Supplementary Material:
Assignment-Space-based Multi-Object Tracking and Segmentation

In this section, we provide additional details and analysis of the proposed approach for Multi-Object Tracking and Segmentation. In Sec. A we elaborate on the parameter learning procedure for MOTS that has been discussed in Sec. 3.3. In Sec. B, we compare the time-complexities of methods working in the detection space with our method operating in the assignment space. In Sec. C, we discuss experimental details: datasets (Sec. C.1), additional ablation study on KITTI-MOTS (Sec. C.2) and some additional qualitative results (Sec. C.3).

Fig. 6 visualizes the tracks formed by our method across multiple video frames from 4 video sequences of the KITTI-MOTS validation dataset. The different colors represent different objects in the respective videos. In the image, x and y are the horizontal and vertical axes of each video frame and t represents the progress of the respective videos from the frame at time t_1 to the frame at time T.

A. Learning of Parameters

In this section we detail the parameter learning procedure that has been discussed in Sec. 3.3. Eq. (6) shows the learning objective. We minimize this objective w.r.t. the tracking parameters (λ) and the deep-net parameters (θ). In Sec. 4, we evaluate how the approach described in Sec. 3 performs.

In all our experiments (described in Sec. 4), “Ours” represents the approach where we optimize objective Eq. (6) only with respect to λ. For Tab. 2 and Tab. 3, we learn λ on the KITTI-MOTS training dataset using stochastic gradient descent with a learning rate of 0.05 for 50 epochs. The optimal $\lambda = [\lambda_{\text{iou}}, \lambda_{\text{app}}, \lambda_{\text{dist}}, \lambda_{\text{iou},2}, \lambda_{\text{app},2}, \lambda_{\text{dist},2}]$ learnt for cars, on PointTrack [58] detections are $[-6.93, 4.96, 0.47, -1.80, -0.45, 0.10]$.

“Ours (JT)” in Tab. 1, and Tab. 3 represents the approach where we optimize the objective given in Eq. (6) with respect to both λ and θ. Here λ is first trained for 40 epochs using stochastic gradient descent with a learning rate of 0.05. θ is initialized with the weights from the pre-trained detection-segmentation network (MaskRCNN [14] and PointTrack [58] respectively for pedestrians and cars). MaskRCNN [14] is previously trained on the COCO and Cityscapes datasets. We use the pre-trained weights from https://dl.fbaipublicfiles.com/detectron2/Misc/cascade_mask_rcnn_X_152_32x8d_FPN_IN5k_gn_dconv/18131413/model_0039999_e76410.pkl. Afterwards, we use the refinement net [29] to improve mask quality, a procedure used in https://motchallenge.net/workshops/bmtt2020/tracking.html.

For PointTrack [58], we use the SpatialEmbedding Network [35] trained on the KINS dataset and fine-tuned on the KITTI-MOTS dataset as described in [58]. We jointly train θ and λ on the KITTI-MOTS training dataset using our learning objective (Eq. (6)) for 10 epochs. We use stochastic gradient descent with a learning rate of 10^{-6} for θ and 0.05 for λ. Note that SpatialEmbedding [35] or PointTrack [58] networks are not available for pedestrians, so we use MaskRCNN for pedestrians instead.

Training data $\mathcal{T}$. As discussed in Sec. 3.3, we are given a set of detections $\mathcal{D}$ and our task is to find (λ, θ) such that the learning objective (Eq. (6)) is minimized. Our training set is $\mathcal{T} = \{(x, y_{\text{GT}})\}$, as mentioned in Sec. 3.3, where x is a video clip of T frames and $y_{\text{GT}} = (y^2_{\text{GT}}, \dots, y^T_{\text{GT}})$ denotes a sequence of elements $y^t_{\text{GT}} \in \mathcal{Y}^t$ that refer to the ground truth assignment $a^t_{y^t_{\text{GT}}}$ of objects $\mathcal{D}^{t-1}$ and $\mathcal{D}^t$ between frames $t-1$ and t. In order to construct $a^t_{y^t_{\text{GT}}}$ from given detections $\mathcal{D}^{t-1}$ and $\mathcal{D}^t$ for frames $t-1$ and t, we first associate the ground truth detections to our given detections $\mathcal{D}^t$ based on mask overlap. Following this, we find the ground truth assignment matrix $a^t_{y^t_{\text{GT}}}$.

