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

Abstract

This paper extends the popular task of multi-object tracking to multi-object tracking and segmentation (MOTS). Towards this goal, we create dense pixel-level annotations for two existing tracking datasets using a semi-automatic annotation procedure. Our new annotations comprise 65,213 pixel masks for 977 distinct objects (cars and pedestrians) in 10,870 video frames. For evaluation, we extend existing multi-object tracking metrics to this new task. Moreover, we propose a new baseline method which jointly addresses detection, tracking, and segmentation with a single convolutional network. We demonstrate the value of our datasets by achieving improvements in performance when training on MOTS annotations. We believe that our datasets, metrics and baseline will become a valuable resource towards developing multi-object tracking approaches that go beyond 2D bounding boxes. We make our annotations, code, and models available at https://www.vision.rwth-aachen.de/page/mots.

1. Introduction

In recent years, the computer vision community has made significant advances in increasingly difficult tasks. Deep learning techniques now demonstrate impressive performance in object detection as well as image and instance segmentation. Tracking, on the other hand, remains challenging, especially when multiple objects are involved. In particular, results of recent tracking evaluations [37, 7, 25] show that bounding box level tracking performance is saturating. Further improvements will only be possible when moving to the pixel level. We thus propose to think of all three tasks – detection, segmentation and tracking – as interconnected problems that need to be considered together.

Datasets that can be used to train and evaluate models for instance segmentation usually do not provide annotations on video data or even information on object identities across different images. Common datasets for multi-object tracking, on the other hand, provide only bounding
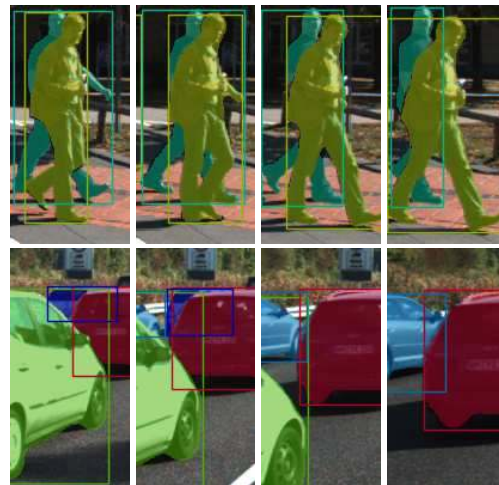

Figure 1: **Segmentations vs. Bounding Boxes.** When objects pass each other, large parts of an object's bounding box may belong to another instance, while per-pixel segmentation masks locate objects precisely. The shown annotations are crops from our KITTI MOTS dataset.

box annotations of objects. These can be too coarse, *e.g.*, when objects are partially occluded such that their bounding box contains more information from other objects than from themselves, see Fig. 1. In these cases, pixel-wise segmentation of the objects results in a more natural description of the scene and may provide additional information for subsequent processing steps. For segmentation masks there is a well-defined ground truth, whereas many different (non-tight) boxes might roughly fit an object. Similarly, tracks with overlapping bounding boxes create ambiguities when compared to ground truth that usually need to be resolved at evaluation time by heuristic matching procedures. Segmentation based tracking results, on the other hand, are by definition non-overlapping and can thus be compared to ground truth in a straightforward manner.

In this paper, we therefore propose to extend the well-known multi-object tracking task to instance segmentation tracking. We call this new task "Multi-Object Tracking

and Segmentation (MOTS)". To the best of our knowledge, there exist no datasets for this task to date. While there are many methods for bounding box tracking in the literature, MOTS requires combining temporal and mask cues for success. We thus propose TrackR-CNN as a baseline method which addresses all aspects of the MOTS task. TrackR-CNN extends Mask R-CNN [14] with 3D convolutions to incorporate temporal information and by an association head which is used to link object identities over time.

In summary, this paper makes the following **contributions:** (1) We provide two new datasets with temporally consistent object instance segmentations based on the popular KITTI [13] and MOTChallenge [37] datasets for training and evaluating methods that tackle the MOTS task. (2) We propose the new soft Multi-Object Tracking and Segmentation Accuracy (sMOTSA) measure that can be used to simultaneously evaluate all aspects of the new task. (3) We present TrackR-CNN as a baseline method which addresses detection, tracking, and segmentation jointly and we compare it to existing work. (4) We demonstrate the usefulness of the new datasets for end-to-end training of pixel-level multi-object trackers. In particular, we show that with our datasets, joint training of segmentation and tracking procedures becomes possible and yields improvements over training only for instance segmentation or bounding box tracking, which was possible previously.

2. Related Work

Multi-Object Tracking Datasets. In the multi-object tracking (MOT) task, an initially unknown number of targets from a known set of classes must be tracked as bounding boxes in a video. In particular, targets may enter and leave the scene at any time and must be recovered after long-time occlusion and under appearance changes. Many MOT datasets focus on street scenarios, for example the KITTI tracking dataset [13], which features video from a vehicle-mounted camera; or the MOTChallenge datasets [26, 37] that show pedestrians from a variety of different viewpoints. UA-DETRAC [57, 35] also features street scenes but contains annotations for vehicles only. Another MOT dataset is PathTrack [36], which provides annotations of human trajectories in diverse scenes. PoseTrack [2] contains annotations of joint positions for multiple persons in videos. None of these datasets provide segmentation masks for the annotated objects and thus do not describe complex interactions like in Fig. 1 in sufficient detail.

