Supplementary Material for MOTS: Multi-Object Tracking and Segmentation

Paul Voigtlaender[1] Michael Krause[1] Aljoša Ošep[1] Jonathon Luiten[1]
Berin Balachandar Gnana Sekar[1] Andreas Geiger[2] Bastian Leibe[1]
[1]RWTH Aachen University [2]MPI for Intelligent Systems and University of Tübingen
{voigtlaender,osep,luiten,leibe}@vision.rwth-aachen.de
{michael.krause,berin.gnana}@rwth-aachen.de andreas.geiger@tue.mpg.de

1. Losses for the Association Head

TrackR-CNN uses association scores based on vectors predicted by an association head to identify the same object across time. In our baseline model, we train this head using a batch hard triplet loss proposed by Hermans *et al*. [3], which we state again here: Let $\mathcal{D}$ denote the set of detections for a video. Each detection $d \in \mathcal{D}$ has a corresponding association vector a_d and is assigned a ground truth track id id_d determined by its overlap with the ground truth objects (we only consider detections which sufficiently overlap with a ground truth object here). For a video sequence of T time steps, the association loss in the batch-hard formulation with margin α is then given by

$$\mathcal{L}_{batch\_hard} = \frac{1}{|D|} \sum_{d \in \mathcal{D}} \max\Big(\max_{\substack{e \in \mathcal{D}: \\ id_e = id_d}} \|a_e - a_d\| - \min_{\substack{e \in \mathcal{D}: \\ id_e \neq id_d}} \|a_e - a_d\| + \alpha, 0\Big). \tag{1}$$

Intuitively, each detection d is selected as an anchor and then the most dissimilar detection with the same id is selected as a hard positive example and the most similar detection with a different id is selected as a hard negative example for this anchor. The margin α and maximum operation ensure that the distance of the anchor to the hard positive is smaller than its distance to the hard negative example by at least α.

In order to justify our choice of the batch-hard loss, we also report results using two alternative loss formulations, namely the batch all loss [3] which considers all pairs of detections, *i.e*.

$$\mathcal{L}_{batch\_all} = \frac{1}{|\mathcal{D}|^2} \sum_{d \in \mathcal{D}} \sum_{e \in D} \max\big(\|a_e - a_d\| - \|a_e - a_d\| + \alpha, 0\big) \tag{2}$$

and the contrastive loss [2]

$$\mathcal{L}_{contrastive} = \frac{1}{|\mathcal{D}|^2}\Big(\sum_{d \in \mathcal{D}} \sum_{\substack{e \in \mathcal{D} \\ id_e = id_d}} \|a_e - a_d\|^2 + \sum_{d \in \mathcal{D}} \sum_{\substack{e \in \mathcal{D} \\ id_e \neq id_d}} \max(\alpha - \|a_e - a_d\|, 0)^2\Big). \tag{3}$$

Table 1 compares the performance of these different variants of the loss function on the KITTI MOTS validation set. It can be seen that the batch hard triplet loss performs better than just considering all pairs of detections (*Batch All Triplet*), or using the conventional contrastive loss (*Contrastive*). Especially for pedestrians performance using the contrastive loss is low.

Association Loss	sMOTSA		MOTSA		MOTSP	
	Car	Ped	Car	Ped	Car	Ped
Batch Hard Triplet	76.2	**46.8**	87.8	**65.1**	**87.2**	**75.7**
Batch All Triplet	75.5	45.3	87.1	63.8	87.1	75.6
Contrastive	**76.4**	43.2	**88.7**	61.5	86.7	75.2

Table 1: **Different Association Losses for TrackR-CNN**. Comparison of results on the KITTI MOTS validation set.