B. Time Complexity Analysis

In this section we discuss the time complexities of a typical detection-space based network flow method discussed in Sec. 2 and our assignment-space based approach. For a total of N objects across T video frames, the time-complexity of a general network flow approach (*e.g.*, [59]) in the detection space is at least $O(TN^3)$. This is because we find N best paths, and optimizing for one best path for T frames has a complexity of $O(TN^2)$. The complexity is higher if the track hypotheses depend on locations, as seen in Multi-Hypothesis Tracking (MHT) discussed in Sec. 2. If n is the number of objects in a single frame (for simplicity, let's assume n is constant across frames), the complexity of our assignment-space based method is Kn^3T for K-best assignments per frame-pair. The Hungarian-Murty [34] algorithm ($O(Kn^3)$) is performed for $T-1$ frame pairs, and a single best path ($O(K^2T)$) is obtained afterwards. The total complexity is $O(K^2T + Kn^3T) = O(K^2n^3T)$. For us $K \leq 20$. Note that $n = N$ only when all objects in the video appear in all frames. Usually, $n \ll N$ making the assignment space more efficient.

Note that many network-flow based methods in the detection space (discussed in Sec. 2) use additional approximations to prune the detection-space, *i.e.*, to reduce the complexity. For the assignment space, we don't use additional approximations beyond focusing on the K best as-

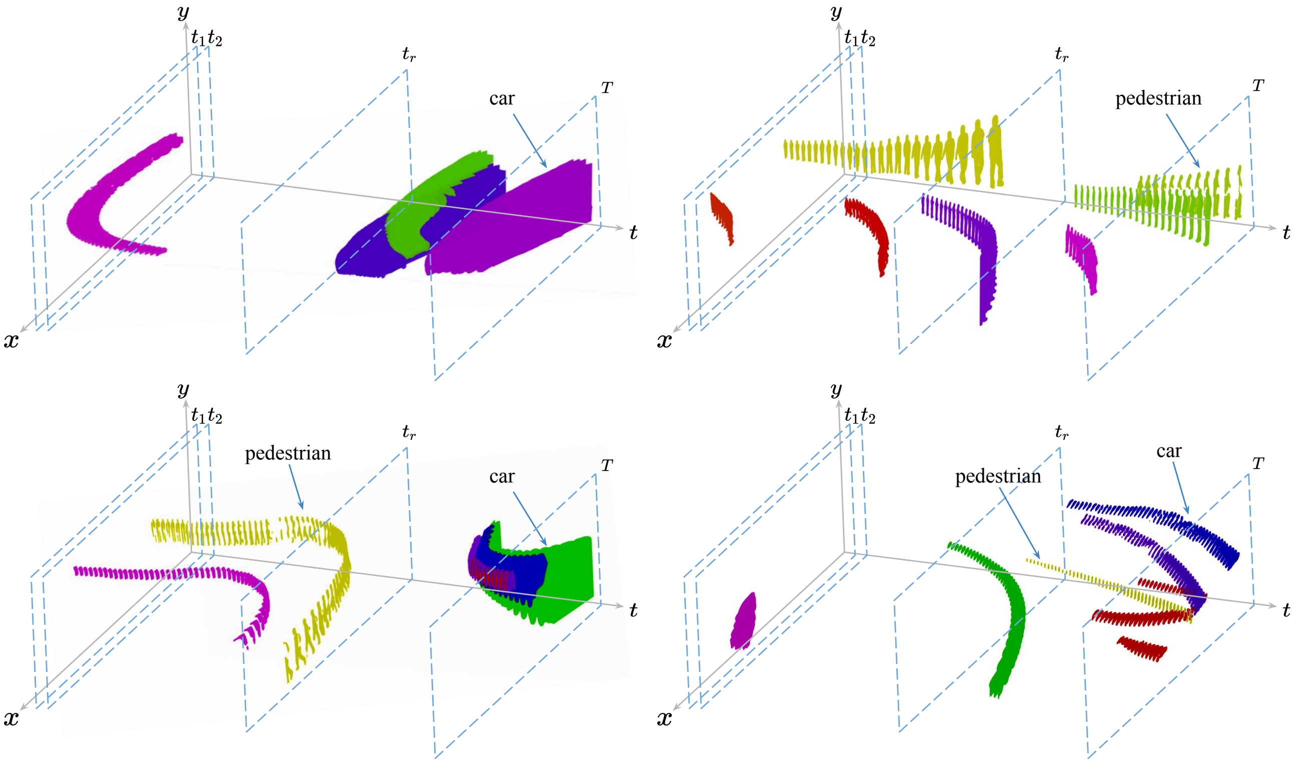

Figure 6. The tracks evaluated by our method as discussed in Sec. 3 across video frames for 4 different sequences of the KITTI-MOTS validation dataset. The different colors represent different objects. x and y are the horizontal and vertical axes of each video frame and t represents the progress of video frames from time t_1 to time T.

signments. The method optimizes for a single path. It is hence possible to optimize even for a large number of temporal steps T. In our experiments, T refers to all the frames in a given video.

C. Additional Experiments

In this section, we provide additional analysis and details to support the results discussed in Sec. 4. In Sec. C.1, we discuss the datasets used to study the approach on the MOTS and MOT tasks. In Sec. C.2 and Sec. C.3 we discuss some additional ablation studies on the KITTI-MOTS dataset and provide some qualitative results.

