.2	60.1	59.2
	Bus	48.2	86.2	67.6	63.7	60.0	65.8	65.3
	Train	0.0	-	0.0	14.4	15.2	28.1	0.0
	Motorcycle	31.6	77.0	58.2	48.5	42.8	56.3	46.6
	Bicycle	30.1	77.1	50.1	46.3	34.7	50.0	54.2
	Average	**39.1**	**70.8**	**53.3**	**50.8**	**47.2**	**52.9**	**52.4**

in ClsA. The improvements are more prominent on common and rare categories, where TETer can achieve over 3 points improvement in TETA. For rare categories, TETer achieves 1.8 points improvement in LocA and over 7 points in AssocA. Even on frequency categories, TETer can still improve AssocA by over 6 points. Table 3 also shows that the major differences between frequent and rare categories lies in classification. The localization and association capabilities of both trackers already generalize very well on rare categories.

C Exemplar-based Classification

Given an example object, exemplar-based classification means classifying objects by comparing with the given example to determine whether they belong to the same class. Given two neighboring frames t_1 and t_2 in a video sequence, all objects in t_1 will be treated as exemplars. For each exemplar, we find all target objects in t_2 that belong to the same class as the exemplar.

In this experiment, we compare our Class Exemplar Matching (CEM) with a hard prior baseline that matches objects with the same predicted class label. We evaluate both methods on the TAO validation set and compute precision-recall (PR) curves for comparison. A true positive (TP) match is a match between two

Table 2: Changing the margin r with a fixed α

Method	r	TETA	LocA	AssocA	ClsA	LocRe	LocPr
QDTrack	50	30.0	50.5	27.4	12.1	53.1	75.8
	75	30.6	52.4	27.4	12.1	53.1	84.7
	90	30.8	53.0	27.4	12.1	53.1	91.1
TETer (ours)	50	33.2	51.6	35.0	13.2	54.3	75.3
	75	33.9	53.6	35.0	13.2	54.3	84.3
	90	34.1	54.1	35.0	13.2	54.3	90.5

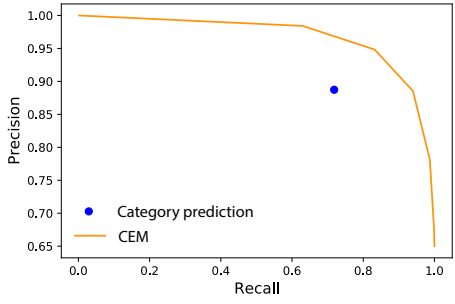


Fig. 1: Precision-Recall (PR) curves of in-sequence instance retrieval based on CEM and category prediction on the TAO validation set. A retrieved result is correct if it has the same labeled category with the target instance

objects that belong to the same category. A false positive (FP) match is a match between two objects that belong to different categories. A false negative (FN) is a non-match between two objects that belong to the same category. To compute the PR curve, we sample 10 thresholds from 0 to 0.99 with a fixed step size.

Fig. 1 shows the results of the experiment. The hard prior baseline takes the argmax of the predictions of a softmax classifier from Faster R-CNN, thus there is only a single value in the PR curve. CEM significantly outperforms the hard prior baseline.

D TETA Details

We provide additional details regarding our TETA metric about how it disentangles classification and how it deals with evaluation on datasets with complete annotations.

D.1 Disentangling Classification

The most direct way to disentangle classification is not to consider per-class performance and evaluate every object class-agnostically. However, on large-scale, long-tailed datasets, such evaluation will be dominated by objects of the few common categories, and the overall performance will not reflect the improvements on rare classes. On the other hand, per-class evaluation requires us to select prediction results for each class, which is sensitive to classification performance. If the

Table 3: Per frequency group results on the TAO validation set using TETA

Method	Freq. Group	TETA	LocA	AssocA	ClsA
QDTrack [8]	Frequent	36.3	52.4	32.0	24.5
	Common	23.9	47.2	21.7	2.9
	Rare	26.7	52.7	27.4	0.0
	All	30.0	50.5	27.4	12.1
TETer (ours)	Frequent	39.4	53.9	38.7	25.7
	Common	27.3	47.3	30.4	4.1
	Rare	30.1	54.5	35.3	0.4
	All	**33.2**	**51.6**	**35.0**	**13.2**

classification is wrong, the contribution in localization and association will be ignored. TETA can naturally deal with this issue with the local cluster evaluation since we select predictions based on their location rather than class. To evaluate a particular class, we access predictions in the local clusters of ground-truth objects belonging to the chosen class. Thus, we can evaluate the localization and association performance even when the class predictions are wrong.