Video Object Segmentation Datasets. In the video object segmentation (VOS) task, instance segmentations for one or multiple generic objects are provided in the first frame of a video and must be segmented with pixel accuracy in all subsequent frames. Existing VOS datasets contain only few objects which are also present in most frames. In addition, the common evaluation metrics for this task (region Jaccard in-

dex and boundary F-measure) do not take error cases like id switches into account that can occur when tracking multiple objects. In contrast, MOTS focuses on a set of pre-defined classes and considers crowded scenes with many interacting objects. MOTS also adds the difficulty of discovering and tracking a varying number of new objects as they appear and disappear in a scene.

Datasets for the VOS task include the DAVIS 2016 dataset [43], which focuses on single-object VOS, and the DAVIS 2017 [45] dataset, which extends the task for multi-object VOS. Furthermore, the YouTube-VOS dataset [59] is available and orders of magnitude larger than DAVIS. In addition, the Segtrackv2 [28] dataset, FBMS [40] and an annotated subset of the YouTube-Objects dataset [46, 19] can be used to evaluate this task.

Video Instance Segmentation Datasets. Cityscapes [12], BDD [61], and ApolloScape [18] provide video data for an automotive scenario. Instance annotations, however, are only provided for a small subset of non-adjacent frames or, in the case of ApolloScape, for each frame but without object identities over time. Thus, they cannot be used for end-to-end training of pixel-level tracking approaches.

Methods. While a comprehensive review of methods proposed for the MOT or VOS tasks is outside the scope of this paper (for the former, see *e.g.* [27]), we will review some works that have tackled (subsets of) the MOTS task or are in other ways related to TrackR-CNN.

Seguin *et al.* [51] derive instance segmentations from given bounding box tracks using clustering on a superpixel level, but they do not address the detection or tracking problem. Milan *et al.* [38] consider tracking and segmentation jointly in a CRF utilizing superpixel information and given object detections. In contrast to both methods, our proposed baseline operates on pixel rather than superpixel level. CAMOT [42] performs mask-based tracking of generic objects on the KITTI dataset using stereo information, which limits its accuracy for distant objects. CDTS [24] performs unsupervised VOS, *i.e.*, without using first-frame information. It considers only short video clips with few object appearances and disappearances. In MOTS, however, many objects frequently enter or leave a crowded scene. While the above mentioned methods are able to produce tracking outputs with segmentation masks, their performance could not be evaluated comprehensively, since no dataset with MOTS annotations existed.

Lu *et al.* [33] tackle tracking by aggregating location and appearance features per frame and combining these across time using LSTMs. Sadeghian *et al.* [50] also combine appearance features obtained by cropped detections with velocity and interaction information using a combination of LSTMs. In both cases, the combined features are input into a traditional Hungarian matching procedure. For our baseline model, we directly enrich detections using temporal in-

formation and learn association features jointly with the detector rather than only "post-processing" given detections.

Semi-Automatic Annotation. There are many methods for semi-automatic instance segmentation, *e.g.* generating segmentation masks from scribbles [49], or clicks [58]. These methods require user input for every object to be segmented, while our annotation procedure can segment many objects fully-automatically, letting annotators focus on improving results for difficult cases. While this is somewhat similar to an active learning setting [11, 56], we leave the decision which objects to annotate with our human annotators to guarantee that all annotations reach the quality necessary for a long-term benchmark dataset (*c.f*. [32]).

Other semi-automatic annotation techniques include Polygon-RNN [9, 1], which automatically predicts a segmentation in form of a polygon from which vertices can be corrected by the annotator. Fluid Annotation [3] allows the annotator to manipulate segments predicted by Mask R-CNN [14] in order to annotate full images. While speeding up the creation of segmentation masks of objects in isolated frames, these methods do not operate on a track level, do not make use of existing bounding box annotations, and do not exploit segmentation masks which have been annotated for the same object in other video frames.

3. Datasets

Annotating pixel masks for every frame of every object in a video is an extremely time-consuming task. Hence, the availability of such data is very limited. We are not aware of any existing datasets for the MOTS task. However, there are some datasets with MOT annotations, *i.e.*, tracks annotated at the bounding box level. For the MOTS task, these datasets lack segmentation masks. Our annotation procedure therefore adds segmentation masks for the bounding boxes in two MOT datasets. In total, we annotated 65,213 segmentation masks. This size makes our datasets viable for training and evaluating modern learning-based techniques.

Semi-automatic Annotation Procedure. In order to keep the annotation effort manageable, we propose a semi-automatic method to extend bounding box level annotations by segmentation masks. We use a convolutional network to automatically produce segmentation masks from bounding boxes, followed by a correction step using manual polygon annotations. Per track, we fine-tune the initial network using the manual annotations as additional training data, similarly to [6]. We iterate the process of generating and correcting masks until pixel-level accuracy for all annotation masks has been reached.

For converting bounding boxes into segmentation masks, we use a fully-convolutional refinement network [34] based on DeepLabv3+ [10] which takes as input a crop of the input image specified by the bounding box with a small context region added, together with an additional input channel

Figure 2: **Sample Images of our Annotations.** KITTI MOTS (top) and MOTSChallenge (bottom).

that encodes the bounding box as a mask. Based on these cues, the refinement network predicts a segmentation mask for the given box. The refinement network is pre-trained on COCO [29] and Mapillary [39], and then trained on manually created segmentation masks for the target dataset.

