DeepAutoTrack (DAT): Vehicle Trajectory Prediction for Autonomous Driving

We would first like to thank the reviewers for the supportive comments and the constructive criticisms. Your valuable feedbacks help us further improve the analysis and the overall presentation of the paper. Because of the significant improvement made to the new version of the paper, we would kindly ask reviewers refer to the new version of the paper. Due to space constraint, we address only the major criticisms below.

Modifications: (1) We have implemented SocialLSTM [2] as requested by R2 and preliminary results on 2D space exhibit that the inclusion of "social interaction" further improves the prediction accuracy. (2) We have restructured our methodology in Sec. 3 and further explain the intuition behind the chosen architectures.

Prior works on sequence prediction: We agree that the first version of the draft is indeed weak in presenting prior works on sequence prediction. We have restructured our related works. Specifically, we have supplemented works on activity forecasting/trajectory prediction works in the "Activity forecasting" part in Sec 2.

SYNTHIA and release of the dataset: Currently, experiments on the SYNTHIA dataset [27, 37] focus on single frame, static images for semantic labeling or scene representation. To our knowledge, our paper is the first attempt to utilize *continuous video streams* for the prediction problem. We would like to release publicly-available dataset for prediction challenge based on SYNTHIA to spur future research upon paper publication.

Instance association via tracking: We would like to clarify the overloaded term "tracking by detection" used in the first version of our manuscript. Most of the modern trackers focus on the problem of "class-agnostic" generic object tracking[1]. In the context of autonomous driving, however, we know in advance the classes of object of interests for tracking (*e.g.*, cars, pedestrians, cyclists). Therefore, with this extra source of information, we propose to use the neural nets object detection template to forcibly update the tacker's model to overcome the problem of model drifting. Hence the tracker is served solely as *instance associator* for the traffic participant. During our qualitative examination, we find out that this simple combination achieves more robust result for instance association than those of state-of-the-art "class-agnostic" trackers.

R1: Explain the intuition of three LSTMs.

R2: Our preliminary result using 2D information shows that social tensor generates slightly better prediction than the isolation case (Tab.4, Fig.7, bottom row). Note that our mean error in unit space is much lower than the pedestrian trajectory prediction in Social-LSTM[2] (0.049 vs. 0.27). We analyze the following two factors: car trajectory is a simpler case: both in time (prediction frame is 8 in our case and 12 in [2]) and in space (car trajectory is more linear and confined in our problem setting). However, the relevant improvement using social tensor is lower for the car trajectory prediction (4% vs. 38%). We reckon it's due to only around 3 cars of interests on average appearing in our high-way driving scene. And we believe in a more complex social scene such as urban driving, the improvement will be more pronounced.

R2: The "avg. disp. error" is essentially the same as the "center error" in Tab.2 of our paper. The subtle difference is that the former is in unit space and the latter is in pixel space. The "avg. non-linear disp. error" considers the non-linear turns from human-human interactions and the heuristic for choosing the non-linear regions of a trajectory is problem dependent. The metrics adopted in our paper consider the coverage of the target from a tracking perspective. Moreover, the 3D occupancy grid is more pertinent in the context of autonomous driving than the 2D co-ordinates.

R2: Semantics in our paper are presented both in the training and the test time and we rectified them as as "auxiliary" information. They are used directly as a tensor embedding for the recurrent network. In the supplementary material we try to avoid extra cluttering, hence we visualized only one object being tracked. Nonetheless, multiple objects are detected and tracked simultaneously in our system and we redraw the figures to make multiple tracking more conspicuous.

R3: We thank the reviewer for the very meticulous review. We have indeed noticed the recently released photo-realistic simulators CARLA and Sim4CV. And we found that CARLA is especially suited as a further extension to verify our proposed system due to the availability of continuous video streams and ground truth segmentation. From the semantic segmentation results of Synthia, we have observed that model trained on synthetic dataset produced good segmentations by itself on real datasets, and dramatically boosted accuracy in combination with real data. Hence, we believe the hybrid of synthetic data and real-world data would be the most data efficient and promising way for the model generalization.

[1] In order to adapt to temporal changes, a continuous learning strategy is applied, where the model is updated rigorously in every frame. This excessive update strategy causes both lower frame-rates and degradation of robustness due to model drifting caused by scale variations, deformations, and out-of-plane rotations.