Egocentric Vision-based Future Vehicle Localization for Intelligent Driving Assistance Systems

Yu Yao[1*], Mingze Xu[2*], Chiho Choi[3], David J. Crandall[2], Ella M. Atkins[1], Behzad Dariush[3]

Abstract*— Predicting the future location of vehicles is essential for safety-critical applications such as advanced driver assistance systems (ADAS) and autonomous driving. This paper introduces a novel approach to simultaneously predict both the location and scale of target vehicles in the first-person (egocentric) view of an ego-vehicle. We present a multi-stream recurrent neural network (RNN) encoder-decoder model that separately captures both object location and scale and pixel-level observations for future vehicle localization. We show that incorporating dense optical flow improves prediction results significantly since it captures information about motion as well as appearance change. We also find that explicitly modeling future motion of the ego-vehicle improves the prediction accuracy, which could be especially beneficial in intelligent and automated vehicles that have motion planning capability. To evaluate the performance of our approach, we present a new dataset of first-person videos collected from a variety of scenarios at road intersections, which are particularly challenging moments for prediction because vehicle trajectories are diverse and dynamic. *Code and dataset have been made available at: https://usa.honda-ri.com/hevi

I. Introduction

Safe driving requires not just accurately identifying and locating nearby objects, but also predicting their *future* locations and actions so that there is enough time to avoid collisions. Precise prediction of nearby vehicles' future locations is thus essential for both autonomous and semi-autonomous (e.g., Advanced Driver Assistance Systems, or ADAS) driving systems as well as safety-related systems [1]. Extensive research [2]–[4] has been conducted on predicting vehicles' future actions and trajectories using overhead (bird's eye view) observations. But obtaining overhead views requires either an externally-mounted camera (or LiDAR), which is not common on today's production vehicles, or aerial imagery that must be transfered to the vehicle over a network connection.

A much more natural approach is to use forward-facing cameras that record the driver's "first-person" or "egocentric" perspective. In addition to being easier to collect, the first-person perspective captures rich information about the object appearance, as well as the relationships and interactions between the ego-vehicle and objects in the environment. Due to these advantages, egocentric videos have been directly used in applications such as action recognition [5], [6], navigation [7]–[9], and end-to-end autonomous driving [10]. For trajectory prediction, some work has simulated bird's eye views by projecting egocentric video frames onto the ground plane [2], [3], but these projections can be incorrect due to road irregularities or other sources of distortion, which prevent accurate vehicle position prediction.

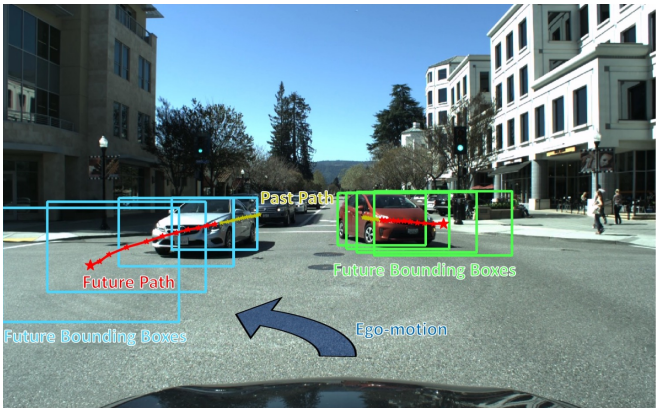

Fig. 1: Illustration of future vehicle localization. Location and scale are represented as bounding boxes in predictions.

This paper considers the challenging problem of predicting relative future locations and scales (represented as bounding boxes in Figure 1) of nearby vehicles with respect to an ego-vehicle equipped with an egocentric camera. We introduce a multi-stream RNN encoder-decoder (RNN-ED) architecture to effectively encode past observations from different domains and generate future bounding boxes. Unlike other work that has addressed prediction in simple scenarios such as freeways [2], [3], we consider urban driving scenarios with a variety of multi-vehicle behaviors and interactions.

The contributions of this paper are three-fold. First, our work present a novel perspective for intelligent driving systems to predict vehicle's future location under egocentric view and challenging driving scenarios such as intersections. Second, we propose a multi-stream RNN-ED architecture for improved temporal modeling and explicitly capturing vehicles' motion as well as their appearance information by using dense optical flow and future ego-motion as inputs. Third, we publish a new first-person video dataset — the Honda Egocentric View - Intersection (HEV-I) dataset — collected in a variety of scenarios involving road intersections. The dataset includes over $2,400$ vehicles (after filtering) in 230 videos. We evaluate our approach on this new proposed dataset, along with the existing KITTI dataset, and achieve the state-of-the-art results.

