A Data-Efficient Approach for Long-Term Human Motion Prediction Using Maps of Dynamics

Yufei Zhu[1], Andrey Rudenko[2], Tomasz Piotr Kucner[3], Achim J. Lilienthal[1,4], Martin Magnusson[1]

***Abstract*— Human motion prediction is essential for the safe and smooth operation of mobile service robots and intelligent vehicles around people. Commonly used neural network-based approaches often require large amounts of complete trajectories to represent motion dynamics in complex semantically-rich spaces. This requirement may complicate deployment of physical systems in new environments, especially when the data is being collected online from onboard sensors. In this paper we explore a data-efficient alternative using *maps of dynamics* (MoD) to represent place-dependent multi-modal spatial motion patterns, learned from prior observations. Our approach can perform efficient human motion prediction in the long-term perspective of up to 60 seconds. We quantitatively evaluate its performance with limited amount of training data in comparison to an LSTM-based baseline, and qualitatively show that the predicted trajectories reflect the natural semantic properties of the environment, e.g. the locations of short- and long-term goals, navigation in narrow passages, around obstacles, etc.**

I. INTRODUCTION

Long-term human motion prediction (LHMP) is important for autonomous robots and vehicles to operate safely in populated environments [1]. Accurately predicting the future trajectories of people in their surroundings over extended time periods is essential for enhancing motion planning, tracking, automated driving, human-robot interaction, intelligent safety monitoring and surveillance.

Human motion is complex and may be influenced by several hard-to-model factors, including social rules and norms, personal preferences, and subtle cues in the environment that are not represented in geometric maps. To address these challenges, popular neural network approaches learn motion dynamics directly from data, with many recent studies developing models based on LSTMs [2], GANs [3], CNNs [4], CVAEs [5] and transformers [6]. Most of these approaches focus on learning to predict stochastic interactions between diverse moving agents in the short-term perspective in scenarios where the effect of the environment topology and semantics is minimal.

When predicting long-term human motion in complex, large-scale environments, the influence of the surrounding space (e.g. passages, stairs, entrances, various objects and semantically-meaningful areas) on human motion goes beyond what is contained in the current state of the moving person or the observed interactions. This impact has to be modelled explicitly, for instance by informing the prediction method with a semantic map [7, 8, 9]. Another effective approach to address this challenge is to use *maps of dynamics* (MoDs). MoDs are maps that encode spatial or spatio-temporal motion patterns as a feature of the environment. MoD-informed long-term human motion prediction (MoD-LHMP) approaches are particularly suited to predict motion in the long-term perspective, where the environment effects become critical for making accurate predictions. MoDs efficiently encode the stochastic local motion patterns over the entire map, informing the predictor in areas which may have no influence on the immediate decisions of the walking people, but become critical in the long-term perspective.

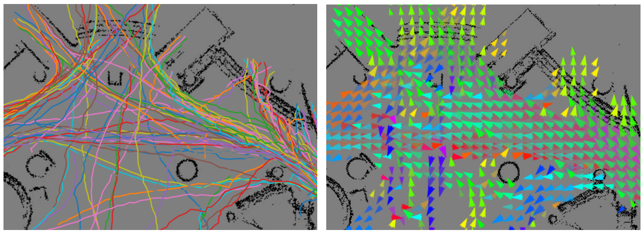

Fig. 1. Maps of dynamics provide an efficient and lightweight encoding of sparse and incomplete velocity data to characterize the motion flows in the environment. We propose a method to predict long-term multi-modal human motion using data-efficient CLiFF maps [10]. **Left:** trajectories from the ATC dataset used for training. **Right:** CLiFF map.

As a proof of concept for MoD-LHMP, we propose to build CLiFF MoDs [10] from training data and use them to bias a constant velocity motion prediction method, generating stochastic trajectory predictions for up to 60 s into the future.

One crucial advantage of the MoD-LHMP approach is its data efficiency. Prior art neural network-based approaches often require large amounts of data for training, and their performance can significantly degrade in absence thereof. Typically, these approaches also need complete sequences of tracked positions for training. The proposed MoD-LHMP approach, on the other hand, allows encoding human motion from sparse and incomplete data, requiring only observed velocities in discrete locations and interpolating the missing motion in between. This property is relevant, for instance, when the deployed robot collects the data in an online fashion from on-board sensors and with a limited field of view.