2. Details of the Annotation Procedure

We noticed that wrong segmentation results often stem from imprecise or wrong bounding box annotations of the original MOT datasets. For example, the annotated bounding boxes for the KITTI tracking dataset [1] are amodal, *i.e*., they extend to the ground even if only the upper body of a person is visible. In these cases, our annotators corrected these bounding boxes instead of adding additional polygon annotations. We also corrected the bounding box

	sMOTSA		MOTSA		MOTSP	
	Car	Ped	Car	Ped	Car	Ped
GT Boxes (orig) + Filling	33.7	-66.1	55.5	-57.7	71.8	54.6
GT Boxes (orig) + Ellipse	52.3	-31.9	74.0	-14.5	74.9	57.4
GT Boxes (orig) + MG	77.3	36.5	90.4	55.7	86.3	75.3
GT Boxes (tight) + Filling	61.3	-1.7	83.9	22.0	75.4	60.5
GT Boxes (tight) + Ellipse	70.9	17.2	91.8	42.4	78.1	64.2
GT Boxes (tight) + MG	**82.5**	**50.0**	**95.3**	**71.1**	**86.9**	**75.4**

Table 2: **Ground Truth Results on KITTI MOTS**. +MG denotes mask generation with a KITTI MOTS fine-tuned Mask R-CNN..

level tracking annotations in cases where they contained errors or missed objects. Finally, we retained ignore regions that were labeled in the source datasets, *i.e.*, image regions that contain unlabeled objects from nearby classes (like vans and buses) or target objects that were to small to be labeled. Hypothesized masks that are mapped to ignore regions are neither counted as true nor as false positives in our evaluation procedure.

3. Ground Truth Experiments

We performed additional experiments to demonstrate the difficulty of generating accurate segmentation masks even when the ground truth bounding boxes are given (see Table 2). As in the main paper, we consider two variants of the ground truth: the original bounding boxes from KITTI (*orig*), which are amodal, *i.e.* if only the upper body of a person is visible, the box will still extend to the ground, and tight bounding boxes (*tight*) derived from our segmentation masks. We created masks for the boxes by simply filling the full box (+*Filling*), by inserting an ellipse (+*Ellipse*), and by generating masks using the KITTI MOTS fine-tuned Mask R-CNN (+*MG*). In each case, instance ids are retained from the corresponding boxes.

Our results show that rectangles and ellipses are not sufficient to accurately localize objects when mask-based matching is used, even with perfect track hypotheses. The problem is amplified when using amodal boxes, which often contain large regions that do not show the object. This further validates our claim that MOT tasks can benefit from pixel-wise evaluation. The relatively low scores for pedestrians also imply a limit to post-hoc masks generation using the KITTI fine-tuned Mask R-CNN.

4. Visualization of Association Vectors

We present a visualization of the association vectors produced by our TrackR-CNN model on a sequence of the KITTI MOTS validation set in Figure 1. Here, all association vectors for detections produced by TrackR-CNN on sequence 18 were used for principal component analysis and then projected onto the two components explaining most of their variance. The resulting two dimensional vectors were used to arrange the crops for the corresponding detections in 2D. The visualization was created using the TensorBoard embedding projector. It can be seen that crops belonging to the same car are in most cases close to each other in the embedding space.

5. Qualitative results

We present further qualitative results of our baseline TrackR-CNN model on the KITTI MOTS and MOTSChallenge validation sets including some illustrative failure cases. See Figures 2, 3, 4 and 5 on the following pages.

References

[1] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 1

[2] R. Hadsell, S. Chopra, and Y. Lecun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 1

[3] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1

Figure 1: **Visualization using PCA on the association vectors of detections generated by TrackR-CNN on sequence 18 of KITTI MOTS.** Detections with similar appearance are grouped together by minimizing the association loss.

Figure 2: **Qualitative Results on MOTSChallenge.** While complex scenes with many occluding objects often work well, there can still be missing detections and id switches during difficult occlusions, as in this example (highlighted by red ellipses).