In the beginning, we annotate (as polygons) two segmentation masks per object in the considered dataset.[1] The refinement network is first trained on all manually created masks and afterwards fine-tuned individually for each object. These fine-tuned variants of the network are then used to generate segmentation masks for all bounding boxes of the respective object in the dataset. This way the network adapts to the appearance and context of each individual object. Using two manually annotated segmentation masks per object for fine-tuning the refinement network already produces relatively good masks for the object's appearances in the other frames, but often small errors remain. Hence, we manually correct some of the flawed generated masks and re-run the training procedure in an iterative process. Our annotators also corrected imprecise or wrong bounding box annotations in the original MOT datasets.

KITTI MOTS. We performed the aforementioned annotation procedure on the bounding box level annotations from the KITTI tracking dataset [13]. A sample of the annotations is shown in Fig. 2. To facilitate training and evaluation, we divided the 21 training sequences of the KITTI tracking dataset[2] into a training and validation set, respectively[3]. Our split balances the number of occurrences of each class – cars and pedestrians – roughly equally across training and validation set. Statistics are given in Table 1.

---

[1]The two frames annotated per object are chosen by the annotator based on diversity.

[2]We are currently applying our annotation procedure to the KITTI test set with the goal of creating a publicly accessible MOTS benchmark.

[3]Sequences 2, 6, 7, 8, 10, 13, 14, 16 and 18 were chosen for the validation set, the remaining sequences for the training set.

| | KITTI MOTS | | MOTSChallenge |
	train	val	
# Sequences	12	9	4
# Frames	5,027	2,981	2,862
# Tracks Pedestrian	99	68	228
# Masks Pedestrian			
Total	8,073	3,347	26,894
Manually annotated	1,312	647	3,930
# Tracks Car	431	151	-
# Masks Car			
Total	18,831	8,068	-
Manually annotated	1,509	593	-

Table 1: **Statistics of the Introduced KITTI MOTS and MOTSChallenge Datasets**. We consider pedestrians for both datasets and also cars for KITTI MOTS.

The relatively high number of manual annotations required demonstrates that existing single-image instance segmentation techniques still perform poorly on this task. This is a major motivation for our proposed MOTS dataset which allows for incorporating temporal reasoning into instance segmentation models.

MOTSChallenge. We further annotated 4 of 7 sequences of the MOTChallenge 2017 [37] training dataset[4] and obtained the MOTSChallenge dataset. MOTSChallenge focuses on pedestrians in crowded scenes and is very challenging due to many occlusion cases, for which a pixel-wise description is especially beneficial. A sample of the annotations is shown in Fig. 2, statistics are given in Table 1.

4. Evaluation Measures

As evaluation measures we adapt the well-established CLEAR MOT metrics for multi-object tracking [4] to our task. For the MOTS task, the segmentation masks per object need to be accommodated in the evaluation metric. Inspired by the Panoptic Segmentation task [23], we require that both the ground truth masks of objects and the masks produced by a MOTS method are non-overlapping, *i.e.*, each pixel can be assigned to at most one object. We now introduce our evaluation measures for MOTS.

Formally, the ground truth of a video with T time frames, height h, and width w consists of a set of N non-empty ground truth pixel masks $M = \{m_1, \dots, m_N\}$ with $m_i \in \{0, 1\}^{h \times w}$, each of which belongs to a corresponding time frame $t_m \in \{1, \dots, T\}$ and is assigned a ground truth track id $id_m \in \mathbb{N}$. The output of a MOTS method is a set of K non-empty hypothesis masks $H = \{h_1, \dots, h_K\}$ with $h_i \in \{0, 1\}^{h \times w}$, each of which is assigned a hypothesized track id $id_h \in \mathbb{N}$ and a time frame $t_h \in \{1, \dots, T\}$.

Establishing Correspondences. An important step for the CLEAR MOT metrics [4] is to establish correspondences between ground truth objects and tracker hypotheses. In the bounding box-based setup, establishing correspondences is non-trivial and performed by bipartite matching, since ground truth boxes may overlap and multiple hypothesized boxes can fit well to a given ground truth box. In the case of MOTS, establishing correspondences is greatly simplified since we require that each pixel is uniquely assigned to at most one object in the ground truth and the hypotheses respectively. Thus, at most one predicted mask can have an Intersection-over-Union (IoU) of more than 0.5 with a given ground truth mask [23]. Hence, the mapping $c : H \to M \cup \{\emptyset\}$ from hypothesis masks to ground truth masks can simply be defined using mask-based IoU as

$$c(h) = \begin{cases} \arg\max_{m \in M} \text{IoU}(h, m), & \text{if } \max_{m \in M} \text{IoU}(h, m) > 0.5 \\ \emptyset, & \text{otherwise.} \end{cases} \quad (1)$$

The set of true positives $\text{TP} = \{h \in H \mid c(h) \neq \emptyset\}$ is comprised of hypothesized masks which are mapped to a ground truth mask. Similarly, false positives are hypothesized masks that are not mapped to any ground truth mask, *i.e.* $\text{FP} = \{h \in H \mid c(h) = \emptyset\}$. Finally, the set $\text{FN} = \{m \in M \mid c^{-1}(m) = \emptyset\}$ of false negatives contains the ground truth masks which are not covered by any hypothesized mask.