D.2 TETA with Complete Annotations

Multiple categories TETA can also work with complete annotations. First, the localization accuracy is not affected. In the case of incomplete annotations, we treat every unmatched predictions in each cluster as false positives. If we have exhaustive annotations, we still treat those unmatched predictions as false positives. The remaining question is how to penalize predictions that are not in any clusters. For such predictions, we know that they are not highly overlapped with any ground truth box, since we have exhaustive annotations. The predictions thus false classify background as one of the foreground classes, and so we treat them as classification false positives.

Single category For single category with exhaustive annotations, the classification term of TETA is meaningless and can be ignored. Also, since we do not need to perform per-class evaluation, the margin of the local cluster does not matter either. Thus, we can set the margin r to 0. With these changes, TETA becomes similar to the HOTA [7] metric with the only difference being that we use arithmetic mean instead of geometric mean.

D.3 Ablation Study of TETA

We provide an ablation study of the local cluster IoU margin r of TETA. We perform this experiment on incomplete dataset TAO.

The results are shown in Table 2. The LocRe and LocPr represent the localization recall and precision. As we can see, with a larger r, the LocPr increases since TETA becomes more conservative regarding identifying FPs. In the mean time, TETA makes fewer mistakes where the objects with no annotations are wrongly identified as FPs. In extreme crowded scenarios with incomplete annotations, it's recommended to set a higher r to avoid false punishment.

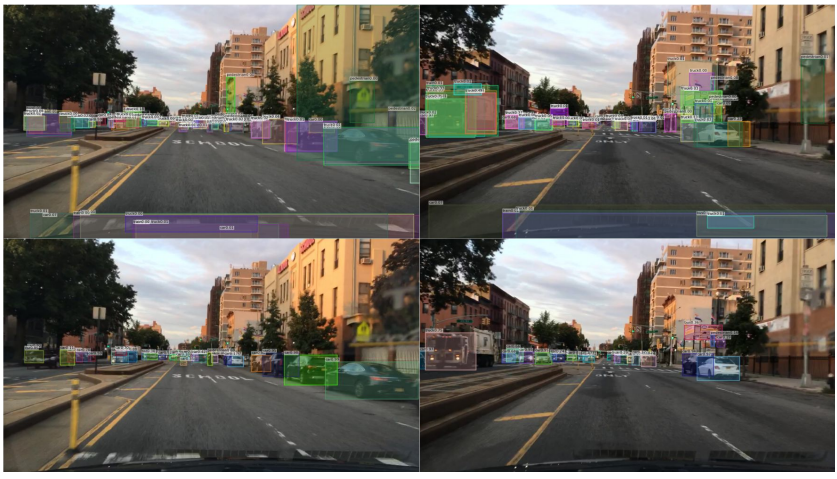


Fig. 2: Qualitative comparison of tracker optimized for the TAO [2] metric (top) and tracker optimized for TETA (bottom) on BDD100K. The tracker optimized for TAO produces more false positives

E Qualitative Results

We provide additional qualitative results of TETer.

E.1 Category label prediction vs. CEM

We first compare the QDTrack [8] which uses class prediction as hard prior to associate objects with TETer which uses CEM. In Fig. 3, we show an example of QDTrack producing ID Switches due to errors in classification, whereas TETer is more robust to such issues.