[1]Robotics Institute, University of Michigan, Ann Arbor, MI 48109, USA. {brianyao,ematkins}@umich.edu
[2]School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47408, USA. {mx6,djcran}@indiana.edu
[3]Honda Research Institute, Mountain View, CA 94043, USA. {cchoi,bdariush}@honda-ri.com
*This work was done when Yu Yao and Mingze Xu were interns at Honda Research Institute, Mountain View, CA 94043, USA.

II. Related Work

Egocentric Vision. An egocentric camera view is often the most natural perspective for observing an ego-vehicle environment, but it introduces additional challenges due to its narrow field of view. The literature in egocentric visual perception has typically focused on activity recognition [5], [6], [11]–[13], object detection [14]–[16], person identification [17]–[19], video summarization [20], and gaze anticipation [21]. Recently, papers have also applied egocentric vision to ego-action estimation and prediction. For example, Park *et al*. [22] proposed a method to estimate the location of a camera wearer in future video frames. Su *et al*. [23] introduced a Siamese network to predict future behaviors of basketball players in multiple synchronized first-person views. Bertasius *et al*. [24] addressed the motion planning problem for generating an egocentric basketball motion sequence in the form of a 12-d camera configuration trajectory.

More directly related to our problem, two recent papers have considered predicting pedestrians' future locations from egocentric views. Bhattacharyya *et al*. [25] model observation uncertainty using Bayesian Long Short-Term Memory (LSTM) networks to predict the distribution of possible future locations. Their technique does not try incorporate image features such as object appearance. Yagi *et al*. [26] use human pose, scale, and ego-motion as cues in a convolution-deconvolution (Conv1D) framework to predict future locations. The specific pose information applies to people but not to other on-road objects like vehicles. Their Conv1D model captures important features of the activity sequences but does not explicitly model temporal updating along each trajectory. In contrast, our paper proposes a multi-stream RNN-ED architecture using past vehicle locations and image features as inputs for predicting vehicle locations from egocentric view.

Trajectory Prediction. Previous work on vehicle trajectory prediction has used motion features and probabilistic models [2], [27]. The probability of specific motions (e.g., lane change) is first estimated, and the future trajectory is predicted using Kalman filtering. Computer vision and deep learning techniques achieved convincing results in several fields [28]–[30], and have been recently investigated for trajectory prediction. Alahi *et al*. [31] proposed Social-LSTM to model pedestrian trajectories as well as their interactions. The proposed social pooling method was then improved by Gupta *et al*. [32] to capture global context for a Generative Adversarial Network (GAN). Social pooling is first applied to vehicle trajectory prediction in Deo *et al*. [3] with multi-modal maneuver conditions. Other work models scene context information using attention mechanisms to assist trajectory prediction [33], [34]. Lee *et al*. [4] incorporate RNN models with conditional variational autoencoders to generate multimodal predictions, and select the best prediction by ranking scores.

However, these methods model trajectories and context information from a bird's eye view in a static camera setting, which significantly simplifies the challenge of measuring distance from visual features. In contrast, in monocular first-person views, physical distance can be estimated only indirectly, through scaling and observations of participant vehicles, and the environment changes dynamically due to ego-motion effects. Consequently, previous work cannot be directly applied to first-person videos. On the other hand, the first-person view provides higher quality object appearance information compared to birds eye view images, in which objects are represented only by the coordinates of their geometric centers. This paper encodes past location, scale, and corresponding optical flow fields of target vehicles to predict their future locations, and we further improve prediction performance by incorporating future ego-motion.

III. Future vehicle localization from first-person views

We now present our approach to predicting future bounding boxes of vehicles in first-person view. Our method differs from traditional trajectory prediction because the distances of object motion in perspective images do not correspond to physical distances directly, and because the motion of the camera (ego-motion) induces additional apparent motion on nearby objects.

Consider a vehicle visible in the egocentric field of view, and let its past bounding box trajectory be $\mathbf{X} = \{X_{t_0-\tau+1}, X_{t_0-\tau+2}, ..., X_{t_0}\}$, where $X_t = [c_t^x, c_t^y, w_t, h_t]$ is the bounding box of the vehicle at time t (i.e., its center location and width and height in pixels, respectively). Similarly, let the future bounding box trajectory be given by $\mathbf{Y} = \{Y_{t_0+1}, Y_{t_0+2}, ..., Y_{t_0+\delta}\}$. Given image evidence observed from the past τ frames, $\mathbf{O} = \{O_{t_0-\tau+1}, O_{t_0-\tau+2}, ..., O_{t_0}\}$, and its corresponding past bounding box trajectory $\mathbf{X}$, our goal is to predict $\mathbf{Y}$.

