TITAN: Future Forecast using Action Priors

Srikanth Malla  Behzad Dariush  Chihho Choi
Honda Research Institute USA
{smalla, bdariush, cchoi}@honda-ri.com

Abstract

We consider the problem of predicting the future trajectory of scene agents from egocentric views obtained from a moving platform. This problem is important in a variety of domains, particularly for autonomous systems making reactive or strategic decisions in navigation. In an attempt to address this problem, we introduce TITAN (Trajectory Inference using Targeted Action priors Network), a new model that incorporates prior positions, actions, and context to forecast future trajectory of agents and future ego-motion. In the absence of an appropriate dataset for this task, we created the TITAN dataset that consists of 700 labeled video-clips (with odometry) captured from a moving vehicle on highly interactive urban traffic scenes in Tokyo. Our dataset includes 50 labels including vehicle states and actions, pedestrian age groups, and targeted pedestrian action attributes that are organized hierarchically corresponding to atomic, simple/complex-contextual, transportive, and communicative actions. To evaluate our model, we conducted extensive experiments on the TITAN dataset, revealing significant performance improvement against baselines and state-of-the-art algorithms. We also report promising results from our Agent Importance Mechanism (AIM), a module which provides insight into assessment of perceived risk by calculating the relative influence of each agent on the future ego-trajectory. The dataset is available at https://usa.honda-ri.com/titan

1. Introduction

The ability to forecast future trajectory of agents (individuals, vehicles, cyclists, etc.) is paramount in developing navigation strategies in a range of applications including motion planning and decision making for autonomous and cooperative (shared autonomy) systems. We know from observation that the human visual system possesses an uncanny ability to forecast behavior using various cues such as experience, context, relations, and social norms. For example, when immersed in a crowded driving scene, we are able to reasonably estimate the intent, future actions, and future location of the traffic participants in the next few seconds. This is undoubtedly attributed to years of prior experience and observations of interactions among humans and other participants in the scene. To reach such human level ability to forecast behavior is part of the quest for visual intelligence and the holy grail of autonomous navigation, requiring new algorithms, models, and datasets.

In the domain of behavior prediction, this paper considers the problem of future trajectory forecast from egocentric views obtained from a mobile platform such as a vehicle in a road scene. This problem is important for autonomous agents to assess risks or to plan ahead when making reactive or strategic decisions in navigation. Several recently reported models that predict trajectories incorporate social
norms, semantics, scene context, etc. The majority of these algorithms are developed from a stationary camera view in surveillance applications, or overhead views from a drone.

The specific objective of this paper is to develop a model that incorporates prior positions, actions, and context to simultaneously forecast future trajectory of agents and future ego-motion. In a related problem, the ability to predict future actions based on current observations has been previously studied in [25, 47, 46, 45, 50]. However, to the best of our knowledge, action priors have not been used in forecasting future trajectory, partly due to a lack of an appropriate dataset. A solution to this problem can help address the challenging and intricate scenarios that capture the interplay of observable actions and their role in future trajectory forecast. For example, when the egocentric view of a mobile agent in a road scene captures a delivery truck worker closing the tailgate of the truck, it is highly probable that the worker’s future behavior will be to walk toward the driver side door. Our aim is to develop a model that uses such action priors to forecast trajectory.

The algorithmic contributions of this paper are as follows. We introduce TITAN (Trajectory Inference using Targeted Action priors Network), a new model that incorporates prior positions, actions, and context to simultaneously forecast future trajectory of agents and future ego-motion. Our framework introduces a new interaction module to handle dynamic number of objects in the scene. While modeling pair-wise interactive behavior from all agents, the proposed interaction module incorporates actions of individuals in addition to their locations, which helps the system to understand the contextual meaning of motion behavior. In addition, we propose to use multi-task loss with aleatoric homoscedastic uncertainty [22] to improve the performance of multi-label action recognition. For ego-future, Agent Importance Mechanism (AIM) is presented to identify objects that are more relevant for ego-motion prediction.