In the following, let $pred : M \to M \cup \{\emptyset\}$ denote the latest tracked predecessor of a ground truth mask, or $\emptyset$ if no tracked predecessor exists. So $q = pred(p)$ is the mask q with the same id ($id_q = id_p$) and the largest $t_q < t_p$ such that $c^{-1}(q) \neq \emptyset$ [5]. The set IDS of id switches is then defined as the set of ground truth masks whose predecessor was tracked with a different id. Formally,

$$\text{IDS} = \{m \in M \mid c^{-1}(m) \neq \emptyset \wedge pred(m) \neq \emptyset \wedge \\ id_{c^{-1}(m)} \neq id_{c^{-1}(pred(m))}\}. \quad (2)$$

Mask-based Evaluation Measures. Additionally, we define a soft version $\widetilde{\text{TP}}$ of the number of true positives by

$$\widetilde{\text{TP}} = \sum_{h \in TP} \text{IoU}(h, c(h)). \quad (3)$$

Given the previous definitions, we define mask-based variants of the original CLEAR MOT metrics [4]. We propose the multi-object tracking and segmentation accuracy (MOTSA) as a mask IoU based version of the box-based MOTA metric, *i.e.*

$$\text{MOTSA} = 1 - \frac{|FN| + |FP| + |IDS|}{|M|} = \frac{|TP| - |FP| - |IDS|}{|M|}, \quad (4)$$

and the mask-based multi-object tracking and segmentation precision (MOTSP) as

---

[4]Sequences 2, 5, 9 and 11 were annotated.

[5]This definition corresponds to the one used by MOTChallenge. Note that the original KITTI tracking benchmark does not count id switches if the target was lost by the tracker.
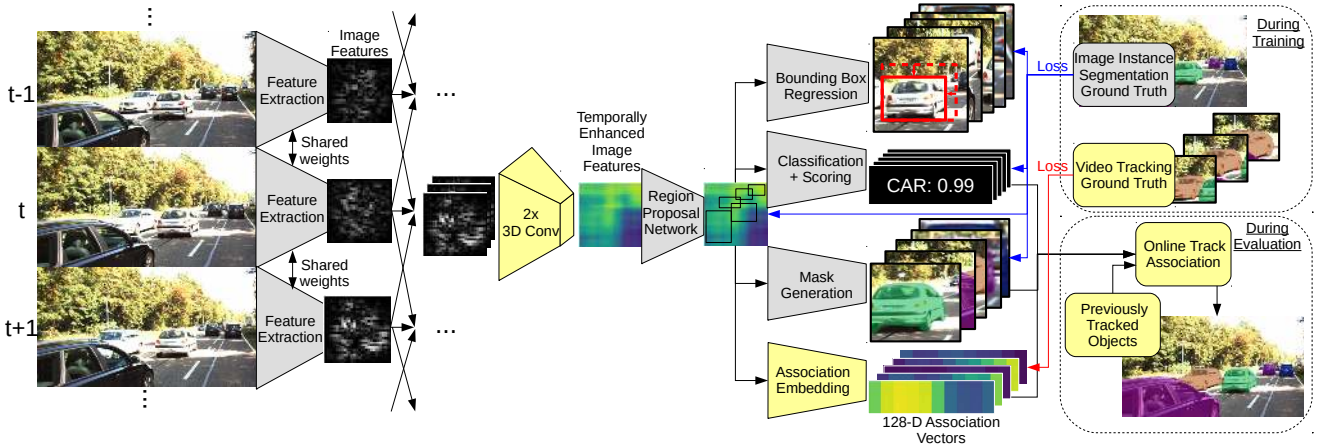
Figure 3: **TrackR-CNN Overview.** We extend Mask R-CNN by 3D convolutions to incorporate temporal context and by an association head that produces association vectors for each detection. The Euclidean distances between association vectors are used to associate detections over time into tracks. Differences to Mask R-CNN are highlighted in yellow.

$$\mathrm{MOTSP} = \frac{\widetilde{TP}}{|TP|}. \qquad (5)$$

Finally, we introduce the soft multi-object tracking and segmentation accuracy (sMOTSA)

$$\mathrm{sMOTSA} = \frac{\widetilde{TP} - |FP| - |IDS|}{|M|}, \qquad (6)$$

which accumulates the soft number $\widetilde{TP}$ of true positives instead of counting how many masks reach an IoU of more than 0.5. sMOTSA therefore measures segmentation as well as detection and tracking quality.

5. Method

In order to tackle detection, tracking, and segmentation, *i.e.* the MOTS task, jointly with a neural network, we build upon the popular Mask R-CNN [14] architecture, which extends the Faster R-CNN [48] detector with a mask head. We propose TrackR-CNN (see Fig. 3) which in turn extends Mask R-CNN by an association head and two 3D convolutional layers to be able to associate detections over time and deal with temporal dynamics. TrackR-CNN provides mask-based detections together with association features. Both are input to a tracking algorithm that decides which detections to select and how to link them over time.

Integrating temporal context. In order to exploit the temporal context of the input video [8], we integrate 3D convolutions (where the additional third dimension is time) into Mask R-CNN on top of a ResNet-101 [15] backbone. The 3D convolutions are applied to the backbone features in order to augment them with temporal context. These augmented features are then used by the region proposal network (RPN). As an alternative we also consider convolutional LSTM [53, 30] layers. Convolutional LSTM retains the spatial structure of the input by calculating its activations using convolutions instead of matrix products.