We propose a multi-stream RNN encoder-decoder (RNN-ED) model to encode temporal information of past observations and decode future bounding boxes, as shown in Figure 2. The past bounding box trajectory is encoded to provide location and scale information, while dense optical flow is encoded to provide pixel-level information about vehicle scale, motion, and appearance changes. Our decoder can also consider information about future ego-motion, which could be available from the planner of an intelligent vehicle. The decoder generates hypothesized future bounding boxes by temporally updating from the encoded hidden state.

A. Temporal Modeling

1) Location-Scale Encoding: One straightforward approach to predict the future location of an object is to extrapolate a future trajectory from the past. However, in perspective images, physical object location is reflected by both its pixel location and scale. For example, a vehicle located at the center of an image could be a nearby lead vehicle or a distant vehicle across the intersection, and such a difference could cause a completely different future motion. Therefore, this paper predicts both the location and scale of participant vehicles, i.e., their bounding boxes. The scale

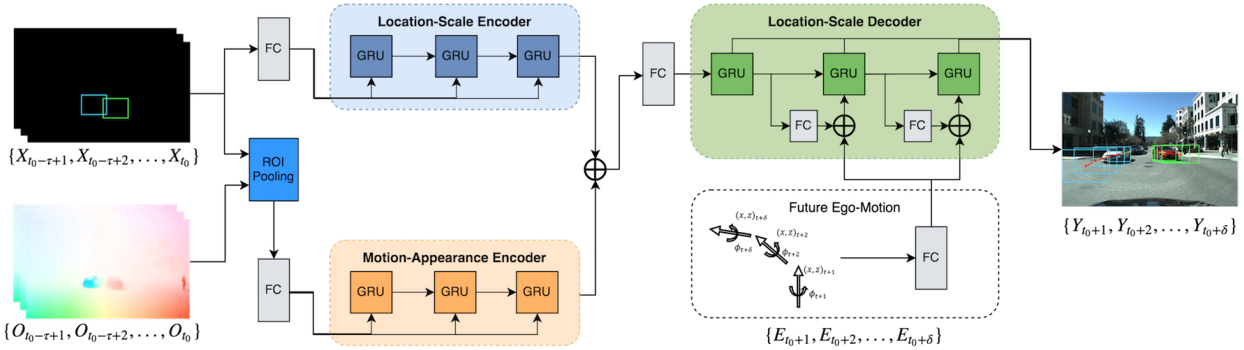

Fig. 2: The proposed future vehicle localization framework (better in color).

information is also able to represent depth (distance) as well as vehicle orientation, given that distant vehicles tend to have smaller bounding boxes and crossing vehicles tend to have larger aspect ratios.

2) Motion-Appearance Encoding: Another important cue for predicting a vehicle's future location is pixel-level information about motion and appearance. Optical flow is widely used as a pattern of relative motion in a scene. For each feature point, optical flow gives an estimate of a vector $[u, v]$ that describes its relative motion from one frame to the next caused by the motion of the object and the camera. Compared to sparse optical flow obtained from traditional methods such as Lucas-Kanade [35], dense optical flow offers an estimate at every pixel, so that moving objects can be distinguished from the background. Also, dense optical flow captures object appearance changes, since different object pixels may have different flows, as shown in the left part of Fig. 2.

In this paper, object vehicle features are extracted by a region-of-interest pooling (ROIPooling) operation using bilinear interpolation from the optical flow map. The ROI region is expanded from the bounding box to contain contextual information around the object, so that its relative motion with respect to the environment is also encoded. The resulting relative motion vector is represented as $O_t = [u_1, v_1, u_2, v_2, ...u_n, v_n]_t$, where n is the size of the pooled region.

We use two encoders for temporal modeling of each input stream and apply the late fusion method:

$$h_t^X = GRU_X(\phi_X(X_{t-1}), h_{t-1}^X; \theta_X) \quad (1a)$$

$$h_t^O = GRU_O(\phi_O(O_{t-1}), h_{t-1}^O; \theta_O) \quad (1b)$$

$$\mathcal{H} = \phi_{\mathcal{H}}(Average(h_{t_0}^X, h_{t_0}^O)) \quad (1c)$$

where GRU represents the gated recurrent units [36] with parameter θ, $\phi(\cdot)$ are linear projections with ReLU activations, and h_t^x and h_t^o are the hidden state vectors of the GRU models at time t.