Apart from algorithmic contributions, we introduce a novel dataset, referred to as TITAN dataset, that consists of 700 video clips captured from a moving vehicle on highly interactive urban traffic scenes in Tokyo. The pedestrians in each clip were labeled with various action attributes that are organized hierarchically corresponding to atomic, simple/complex contextual, transportive, and communicative actions. The action attributes were selected based on commonly observed actions in driving scenes, or those which are important for inferring intent (e.g., waiting to cross). We also labeled other participant categories, including vehicle category (4 wheel, 2 wheel), age-groups, and vehicle state. The dataset contains synchronized ego-motion information from an IMU sensor. To our knowledge, this is the only comprehensive and large scale dataset suitable for studying action priors for forecasting the future trajectory of agents from ego-centric views obtained from a moving platform. Furthermore, we believe our dataset will contribute to advancing research for action recognition in driving scenes.

2. Related Work

2.1. Future Trajectory Forecast

**Human Trajectory Forecast** Encoding interactions between humans based on their motion history has been widely studied in the literature. Focusing on input-output time sequential processing of data, recurrent neural network (RNN)-based architectures have been applied to the future forecast problem in the last few years [2, 26, 17, 56, 60]. More recently, RNNs are used to formulate a connection between agents with their interactions using graph structures [54, 30]. However, these methods suffer from understanding of environmental context with no or minimal considerations of scene information. To incorporate models of human interaction with the environment, [57] takes local to global scale image features into account. More recently, [10] visually extracts relational behavior of humans interacting with other agents as well as environments.

**Vehicle Trajectory Forecast** Approaches for vehicle motion prediction have developed following the success of interaction modeling using RNNs. Similar to human trajectory forecast, [13, 35, 30, 29] only consider the past motion history. These methods perform poorly in complex road environments without the guidance of structured layouts. Although the subsequent approaches [40, 28, 11] partially overcome these issues by using 3D LiDAR information as inputs to predict future trajectories, their applicability to current production vehicles is limited due to the higher cost. Recent methods [3, 58, 31] generate trajectories of agents from an egocentric view. However, they do not consider interactions between road agents in the scene and the potential influence to the ego-future. In this work, we explicitly model pair-wise interactive behavior from all agents to identify objects that are more relevant for the target agent.

2.2. Action Recognition

With the success of 2D convolutions in image classification, frame-level action recognition has been presented in [20]. Subsequently, [44] separates their framework into two streams: one to encode spatial features from RGB images and the other to encode temporal features from corresponding optical flow. Their work motivated studies that model temporal motion features together with spatial image features from videos. A straightforward extension has been shown in [51, 52], replacing 2D convolutions by 3D convolutions. To further improve the performance of these models, several research efforts have been provided such as I3D [7] that inflates a 2D convolutional network into 3D to benefit from the use of pre-trained models and 3D ResNet [18] that adds residual connections to build a very
deep 3D network. Apart from them, other approaches capture pair-wise relations between actor and contextual features [49] or those between pixels in space and in time [55]. More recently, Timeception [19] models long range temporal dependencies, particularly focusing on complex actions.

2.3. Datasets

Future Trajectory Several influential RGB-based datasets for pedestrian trajectory prediction have been reported in the literature. These datasets are typically created from a stationary surveillance camera [27, 37, 34], or from aerial views obtained from a static drone-mounted camera [41]. In driving scenes, the 3D point cloud-based datasets [15, 36, 23, 5, 1, 9] were originally introduced for detection, tracking, etc., but recently used for vehicle trajectory prediction as well. Also, [58, 8] provide RGB images captured from an egocentric view of a moving vehicle and applied to future trajectory forecast problem. The JAAD [39], CMU-UAH [33], and PIE [38] datasets are most similar to our TITAN dataset in the sense that they are designed to study the intentions and actions of objects from on-board vehicles. However, their labels are limited to simple actions such as walking, standing, looking, and crossing. These datasets, therefore, do not provide an adequate number of actions to use as priors in order to discover contextual meaning of agents’ motion behavior. To address these limitations, our TITAN dataset provides 50 labels including vehicle states and actions, pedestrian age groups, and targeted pedestrian action attributes that are hierarchically organized as illustrated in the supplementary material.

