Looking to Relations for Future Trajectory Forecast

Chiho Choi
Honda Research Institute USA
cchoi@honda-ri.com

Behzad Dariush
Honda Research Institute USA
bdariush@honda-ri.com

Abstract

Inferring relational behavior between road users as well as road users and their surrounding physical space is an important step toward effective modeling and prediction of navigation strategies adopted by participants in road scenes. To this end, we propose a relation-aware framework for future trajectory forecast. Our system aims to infer relational information from the interactions of road users with each other and with the environment. The first module involves visual encoding of spatio-temporal features, which captures human-human and human-space interactions over time. The following module explicitly constructs pair-wise relations from spatio-temporal interactions and identifies more descriptive relations that highly influence future motion of the target road user by considering its past trajectory. The resulting relational features are used to forecast future locations of the target, in the form of heatmaps with an additional guidance of spatial dependencies and consideration of the uncertainty. Extensive evaluations on the public benchmark datasets demonstrate the robustness and efficacy of the proposed framework as observed by performances higher than the state-of-the-art methods.

1. Introduction

Forecasting future trajectories of moving participants in indoor and outdoor environments has profound implications for execution of safe and naturalistic navigation strategies in partially and fully automated vehicles [3, 42, 41, 10] and robotic systems [49, 19, 18, 4]. While autonomous navigation of robotic systems in dynamic indoor environments is an increasingly important application that can benefit from such research, the potential societal impact may be more consequential in the transportation domain. This is particularly apparent considering the current race to deployment of automated driving and advanced driving assistance systems on public roads. Such technologies require advanced decision making and motion planning systems that rely on estimates of the future position of road users in order to realize safe and effective mitigation and navigation strategies.

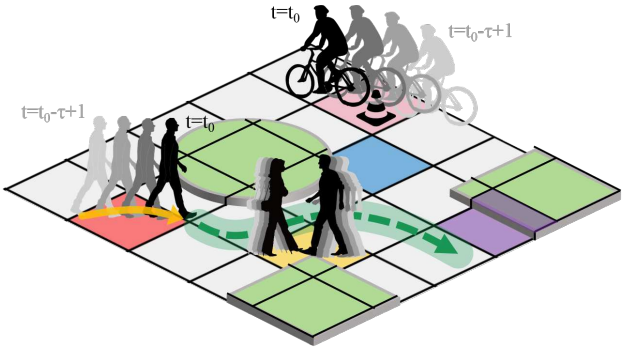


Figure 1: Spatio-temporal features are visually encoded from discretized grid to locally discover (i) human-human (■: woman↔man) and (ii) human-space interactions (■: man↔ground, ■: cyclist↔cone) over time. Then, their pair-wise relations (*i.e.*, ■↔■, ■↔■, ■↔■, ■↔■, ■↔■, ...) with respect to the past motion of the target (→) are investigated from a global perspective for trajectory forecast.

Related research [46, 1, 36, 23, 37, 12, 13, 43, 45, 32, 33, 47] has attempted to predict future trajectories by focusing on social conventions, environmental factors, or pose and motion constraints. They have shown to be more effective when the prediction model learns to extract these features by considering human-human (*i.e.*, between road agents) or human-space (*i.e.*, between a road agent and environment) interactions. Recent approaches [20, 44] have incorporated both interactions to understand behavior of agents toward environments. However, they restrict human interactions to nearby surroundings and overlook the influence of distant obstacles in navigation, which is not feasible in real-world scenarios. In this view, we present a framework where such interactions are not limited to nearby road users nor surrounding medium. The proposed relation-aware approach *fully* discovers human-human and human-space interactions from local scales and learns to infer relations from these interactions from global scales for future trajectory forecast.

Inferring relations of interactive entities has been researched for many years, but the focus is on the implications of relations between the object pair as in [35, 22]. Recently,

[34] introduced the relation network pipeline where 'an object' is a visual encoding of spatial features computed using a convolutional kernel within a receptive field. Our work further expands [34] in the sense that the word 'object' incorporates spatial behavior of entities (road users, if they exist) and environmental representations (road structures or layouts) together with their temporal interactions over time, which naturally corresponds to human-human and human-space interactions (see Figure 1). On top of this, we consider learning to infer relational behavior between objects (*i.e.*, spatio-temporal interactions) for trajectory prediction.

In practice, the relations between all object pairs do not equally contribute to understanding the past and future motion of a specific road user. For example, a distant building behind a car does not have meaningful relational information with the ego-vehicle that is moving forward to forecast its future trajectory. To address the different importance of relations, the prediction model should incorporate a function to selectively weight pair-wise relations based on their potential influence to the future path of the target. Thus, we design an additional relation gate module (RGM) which is inspired by an internal gating process of a long-short term memory (LSTM) unit. Our RGM shares the same advantages of control of information flow through multiple switch gates. While producing relations from spatio-temporal interactions, we enforce the module to identify more descriptive relations that highly influence the future motion of the target by further conditioning on its past trajectory.

An overview of the proposed approach is presented in Figure 2. Our system visually encodes spatio-temporal features (*i.e.*, objects) through the spatial behavior encoder and temporal interaction encoder using a sequence of past images (see Figure 3). The following RGM first infers relational behavior of all object pairs and then focuses on looking at which pair-wise relations will be potentially meaningful to forecast the future motion of the target agent under its past behavior (see Figure 4). As a result, the gated relation encoder (GRE) produces more informative relational features from a target perspective. The next stage of our system is to forecast future trajectory of the target over the next few seconds using the aggregated relational features. Here, we predict future locations in the form of heatmaps to generate a pixel-level probability map which can be (i) further refined by considering spatial dependencies between the predicted locations and (ii) easily extended to learn the uncertainty of future forecast at test time.

The main contributions of this paper are as follows:

1. Encoding of spatio-temporal behavior of agents and their interactions toward environments, corresponding to human-human and human-space interactions.
2. Design of relation gating process conditioned on the past motion of the target to capture more descriptive relations with a high potential to affect its future.
3. Prediction of a pixel-level probability map that can be penalized with the guidance of spatial dependencies and extended to learn the uncertainty of the problem.
4. Improvement of model performance by $14-15\%$ over the best state-of-the-art method using the proposed framework with aforementioned contributions.

2. Related Work

This section provides a review of deep learning based trajectory prediction. We refer the readers to [11, 17] for a review on recognition and prediction of human action, motion, and intention, and [26, 29] for a review on human interaction, behavior understanding, and decision making.

***Human-human interaction* oriented approaches** Discovering social interactions between humans has been a mainstream approach to predict future trajectories [31, 2, 48, 23, 43, 39]. Following the pioneering work [14] on modeling human-human interactions, similar social models have been presented for the data-driven methods. A social pooling layer was proposed in [1] in between LSTMs to share intermediate features of neighboring individuals across frames, and its performance was efficiently improved in [12]. While successful in many cases, they may fail to provide acceptable future paths in a complex road environment without the guidance of scene context.

***Human-space interaction* oriented approaches** Modeling scene context of humans interacting with environments has been introduced as an additional modality to their social interactions. [20] modeled human-space interactions using deep learned scene features of agents' neighborhood, assuming only local surroundings of the target affect its future motion. However, such restriction of the interaction boundary is not feasible in real-world scenarios and may cause failures of the model toward far future predictions. More recently, [44] expanded local scene context through additional global scale image features. However, their global features rather implicitly provide information about road layouts than explicitly model interactive behavior of humans against road structures and obstacles. In contrast, our framework is designed to discover local human-human and human-space interactions from global scales. We locally encode spatial behavior of road users and environmental representations together with their temporal interactions over time. Then, our model infers relations from a global perspective to understand past and future behavior of the target against other agents and environments.

***Human action* oriented approaches** These approaches rely on action cues of individuals. To predict a future trajectory of pedestrians from first-person videos, temporal changes of orientation and body pose are encoded as one of the features in [45]. In parallel, [13] uses head pose as a proxy to build a better forecasting model. Both methods find that gaze, inferred by the body or head orientation, and

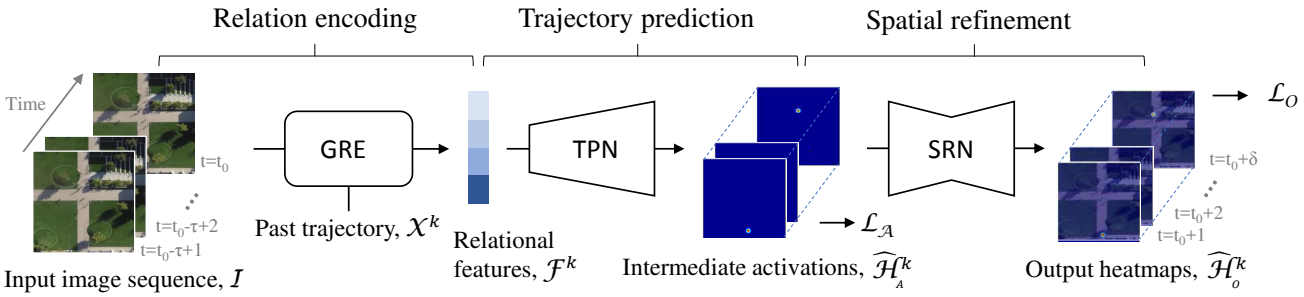


Figure 2: Given a sequence of images, the GRE visually analyzes spatial behavior of road users and their temporal interactions with respect to environments. The subsequent RGM of GRE infers pair-wise relations from these interactions and determines which relations are meaningful from a target agent's perspective. The aggregated relational features are used to generate initial heatmaps through the TPN. Then, the following SRN further refines these initial predictions with a guidance of their spatial dependencies. We additionally embed the uncertainty of the problem into our system at test time.

the person's destination are highly correlated. However, as with human-human interaction oriented approaches, these methods may not generalize well to unseen locations as the model does not consider the road layout.

3. Relational Inference

We extend the definition of 'object' in [34] to a spatio-temporal feature representation extracted from each region of the discretized grid over time. It enables us to visually discover (i) *human-human interactions* where there exist multiple road users interacting with each other over time, (ii) *human-space interactions* from their interactive behavior with environments, and (iii) *environmental representations* by encoding structural information of the road. The pair-wise relations between objects (*i.e.*, local spatio-temporal features) are inferred from a global perspective. Moreover, we design a new operation function to control information flow so that the network can extract descriptive relational features by looking at relations that have a high potential to influence the future motion of the target.

3.1. Spatio-Temporal Interactions

Given τ past images $\mathcal{I} = \{I_{t_0-\tau+1}, I_{t_0-\tau+2}, ..., I_{t_0}\}$, we visually extract spatial representations of the static road structures, the road topology, and the appearance of road users from individual frames using the spatial behavior encoder with 2D convolutions. The concatenated features along the time axis are spatial representations $S \in \mathbb{R}^{\tau \times d \times d \times c}$. As a result, each entry $s_i \in \mathbb{R}^{\tau \times 1 \times 1 \times c}$ of $S = \{s_1, ..., s_n\}$ contains frame-wise knowledge of road users and road structures in i-th region of the given environment. Therefore, we individually process each entry s_i of S using the temporal interaction encoder with a 3D convolution to model sequential changes of road users and road structures with their temporal interactions as in Figure 3. We observed that the joint use of 2D convolutions for spatial modeling and 3D convolution for temporal modeling extracts more discriminative spatio-temporal features as compared to alternative methods such as 3D convolutions as a whole or 2D convolutions with an LSTM. Refer to Section 5.2 for detailed description and empirical validation. The resulting spatio-temporal features $O \in \mathbb{R}^{d \times d \times c}$ contains a visual interpretation of spatial behavior of road users and their temporal interactions with each other and with environments. We decompose O into a set of objects $\{o_1, ..., o_n\}$, where $n = d^2$ and an object $o_i \in \mathbb{R}^{1 \times 1 \times c}$ is a c-dimensional feature vector.

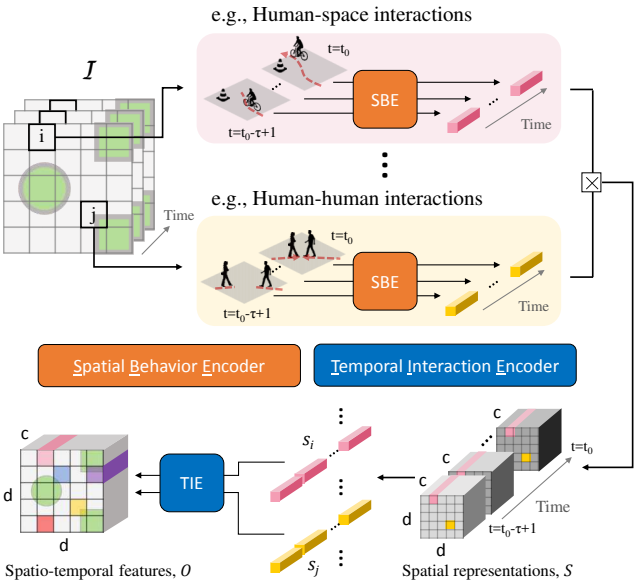


Figure 3: We model human-human and human-space interactions by visually encoding spatio-temporal features from each region of the discretized grid.

3.2. Relation Gate Module

Observations from actual prediction scenarios in road scenes suggest that humans focus on only few important relations that may potentially constrain the intended path,

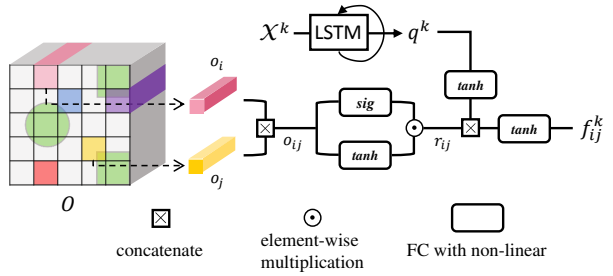


Figure 4: The relation gate module controls information flow through multiple switches and determines not only whether the given object pair has meaningful relations from a spatio-temporal perspective, but also how important their relations are with respect to the motion context of the target.

instead of inferring every relational interactions of all road users. In this view, we propose a module which is able to address the benefits of discriminatory information process with respect to their relational importance.

We focused on the internal gating process of an LSTM unit that controls information flow through multiple switch gates. Specifically, the LSTM employs a sigmoid function with a tanh layer to determine not only which information is useful, but also how much weight should be given. The efficacy of their control process leads us to design a relation gate module (RGM) which is essential to generate more descriptive relational features from a target perspective. The structure of the proposed RGM is displayed in Figure 4.

Let $g_\theta(\cdot)$ be a function which takes as input a pair of two objects (o_i, o_j) and spatial context q^k. Note that q^k is an m-dimensional feature representation extracted from the past trajectory $\mathcal{X}^k = \{X^k_{t_0-\tau+1}, X^k_{t_0-\tau+2}, ..., X^k_{t_0}\}$ of the k-th road user observed in $\mathcal{I}$. Then, the inferred relational features $\mathcal{F}^k$ are described as follows:

$$\mathcal{F}^k = \sum_{i,j} g_\theta(o_i, o_j, q^k), \tag{1}$$

where $\theta = \{\alpha, \beta, \mu, \lambda\}$ is the learnable parameters of $g(\cdot)$. Through the function $g_\theta(\cdot)$, we first determine whether the given object pair has meaningful relations from a spatio-temporal perspective by computing $r_{ij} = tanh_\alpha(o_{ij}) \odot \sigma_\beta(o_{ij})$, where $o_{ij} = o_i \boxtimes o_j$ is the concatenation of two objects. Note that we add $\alpha, \beta, \mu, \lambda$ as a subscript of tanh and sigmoid function to present that these functions come after a fully connected layer. Then, we identify how their relations can affect the future motion of the target k based on its past motion context q^k by $f^k_{ij} = tanh_\lambda(r_{ij} \boxtimes tanh_\mu(q^k))$. This step is essential in (i) determining whether the given relations r_{ij} would affect the target road user's potential path and (ii) reasoning about the best possible route, given the motion history q^k of the target. We subsequently collect all relational information from every pair and perform element-wise sum to produce relational features $\mathcal{F}^k \in \mathbb{R}^{1 \times w}$. Note that the resulting $\mathcal{F}^k$ is target-specific, and hence individual road users generate unique relational features using the same set of objects O with a distinct motion context q^k.

4. Future Trajectory Prediction

The proposed approach aims to predict δ number of future locations $\mathcal{Y}^k = \{Y^k_{t_0+1}, Y^k_{t_0+2}, ..., Y^k_{t_0+\delta}\}$ for the target road user k using $\mathrm{X}^k = \{\mathcal{I}, \mathcal{X}^k\}$. Rather than regressing numerical coordinates of future locations, we generate a set of likelihood heatmaps following the success of human pose estimation in [38, 25, 5]. The following section details how the proposed method learns future locations.

4.1. Trajectory Prediction Network

To effectively identify the pixel-level probability map, we specifically design a trajectory prediction network $a_\psi(\cdot)$ with a set of deconvolutional layers. Details of the network architecture are described in the supplementary material. We first reshape the relational features $\mathcal{F}^k$ extracted from GRE to be the dimension $1 \times 1 \times w$ before running the proposed trajectory prediction network (TPN). The reshaped features are then incrementally upsampled using six deconvolutional layers, each with a subsequent ReLU activation function. As an output, the network $a_\psi(\cdot)$ predicts a set of activations in the form of heatmaps $\widehat{\mathcal{H}}^k_A \in \mathbb{R}^{W \times H \times \delta}$ through the learned parameters ψ. At training time, we minimize the sum of squared error between the ground-truth heatmaps $\mathcal{H}^k \in \mathbb{R}^{W \times H \times \delta}$ and the prediction $\widehat{\mathcal{H}}^k_A$, all over the 2D locations (u, v). The L2 loss $\mathcal{L}_\mathcal{A}$ is as follows: $\mathcal{L}_\mathcal{A} = \sum_\delta \sum_{u,v} \left(\mathcal{H}^k_{(\delta)}(u,v) - \widehat{\mathcal{H}}^k_{A(\delta)}(u,v)\right)^2$. Note that $\mathcal{H}^k$ is generated using a Gaussian distribution with a standard deviation (1.8 in practice) on the ground-truth coordinates $\mathcal{Y}^k$ in a 2D image space. Throughout the experiments, we use heatmaps with $W = H = 128$ which balances computational time, quantization error, and prediction accuracy from the proposed network structures.

4.2. Refinement with Spatial Dependencies

The TPN described in the previous section is designed to output a set of heatmaps, where predicted heatmaps correspond to the future locations over time. In practice, however, the output trajectory is sometimes unacceptable for road users as shown in Figure 5. Our main insight for the cause of this issue is a lack of *spatial dependencies* [28, 40][1] among heatmap predictions. Since the network independently predicts δ number of pixel-level probability maps, there is no constraint to enforce heatmaps to be spatially aligned across predictions. In the literature, [28, 40] have shown that inflating receptive fields enables the network to

[1] Although [28, 40] used the term for kinematic dependencies of human body joints, we believe future locations have similar spatial dependencies between adjacent locations as one follows the other.

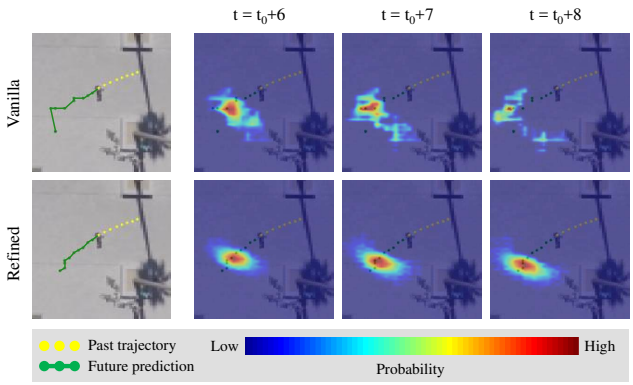


Figure 5: Visual analysis of spatial refinement. The first row shows the predicted future locations from the vanilla trajectory prediction network as presented in Section 4.1. Heatmap predictions are ambiguous, and hence the trajectory is unrealistic. The second row shows the refined locations by considering spatial dependencies as in Section 4.2.

learn *implicit spatial dependencies* in a feature space without the use of hand designed priors or specific loss function. Similarly, we design a spatial refinement network (SRN) with large kernels, so the network can make use of rich contextual information between the predicted locations.

We first extract intermediate activations h_{D5} from the TPN and let through a set of convolutional layers with stride 2 so that the output feature map h_{C17} to be the same size as h_{D2} (earlier activation of TPN). Then, we upsample the concatenated features $h_{\text{C17}} \boxtimes h_{\text{D2}}$ using four deconvolutional layers followed by a 7 x 7 and 1 x 1 convolution. By using large receptive fields and increasing the number of layers, the network is able to effectively capture dependencies [40], which results in less confusion between heatmap locations. In addition, the use of a 1 x 1 convolution enforces our refinement process to further achieve pixel-level correction in the filter space. See the supplementary material for structural details. Consequently, the output heatmaps $\widehat{\mathcal{H}}^k_{\mathcal{O}}$ with spatial dependencies between heatmap locations show improvement in prediction accuracy as shown in Figure 5.

To train our SRN together with optimizing the rest of the system, we define another L2 loss: $\mathcal{L}_{\mathcal{O}} = \sum_{\delta} \sum_{u,v} \left(\mathcal{H}^k_{(\delta)}(u,v) - \widehat{\mathcal{H}}^k_{\mathcal{O}(\delta)}(u,v) \right)^2$. Then the total loss can be drawn as follows: $\mathcal{L}_{optimize} = \zeta \mathcal{L}_{\mathcal{A}} + \eta \mathcal{L}_{\mathcal{O}}$. We observe that the loss weights $\zeta = \eta = 1$ properly optimize our SRN with respect to the learned TPN and GRE.

4.3. Uncertainty of Future Prediction

Forecasting future trajectory can be formulated as an uncertainty problem since several plausible trajectories may exist with the given information. Its uncertainty has been often addressed in the literature [20, 12, 32] by generating multiple prediction hypotheses. Specifically, these approaches mainly focus on building their system based on deep generative models such as variational autoencoders [20] and generative adversarial networks [12, 32]. As the prediction models are trained to capture the future trajectory distributions, they sample multiple trajectories from the learned data distributions with noise variations, addressing multi-modal predictions. Unlike these methods, the proposed approach is inherently deterministic and generates a single trajectory prediction. Thus, our framework technically embeds the uncertainty of future prediction by adopting Monte Carlo (MC) dropout.

Bayesian neural networks (BNNs) [6, 24] are considered to tackle the uncertainty[2] of the network's weight parameters. However, the difficulties in performing inference in BNNs often led to perform approximations of the parameters' posterior distribution. Recently, [8, 9] found that inference in BNNs can also be approximated by sampling from the posterior distribution of the deterministic network's weight parameters using dropout. Given a dataset $\mathbf{X} = \{\mathrm{X}_1, ..., \mathrm{X}_N\}$ and labels $\mathbf{Y} = \{\mathcal{Y}_1, ..., \mathcal{Y}_N\}$, the posterior distribution about the network's weight parameters $\boldsymbol{\omega}$ is as follows: $p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})$. Since it cannot be evaluated analytically, a simple distribution $q^*(\boldsymbol{\omega})$ which is tractable is instead used. In this way, the true model posterior can be approximated by minimizing the Kullback-Leibler divergence between $q^*(\boldsymbol{\omega})$ and $p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})$, which results in performing variational inference in Bayesian modeling [8]. Dropout variational inference is a practical technique [15, 16] to approximate variational inference using dropout at training time to update model parameters and at test time to sample from the dropout distribution $q(\boldsymbol{\omega})$. As a result, the predictive distribution with Monte Carlo integration is as follows:

$$p(\mathcal{Y}|\mathrm{X}, \mathbf{X}, \mathbf{Y}) \approx \frac{1}{L} \sum_{l=1}^{L} p(\mathcal{Y}|\mathrm{X}, \hat{\boldsymbol{\omega}}) \qquad \hat{\boldsymbol{\omega}} \sim q(\boldsymbol{\omega}), \tag{2}$$

where L is the number of samples with dropout at test time.

The MC sampling technique enables us to capture multiple plausible trajectories over the uncertainties of the learned weight parameters. For evaluation, however, we use the mean of L samples as our prediction, which best approximates variational inference in BNNs as in Eqn. 2. The efficacy of the uncertainty embedding is visualized in Figure 6. We compute the variance of $L = 5$ samples to measure the uncertainty (second row) and their mean to output future trajectory (third row). At training and test time, we use dropout after C6 (with drop ratio $r = 0.2$) and C8 ($r = 0.5$) of the spatial behavior encoder and fully connected layers ($r = 0.5$) of the RGM, which seems reasonable to balance regularization and model accuracy.

[2]Uncertainty can be categorized into two types [7]: (i) *epistemic* caused by uncertainty in the network parameters and (ii) *aleatoric* captured by inherent noise. We focus on epistemic uncertainty in this paper.

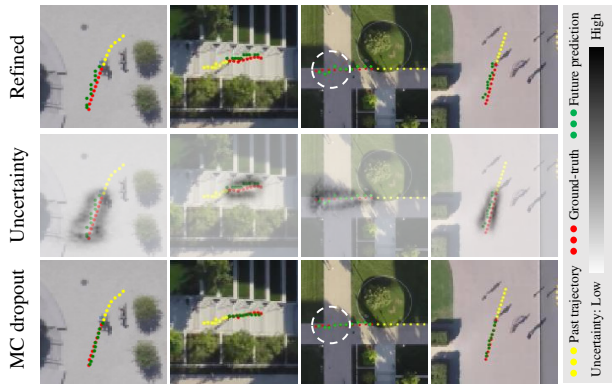


Figure 6: The efficacy of the uncertainty embedding into our framework. We observe that the performance of our model (first row) can be improved with MC dropout (third row). The uncertainty is visualized in the second row.

5. Experiments

We mainly use the SDD dataset [30] to evaluate our approach and use ETH [27] and UCY [21] to additionally compare the performance with the state-of-the-art methods.

5.1. Dataset and Preprocessing

The proposed approach aims to infer relational behavior of agents toward the environment, in addition to that against other agents. For this purpose, SDD [30] fits well due to its diverse scenarios with different types of road obstacles and layouts, captured from a static platform. We exclude outliers following the preprocessing step in [20]. As a result, 19.5 K instances[3] are used to train and test our model. Next, we find a center coordinate of each bounding box and use it to locate a corresponding road user onto images. Note that all RGB images are resized to fit in a 256x256 image template, and the corresponding center coordinates are rescaled to the 128x128 pixel space. Finally, we generate ground-truth heatmaps $\mathcal{H}$ of size 128x128 using the rescaled center coordinates. At training and test time, we use 3.2 sec of past images $\mathcal{I}$ and coordinates $\mathcal{X}^k$ of the target road user k as input and predict 4.0 sec of future frames as heatmaps $\widehat{\mathcal{H}}^k$. For evaluation, we first find a coordinate of a point with a maximum likelihood from each heatmap and further process the coordinates to be the same scale as original images. Then, the distance error between the ground-truth future locations $\mathcal{Y}^k$ and our predictions $\widehat{\mathcal{Y}}^k$ is calculated. We report our performance at 1 / 5 scale as proposed in [20].

5.2. Comparison to Baselines

We conduct extensive evaluations to verify our design choices. Table 1 quantitatively compares the self-generated baseline models by measuring average distance error (ADE) during a given time interval and final distance error (FDE) at a specific time frame in *pixels*.

Spatio-temporal interactions: Encoding spatio-temporal features from images is crucial to discover both human-human and human-space interactions, which makes our approach distinct from others. We first conduct ablative tests to demonstrate the rationale of using spatio-temporal representations for understanding the relational behavior of road users. For this, we compare four baselines[4]: (i) *RE_Conv2D* which discovers only spatial interactions from τ past images using 2D convolutions; (ii) *RE_Conv3D* which extracts both spatial and temporal interactions using a well-known technique, 3D convolutions; (iii) *RE_Conv2D+LSTM* which first extracts spatial behavior using 2D convolutions and then build temporal interactions using LSTM; and (iv) *RE_Conv2D+Conv3D* where we infer spatio-temporal interactions as discussed in Section 3.1. As shown in the second section of Table 1, the performance of the *RE_Conv2D+LSTM* baseline is dramatically improved against *RE_Conv2D* by replacing the final convolutional layer with LSTM. The result indicates that discovering spatial behavior of road users and their temporal interactions is essential to learn descriptive relations. It is further enhanced by using 3D convolutions instead of LSTM, as *RE_Conv2D+Conv3D* achieves lower prediction error than does the *RE_Conv2D+LSTM* baseline. This comparison validates the rationale of our use of 2D and 3D convolutions together to model more discriminative spatio-temporal features from a given image sequence. Interestingly, the *RE_Conv3D* baseline shows similar performance to *RE_Conv2D* that is trained to extract only spatial information. For *RE_Conv3D*, we gradually decrease the depth size from τ to 1 through 3D convolutional layers for a consistent size of spatio-temporal features O over all baselines. In this way, the network observes temporal information from nearby frames in the early convolutional layers. However, it might not propagate those local spatio-temporal features to the entire sequence in the late layers.

Relation gate module: To demonstrate the efficacy of the proposed RGM, we train an additional model *GRE_Vanilla* as a baseline which simply replaces the fully connected layers of *RE_Conv2D+Conv3D* with the proposed RGM pipeline. Note that we match its number of parameters to *RE_Conv2D+Conv3D* for a fair comparison. The third section of Table 1 validates the impact of the RGM, showing the improvements of both ADE and FDE by a huge margin in comparison to the *RE_Conv2D+Conv3D* baseline. The internal gating process of our RGM explicitly determines which objects are more likely to affect the future target motion and allows the network to focus on exploring their relations to the target road user based on the given context. The

[3] [20] might be more aggressively found those of unstabilized images, but we were not able to further remove outliers to match their number.

[4] The baselines with a prefix *RE_* do not employ the proposed gating process but assume equal importance of relations similarly to [34].

Category	Method	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art	S-LSTM [1]	1.93 / 3.38	3.24 / 5.33	4.89 / 9.58	6.97 / 14.57
	DESIRE [20]	- / **2.00**	- / 4.41	- / 7.18	- / 10.23
Spatio-temporal Interactions	*RE_Conv2D*	2.42 / 3.09	3.50 / 5.23	4.72 / 8.16	6.19 / 11.92
	RE_Conv3D	2.58 / 3.24	3.62 / 5.29	4.83 / 8.25	6.27 / 11.92
	RE_Conv2D+LSTM	2.51 / 3.19	3.54 / 5.08	4.60 / 7.54	5.81 / 10.52
	RE_Conv2D+Conv3D	2.36 / 2.99	3.33 / 4.80	4.37 / 7.26	5.58 / 10.27
Relation Gate	*GRE_Vanilla*	1.85 / 2.41	2.77 / 4.27	3.82 / 6.70	5.00 / 9.58
Spatial Refine	*GRE_Deeper*	2.19 / 2.84	3.24 / 4.88	4.36 / 7.44	5.63 / 10.54
	GRE_Refine	1.71 / 2.23	2.57 / 3.95	3.52 / 6.13	4.60 / 8.79
Uncertainty (Ours)	*GRE_MC-2*	1.66 / 2.17	2.51 / 3.89	3.46 / 6.06	4.54 / 8.73
	GRE_MC-5	1.61 / 2.13	**2.44** / 3.85	**3.38** / 5.99	**4.46** / 8.68
	GRE_MC-10	**1.60** / 2.11	2.45 / **3.83**	3.39 / **5.98**	4.47 / **8.65**

Table 1: Quantitative comparison (ADE / FDE in *pixels*) of our approach with the self-generated baselines as well as state-of-the-art methods [1, 20] using SDD [30]. Note that we report our performance at 1 / 5 resolution as proposed in [20].

	ETH_hotel	ETH_eth	UCY_univ	UCY_zara01	UCY_zara02	Average
State-of-the-art						
S-LSTM [1]	0.076 / 0.125	0.195 / 0.366	0.196 / 0.235	0.079 / 0.109	0.072 / 0.120	0.124 / 0.169
SS-LSTM [44]	0.070 / 0.123	0.095 / 0.235	0.081 / 0.131	0.050 / **0.084**	**0.054 / 0.091**	0.070 / 0.133
Ours						
GRE_Vanilla	0.020 / 0.036	0.054 / 0.113	0.067 / 0.129	0.055 / 0.112	0.076 / 0.152	0.048 / 0.094
GRE_Refine	0.019 / 0.034	0.052 / 0.100	0.066 / 0.128	0.054 / 0.100	0.073 / 0.136	0.046 / 0.087
GRE_MC-2	**0.018 / 0.033**	**0.052 / 0.100**	**0.065 / 0.128**	**0.050** / 0.099	0.071 / 0.134	**0.045 / 0.086**

Table 2: Quantitative comparison (ADE / FDE in normalized *pixels*) of the proposed approach with the state-of-the-art methods [1, 44] using the ETH [27] and UCY [21] dataset.

implication is that the use of the RGM is more beneficial for relational inference, and its generalization in other domains is being considered as our future work.

Spatial refinement: In addition to the qualitative evaluation in Figure 5, we quantitatively explore how the proposed spatial refinement process helps to produce more acceptable future trajectory. The *GRE_Refine* baseline is trained using the additional spatial refinement network on top of the *GRE_Vanilla* structure. In Table 1, *GRE_Refine* significantly outperforms *GRE_Vanilla* both in terms of ADE and FDE all over time. It validates that the proposed network effectively acquires rich contextual information about dependencies between future locations from initial activations $\widehat{\mathcal{H}}_A$ in a feature space. To further validate the use of the separate SRN structure, we additionally design a single end-to-end network (*GRE_Deeper*), replacing the shallow TPN of *GRE_Vanilla* with larger receptive fields and adding more layers (D1-D2 and C18-C25). Its performance is even worse than *GRE_Vanilla*. The *GRE_Deeper* baseline experiences the difficulties in training, which can be interpreted as vanishing gradient. Thus, we conclude that the proposed approach with the separate SRN takes advantage of the intermediate supervision with two loss functions ($\mathcal{L}_{\mathcal{A}}$ and $\mathcal{L}_{\mathcal{O}}$), preventing the vanishing gradient problem [40].

Monte Carlo dropout: To validate our uncertainty strategy for future trajectory forecast, we generate a set of *GRE_MC* baselines with a different suffix *-L*, where *L* denotes the number of samples drawn at test time. The fact that any *GRE_MC-L* baselines performs better than *GRE_Refine* certainly indicates the efficacy of the presented uncertainty embedding. By operating along with heatmap prediction, the presented approach eventually helps us to choose the points with the global maximum over the samples. Therefore, the experiments consistently show the decrease in error rate for both near and far future prediction. It is also worth noting that the use of more samples gradually increases the overall performance but introduces a bottleneck at some point as the error rate of *GRE_MC-10* is not significantly improved from *GRE_MC-5*.

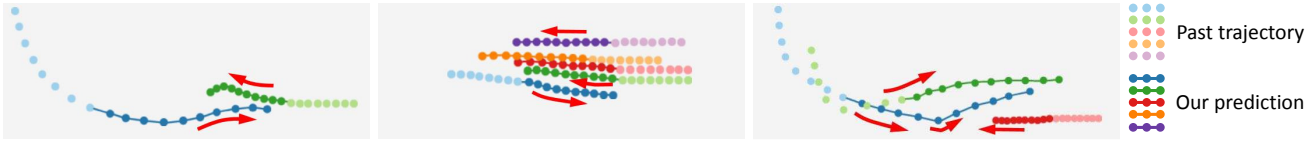
5.3. Comparison with Literature

We quantitatively compare the performance of our models to the state-of-the-art methods using a publicly available SDD dataset [30]. Two different methods are used for fair comparisons, one from *human-human interaction* oriented approaches (S-LSTM [1]) and the other from *human-space interaction* oriented approaches (DESIRE[5] [20]). In Table 1, both ADE and FDE are examined from four dif-

[5]We use *DESIRE-SI-IT0 Best* which shows the best performance among those without using the oracle error metric.

(a) (b) (c) (d)

Figure 7: The proposed approach properly encodes (a) human-human and (b) human-space interactions by inferring relational behavior from a physical environment (highlighted by a dashed arrow). However, we sometimes fail to predict a future trajectory when a road user (c) unexpectedly changes the direction of its motion or (d) does not consider the interactions with an environment. (Color codes: Yellow - given past trajectory, Red - ground-truth, and Green - our prediction)



(a) (b) (c)

Figure 8: Illustrations of our prediction during complicated human-human interactions. (a) A cyclist (●●●) interacts with a person moving slow (●●●). (b) A person (●●●) meets a group of people. (c) A cyclist (●●●) first interacts with another cyclist in front (●●●) and then considers the influence of a person (●●●). The proposed approach socially avoids potential collisions.

ferent time steps. The results indicate that incorporating scene context is crucial to successful predictions as our methods and [20] show a lower error rate than that of [1]. Moreover, all of our models with *GRE* generally outperform [20], validating the robustness of the proposed spatio-temporal interactions encoding pipeline which is designed to discover the entire human-human and human-space interactions from local to global scales. Note that the effectiveness of our approach is especially pronounced toward far future predictions. As discussed in Section 2, the state-of-the-art methods including [1, 20] restrict human interactions to nearby surroundings and overlook the influence of distant road structures, obstacles, and road users. By contrast, the proposed approach does not limit the interaction boundary but considers interactions of distant regions, which results in more accurate predictions toward the far future. Note that ADE / FDE at 4 sec is 5.93 / 10.56 without interactions of distant regions (worse than 5.00 / 9.58 of *GRE_Vanilla*).