B. Future Ego-Motion Cue

Awareness of future ego-motion is essential to predicting the future location of participant vehicles. For autonomous vehicles, it is reasonable to assume that motion planning (e.g. trajectory generation) is available [37], so that the future pose of the ego vehicle can be used to aid in predicting the relative position of nearby vehicles. Planned ego-vehicle motion information may also help anticipate motion caused by interactions between vehicles: the ego-vehicle turning left at intersection may result in other vehicles stopping to yield or accelerating to pass, for example.

In this paper, the future ego motion is represented by 2D rotation matrices $R_t^{t+1} \in \mathbb{R}^{2\times2}$ and translation vectors $T_t^{t+1} \in \mathbb{R}^2$ [26], which together describe the transformation of the camera coordinate frame from time t to $t+1$. The relative, pairwise transformations between frames can be composed to estimate transformations across the prediction horizon from the current frame:

$$R_{t_0}^{t_0+i} = \prod_{t=t_0}^{t_0+i-1} R_t^{t+1} \quad (2a)$$

$$T_{t_0}^{t_0+i} = T_{t_0}^{t_0+i-1} + R_{t_0}^{t_0+i-1} T_{t_0+i-1}^{t_0+i} \quad (2b)$$

The future ego-motion feature is represented by a vector $E_t = [\psi_{t_0}^t, x_{t_0}^t, z_{t_0}^t]$, where $t > t_0$, $\psi_{t_0}^t$ is the yaw angle extracted from $R_{t_0}^t$, and $x_{t_0}^t$ and $z_{t_0}^t$ are translations from the coordinate frame at time t_0. We use a right-handed coordinate fixed to ego vehicle, where vehicle heading aligns with positive x. Estimated future motion is then used as input to the trajectory decoding model.

C. Future Location-Scale Decoding

We use another GRU for decoding future bounding boxes. The decoder hidden state is initialized from the final fused hidden state of the past bounding box encoder and the optical flow encoder:

$$h_{t+1}^Y = GRU_Y(f(h_t^Y, E_t), h_t^Y; \theta_Y) \quad (3a)$$

$$Y_{t_0+i} - X_{t_0} = \phi_{out}(h_{t_0+i}^Y) \quad (3b)$$

$$f(h_t^Y, E_t) = Average(\phi_Y(h_t^Y), \phi_e(E_t)) \quad (3c)$$

where h_t^Y is the decoder's hidden state, $h_{t_0}^Y = \mathcal{H}$ is the initial hidden state of the decoder, and $\phi(\cdot)$ are linear projections with ReLU activations applied for domain transfer. Instead of directly generating the future bounding boxes $\mathbf{Y}$, our RNN decoder generates the relative location and scale of the future bounding box from the current frame as in (3b), similar to [26]. In this way, the model output is shifted to have zero initial, which improves the performance.

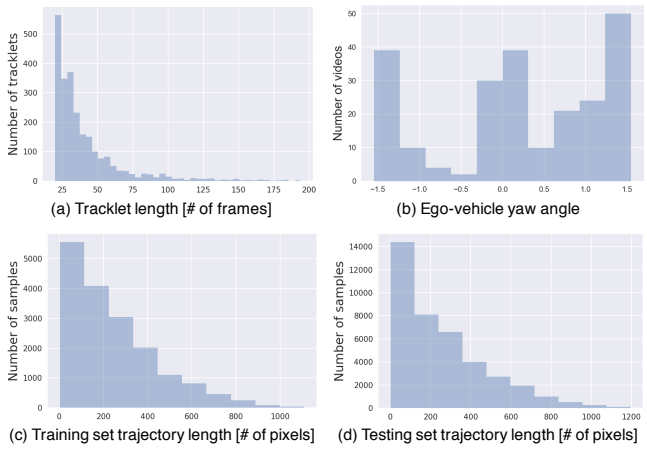

Fig. 3: HEV-I dataset statistics.

IV. Experiments

A. Dataset

The problem of future vehicle localization in egocentric cameras is particularly challenging when multiple vehicles execute different motions (e.g. ego-vehicle is turning left but yields to another moving car). However, to the best of our knowledge, most existing autonomous driving datasets are proposed for scene understanding tasks [38], [39] that do not contain much diverse motion. This paper introduces a new egocentric vision dataset, the *Honda Egocentric View-Intersection* (HEV-I) data, that focuses on intersection scenarios where vehicles exhibit diverse motions due to complex road layouts and vehicle interactions. HEV-I was collected from different intersection types in the San Francisco Bay Area, and consists of 230 videos each ranging between 10 to 60 seconds. Videos were captured by an RGB camera mounted on the windshield of the car, with 1920×1200 resolution (reduced to 1280×640 in this paper) at 10 frames per second (fps).