Association Head. In order to be able to associate detections over time, we extend Mask R-CNN by an association head which is a fully connected layer that gets region proposals as inputs and predicts an association vector for each proposal. The association head is inspired by the embedding vectors used in person re-identification [17, 5, 31, 55, 62]. Each association vector represents the identity of a car or a person. They are trained in a way that vectors belonging to the same instance are close to each other and vectors belonging to different instances are far away from each other. We define the distance $d(v, w)$ between two association vectors v and w as their Euclidean distance, *i.e.*

$$d(v, w) := \|v - w\|. \qquad (7)$$

We train the association head using the batch hard triplet loss proposed by Hermans *et al.* [17] adapted for video sequences. This loss samples hard positives and hard negatives for each detection. Formally, let $\mathcal{D}$ denote the set of detections for a video. Each detection $d \in \mathcal{D}$ consists of a mask $mask_d$ and an association vector a_d, which come from time frame t_d, and is assigned a ground truth track id id_d determined by its overlap with the ground truth objects. For a video sequence of T time steps, the association loss in the batch-hard formulation with margin α is then given by

$$\frac{1}{|D|} \sum_{d \in \mathcal{D}} \max \big(\max_{\substack{e \in \mathcal{D}: \\ id_e = id_d}} \|a_e - a_d\| - \min_{\substack{e \in \mathcal{D}: \\ id_e \neq id_d}} \|a_e - a_d\| + \alpha, 0 \big).$$
$$(8)$$

Mask Propagation. Mask-based IoU together with optical flow warping is a strong cue for associating pixel masks over time [42, 34]. Hence, we also experiment with mask warping as an alternative cue to association vector similarities. For a detection $d \in \mathcal{D}$ at time $t - 1$ with mask $mask_d$ and a detection $e \in \mathcal{D}$ with mask $mask_e$ at time t, we define

the mask propagation score [34, 42] as

$$\text{maskprop}(mask_d, mask_e) = \text{IoU}(\mathcal{W}(mask_d), mask_e), \quad (9)$$

where $\mathcal{W}(m)$ denotes warping mask m forward by the optical flow between frames $t-1$ and t.

Tracking. In order to produce the final result, we still need to decide which detections to report and how to link them into tracks over time. For this, we extend existing tracks with new detections based on their association vector similarity to the most recent detection in that track.

More precisely, for each class and each frame t, we link together detections at the current frame that have detector confidence larger than a threshold γ with detections selected in the previous frames using the association vector distances from Eq. 7. We only choose the most recent detection for tracks from up to a threshold of β frames in the past. Matching is done with the Hungarian algorithm, while only allowing pairs of detections with a distance smaller than a threshold δ. Finally, all unassigned high confidence detections start new tracks.

The resulting tracks can contain overlapping masks which we do not allow for the MOTS task (*c.f.* Section 4). In such a case, pixels belonging to detections with a higher confidence (given by the classification head of our network) take precedence over detections with lower confidence.

6. Experiments

Experimental Setup. For Mask R-CNN we use a ResNet-101 backbone [15] and pre-train it on COCO [29] and Mapillary [39]. Afterwards, we construct TrackR-CNN by adding the association head and integrating two depthwise separable 3D convolution layers with $3 \times 3 \times 3$ filter kernels each (two dimensions are spatial and the third is over time), ReLU activation, and 1024 feature maps between the backbone and the region proposal network. The 3D convolutions are initialized to an identity function after which the ReLU is applied. When using convolutional LSTM, weights are initialized randomly and a skip connection is added to preserve activations for the pretrained weights of subsequent layers during the initial steps of training. TrackR-CNN is then trained on the target dataset, *i.e.* KITTI MOTS or MOTSChallenge, for 40 epochs with a learning rate of $5 \cdot 10^{-7}$ using the Adam [22] optimizer. During training, mini-batches which consist of 8 adjacent frames of a single video are used, where 8 was the maximum possible number of frames to fit into memory with a Titan X (Pascal) graphics card. At batch boundaries, the input to the 3D convolution layer is zero padded in time. When using convolutional LSTM, gradients are backpropagated through all 8 frames during training and at test time the recurrent state is propagated over the whole sequence. The vectors produced by the association head have 128 dimensions and the association loss defined in Eq. 8 is computed over the detections

	sMOTSA		MOTSA		MOTSP	
	Car	Ped	Car	Ped	Car	Ped
TrackR-CNN (ours)	76.2	46.8	87.8	65.1	**87.2**	**75.7**
Mask R-CNN + maskprop	75.1	45.0	86.6	63.5	87.1	75.6
TrackR-CNN (box orig) + MG	75.0	41.2	87.0	57.9	86.8	76.3
TrackR-CNN (ours) + MG	76.2	**47.1**	87.8	**65.5**	87.2	75.7
CAMOT [42] (our det)	67.4	39.5	78.6	57.6	86.5	73.1
CIWT [41] (our det) + MG	68.1	42.9	79.4	61.0	86.7	75.7
BeyondPixels [52] + MG	**76.9**	-	**89.7**	-	86.5	-
GT Boxes (orig) + MG	77.3	36.5	90.4	55.7	86.3	75.3
GT Boxes (tight) + MG	**82.5**	**50.0**	**95.3**	**71.1**	**86.9**	**75.4**

Table 2: **Results on KITTI MOTS.** +MG denotes mask generation with a KITTI MOTS fine-tuned Mask R-CNN. BeyondPixels is a state-of-the-art MOT method for cars and uses a different detector than the other methods.

obtained in one batch. We choose a margin of $\alpha = 0.2$, which proved useful in [17]. For the mask propagation experiments, we compute optical flow between all pairs of adjacent frames using PWC-Net [54]. Our whole tracker achieves a speed of around 2 frames per second at test time. When using convolutional LSTM, it runs online and when using 3D convolutions in near-online fashion due to the two frames look-ahead of the 3D convolutions.