E.2 Class-agnostic vs. CEM

We further show the comparison between the class-agnostic association (AET baseline in Section 5.5) and association with CEM. We observe that most class agnostic association errors happen in rare classes where there are not enough videos to train the class-agnostic instance association module well. For instance, Fig. 5 (a) shows the bicycle (16) is wrongly associated with the car with class-agnostic association, while using the CEM module helps to avoid the mistake. The CEM module utilizes the supervision from large-scale object detection datasets to learn fine-grained class appearance differences, which helps the association on rare classes.

E.3 Tracking results comparison: TAO metric vs. TETA

We provide results for cross-dataset analysis. In Fig. 2, we show predictions from trackers that are optimized either for the TAO metric or TETA. The tracker optimized for the TAO metric generates more false positives that highly overlap, producing results that are difficult to use in practice. On the other hand, the tracker optimized for TETA produces cleaner results.

Fig. 3: Qualitative comparison of QDTrack [8] (top) and TETer (bottom) on BDD100K. QDTrack has ID Switches due to classification errors. (Same color represent the same track)

E.4 Rare class retrieval

We perform the class retrieval experiments in the rare classes on TAO to show the effectiveness of the CEM embeddings. We take the objects in the first frame of each ground truth track and use them as the retrieval templates to retrieve ground truth objects in the whole TAO validation set. The softmax prediction means we use the softmax confidence to retrieve objects that are predicted as the same class as the template. The CEM means we use the CEM embedding similarity to perform the retrieval. Fig. 6 shows the CEM embedding can successfully retrieve the examples in the rare classes, while its softmax fails. Fig. 7 shows some failures cases where the CEM module retrieves the wrong class due to occlusion or high visual similarities.

E.5 TETer failure cases

We also show some common failure cases of TETer on TAO in Fig. 4. Note that TAO is annotated at 1 FPS. Thus, fast-moving objects usually have huge appearance changes in neighboring frames. Due to the large appearance and location variations, tracking is challenging on TAO. Also, TETer suffers from localization errors caused by occlusion.

F More Implementation Details

We provide more implementation and training details of our method and evaluation setup in different benchmarks.

Fig. 4: Failure cases of TETer on TAO. Big appearance changes (top) and occlusions (bottom)

F.1 Network architecture

We use the popular Faster R-CNN [9] with ResNet as the backbone. Specifically, we use the ResNet-101 [5] for TETer on TAO and ResNet-50 on BDD100K. For the exemplar encoder, we use *4conv-3fc* head with group normalization [10]. The final output channel numbers are 1230 for TETer on TAO [2] and 256 for BDD100K [11]. We use the same network architecture for the instance appearance encoder but with only *1fc* layers for the final output. The channel number of the instance appearance encoder is 256 by default on both datasets.

F.2 Training

TAO We train TETer following the TAO [3] set up with a mixed LVISv0.5 [4] and COCO [6] dataset. We set the batch size to 16 and the learning rate to 0.02. We train 24 epochs in total and decrease the learning by 0.1 after 16 and 22 epochs. For data augmentation, we randomly flip the images horizontally with a 0.5 ratio. We randomly resize the training images to keep their short edges between 640 to 800. We randomly sample images to form mini-batches with additionally repeat sampling for rare classes [4]. We set the repeat factor to 0.001. We train the instance appearance encoder on the TAO training set following the same setting as QDTrack [8].

BDD100K We use the same object detector as QDTrack [8]. For training the exemplar encoder, we freeze the object detector above and train with 8 BDD100K MOT categories using the BDD100K Detection set, which contains 70K images. For data augmentation, we randomly flip the images horizontally with a 0.5 ratio. We randomly resize the training images to keep their short edges between 640 to 800. We randomly sampled images to form mini-batches. We set the batch size to 128.