C.1. Datasets

KITTI-MOTS: The KITTI dataset [12] contains sequences of traffic scenes captured from a moving car. For the task of MOTS, Voigtlaender *et al.* [54] created pixel-level mask annotations for every frame for 21 video sequences from the KITTI training data. The annotations are for cars and pedestrians. The 21 training sequences from KITTI are split into training and validation sets of the KITTI-MOTS dataset. The split balances the number of cars and pedestrians roughly equally across the training and validation sets. More specifically, there are 12 training and 9 validation sequences in the KITTI-MOTS dataset, consisting of $5,027$ and $2,981$ frames, 99 and 66 distinct pedestrian IDs and 431 and 151 distinct car IDs respectively.

MOTSChallenge: This data consists of 4 out of 7 sequences of pedestrians in crowded scenes from the MOTChallenge [31] training data. Voigtlaender *et al.* [54] provide pixel-level annotations for each frame on all 4 sequences for a total of $2,862$ frames in which there are 288 distinct pedestrian IDs. This is a particularly challenging dataset due to heavy occlusions.

MOT17: MOT17 is a MOTChallenge tracking benchmark consisting of challenging pedestrian tracking sequences, with frequent occlusions and crowded scenes. They contain sequences with varying viewing angle, size and number of objects, camera motion and frame rate. Note that although MOT20 is more recent, MOT17 is more widely used; there are many more MOT17 submissions on the leaderboard. We test our approach on MOT17 using the detections provided by the MOTChallenge to ensure a fair comparison with other MOT batch-methods.

C.2. Additional Ablation Study on KITTI-MOTS

In Tab. 3, Tab. 4 and Tab. 5, we analyze the effects of different parameters and components on our tracking ap-