TABLE I: Comparison with KITTI dataset. The number of vehicles is tallied after filtering out short sequences.

Dataset	# videos	# vehicles	scene types
KITTI	38	541	residential, highway, city road
HEV-I	230	2477	urban intersections

Following prior work [26], we first detected vehicles by using Mask-RCNN [28] pre-trained on the COCO dataset. We then used Sort [40] with a Kalman filter for multiple object tracking over each video. In first-person videos, the duration of vehicles can be extremely short due to high relative motion and limited fields of view. On the other hand, vehicles at stop signs or traffic lights do not move at all over a short period. In our dataset, we found a sample of 2 seconds length is reasonable for including many vehicles while maintaining reasonable travel lengths. We use the past 1 second of observation data as input to predict the bounding boxes of vehicles for the next 1 second. We randomly split the training (70%) and testing (30%) videos, resulting in $\sim 40,000$ training and $\sim 17,000$ testing samples.

TABLE II: Quantitative results of proposed methods and baselines on HEV-I dataset with metrics FDE/ADE/FIOU.

Models	Easy Cases	Challenging Cases	All Cases
Linear	31.49 / 17.04 / 0.68	107.93 / 56.29 / 0.33	72.37 / 38.04 / 0.50
ConstAccel	20.82 / 13.86 / 0.74	90.33 / 49.06 / 0.35	58.00 / 28.05 / 0.53
Conv1D [26]	18.84 / 12.09 / 0.75	37.95 / 20.97 / 0.64	29.06 / 16.84 / 0.69
RNN-ED-X	23.57 / 11.96 / 0.74	43.15 / 22.24 / 0.60	34.04 / 17.46 / 0.67
RNN-ED-XE	22.28 / 11.60 / 0.74	42.27 / 22.39 / 0.61	32.97 / 17.37 / 0.67
RNN-ED-XO	17.45 / 8.68 / 0.78	32.61 / 16.72 / 0.66	25.56 / 12.98 / 0.72
RNN-ED-XOE	**16.72 / 8.52 / 0.80**	**32.05 / 16.63 / 0.66**	**24.92 / 12.86 / 0.73**

TABLE III: Quantitative results on KITTI dataset. We compare our best model with baselines for simplicity.

Models	FDE	ADE	FIOU
Linear	78.19	38.21	0.33
ConstAccel	55.66	25.78	0.39
Conv1D [26]	44.13	24.38	0.49
Ours	**37.11**	**17.88**	**0.53**

Statistics of HEV-I are shown in Fig. 3. As shown, most vehicle tracklets are short in Fig. 3 (a) because vehicles usually drive fast and thus leave the field of the first-person view quickly. Fig. 3 (b) shows the distribution of ego vehicle yaw angle (in rad) across all videos, where positive indicates turning left and negative indicates turning right. It can be seen that HEV-I contains a variety of different ego motions. Distributions of training and test sample trajectory lengths (in pixels) are presented in Fig. 3 (c) and (d). Although most lengths are shorter than 100 pixels, the dataset also contains plenty of longer trajectories. This is important since usually the longer the trajectory is, the more difficult it is to predict. Compared to existing data like KITTI, the HEV-I dataset contains more videos and vehicles, as shown in Table I. Most object vehicles in KITTI are parked on the road or driving in the same direction on highways, while in HEV-I, all vehicles are at intersections and performing diverse maneuvers.

B. Implementation Details

We compute dense optical flow using Flownet2.0 [29] and use a 5×5 ROIPooling operator to produce the final flattened feature vector $O_t \in \mathbb{R}^{50}$. ORB-SLAM2 [41] is used to estimate ego-vehicle motion from first-person videos.

We use Keras with TensorFlow backend [42] to implement our model and perform training and experiments on a system with Nvidia Tesla P100 GPUs. We use the gated recurrent unit (GRU) [43] as basic RNN cell. Compared to long short-term memory (LSTM) [44], GRU has fewer parameters, which makes it faster without affecting performance [45]. The hidden state size of our encoder and decoder GRUs is 512. We use the Adam [46] optimizer with fixed learning