Action Recognition A variety of datasets have been introduced for action recognition with a single action label [24, 48, 20, 32, 21] and multiple action labels [43, 59, 4] in videos. Recently released datasets such as AVA [16], READ [14], and EPIC-KITCHENS [12] contain actions with corresponding localization around a person or object. Our TITAN dataset is similar to AVA in the sense that it provides spatio-temporal localization for each agent with multiple action labels. However, the labels of TITAN are organized hierarchically from primitive atomic actions to complicated contextual activities that are typically observed from on-board vehicles in driving scenes.

3. TITAN Dataset

In the absence of an appropriate dataset suitable for our task, we introduce the TITAN dataset for training and evaluation of our models as well as to accelerate research on trajectory forecast. Our dataset is sourced from 10 hours of video recorded at 60 FPS in central Tokyo. All videos are captured using a GoPro Hero 7 Camera with embedded IMU sensor which records synchronized odometry data at 100 HZ for ego-motion estimation. To create the final annotated dataset, we extracted 700 short video clips from the original (raw) recordings. Each clip is between 10-20 seconds in duration, image size width:1920px, height:1200px and annotated at 10 HZ sampling frequency. The characteristics of the selected video clips include scenes that exhibit a variety of participant actions and interactions.

The taxonomy and distribution of all labels in the dataset are depicted in Figure 2. The total number of frames annotated is approximately 75,262 with 395,770 persons, 146,840 4-wheeled vehicles and 102,774 2-wheeled vehicles. This includes 8,592 unique persons and 5,504 unique vehicles. For our experiments, we use 400 clips for training, 200 clips for validation and 100 clips for testing. As mentioned in Section 2.3, there are many publicly available datasets related to mobility and driving, many of which include ego-centric views. However, since those datasets do not provide action labels, a meaningful quantitative comparison of the TITAN dataset with respect to existing mobility datasets is not possible. Furthermore, a quantitative comparison with respect to action localization datasets such as AVA is not warranted since AVA does not include ego-
centric views captured from a mobile platform.

In the TITAN dataset, every participant (individuals, vehicles, cyclists, etc.) in each frame is localized using a bounding box. We annotated 3 labels (person, 4-wheeled vehicle, 2-wheeled vehicle), 3 age groups for person (child, adult, senior), 3 motion-status labels for both 2 and 4-wheeled vehicles, and door/trunk status labels for 4-wheeled vehicles. For action labels, we created 5 mutually exclusive person action sets organized hierarchically (Figure 2). In the first action set in the hierarchy, the annotator is instructed to assign exactly one class label among 9 atomic whole body actions/postures that describe primitive action poses such as sitting, standing, bending, etc. The second action set includes 13 actions that involve single atomic actions with simple scene context such as jay-walking, waiting to cross, etc. The third action set includes 7 complex contextual actions that describe a sequence of atomic actions with higher contextual understanding, such as getting in/out of a 4-wheel vehicle, loading/unloading, etc. The fourth action set includes 4 transportive actions that describe the act of manually transporting an object by carrying, pulling or pushing. Finally, the fifth action set includes 4 communicative actions observed in traffic scenes such as talking on the phone, looking at phone, or talking in groups. In each action sets 2-5, the annotators were instructed to assign ‘None’ if there is no label. This hierarchical strategy was designed to produce unique (unambiguous) action labels while reducing the annotators’ cognitive workload and thereby improving annotation quality. The tracking ID’s of all localized objects are associated within each video clip. Example scenarios are displayed in Figure 3.