In this work, we evaluate the efficiency of MoD-based motion encoding for making accurate long-term predictions. In our experiments we sample few trajectories from the ATC dataset and use them to build a CLiFF map and train LSTM-

[1] AASS MRO lab, Örebro University, Sweden yufei.zhu@oru.se

[2] Bosch Corporate Research, Robert Bosch GmbH, Stuttgart, Germany andrey.rudenko@bosch.com

[3] Aalto University, Finland

[4] Chair "Perception for Intelligent Systems", TU Munich, Germany

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017274 (DARKO).

based baselines. We then compare these methods using the ADE/FDE prediction accuracy metrics. Furthermore, we qualitatively demonstrate that the CLiFF-LHMP approach has the ability to predict human motion in complex environments over very long time horizons, implicitly inferring common goal points and correctly predicting trajectories that follow the complex topology of the environment, e.g. navigating around corners or obstacles or passing through narrow passages such as doors.

II. METHOD

A. Maps of Dynamics

In the proposed approach for human motion prediction, we exploit Maps of Dynamics (MoD) which encode human dynamics as a feature of the environment. By using velocity observations, human dynamics can be represented through flow models. In this work, we employ Circular-Linear Flow Field map (CLiFF-map) [10] to represent the flow of human motion. CLiFF-map represents local flow patterns as a multi-modal, continuous joint distribution of speed and orientation. As the orientation of velocity is a circular variable, and speed of velocity is a linear variable, CLiFF-map associates a semi-wrapped Gaussian mixture model (SWGMM) with each location, describing flow patterns around the given location, see Fig. 1. By using SWGMM, CLiFF-map is able to properly address multimodality in the data, thereby enhancing its capability to predict long-term human motion. A CLiFF-map represents motion patterns based on local observations and estimates the likelihood of motion at a given query location. As it can be built from incomplete or spatially sparse data, CLiFF-map efficiently captures human motion patterns without requiring large amounts of data. This characteristic makes CLiFF-LHMP a data-efficient approach for predicting human motion.

B. Motion Prediction

We frame the task of predicting a person's future trajectory as using a short observed trajectory to infer a sequence of future states. The length of the observation history is $O_s \in \mathbb{R}^+$ s, equivalent to an integer $O_p > 0$ observation time steps. With the current time-step denoted as the integer $t_0 \geq 0$, the sequence of observed states is $\mathcal{H} = \langle s_{t_0-1}, ..., s_{t_0-O_p} \rangle$, where s_t is the state of a person at time-step t. A state is represented by 2D Cartesian coordinates (x, y), speed ρ and orientation θ: $s = (x, y, \rho, \theta)$.

From the observed sequence $\mathcal{H}$, we derive the observed speed ρ_{obs} and orientation θ_{obs} at time-step t_0. Then the current state becomes $s_{t_0} = (x_{t_0}, y_{t_0}, \rho_{\text{obs}}, \theta_{\text{obs}})$. The values of ρ_{obs} and θ_{obs} are calculated as a weighted sum of the finite differences in the observed states, as in the popular ATLAS benchmark [11], such that $\rho_{\text{obs}} = \sum_{t=1}^{O_p} v_{t_0-t} g(t)$ and $\theta_{\text{obs}} = \sum_{t=1}^{O_p} \theta_{t_0-t} g(t)$, where $g(t) = (\sigma\sqrt{2\pi} e^{\frac{1}{2}(\frac{t}{\sigma})^2})^{-1}$.

Given the current state s_{t_0}, the goal is to estimate a sequence of future states. Future states are predicted for a given horizon $T_s \in \mathbb{R}^+$ s. T_s is equivalent to $T_p > 0$ prediction time steps assuming the constant time interval Δt between two predictions. Thus, the prediction horizon is $T_s = T_p \Delta t$. The future sequence is then denoted as $\mathcal{T} = \langle s_{t_0+1}, s_{t_0+2}, ..., s_{t_0+T_p} \rangle$.

Algorithm 1: CLiFF-LHMP

```
Input: $\mathcal{H}$, $x_{t_0}$, $y_{t_0}$
Output: $\mathcal{T}$
$\mathcal{T} = \{\}$
$\rho_{\text{obs}}, \theta_{\text{obs}} \leftarrow$ getObservedVelocity($\mathcal{H}$)
$s_{t_0} = (x_{t_0}, y_{t_0}, \rho_{\text{obs}}, \theta_{\text{obs}})$
for $t = t_0 + 1, ..., t_0 + T_p$ do
    $x_t, y_t \leftarrow$ getNewPosition($s_{t-1}$)
    $\theta_s \leftarrow$ sampleVelocityFromCLiFFmap($x_t, y_t$)
    ($\rho_t, \theta_t$) $\leftarrow$ predictVelocity($\theta_s, \rho_{t-1}, \theta_{t-1}$)
    $s_t \leftarrow (x_t, y_t, \rho_t, \theta_t)$
    $\mathcal{T} \leftarrow \mathcal{T} \cup s_t$
return $\mathcal{T}$
```

The CLiFF-LHMP algorithm is presented in Alg. 1. With the input of a CLiFF-map and past states of a person, the algorithm predicts a sequence of future states. To estimate $\mathcal{T}$, for each prediction time step, we sample a velocity from the CLiFF-map at the current position (x_t, y_t) to bias the prediction with the learned motion patterns represented by the CLiFF-map. To sample a velocity at a given location (x, y), we first get the SWGMMs Ξ_{near} whose distances to (x, y) are less than r_s, where r_s is the sampling radius. After getting the sampled velocity, the velocity (ρ_t, θ_t) is predicted by assuming that a person will continue walking with the same speed as in the last time step, $\rho_t = \rho_{t-1}$, and biasing the direction of motion with the sampled orientation θ_s as:

$$\theta_t = \theta_{t-1} + (\theta_s - \theta_{t-1}) \cdot K(\theta_s - \theta_{t-1}), \tag{1}$$

where $K(\cdot)$ is a kernel function that defines the degree of impact of the CLiFF-map. We use a Gaussian kernel with a parameter β that represents the kernel width:

$$K(x) = e^{-\beta \|x\|^2}. \tag{2}$$

With kernel K, we scale the CLiFF-map term by the difference between the velocity sampled from the CLiFF-map and the current velocity according to a constant velocity model (CVM). The sampled velocity is trusted less if it deviates more from the current velocity. A larger β value makes the method behave more like a CVM, and a smaller β makes it more closely follow the CLiFF-map.

III. EVALUATION

In this section, we evaluate the performance of the proposed CLiFF-LHMP approach. We compare our method against LSTM-based human motion prediction methods. Vanilla LSTM [12] is used as the baseline representative of LSTM-based methods.

A. Implementation Details

We evaluate the prediction performance using the ATC dataset [13], which contains trajectories recorded in a shopping mall in Japan. The dataset covers a large indoor environment with a total area of around $900\,\text{m}^2$. The ATC dataset consists of 92 days in total. Given the immense size of the ATC dataset, a subset covering the first four days can be considered representative. We use the subset in the experiments, with the first day (Oct.24) for training, and the remaining 3 days for testing. Both the LSTM and CLiFF-LHMP approaches are trained with same data and also evaluated with same data to ensure a fair comparison.

In ATC dataset, the original detection rate is $30\,\mathrm{Hz}$. We downsample the data to $2.5\,\mathrm{Hz}$ to align with $0.4\,\mathrm{s}$ observation time interval, as commonly used in human motion prediction. For each trajectory, we take $3.2\,\mathrm{s}$ (the first 8 positions) as the observation history and use the remaining trajectory (up to the maximum prediction horizon) as the prediction ground truth. Instead of using a fixed prediction horizon, we explore a wider range of values T_s up to a maximum value in our evaluation. The maximum prediction horizon is determined based on the length distribution of the dataset. We use the 90th percentile value, which is $60\,\mathrm{s}$, as maximum prediction horizon for experiments of ATC dataset. As LSTM-based approaches require complete trajectories for training, we use for all compared approaches trajectories equal or longer than $60\,\mathrm{s}$ for both training and testing.

Given the area and trajectory lengths distribution in the ATC dataset, when evaluating CLiFF-LHMP, we set prediction time step Δt to $1\,\mathrm{s}$, CLiFF-map resolution to $1\,\mathrm{m}$, sampling radius r_s to $1\,\mathrm{m}$ and kernel parameter β to 1. For training vanilla LSTM model, we set the dimension of hidden state of the LSTM model set to 128 and the learning rate set to 0.003.

For the evaluation of the predictive performance we use the *Average* and *Final Displacement Errors* (ADE and FDE) metrics. ADE describes the error between points on the predicted trajectories and the respective ground truth at the same time step. FDE describes the error at the last prediction time step.

We stop predicting when the sample reaches an area outside of the MoD, in case of the CLiFF map, i.e. when no SWGMMs are available within the radius r_s around the sampled location. Predicted trajectories that end before T_s will only be included in the ADE/FDE evaluation up to the last predicted point. When predicting for each ground truth trajectory, the prediction horizon T_s is set either equal to its length or $60\,\mathrm{s}$ for longer trajectories.

B. Experiments and Results

1) Efficiency of motion prediction with limited data: To evaluate the data efficiency of the CLiFF-LHMP method, we run a series of experiments with varying amount of randomly sampled training data, between 100 and 1000 trajectories, to build the CLiFF-map and train the LSTM model.

Figure 2 shows the ADE and FDE results for CLiFF-LHMP and vanilla LSTM for prediction horizon of $60\,\mathrm{s}$, with the number of training trajectories ranging from 100 to 1000. CLiFF-LHMP consistently outperforms LSTM when predicting long-term human motion in these cases. When more than 200 training trajectories are used, the standard deviation of ADE and FDE of CLiFF-LHMP is also lower than for LSTM. While the performance of LSTM drops substantially for smaller training data sets, especially when training with fewer than 200 trajectories, CLiFF-LHMP has a stable performance even with only 100 training trajectories. Compared with using 1000 training trajectories, when training with only 100 trajectories, the error merely increases 2% in ADE and 1% in FDE for CLiFF-LHMP, while for LSTM the ADE increases by 61% and the FDE by 49%. Figure 3 shows a comparison on different prediction horizons from $10\,\mathrm{s}$ to $60\,\mathrm{s}$ for three sizes of the training dataset (200, 600, 1000). When the prediction horizon increases, CLiFF-LHMP becomes slightly more sensitive to the amount of training data.

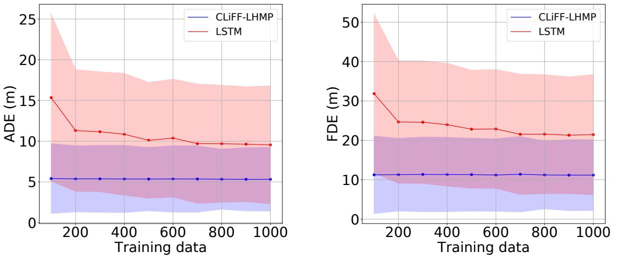

Fig. 2. ADE/FDE of CLiFF-LHMP and LSTM in the ATC dataset, using different amounts of trajectories (100–1000) as training data. The prediction horizon is $60\,\mathrm{s}$. The shade represents one std. dev.

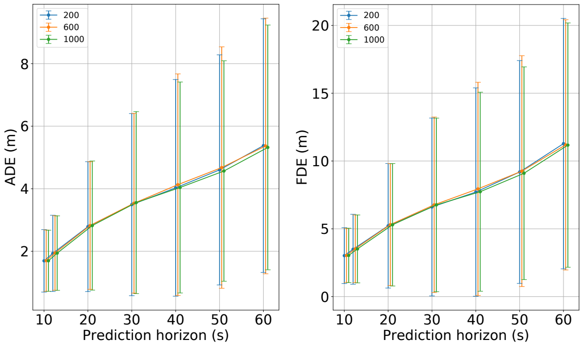

Fig. 3. ADE/FDE of CLiFF-LHMP in the ATC dataset with training dataset of 200, 600, 1000 trajectories and with prediction horizon 10–$60\,\mathrm{s}$, and $12\,\mathrm{s}$.

2) Efficiency of motion representation: To compare the quality of the underlying CLiFF-map itself, trained with different amounts of data, we compute the Kullback-Leibler (KL) divergence [14] between the distributions represented in the CLiFF-maps. The KL divergence results are shown as heatmaps in Figure 4. CLiFF-map associates a Gaussian Mixture Model to each location, and we use a KL divergence heatmap to visualize the changes between two different CLiFF-maps. The first image in Figure 4 shows the changes of CLiFF-maps built with 100 and 1000 trajectories, respectively. It is evident that as the number of training trajectories increases, the primary alterations in the CLiFF-map occur predominantly along the boundary regions. Moreover, in highly constrained environments, such as the eastern corridor of the ATC map, the velocity distributions exhibit comparatively minimal variations. The other three figures in

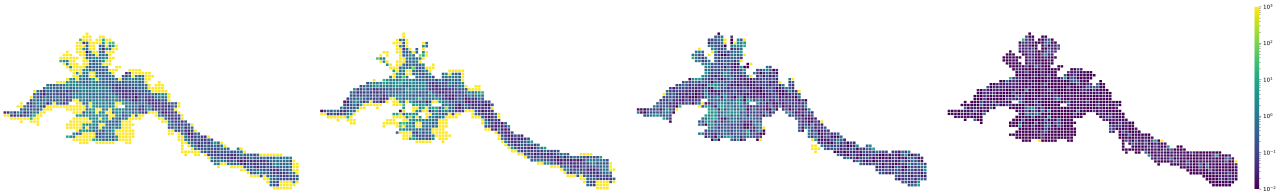

Fig. 4. Heatmap of KL divergence of three pairs of CLiFF-map snapshots built with different amount of training data. **Left:** KL divergence of between CLiFF-maps trained with 1000 and 100 trajectories, **Middle left:** 200 and 100 trajectories, **Middle right:** 600 and 500 trajectories, and **Right:** 1000 and 900 trajectories. The color scale indicates the magnitude of KL divergence, with warmer colors indicating higher values, meaning larger differences between the CLiFF-map pair. When the size of training dataset increases, the impact on the CLiFF-map model becomes less significant, showing that the sensitivity of the CLiFF-map to the amount of training data decreases.

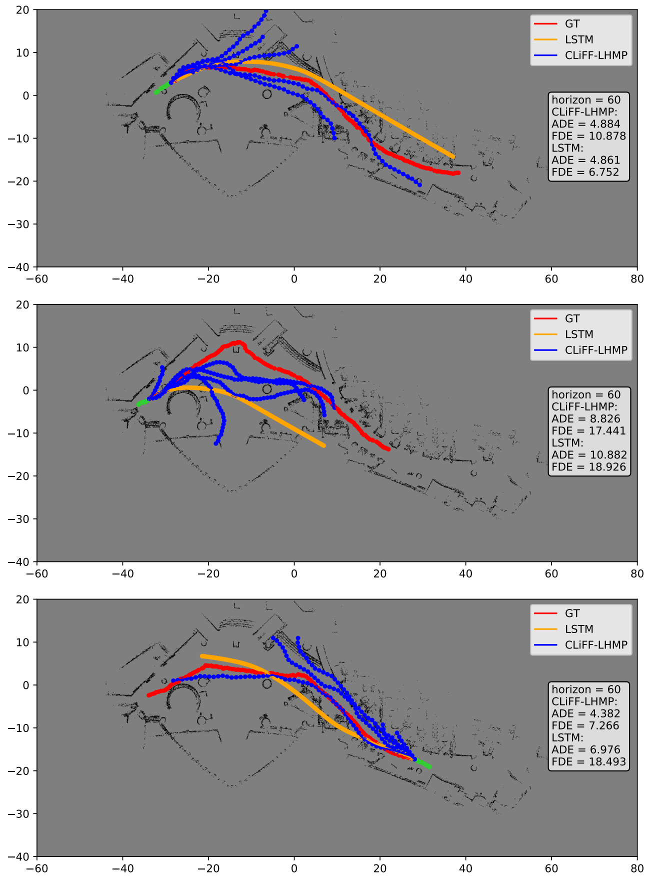

Fig. 5. Predictions in ATC with $T_s = 60$ s. **Red** lines show the ground truth trajectory and **green** line show the observed tracklet. Prediction trajectories of CLiFF-LHMP and LSTM approaches are shown in **blue** and **orange**, respectively. When the trajectory predicted by LSTM is unfeasible by crossing the walls, CLiFF-LHMP make predictions along the corridor.

Figure 4 shows the sensitivity of CLiFF-map to the input data. When the number of training data increases from 900 to 1000 (see the fourth image in Figure 4), the CLiFF-map changes less than when the number of training data increases from 100 to 200 (see the second image in Figure 4). This shows that the CLiFF-map can capture major human motion patterns already with small amounts of training data.

3) Descriptive power of compact motion representation models: Figure 5 shows qualitative examples of predicted trajectories using Maps of Dynamics in the long-term perspective. As no explicit knowledge is given about obstacle layout, LSTM predicts unfeasible trajectory which crosses the walls. In contrast, by exploiting learned motion patterns encoded in the CLiFF-map, our method predict trajectories that follow the complex topology of the environment e.g. navigating around corners or obstacles or passing through narrow passages such as doors, stairs (in the top part of the map) and exits (in the left part).

IV. CONCLUSIONS

In this paper, we present the idea to exploit *Maps of Dynamics* (MoDs) for long-term human motion prediction. As a proof of concept for MoD-LHMP, we propose CLiFF-LHMP. Our method uses the CLiFF-map, a specific MoD that probabilistically represents human motion patterns within a velocity field. Our approach involves sampling velocities from the CLiFF-map to bias constant velocity predictions, generating stochastic trajectory predictions for up to 60 s into the future. We evaluate CLiFF-LHMP using the ATC dataset, with a vanilla LSTM as the baseline approach. The experiments highlight the data efficiency advantage of our method. CLiFF-LHMP is only affected to a minor degree when using less than 200 trajectories as training data, while LSTM requires about three times as many trajectories to reach its optimal performance. The results also demonstrate that our approach consistently outperforms the LSTM method at the long prediction horizon of 60 s. By exploiting learned motion patterns encoded in the CLiFF-map, our method implicitly accounts for the obstacle layouts and predicts trajectories that follow the complex topology of the environment.

A current limitation of CLiFF-LHMP lies in the fact that CLiFF-maps represents spatial motion patterns and are built offline based on past observations. One future direction is the evaluation of additional types of MoDs for long-term human motion prediction, including those capturing temporally-conditioned motion patterns. Another future direction is to learn MoDs online for life-long learning enabling updates based on live motion observations. Additionally, in future work, we aim to formally describe and analyze the MoD-LHMP methodology, include further datasets [15, 16] in the evaluation.

REFERENCES

[1] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. “Human motion trajectory prediction: A survey”. In: *Int. J. of Robotics Research* 39.8 (2020), pp. 895–935.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. “Social LSTM: Human trajectory prediction in crowded spaces”. In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2016, pp. 961–971.

[3] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. “SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints”. In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2019, pp. 1349–1358.

[4] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. “Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction”. In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2020, pp. 14424–14432.

[5] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. “Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 683–700.

[6] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso. “Transformer networks for trajectory forecasting”. In: *Proc. of the IEEE Int. Conf. on Pattern Recognition*. IEEE. 2021, pp. 10335–10342.

[7] J. Wu, J. Ruenz, and M. Althoff. “Probabilistic Map-based Pedestrian Motion Prediction Taking Traffic Participants into Consideration”. In: *Proc. of the IEEE Intell. Veh. Symp. (IV)*. 2018, pp. 1285–1292.

[8] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu. “Multi-Agent Tensor Fusion for Contextual Trajectory Prediction”. In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2019, pp. 12126–12134.

[9] A. Rudenko, L. Palmieri, J. Doellinger, A. J. Lilienthal, and K. O. Arras. “Learning Occupancy Priors of Human Motion From Semantic Maps of Urban Environments”. In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 3248–3255.

[10] T. P. Kucner, M. Magnusson, E. Schaffernicht, V. H. Bennetts, and A. J. Lilienthal. “Enabling Flow Awareness for Mobile Robots in Partially Observable Environments”. In: *IEEE Robotics and Automation Letters* 2.2 (2017), pp. 1093–1100.

[11] A. Rudenko, L. Palmieri, W. Huang, A. J. Lilienthal, and K. O. Arras. “The Atlas Benchmark: an Automated Evaluation Framework for Human Motion Prediction”. In: *Proc. of the IEEE Int. Symp. on Robot and Human Interactive Comm. (RO-MAN)*. 2022.

[12] S. Hochreiter and J. Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–80.

[13] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita. “Person tracking in large public spaces using 3-D range sensors”. In: *IEEE Trans. on Human-Machine Systems* 43.6 (2013), pp. 522–534.

[14] S. Kullback and R. A. Leibler. “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[15] B. Majecka. “Statistical models of pedestrian behaviour in the forum”. In: *Master’s thesis, School of Informatics, University of Edinburgh* (2009).

[16] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal. “THÖR: Human-Robot Navigation Data Collection and Accurate Motion Trajectories Dataset”. In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 676–682.