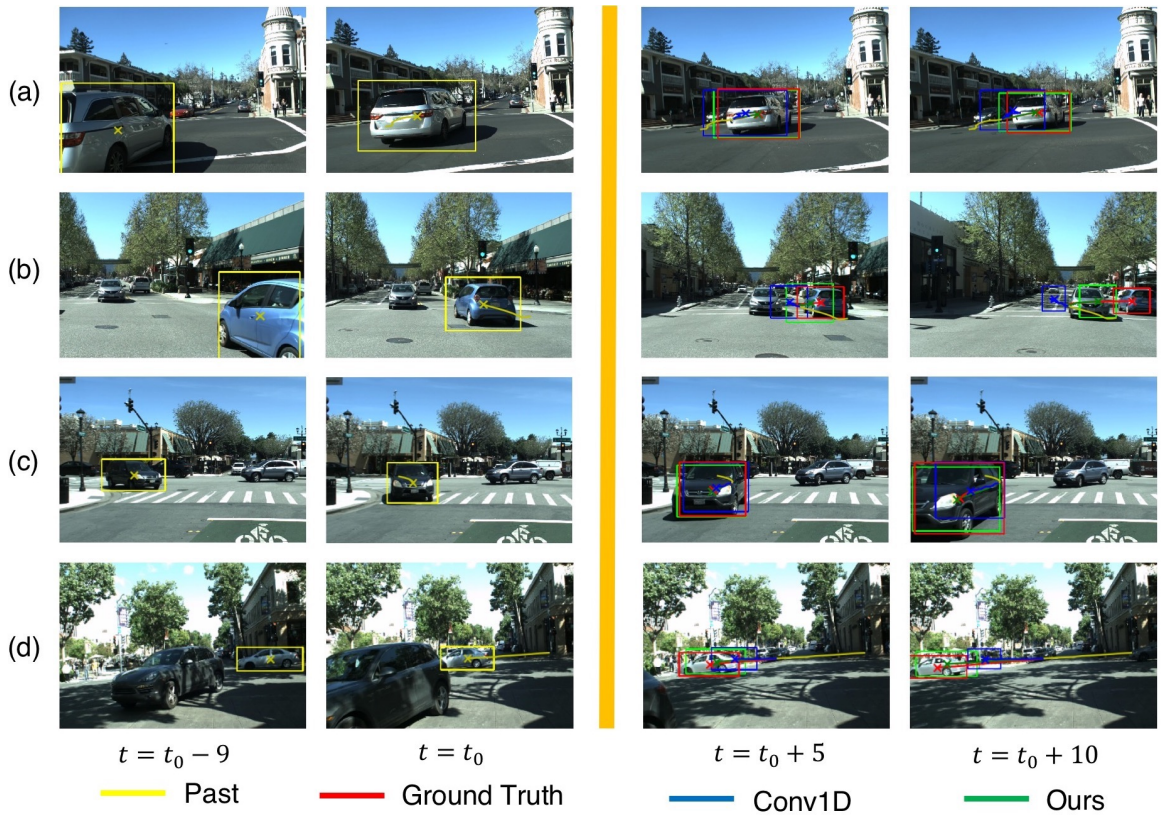

Fig. 4: Qualitative results on HEV-I dataset (better in color).

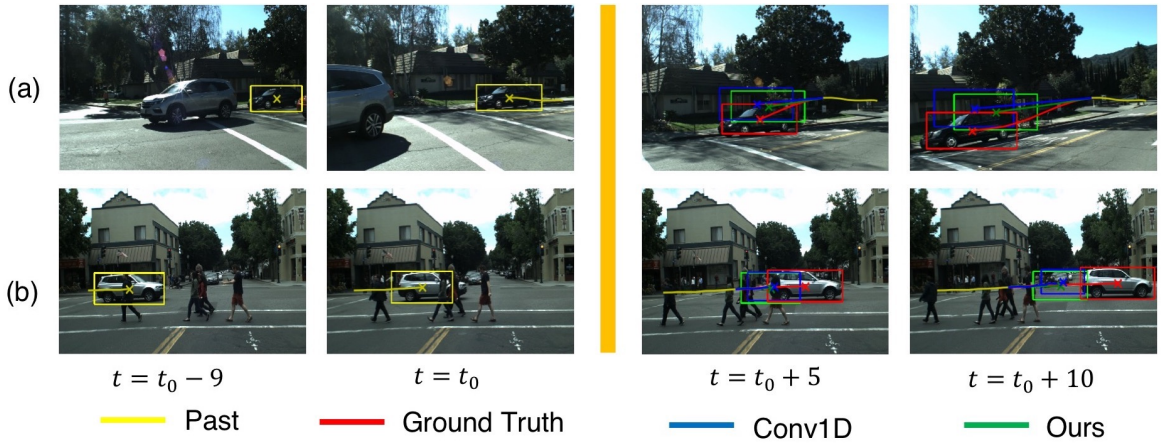

Fig. 5: Failure cases on HEV-I dataset (better in color).

rate 0.0005 and batch size 64. Training is terminated after 40 epochs and the best models are selected.