4. Methodology

Figure 4 shows the block diagram of the proposed TITAN framework. A sequence of image patches \( T_{i=1:T_{obs}} \) is obtained from the bounding box\(^1\) \( x^i = \{c_u, c_v, l_u, l_v\} \) of agent \( i \) at each past time step from 1 to \( T_{obs} \), where \( (c_u, c_v) \) and \( (l_u, l_v) \) represent the center and the dimension of the bounding box, respectively. The proposed TITAN framework requires three inputs as follows: \( \tau_{t=1:T_{obs}} \) for the action detector, \( x^i_t \) for both the interaction encoder and past object location encoder, and \( e_t = \{\alpha_t, \omega_t\} \) for the ego-motion encoder where \( \alpha_t \) and \( \omega_t \) correspond to the acceleration and yaw rate of the ego-vehicle at time \( t \), respectively. During inference, the multiple modes of future bounding box locations are sampled from a bi-variate Gaussian generated by the noise parameters, and the future ego-motions \( \hat{e}_t \) are accordingly predicted, considering the multi-modal nature of the future prediction problem.

Henceforth, the notation of the feature embedding function using multi-layer perceptron (MLP) is as follows: \( \Phi \) is without any activation, and \( \Phi_r, \Phi_t, \) and \( \Phi_s \) are associated with ReLU, tanh, and a sigmoid function, respectively.

4.1. Action Recognition

We use the existing state-of-the-art method as backbone for the action detector. We finetune single-stream I3D [7] and 3D ResNet [18] architecture pre-trained on Kinetics-600 [6]. The original head of the architecture is replaced by a set of new heads (8 action sets of TITAN except age group and type) for multi-label action outputs. The action detector takes \( T_{i=1:T_{obs}} \) as input, which is cropped around the agent \( i \). Then, each head outputs an action label including a ‘None’ class if no action is shown. From our experiments, we observed that certain action sets converge faster than others. This is due in part because some tasks are relatively easier to learn, given the shared representations. Instead of tuning the weight of each task by hand, we adopt the multi-task loss in [22] to further boost performance of our action detector. Note that each action set of the TITAN dataset is mutually exclusive, thus we consider the outputs are independent to each other as follows:

\[
p(y_{m}, \ldots, y_n | f(I)) = \prod_{i=m}^{n} p(y_i | f(I)), \quad (1)
\]

\(^1\)We assume that the bounding box detection using past images is provided by the external module since detection is not the scope of this paper.
Then, we aggregate the hidden states to generate interaction messages of instant interaction with each agent at time $t$ where $\triangledown$ actions $v$ are evaluated through the dynamic RNN with $\{e\}$ with respect to the target agent, $h_{t+1}^{i} = GRU(v_{t}^{i}, h_{t}^{i}, W_{INT})$, where $W_{INT}$ are the weight parameters. Note that we pass the messages of instant interaction with each agent at time $t$, which enables us to find their potential influence at that moment. Then, we aggregate the hidden states to generate interaction features $v_{t}^{i} = \frac{1}{n} \sum_{j} h_{t}^{ij}$ for the target agent $i$, computed from all other agents in the scene at time $t$ as in Figure 5.

The past ego motion encoder takes $e_{l} = (\omega_{l}, \omega_{u})$ as input and embeds the motion history of ego-vehicle using the GRU. We use each hidden state output $h_{t}^{e}$ to compute future locations of other agents. The past object location encoder uses the GRU to embed the history of past motion into a feature space. The input to this module is a bounding box $x^{i}$ of the target agent $i$ at each past time step, and we use the embedding $\Phi(x_{t}^{i})$ for the GRU. The output hidden state $h_{t}^{p}$ of the encoder is updated by $h_{t+1}^{p} = GRU(h_{t}^{p}, \Phi(x_{t}^{i}), W_{POL})$, where $W_{POL}$ are the weight parameters. We use its final hidden state as an initial hidden state input of the future object location decoder.

The future bounding boxes of the target agent $i$ are decoded using the GRU-based future object location decoder from time step $T_{obs} + 1$ to $T_{pred}$. At each time step, we output a 10-dimensional vector where the first 5 values are the center $\mu_{c} = (c_{u}, c_{v})$, variance $\sigma_{c} = (\sigma_{cu}, \sigma_{cv})$, and its correlation $\rho_{c}$ and the rest 5 values are the dimension $\mu_{d} = (l_{u}, l_{v})$, variance $\sigma_{d} = (\sigma_{lu}, \sigma_{lv})$, and its correlation $\rho_{d}$. We use two bi-variate Gaussians for bounding box centers and dimensions, so that they can be independently sampled. We use the negative log-likelihood loss function as:

$$L_{a} = \frac{1}{T} \sum_{t=T_{obs}+1}^{T_{pred}} \log p(c_{t}^{i}, \sigma_{t}^{i}, \rho_{c})p(l_{t}^{i}, \sigma_{l}^{i}, \rho_{l}).$$