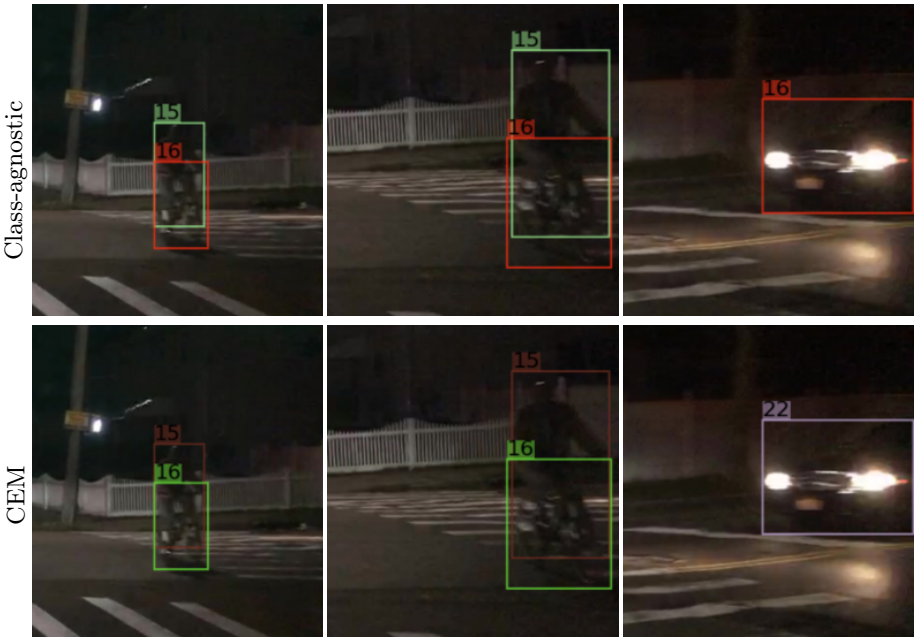
F.3 Inference and evaluation

TAO We evaluate our model on the TAO validation set with TETA. For the close-set setting, the TAO validation set contains 988 videos with 302 classes, a subset with LVIS classes. For the open-set setting, we merge the additional free-form classes [2] as one unknown class. During inference, we use the fixed image scale with 800 at the short edge. We initialize a new track if the object has detection confidence higher than 0.0001.

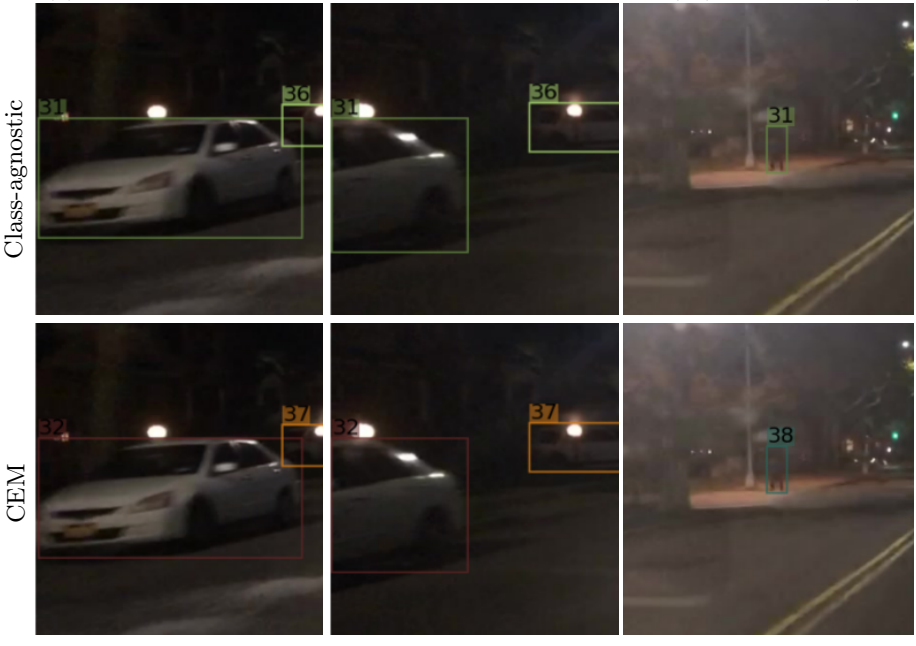
BDD100K The BDD100K contains 200 videos (40k) for validation and 400 videos (80k) for testing. We use both the BDD100K validation set and the test set for evaluation. For the inference, we use the $(1296, 720)$ image scale. We use the best performed model in the validation set which is saved at 2 epoch. We initialize a new track if the detection confidence is higher than 0.7.