We tune the thresholds for our tracking system (δ, β, γ) for each class separately on the target training set with random search using 1000 iterations per experiment.

Main Results. Table 2 shows our results on the KITTI MOTS validation set. We achieve competitive results, beating several baselines. *Mask R-CNN + maskprop* denotes a simple baseline for which we fine-tuned the COCO and Mapillary pre-trained Mask R-CNN on the frames of the KITTI MOTS training set. We then evaluated it on the validation set and linked the mask-based detections over time using mask propagation scores (*c.f.* Section 5). Compared to this baseline, TrackR-CNN achieves higher sMOTSA and MOTSA scores, implying that the 3D convolution layers and the association head help with identifying objects in video. MOTSP scores remain similar.

TrackR-CNN (box orig) denotes a version of our model trained without mask head on the original bounding box annotations of KITTI. We then tuned for MOTA scores according to the original KITTI tracking annotations on our training split. We evaluate this baseline in our MOTS setting by adding segmentation masks as a post-processing step (denoted by +MG) with the mask head of the KITTI fine-tuned Mask R-CNN. sMOTSA and MOTSA scores for this setup are worse than for our method and the previous baseline, especially when considering pedestrians, adding to our observation that non-tight bounding boxes are not
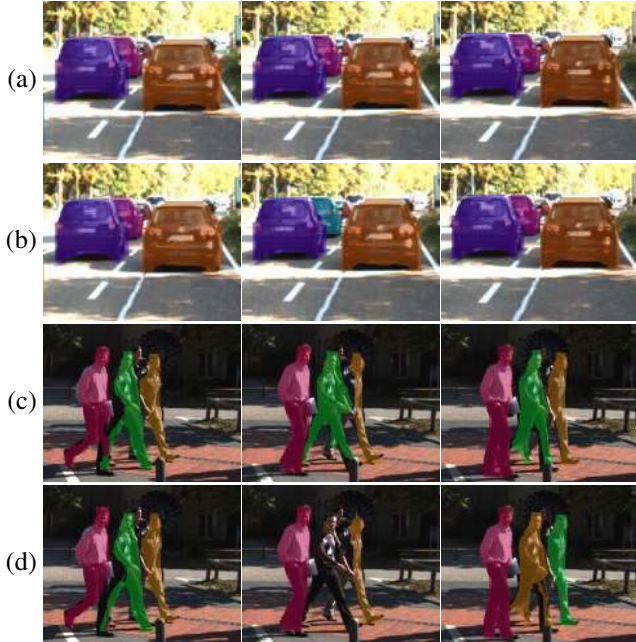
Figure 4: **Qualitative Results on KITTI MOTS.** (a)+(c) Our TrackR-CNN model evaluated on validation sequences of KITTI MOTS. (b)+(d) TrackR-CNN (box orig) + MG evaluated on the same sequences. Training with masks on our data avoids confusion between similar near-by objects.

Temporal component	sMOTSA		MOTSA		MOTSP	
	Car	Ped	Car	Ped	Car	Ped
1xConv3D	76.1	46.3	87.8	64.5	87.1	**75.7**
2xConv3D	76.2	**46.8**	87.8	**65.1**	87.2	**75.7**
1xConvLSTM	75.7	45.0	87.3	63.4	87.2	75.6
2xConvLSTM	76.1	44.8	**87.9**	63.3	87.0	75.2
None	**76.4**	44.8	**87.9**	63.2	**87.3**	75.5

Table 3: **Different Temporal Components for TrackR-CNN**. Comparison of results on KITTI MOTS.

an ideal cue for tracking and that simply using an instance segmentation method on top of bounding box predictions is not sufficient to solve the MOTS task. We show qualitative results for this baseline in Figure 4. The box-based model often confuses similar occluding objects for one another, leading to missed masks and id switches. In contrast, our model hypothesizes consistent masks.

To show that adding segmentation masks as done above does not give an unfair (dis)advantage, we also use the Mask R-CNN mask head to replace the masks generated by our method (*TrackR-CNN (ours) + MG*). The results stay roughly similar, so no major (dis)advantage incurs.

In combination, our baseline experiments show that training on temporally consistent instance segmentation data for video gives advantages both over training on instance segmentation data without temporal information and over training just on bounding box tracking data. Joint training on both was not possible before, which underlines the usefulness of our proposed MOTS datasets.