### 4.2. Future Object Localization

Unlike existing methods, we model the interactions using the past locations of agents conditioned on their actions, which enables the system to explicitly understand the contextual meaning of motion behavior. At each past time step $t$, the given bounding box $x_{t}^{i} = \{c_{u}, c_{v}, l_{u}, l_{v}\}$ is concatenated with the multi-label action vector $a_{t}^{i}$. We model the pairwise interactions between the target agent $i$ and all other agents $j$ through MLP, $v_{t}^{ij} = \Phi_{r}(x_{t}^{i} \otimes a_{t}^{j} \otimes x_{t}^{j} \otimes a_{t}^{i})$ where $\otimes$ is a concatenation operator. The resulting interactions $v_{t}^{ij}$ are evaluated through the dynamic RNN with GRUs to leave more important information with respect to the target agent, $h_{t+1}^{i} = GRU(v_{t}^{ij}, h_{t}^{ij}, W_{INT})$, where $W_{INT}$ are the weight parameters. Note that we pass the messages of instant interaction with each agent at time $t$, which enables us to find their potential influence at that moment. Then, we aggregate the hidden states to generate interaction features $v_{t}^{i} = \frac{1}{n} \sum_{j} h_{t}^{ij}$ for the target agent $i$, computed from all other agents in the scene at time $t$ as in Figure 5.

The past ego motion encoder takes $e_{l} = (\omega_{l}, \omega_{u})$ as input and embeds the motion history of ego-vehicle using the GRU. We use each hidden state output $h_{t}^{e}$ to compute future locations of other agents. The past object location encoder uses the GRU to embed the history of past motion into a feature space. The input to this module is a bounding box $x^{i}$ of the target agent $i$ at each past time step, and we use the embedding $\Phi(x_{t}^{i})$ for the GRU. The output hidden state $h_{t}^{p}$ of the encoder is updated by $h_{t+1}^{p} = GRU(h_{t}^{p}, \Phi(x_{t}^{i}), W_{POL})$, where $W_{POL}$ are the weight parameters. We use its final hidden state as an initial hidden state input of the future object location decoder.

The future bounding boxes of the target agent $i$ are decoded using the GRU-based future object location decoder from time step $T_{obs} + 1$ to $T_{pred}$. At each time step, we output a 10-dimensional vector where the first 5 values are the center $\mu_{c} = (c_{u}, c_{v})$, variance $\sigma_{c} = (\sigma_{cu}, \sigma_{cv})$, and its correlation $\rho_{c}$ and the rest 5 values are the dimension $\mu_{d} = (l_{u}, l_{v})$, variance $\sigma_{d} = (\sigma_{lu}, \sigma_{lv})$, and its correlation $\rho_{d}$. We use two bi-variate Gaussians for bounding box centers and dimensions, so that they can be independently sampled. We use the negative log-likelihood loss function as:

$$L_{a} = \frac{1}{T} \sum_{t=T_{obs}+1}^{T_{pred}} \log p(c_{t}^{i}, \sigma_{t}^{i}, \rho_{c})p(l_{t}^{i}, \sigma_{l}^{i}, \rho_{l}).$$