References

1. 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. 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) 5, 7, 8
3. 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) 7
4. 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) 7
5. 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) 7
6. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 7
7. 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) 4
8. 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) 1, 2, 4, 5, 6, 7
9. 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) 7
10. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 7
11. 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, 7



(a) The class agnostic tracker false associates a bicycle (16) to a car (16).



(b) The class agnostic tracker false associates a car (31) to a pedestrian (31).

Fig. 5: The qualitative comparison between CEM and class-agnostic association. CEM can exploit the semantic supervision offered by large scale datasets to learn fine-grained class appearance differences. Therefore, it can help trackers to avoid the association mistakes in class-agnostic association

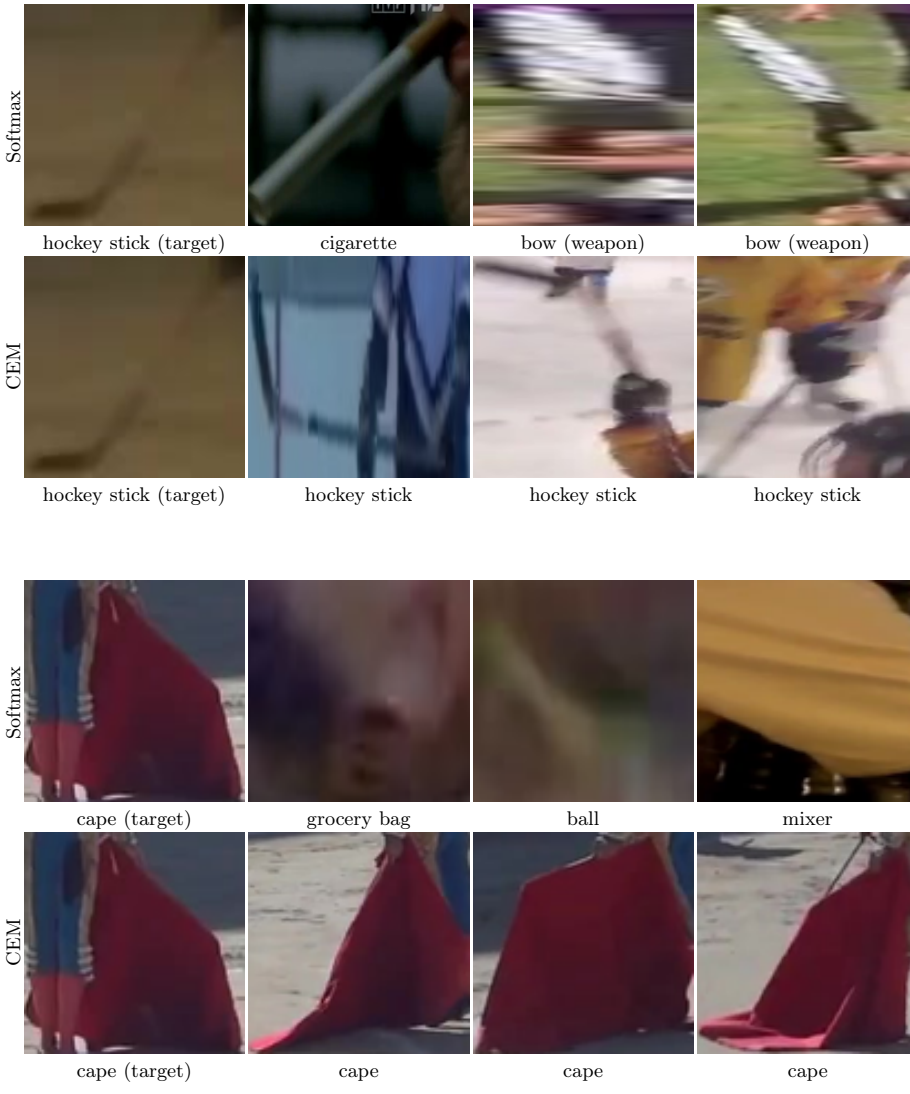


Fig. 6: Rare class retrieval. Objects in the most left column are the retrieval targets. The rest columns are the retrieval examples using softmax classifier or CEM. The retrieval results are ranked by confidence. Columns from left to right are corresponding to the results from high to low confidence. The class name under each example is the ground truth class for the example. The CEM module generalizes well on rare classes and can successfully retrieve them

