LaTeX Guidelines for Author Response

1. Introduction

We would first like to thank the reviewers for your valuable feedback which helped us to improve the analysis of the proposed method and the overall presentation of the paper. Due to space constraints with significant improvement made to the new version of the paper, we would kindly ask reviewers refer the new version of the paper. we address only the major criticisms below.

1.1. Common

1) Unlike stated in the paper, vehicle tracking in images has been studied for years. There are summary papers (e.g. VEHICLE DETECTION AND TRACKING TECHNIQUES: A CONCISE REVIEW Raad Ahmed Hadi1,2 , Ghazali Sulong1 and Loay Edwar George, SIPIJ). There is no mention of prior art.

Where are comparisons with other activity forecasting/trajectory prediction works?

I think the authors should reimplement SocialLSTM for 2D data (Stanford UAV dataset seems to be a 2D standard prediction task – just throw away images and just use the xy plane coordinates, e.g. center of detected bounding box). There is no comparison to existing methods, the authors mention that this is due to the lack of existing datasets.

Who else has experimented on the SYNTHIA dataset? Or are the authors the first ones? Are the authors releasing a specific publicly-available prediction challenge based on SYNTHIA? It is not clear if the dataset is released with the paper.

Currently, experiments on the SYNTHIA dataset [1, 2] 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 problems. We would like to release publicly-available prediction challenge based on Synthia to spur future research upon paper publication.

1.2. R1

They adopt fDSST[6] and use it to track detections. What is the novelty, except for the implementation with specific parameters?

3) 3 very similar LSTMs with different input

4) The one LSTM with more information performs better. That is not surprising and not a real insight.

There are a number of inaccuracies e.g. line 044, where the authors state that automated driving models are either neural networks or controllers with if-then-else rules or line 304, that states that the closest related work to their tracking is an ego-motion regression by X-u,Gao and Darrell Line 471 calls Synthia images photo-realistic.

The actual contributions, and evaluation could be clearer. It took some time to understand the different metrics for detection and tracking as they are supposed to be fused.

The actual contribution is not significant enough and could be described more clearly.

1.3. R2

SocialLSTM used "Avg. disp. Error", "Avg. non-linear disp. error" , and "final disp. Error" What was the justification for the metrics the authors chose?

The "Average displacement error" is

Line 124 - The authors discuss a "lack of proper metrics for evaluating the temporal prediction result." How are the current metrics poorly suited to the task? What would be better?

In the SocialLSTM papers Related Work section, there are about 20 sources that focus on Activity forecasting

The details about the fusion of location and semantic information into the LSTM hidden state seemed missing (was a bit vague). A figure of the architecture of the SEG-LSTM would go a long way in clarifying this (or are the authors using Xu et al.'s "End-to-end Learning of Driving Models from Large-scale Video Datasets" FCN-LSTM from Trevor Darrell's group?).

How is this more than just an engineering pipeline, combining tracking (distilling bounding boxes to coordinates), and then adding an LSTM? SocialLSTM already did the second part in principle.

I was a bit confused about why the tracking-by-detection framework was novel – how is it different from Tracking-learning-detection (TLD),from Z. Kalal, K. Mikolajczyk, and J. Matas. In TPAMI 2012?

Why are scene semantics privileged (which you also call auxiliary) information? If the pipeline relies on fusing the location and semantic stream, if privileged information is missing at test time, then wont the pipeline break? (privileged information is usually present at train time, but absent at test time).

Can multiple objects be tracked at once, or just one? The video in the supplementary material only showed one object being tracked. What was the average number of objects in each SYNTHIA scene frame?

The focus of the paper seems a bit unclear: apparently, the goal is to predict cars future odometry given previous egomotion visual input. But the paper is about pre-

diction of other vehicles trajectories, not about predicting ones own car odometry (which is different).
I felt like the detection/tracking theme slightly bogs down the main focus of prediction in the paper. I think refocusing it on prediction would strengthen the argument.

1.4. R3

There are too many contributions. For example, I would combine 1 and 3 and maybe also 2 and 4. In general, it is better to have a few strong contributions than many weak ones. Section 3 is quite short and it seems that the tracking methodology would fit well here. In general a lot of methodology detail is actually in section 4 (implementation details). Maybe consider reorganizing a little bit.
In a real car accurate depth information might not be available. Also, tracking and semantic segmentation results in the real world will probably be noisier than for synthetic data. Experiments on real data would make the results much stronger.
Figure 6: It would be nice to have some state-of-the-art tracking results as reference e.g. ECO, STAPLE CA, SiamFC
Do you think the results you obtained with the Synthia data will transfer to real-world data?
Not exactly a request, but maybe something to consider for future work: There are photo-realistic simulators available now (CARLA, Sim4CV, etc.) which allow simulation of autonomous driving with real physics and photo-realism similar to Synthia. It would be interesting to implement this system and evaluate real-world performance.

References

[1] D. Hernandez-Juarez, L. Schneider, A. Espinosa, D. Vazquez, A. M. Lopez, U. Franke, M. Pollefeys, and J. C. Moure. Slanted stixels: Representing san franciscos steepest streets. In *British Machine Vision Conference*, 2017. 1

[2] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1