C. Baselines and Metrics

Baselines. We compare the performance of the proposed method with several baselines:

Linear Regression (Linear) extrapolates future bounding boxes by assuming the location and scale change are linear.

Constant Acceleration (ConstAccel) assumes the object has constant horizontal and vertical acceleration in the camera frame, i.e. that the second-order derivatives of X are constant values.

Conv1D is adapted from [26], by replacing the location-scale and pose input streams with past bounding boxes and dense optical flow.

To evaluate the contribution of each component of our model, we also implemented multiple simpler baselines for ablation studies:

RNN-ED-X is an RNN encoder-decoder with only past bounding boxes as inputs.

RNN-ED-XE builds on *RNN-ED-X* but also incorporates future ego-motion as decoder inputs.

RNN-ED-XO is a two-stream RNN encoder-decoder model with past bounding boxes and optical flow as inputs.

RNN-ED-XOE is our best model as shown in Fig.2 with

awareness of future ego-motion.

Evaluation Metrics. To evaluate location prediction, we use final displacement error (FDE) [26] and average displacement error (ADE) [31], where ADE emphasizes more on the overall prediction accuracy along the horizon. To evaluate bounding box prediction, we propose a final intersection over union (FIOU) metric that measures overlap between the predicted bounding box and ground truth at the final frame.

D. Results on HEV-I Dataset

Quantitative Results. As shown in Table II, we split the testing dataset into easy and challenging cases based on the FDE performance of the *ConstAccel* baseline. A sample is classified as easy if the *ConstAccel* achieves FDE lower than the average FDE (58.00), otherwise it is classified as challenging. Intuitively, easy cases include target vehicles that are stationary or whose future locations can be easily propagated from the past, while challenging cases usually involve diverse and intense motion, e.g. the target vehicle suddenly accelerates or brakes. In evaluation, we report the results of easy and challenging cases, as well as the overall results on all testing samples.

Our best method (*RNN-ED-XOE*) significantly outperforms naive baselines including *Linear* and *ConstAccel* on all cases (FDE of **24.92** vs. 72.37 vs. 58.00). It also improves about 15% from the state-of-the-art *Conv1D* baseline. The improvement on challenging cases is more significant since future trajectories are complex and temporal modeling is more difficult. To more fairly compare the capability of RNN-ED and convolution-deconvolution models, we compare *RNN-ED-XO* with *Conv1D*. These two methods use the same features as inputs to predict future vehicle bounding boxes, but rely on different temporal modeling frameworks. The results (FDE of **25.56** vs 29.06) suggest that the RNN-ED architecture offers better temporal modeling compared to *Conv1D*, because the convolution-deconvolution model generates future trajectory in one shot while the RNN-ED model generates a new prediction based on the previous hidden state. Ablation studies also show that dense optical flow features are essential to accurate prediction of future bounding boxes, especially for challenging cases. The FDE is reduced from 34.04 to 25.56 by adding optical flow stream (*RNN-ED-XO*) to *RNN-ED-X* model. By using future ego-motion, performance can be further improved as shown in the last row of Table II.

Qualitative Results. Fig. 4 shows four sample results of our best model (in green) and the *Conv1D* baseline (in blue). Each row represents one test sample and each column corresponds to each time step. The past and prediction views are separated by the yellow vertical line. Example (a) shows a case where the initial bounding box is noisy because it is close to the image boundary, and our results are more accurate than those of *Conv1D*. Example (b) shows how our model, with awareness of future ego-motion, can predict object future location more accurately while the baseline model predicts future location in the wrong direction. Examples (c) and (d) show that for a curved or long trajectory, our model provides better temporal modelling than *Conv1D*. These results are consistent with our evaluation observations.

Failure Cases. Although our proposed method generally performs well, there are still limitations. Fig.5 (a) shows a case when the ground truth future path is curved due to uneven road surface, which our method fails to consider. In Fig.5 (b), the target vehicle is occluded by pedestrians moving in the opposite direction, which creates misleading optical flow that leads to an inaccurate bounding box (especially in $t = t_0$ frame). Future work could avoid this type of error by better modeling the entire traffic scene as well as relations between traffic participants.

E. Results on KITTI Dataset

We also evaluate our method on a 38-video subset of the KITTI raw dataset, including city, road and residential scenarios. Compared to HEV-I, the road surface of KITTI is more uneven and vehicles are mostly parked on the side of the road with occlusions. Another difference is that in HEV-I, the ego-vehicle often stops at intersections to yield to other vehicles, resulting in static samples with no motion at all. We did not remove static samples from the dataset since predicting a static object is also valuable.