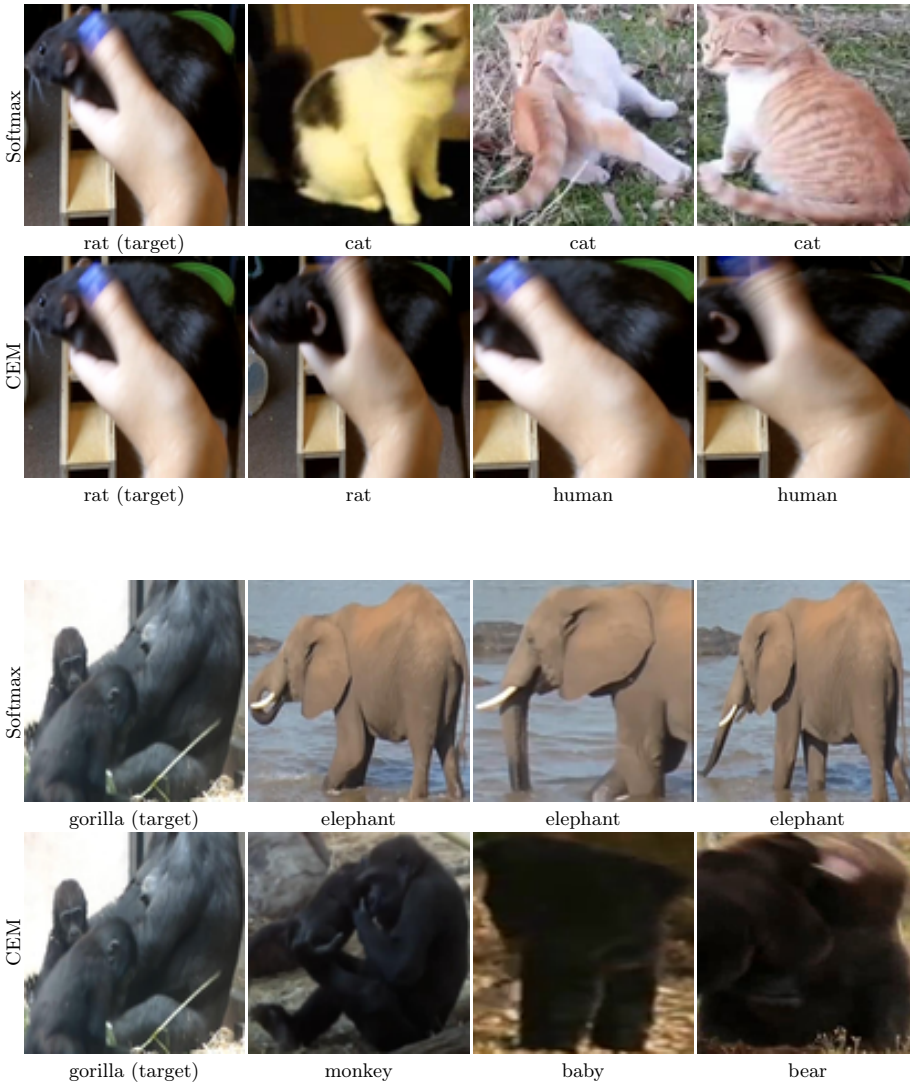


Fig. 7: Failure retrieval cases. The first two rows show that the CEM module retrieve the human hand as rat due to occlusion. The last two rows show the CEM retrieve monkey, bear as the gorilla due to the high visual similarities