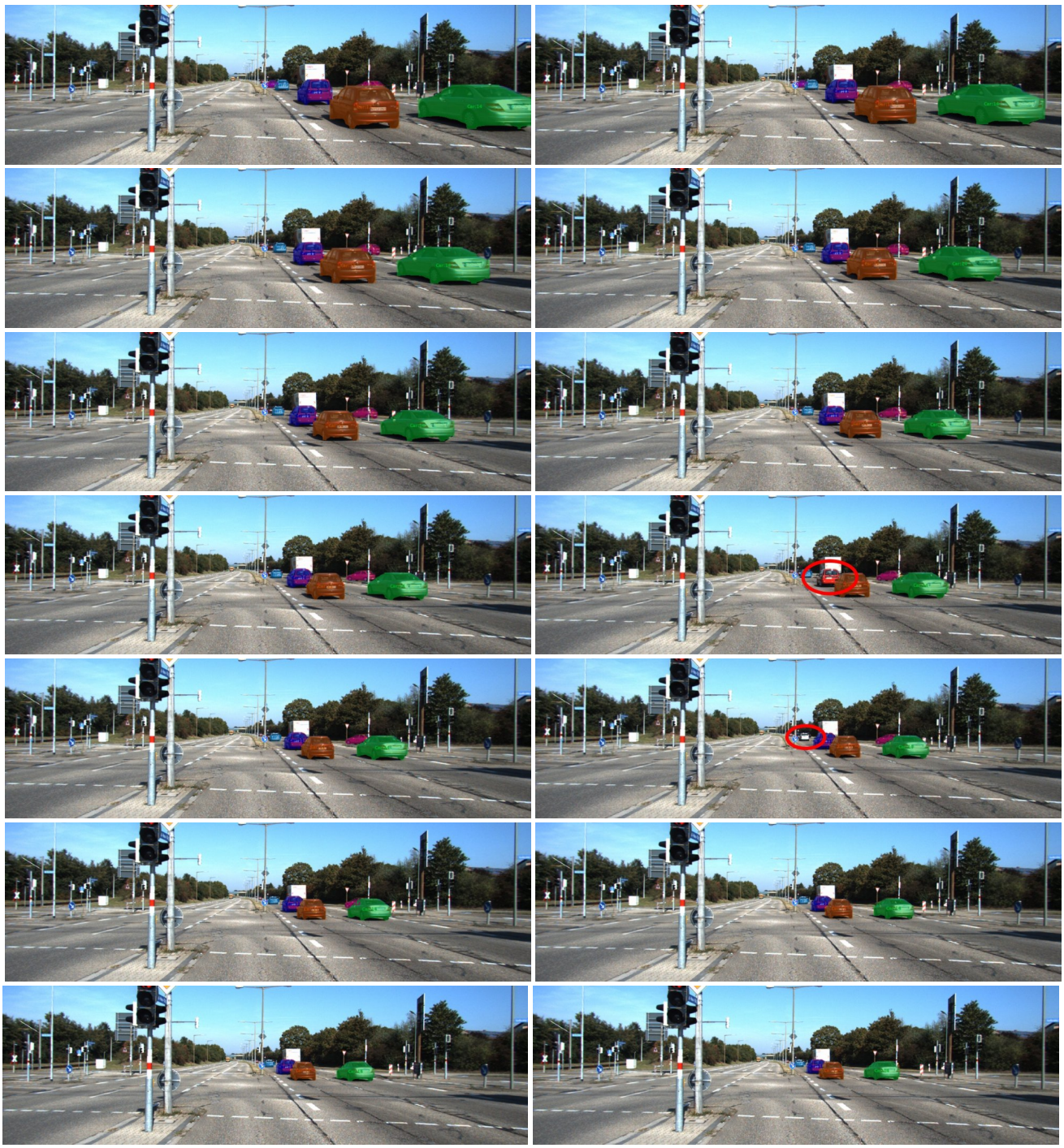

Figure 3: **Qualitative Results on KITTI MOTS.** In simpler scenes, the model is able to continue a track with the same ID after a missing detection (highlighted by red ellipses).

Figure 4: **Qualitative Results on KITTI MOTS.** In a rare failure case, pylons are confused for pedestrians (highlighted by red ellipses). In most cases, detections correspond to real instances of the class.