CAMOT [42] is a mask-based tracker which can track both objects from pre-defined classes and generic objects using 3D information from the stereo setup in KITTI. In the original version, *CAMOT* takes as input generic object proposals from SharpMask [44]. For better comparability, we used the detections from our TrackR-CNN (obtained by running it as a normal detector without association) as inputs instead. Note that *CAMOT* can only track regions for

which depth from stereo is available which limits its recall. The results show that our proposed tracking method performs significantly better than *CAMOT* when using the same set of input detections.

Since there are not many mask-based trackers with source code available, we also considered the bounding box-based tracking methods *CIWT* [41] and *BeyondPixels* [52] and again converted their results to segmentation masks using the KITTI fine-tuned Mask R-CNN mask head. Note that these methods were tuned to perform well on the original bounding box based task.

CIWT [41] combines image-based information with 3D information from stereo for tracking jointly in image and world space. Once more, detections from our TrackR-CNN were used for comparability. Our proposed tracking system which tackles tracking and mask generation jointly performs better than *CIWT* when generating masks post-hoc.

BeyondPixels [52] is one of the strongest tracking methods for cars on the original KITTI tracking dataset. It combines appearance information with 3D cues. We were not able to run their method with our detections since their code for extracting appearance features is not available. Instead we used their original detections which are obtained from RRC [47], a very strong detector. RRC achieves precise localization on KITTI in particular, while the more conventional Mask R-CNN detector was designed for general object detection. The resulting sMOTSA and MOTSA scores are higher than for our method, but still show that there are limits to state-of-the-art bounding box tracking methods on MOTS when simply segmenting boxes using Mask R-CNN.

MOTS Using Ground Truth Boxes. For comparison, we derived segmentation results based on bounding box ground truth and evaluated it on our new annotations. Here, 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. Again, we generated masks using the KITTI MOTS fine-tuned Mask R-CNN. Our results show that even with perfect track hypotheses generating accurate masks remains challenging, especially for pedestrians. This is even more the case when using amodal

Association Mechanism	sMOTSA		MOTSA		MOTSP	
	Car	Ped	Car	Ped	Car	Ped
Association head	**76.2**	**46.8**	**87.8**	**65.1**	**87.2**	**75.7**
Mask IoU	75.5	46.1	87.1	64.4	**87.2**	**75.7**
Mask IoU (train w/o assoc.)	74.9	44.9	86.5	63.3	87.1	75.6
Bbox IoU	75.4	45.9	87.0	64.3	**87.2**	**75.7**
Bbox Center	74.3	43.3	86.0	61.7	**87.2**	**75.7**

Table 4: **Different Association Mechanisms for TrackR-CNN**. Comparison of results on KITTI MOTS.

	sMOTSA	MOTSA	MOTSP
TrackR-CNN (ours)	**52.7**	**66.9**	**80.2**
MHT-DAM [21] + MG	48.0	62.7	79.8
FWT [16] + MG	49.3	64.0	79.7
MOTDT [31] + MG	47.8	61.1	80.0
jCC [20] + MG	48.3	63.0	79.9
GT Boxes (tight) + MG	**55.8**	**74.5**	**78.6**

Table 5: **Results on MOTSChallenge**. +MG denotes mask generation with a domain fine-tuned Mask R-CNN.

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. Further baselines, where we fill the ground truth boxes with rectangles or ellipses can be found in the supplemental material.

Temporal Component. In Table 3, we compare different variants of temporal components for TrackR-CNN. *1xConv3D* and *2xConv3D* means using either one or stacking two depthwise separable 3D convolutional layers between backbone and region proposal network, each with 1024 dimensions. Similarly, *1xConvLSTM* and *2xConvLSTM* denotes one or two stacked convolutional LSTM layers at the same stage with 128 feature channels each. The number of parameters per feature channel in a convolutional LSTM is higher due to gating. Using more feature channels did not seem to be helpful during initial experiments. Finally, *None* denotes adding no additional layers as temporal component. Compared to the *None* baseline, adding two 3D convolutions significantly improves sMOTSA and MOTSA scores for pedestrians, while performance for cars remains comparable. Surprisingly, using convolutional LSTM does not yield any significant gains over the baseline.

Association Mechanism. In Table 4, we compare different mechanism used for association between detections. Each line follows the proposed tracking system explained in Section 5, but different scores are used for the Hungarian matching step. When using the association head, association vectors may match with detections up to β frames in the past. For the remaining association mechanisms, only matching between adjacent frames is sensible.

For *Mask IoU* we only use mask propagation scores from Eq. 9, which degrades sMOTSA and MOTSA scores. This underlines the usefulness of our association head which can outperform an optical flow based cue using embeddings provided by a single neural network. Here, we also try training without the association loss (*Mask IoU (train w/o assoc.)*), which degrades MOTSA scores even more. Therefore, the association loss also has a positive effect on the detector itself. Surprisingly, using bounding box IoU (where the boxes were warped with the median of the optical flow values inside the box, *Bbox IoU*) performs almost the same as mask IoU. Finally, using only distances of bounding box

centers (*Bbox Center*) for association, *i.e.* doing a nearest neighbor search, significantly degrades performance.

MOTSChallenge. Table 5 shows our results on the MOTSChallenge dataset. Since MOTSChallenge only has four video sequences, we trained our method (*TrackR-CNN (ours)*) in a leaving-one-out fashion (evaluating each sequence with a model trained and tuned on the three others).

For comparison, we took pre-computed results of four methods that perform well on the MOT17 benchmark and generated masks using a Mask R-CNN fine-tuned on MOTSChallenge (in a leaving-one-out fashion) to evaluate them on our data. We note that all four sets of results use the strongest set of public detections generated with SDP [60], while TrackR-CNN generates its own detections. It is also unclear how much these methods were trained to perform well on the MOTChallenge training set, on which MOTSChallenge is based. Despite these odds, TrackR-CNN outperforms all other methods. The last line demonstrates that even with the tight ground truth bounding boxes including track information over time, segmenting all pedestrians accurately remains difficult.

7. Conclusion

Until now there has been no benchmark or dataset to evaluate the task of multi-object tracking and segmentation and to directly train methods using such temporally consistent mask-based tracking information. To alleviate this problem, we introduce two new datasets based on existing MOT datasets which we annotate using a semi-automatic annotation procedure. We further introduce the MOTSA and sMOTSA metrics, based on the commonly used MOTA metric, but adapted to evaluate all aspects of mask-based tracking. We finally develop a baseline model that was designed to take advantage of this data. We show that through training on our data, the method is able to outperform comparable methods which are only trained with bounding box tracks and single image instance segmentation masks. Our new datasets now make such joint training possible, which opens up many opportunities for future research.

References

[1] D. Acuna, H. Ling, A. Kar, and S. Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*, 2018. 3

[2] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 2

[3] M. Andriluka, J. Uijlings, and V. Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. *arXiv preprint arXiv:1806.07527*, 2018. 3

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image and Video Processing*, 2008. 4

[5] L. Beyer, S. Breuers, V. Kurin, and B. Leibe. Towards a principled integration of multi-camera re-identification and tracking through optimal bayes filters. In *CVPRW*, 2017. 5

[6] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 3

[7] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 1

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 5

[9] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017. 3

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3

[11] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. In *ECCV*, 2008. 3

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 2, 3

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 2, 3, 5

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[16] R. Henschel, L. Leal-Taix, D. Cremers, and B. Rosenhahn. Fusion of head and full-body detectors for multi-object tracking. In *CVPRW*, 2018. 8

[17] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5, 6

[18] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. *arXiv preprint arXiv:1803.06184*, 2018. 2

[19] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *ECCV*, 2014. 2

[20] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *PAMI*, 2018. 8

[21] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *ICCV*, 2015. 8

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[23] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. *arXiv preprint arXiv:1801.00868*, 2018. 4

[24] Y. J. Koh and C. S. Kim. CDTS: Collaborative detection, tracking, and segmentation for online multiple object segmentation in videos. In *ICCV*, 2017. 2

[25] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin. A novel performance evaluation methodology for single-target trackers. *PAMI*, 2016. 1

[26] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2

[27] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth. Tracking the trackers: an analysis of the state of the art in multiple object tracking. *arXiv preprint arXiv:1704.02781*, 2017. 2

[28] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 2

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3, 6

[30] M. Liu and M. Zhu. Mobile video object detection with temporally-aware feature maps. *CVPR*, 2018. 5

[31] C. Long, A. Haizhou, Z. Zijie, and S. Chong. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *ICME*, 2018. 5, 8

[32] D. Lowell, Z. C. Lipton, and B. C. Wallace. How transferable are the datasets collected by active learners? *arXiv preprint arXiv:1807.04801*, 2018. 3

[33] Y. Lu, C. Lu, and C.-K. Tang. Online video object detection using association LSTM. In *ICCV*, 2017. 2

[34] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 3, 5, 6

[35] S. Lyu, M.-C. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. Del Coco, et al. UA-DETRAC 2017: Report of AVSS2017 & IWT4S challenge on advanced traffic monitoring. In *Advanced Video and Signal Based Surveillance (AVSS)*, 2017. 2

[36] S. Manen, M. Gygli, D. Dai, and L. Van Gool. Pathtrack: Fast trajectory annotation with path supervision. In *ICCV*, 2017. 2

[37] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 4

[38] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid. Joint tracking and segmentation of multiple targets. In *CVPR*, 2015. 2

[39] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 3, 6

[40] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *PAMI*, 2014. 2

[41] A. Ošep, W. Mehner, M. Mathias, and B. Leibe. Combined image- and world-space tracking in traffic scenes. In *ICRA*, 2017. 6, 7

[42] A. Ošep, W. Mehner, P. Voigtlaender, and B. Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. *ICRA*, 2018. 2, 5, 6, 7

[43] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2

[44] P. Pinheiro, T. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *ECCV*, 2016. 7

[45] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2

[46] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 2

[47] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu. Accurate single stage detector using recurrent rolling convolution. In *CVPR*, 2017. 7

[48] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5

[49] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004. 3

[50] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *ICCV*, 2017. 2

[51] G. Seguin, P. Bojanowski, R. Lajugie, and I. Laptev. Instance-level video segmentation from object tracks. In *CVPR*, 2016. 2

[52] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *ICRA*, 2018. 6, 7

[53] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 5

[54] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 6

[55] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multi people tracking with lifted multicut and person re-identification. In *CVPR*, 2017. 5

[56] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *NeurIPS*, 2011. 3

[57] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, and S. Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint arXiv:1511.04136*, 2015. 2

[58] N. Xu, B. L. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *CVPR*, 2016. 3

[59] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 2

[60] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *CVPR*, 2016. 8

[61] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv: 1805.04687*, 2018. 2

[62] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2017. 5