To evaluate our method on KITTI, we first generate the input features following the same process of HEV-I dataset, resulting in $\sim$ 8000 training and $\sim$ 2700 testing samples. Performance of baselines and our best model are shown in Table III. Both learning-based models are trained for 40 epoches and the best models are selected. The results show that our method outperforms all baselines including the state-of-the-art Conv1D (FDE of **37.11** vs 78.19 vs 55.66 vs 44.13). We also observe that both learning-based methods did not perform as well as they did on HEV-I. One possible reason is that KITTI is much smaller so that the models are not fully trained. In general, we conclude that the use of the proposed framework results in more robust future vehicle localization across different datasets.

V. CONCLUSION

We proposed the new problem of predicting the relative location and scale of target vehicles in first-person video. We presented a new dataset collected from intersection scenarios to include as many vehicles and motion as possible. Our proposed multi-stream RNN encoder-decoder structure with awareness of future ego motion shows promising results compared to other baselines on our dataset as well as on KITTI, and we tested how each component contributed to the model through an ablation study.

Future work includes incorporating evidence from scene context, traffic signs/signals, depth data, and other vehicle-environment interactions. Social relationships such as vehicle-to-vehicle and vehicle-to-pedestrian interactions could also be considered.

REFERENCES

[1] Y. Yao and E. Atkins, "The smart black box: A value-driven automotive event data recorder," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 973–978.

[2] N. Deo, A. Rangesh, and M. M. Trivedi, "How would surround vehicles move? a unified framework for maneuver classification and motion prediction," *T-IV*, 2018.

[3] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," *arXiv:1805.06771*, 2018.

[4] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *CVPR*, 2017.

[5] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *CVPR*, 2015.

[6] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *CVPR*, 2016.

[7] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," *arXiv:1711.07280*, 2017.

[8] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," *arXiv:1711.11543*, 2017.

[9] X. Ye, Z. Lin, H. Li, S. Zheng, and Y. Yang, "Active object perceiver: Recognition-guided policy learning for object searching on mobile robots," *arXiv:1807.11174*, 2018.

[10] F. Codevilla, M. Müller, A. Dosovitskiy, A. López, and V. Koltun, "End-to-end driving via conditional imitation learning," *arXiv:1710.02410*, 2017.

[11] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *ICCV*, 2011.

[12] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in *CVPR*, 2011.

[13] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.

[14] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *IJCV*, 2015.

[15] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "First person action-object detection with egonet," *arXiv:1603.04908*, 2016.

[16] M. Gao, A. Tawari, and S. Martin, "Goal-oriented object importance estimation in on-road driving videos," *ICRA*, 2019.

[17] S. Ardeshir and A. Borji, "Ego2Top: Matching viewers in egocentric and top-view videos," in *ECCV*, 2016.

[18] C. Fan, J. Lee, M. Xu, K. K. Singh, Y. J. Lee, D. J. Crandall, and M. S. Ryoo, "Identifying first-person camera wearers in third-person videos," *arXiv:1704.06340*, 2017.

[19] M. Xu, C. Fan, Y. Wang, M. S. Ryoo, and D. J. Crandall, "Joint person segmentation and identification in synchronized first-and third-person videos," *arXiv:1803.11217*, 2018.

[20] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *CVPR*, 2012.

[21] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *ICCV*, 2013.

[22] H. Soo Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *CVPR*, 2016.

[23] S. Su, J. P. Hong, J. Shi, and H. S. Park, "Predicting behaviors of basketball players from first person videos," in *CVPR*, 2017.

[24] G. Bertasius, A. Chan, and J. Shi, "Egocentric basketball motion planning from a single first-person image," in *CVPR*, 2018.

[25] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *CVPR*, 2018.

[26] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *CVPR*, 2018.

[27] J. Wiest, M. Höffken, U. Kreßel, and K. Dietmayer, "Probabilistic trajectory prediction with gaussian mixture models," in *IV*, 2012.

[28] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.

[29] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017.

[30] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, and D. J. Crandall, "Temporal recurrent networks for online action detection," *arXiv preprint arXiv:1811.07391*, 2018.

[31] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *CVPR*, 2016.

[32] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[33] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese, "Car-Net: Clairvoyant attentive recurrent network," in *ECCV*, 2018.

[34] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," *arXiv:1806.01482*, 2018.

[35] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.

[36] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv:1406.1078*, 2014.

[37] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles." *ITSM*, 2016.

[38] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.

[39] 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.

[40] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *ICIP*, 2016.

[41] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-d cameras," *T-RO*, 2017.

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016.

[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *ICML*, 2015.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[45] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555*, 2014.

[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.