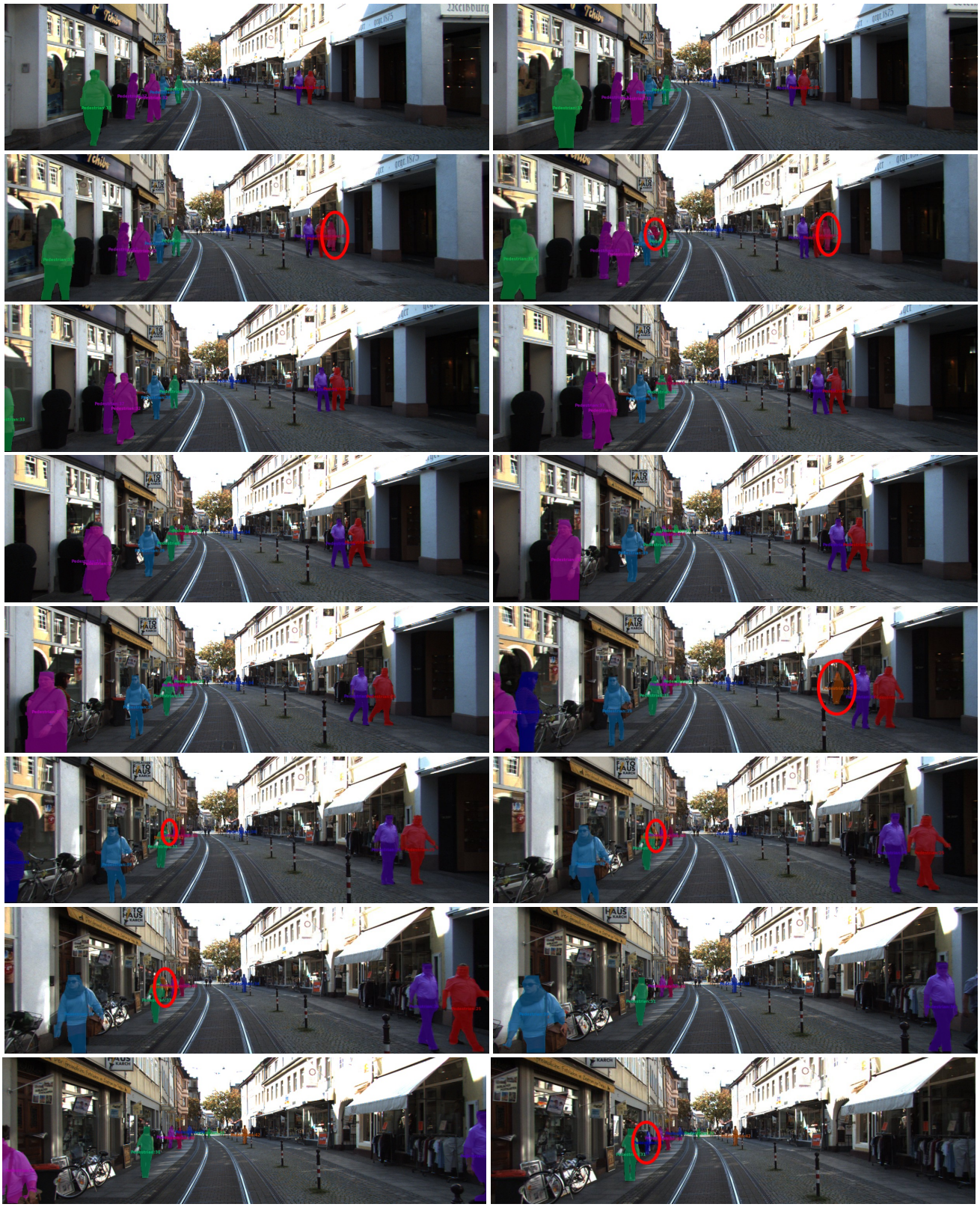

Figure 5: **Qualitative Results on KITTI MOTS.** In less crowded scenes, distinguishing objects works well but some erroneous detections (highlighted by red ellipses) might still happen.