### 4.3. Future Ego-motion prediction

We first embed the predicted future bounding box of all agents $\hat{X} = \{\hat{x}^{1}, \ldots, \hat{x}^{N}\}$ through MLP at each future time step $T_{obs} + 1$ to $T_{pred}$. We further condition it on the previously computed action labels in a feature space through $H_{t}^{e} = \Phi_{r}(r_{T_{obs}}^{e} \otimes \Phi_{r}(\hat{x}^{i}))$, where $r_{T_{obs}}^{e} = \Phi_{r}(a_{T_{obs}}^{e})$. By using the action labels as a prior constraint, we explicitly
lead the model to understand about the contextual meaning of locations. The resulting features of each agent i are weighted using the AIM module $H_{t_i}^i = w_i^1 * H_{t_i}^i$, where the weights $w_i^1 = \Phi_t(H_{t_i}^i)$, similar to self-attention [53]. Then, we sum all features $H_t^i = \sum_i H_{t_i}^i$ for each future time step. This procedure is detailed in Figure 6. Note that our AIM module is simultaneously learned with the future ego-motion prediction, which results in weighting other agents more or less based on their influence/importance to the ego-vehicle. It thus provides insight into assessment of perceived risk while predicting the future motion. We qualitatively evaluate it in Sec. 5.

The last hidden state $h_T^i$ of the past ego motion encoder is concatenated with $H_t^i$ through $h_T^i = \Phi(H_t^i \otimes h_T^i)$ and fed into the future ego motion decoder. The intermediate hidden state $h_t^i$ is accordingly updated by $H_t^i$ at each future time step for recurrent update of the GRU. We output the ego-future using each hidden state $h_t^i$ through $\hat{e}_t^i = \Phi(h_t^i)$ at each future time $T_{obs} + 1$ to $T_{pred}$. For training, we use task dependent uncertainty with L2 loss for regressing both acceleration and angular velocity as shown below:

$$L_E = \frac{||\alpha_t - \hat{\alpha}_t||^2}{\sigma_1^2} + \frac{||\omega_t - \hat{\omega}_t||^2}{\sigma_2^2} + log\sigma_1\sigma_2.$$  (4)

Note that the predicted future ego-motion is deterministic in its process. However, its multi-modality comes from sampling of the predicted future bounding boxes of other agents. In this way, we capture their influence with respect to the ego-vehicle, and AIM outputs the importance weights consistent with the agents’ action and future motion.

5. Experiments

In all experiments performed in this work, we predict up to 2 seconds into the future while observing 1 second of past observations as proposed in [31]. We use average distance error (ADE), final distance error (FDE), and final intersection over union (FIOU) metrics for evaluation of future object localization. We include FIOU in our evaluation since ADE/FDE only capture the localization error of the final bounding box without considering its dimensions. For action recognition, we use per frame mean average precision (mAP). Finally, for ego-motion prediction, we use root mean square error (RMSE) as an evaluation metric.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone</td>
<td>InceptionV1</td>
<td>ResNet50</td>
</tr>
<tr>
<td>atomic</td>
<td>0.9219</td>
<td>0.7552</td>
</tr>
<tr>
<td>simple</td>
<td>0.5318</td>
<td>0.3173</td>
</tr>
<tr>
<td>complex</td>
<td>0.9881</td>
<td>0.9880</td>
</tr>
<tr>
<td>communicative</td>
<td>0.8649</td>
<td>0.8648</td>
</tr>
<tr>
<td>transportive</td>
<td>0.9080</td>
<td>0.9081</td>
</tr>
<tr>
<td>overall</td>
<td>0.8429</td>
<td>0.7664</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.9918</td>
<td>0.7132</td>
</tr>
<tr>
<td>trunk</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>doors</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>overall*</td>
<td>0.8946</td>
<td>0.8128</td>
</tr>
</tbody>
</table>

Table 1. Action recognition results (mAP) on TITAN.

5.1. Action Recognition

We evaluate two state-of-the-art 3D convolution-based architectures, I3D with InceptionV1 and 3D ResNet with ResNet50 as backbone. Both models are pre-trained on Kinetics-600 and finetuned using TITAN with the multi-task loss in Eqn. 2. As detailed in Sec. 4.1, we modify the original structure using new heads that corresponds to the 8 action sets of the TITAN dataset. Their per frame mAP results are compared in Table 1 for each action set. We refer to the supplementary material for the detailed comparison on individual action categories. Note that we use the I3D-based action detector for the rest of our experiments.