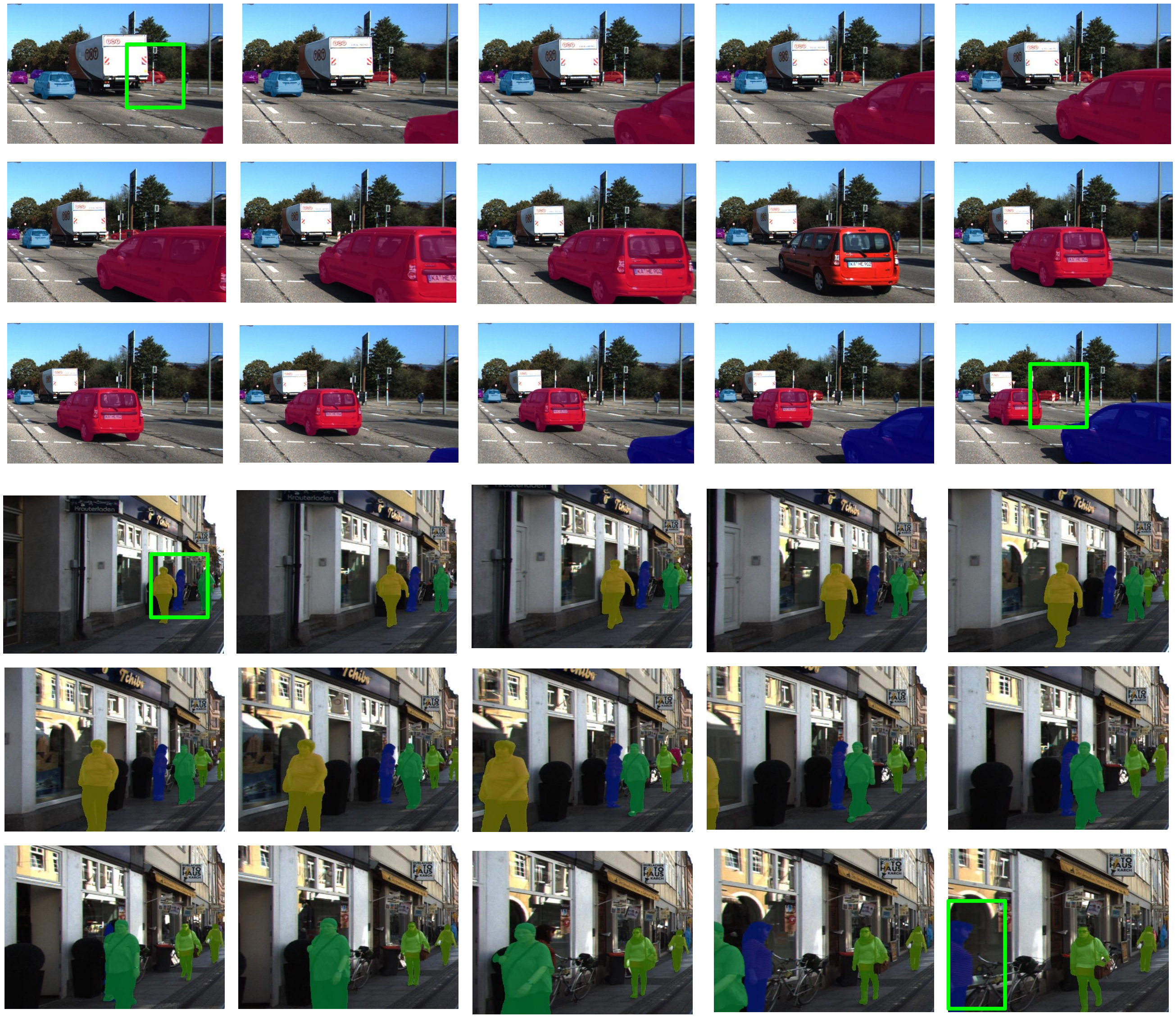

Figure 7. Two difficult cases where our method works well. The first 3 and the last 3 rows each show a sequence of 15 video frames placed one after the other. The first 3 rows represent a sequence of video frames showing a few occluded cars. The car marked with green boxes in the first and last frames is successfully tracked across all the frames, despite heavy occlusion from the large red car. A similar situation is observed in the last 3 rows. The person marked with the green box is tracked successfully after multiple video frames.

proach. Tab. 8 provides additional analysis. We study the effects of using bounding box based (as opposed to mask-based) intersection over union (IoU) in f_{iou} and $f_{\text{iou},2}$ as described in Eq. (1) and Eq. (4). This is represented by "Ours (BB+WL)". "WL" represents the 'without long range' configuration as described in the ablation study of Sec. 4.2. Specifically, this refers to the configuration without the long range assignments described in Sec. 3.4. We see that mask-based IoU ("Ours (WL)") performs better than bounding box based IoU at preserving the respective identities of objects. For mask-based IoU ("Ours (WL)"), the HOTA score is 81.3% as compared to only 79.2% for bounding-box-based IoU ("Ours (BB+WL)"). This establishes the importance of using segmentations for tracking. We also study the effects of the different modalities used in the long range assignments as described in Sec. 3.4. "Ours (LR: Dist)" represents the configuration where we only use the Euclidean distance between the bounding box centers for the long range assignments (Sec. 3.4). "Ours (LR: App)" represents the configuration where we only use the Euclidean distance between the appearance feature vectors for long range assignments (Sec. 3.4). We notice that the performance remains unchanged if we remove distance between the bounding box centers as a modality for long range assignments. This indi-

Method	Dets.	HOTA	DetA	AssA	LocA	IDF1	sMOTSA	IDS$\downarrow$
Ours(BB+WL)	Pt. [58]	79.2	**85.6**	74.0	**91.1**	81.4	84.0	145
Ours (WL)	Pt. [58]	**81.3**	**85.6**	**77.8**	**91.1**	**85.3**	**84.9**	**77**
Ours (LR:Dist)	Pt. [58]	81.7	**85.6**	78.4	**91.1**	86.1	84.9	72
Ours (LR:App)	Pt. [58]	83.3	**85.6**	81.6	**91.1**	**89.3**	**85.4**	**22**
Ours	Pt. [58]	**83.4**	**85.6**	**81.8**	**91.1**	**89.4**	**85.4**	**22**

Table 8. Additional Ablation Study on the KITTI-MOTS validation set (cars).

cates that the discriminative appearance features, obtained from PointTrack [58] perform very well in re-identifying objects after multiple frames. Bounding box center distances are redundant when using these appearance features. However, if we do not have access to discriminative appearance features, bounding-box-based long range assignments help in reducing the identity switches. "Ours (LR: Dist)" performs better at preserving identities than "Ours (WL)", validating this claim. The last row in Tab. 8 (Ours) represent the configuration where both the distance modality and appearance modality are used. Here, the values of $\lambda_{\mathrm{dist,lr}}$ and $\lambda_{\mathrm{app,lr}}$ (Sec. 3.4) are 0.05 and 0.98 respectively. These values have been used for long range assignments in all experiments in this paper.

C.3. Additional Qualitative Results

In this subsection, we provide some examples of difficult cases where our approach is particularly effective in preserving the identities of objects.

Fig. 7 shows two cases with heavy occlusion where our method can effectively preserve the identities of objects. The first 3 and last 3 rows each show a sequence of 15 video frames placed one after the other. The first 3 rows represent a sequence of video frames showing a few occluded cars. The car marked with a green box in the first and last frames is successfully tracked across the frames, despite heavy occlusion from the large red car. A similar situation is observed in the last 3 rows. The person marked with the green box is tracked successfully despite being occluded for multiple video frames.