In addition to the evaluation using SDD, we perform the experiments on the ETH [27] and UCY [21] dataset, comparing with S-LSTM [1] and SS-LSTM [44]. In Table 2, both ADE and FDE at 4.8 sec are examined in normalized $pixels$ as proposed in [44]. Our approach mostly improves the performance over these methods, further validating our capability of interaction modeling and relational inference.

5.4. Qualitative Evaluation

Figure 7 qualitatively evaluates how inferred relations encourage our model to generate natural motion for the target with respect to the consideration of human-human interactions (7a) and human-space interactions (7b). Both cases clearly show that spatio-temporal relational inferences adequately constrain our future predictions to be more realistic. We also present prediction failures in Figure 7c where the road user suddenly changes course and 7d where the road user is aggressive to interactions with an environment. Extension to incorporate such human behavior is our next plan. In Figure 8, we specifically illustrate more complicated human-human interaction scenarios. As validated in these examples, the proposed approach visually infers relational interactions based on the potential influence of others toward the future motion of the target.

6. Conclusion

We proposed a relation-aware framework which aims to forecast future trajectory of road users. Inspired by the human capability of inferring relational behavior from a physical environment, we introduced a system to discover both human-human and human-space interactions. The proposed approach first investigates spatial behavior of road users and structural representations together with their temporal interactions. Given spatio-temporal interactions extracted from a sequence of past images, we identified pair-wise relations that have a high potential to influence the future motion of the target based on its past trajectory. To generate a future trajectory, we predicted a set of pixel-level probability maps and find the maximum likelihood. We further refined the results by considering spatial dependencies between initial predictions as well as the nature of uncertainty in future forecast. Evaluations show that the proposed framework is powerful as it achieves state-of-the-art performance.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[2] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014.

[3] Samer Ammoun and Fawzi Nashashibi. Real time trajectory prediction for collision risk estimation between vehicles. In *2009 IEEE 5th International Conference on Intelligent Computer Communication and Processing*, pages 417–422. IEEE, 2009.

[4] Chao Cao, Peter Trautman, and Soshi Iba. Dynamic channel: A planning framework for crowd navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310. IEEE, 2017.

[6] John S Denker and Yann Lecun. Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems*, pages 853–859, 1991.

[7] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009.

[8] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *4th International Conference on Learning Representations (ICLR) workshop track*, 2016.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[10] Mingfei Gao, Ashish Tawari, and Sujitha Martin. Goal-oriented object importance estimation in on-road driving videos. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[11] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[12] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[15] Alex Kendall, Vijay Badrinarayanan, , and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[16] A Kendall and R Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *Proceedings-IEEE International Conference on Robotics and Automation*, volume 2016, pages 4762–4769, 2016.

[17] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018.

[18] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[19] Chi-Pang Lam, Chen-Tun Chou, Kuo-Hung Chiang, and Li-Chen Fu. Human-centered robot navigationtowards a harmoniously human–robot coexisting environment. *IEEE Transactions on Robotics*, 27(1):99–112, 2011.

[20] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

[21] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[22] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[23] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4636–4644. IEEE, 2017.

[24] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[26] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S Huang. Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007.

[27] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE.

[28] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.

[29] Amir Rasouli and John K Tsotsos. Joint attention in driver-pedestrian interaction: from theory to practice. *arXiv preprint arXiv:1802.02522*, 2018.

[30] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[31] Mikel Rodriguez, Josef Sivic, Ivan Laptev, and Jean-Yves Audibert. Data-driven crowd analysis in videos. In *ICCV 2011-13th International Conference on Computer Vision*, pages 1235–1242. IEEE, 2011.

[32] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[33] Amir Sadeghian, Ferdinand Legros, Maxime Voisin, Ricky Vesel, Alexandre Alahi, and Silvio Savarese. Car-net: Clairvoyant attentive recurrent network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018.

[34] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[35] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[36] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016.

[37] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1501–1510, 2017.

[38] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[39] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[40] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[41] Thomas Weisswange, Sven Rebhan, Bram Bolder, Nico Steinhardt, Frank Joublin, Jens Schmüdderich, and Christian Goerick. intelligent traffic flow assist: Optimized highway driving using conditional behavior prediction. *IEEE Intelligent Transportation Systems Magazine*, in press, April 2019.

[42] Wenda Xu, Jia Pan, Junqing Wei, and John M Dolan. Motion planning under uncertainty for on-road autonomous driving. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2507–2512. IEEE, 2014.

[43] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5275–5284, 2018.

[44] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194. IEEE, 2018.

[45] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[46] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352, 2011.

[47] Yu Yao, Mingze Xu, Chiho Choi, David J Crandall, Ella M Atkins, and Behzad Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[48] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496, 2015.

[49] Qiuming Zhu. Hidden markov model for dynamic obstacle avoidance of mobile robot navigation. *IEEE Transactions on Robotics and Automation*, 7(3):390–397, 1991.