5.2. Future Object Localization

The results of future object localization performance is shown in Table 2. The constant velocity (Const-Vel [42]) baseline is computed using the last two observations for linearly interpolating future positions. Since the bounding box dimensions error is not captured by ADE or FDE, we evaluate on FIOU using two baselines: 1) without scaling the box dimensions, and 2) with scaling linearly the box dimensions. Titan_vanilla is an encoder and decoder RNN without any priors or interactions. It shows better performance than linear models. Both Social-GAN [17] and Social-LSTM [2] improve the performance in ADE and FDE compared to the simple recurrent model (Titan_vanilla) or linear approaches. Note that we do not evaluate FIOU for Social-GAN and Social-LSTM since their original method is not designed to predict dimensions. Titan_AP adds action priors to the past positions and performs better than Titan_vanilla, which shows that the model better understands contextual meaning of the past motion. However, its performance is worse than Titan_EP that includes ego-motion as priors. This is because Titan_AP does not consider the motion behavior of other agents in egocentric view. Titan_IP includes interaction priors as shown in Figure 5 without concatenating actions. Interestingly, its performance is better than Titan_AP (action priors) and Titan_EP (ego priors) as well as Titan_EP+AP (both ego and action priors). It validates the efficacy of our interaction encoder that aims to pass the interactions over
5.3. Future Ego-Motion Prediction

The quantitative results for future ego-motion prediction are shown in Table 3. Between Const-Vel [42] and Const-

Using ground-truth actions as a prior, we observed further improvement in overall ADE by 2 pixels and overall FDE by 3.5 pixels.

...
Figure 9. The importance (or degree of influence) of each agent toward the ego-vehicle’s future trajectory is illustrated by the proportion of red bar relative to the blue bar displayed across the top width of the agent’s bounding box. A red bar spanning across the top width represents the maximum importance derived from the AIM module, while a blue bar spanning across the top width represents minimum importance. (top row) images from same sequence. (bottom row) images from different sequences.

Table 3. Comparison of Future ego motion prediction. acceleration error in \( m/s^2 \) and yaw rate error in \( rad/s \).

<table>
<thead>
<tr>
<th>Method</th>
<th>acc RMSE ↓</th>
<th>yaw rate RMSE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const-Vel [42]</td>
<td>1.745</td>
<td>0.1249</td>
</tr>
<tr>
<td>Const-Acc</td>
<td>1.569</td>
<td>0.1549</td>
</tr>
<tr>
<td>Titan_vanilla</td>
<td>1.201</td>
<td>0.1416</td>
</tr>
<tr>
<td>Titan_FP</td>
<td>1.236</td>
<td>0.1438</td>
</tr>
<tr>
<td>Titan_FP+AP</td>
<td>1.182</td>
<td>0.1061</td>
</tr>
<tr>
<td>Titan_AIM_FP</td>
<td>1.134</td>
<td>0.0921</td>
</tr>
<tr>
<td>Titan_AIM (ours)</td>
<td>1.081</td>
<td>0.0824</td>
</tr>
</tbody>
</table>

6. Conclusion

We presented a model that can reason about the future trajectory of scene agents from egocentric views obtained from a mobile platform. Our hypothesis was that action priors provide meaningful interactions and also important cues for making future trajectory predictions. To validate this hypothesis, we developed a model that incorporates prior positions, actions, and context to simultaneously forecast future trajectory of agents and future ego-motion. For evaluation, we created a novel dataset with over 700 video clips containing labels of a diverse set of actions in urban traffic scenes from a moving vehicle. Many of those actions implicitly capture the agent’s intentions. Comparative experiments against baselines and state-of-art prediction algorithms showed significant performance improvement when incorporating action and interaction priors. Importantly, our framework introduces an Agent Importance Mechanism (AIM) module to identify agents that are influential in predicting the future ego-motion, providing insight into assessment of perceived risk in navigation. For future work, we plan to incorporate additional scene context to capture participant interactions with the scene or infrastructure.

Acknowledgement We thank Akira Kanehara for supporting our data collection and Yuji Yasui, Rei Sakai, and Isht Dwivedi for insightful discussions.
References

[26] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Mannohman Chandraker. Desire:


