NeMo-map: Neural Implicit Motion Fields as Spatio-Temporal Prior Map for Shared Human-Robot Environments

Anonymous Authors

***Abstract*— Prior maps such as floor plans, occupancy grids, and semantic maps are widely used to provide geometric and semantic priors for robot autonomy. In environments where robots coexist with people, prior maps can also encode human motion patterns, capturing where and how people typically move, which is equally important for safe and efficient robot operation. Maps of Dynamics (MoDs) provide such motion-pattern priors, but existing representations rely on discrete spatial sampling and struggle to capture smooth spatio-temporal variations. We propose NeMo-map, a continuous spatio-temporal motion-pattern prior map based on implicit neural functions that directly maps coordinates to the parameters of a Semi-Wrapped Gaussian Mixture Model. This formulation removes the need for spatial discretization, enables smooth generalization across both space and time, and preserves multimodal human motion patterns. Evaluated on two public datasets with real-world pedestrian tracking data, NeMo-map achieves higher accuracy in motion representation while remaining computationally efficient. Furthermore, NeMo-map improves long-term human trajectory prediction as a downstream task, demonstrating its value as a complementary motion-pattern prior map alongside conventional geometric and semantic maps for robot autonomy in human-populated environments.**

I. Introduction

Safe and efficient operation in complex, dynamic human environments is a critical prerequisite for deploying robots in various tasks. Prior maps, such as occupancy grids, floor plans, and semantic maps, provide rich geometric and semantic priors that support robot perception, navigation, and planning [1]. However, these conventional prior maps (CPMs) mainly capture the *static* structure of an environment, lacking information about *how people typically move* through a space. In environments where pedestrian flow significantly shapes robot behavior, such as shopping malls, airports, and hospital corridors, knowledge of human motion patterns is essential for scene understanding, accurate long-term prediction and efficient navigation [2].

Maps of Dynamics (MoDs) [3] address this gap by encoding statistical motion patterns as a property of the environment. As such, they provide a complementary type of prior map: while occupancy maps encode *where* agents can move, MoDs encode *how* they typically move. Incorporating MoDs into motion planning provides substantial benefits in crowded environments, since they encode information about expected motion outside of the robot's sensor range, enabling less reactive and more socially aware behavior [4, 5, 6]. MoDs also serve as powerful priors for long-term human motion prediction [7, 8, 9], producing trajectory predictions that implicitly respect the complex topology of the environment, such as navigating around corners and avoiding obstacles, showing capabilities that are difficult to achieve with purely data-driven methods.

Several approaches have been proposed for constructing MoDs. CLiFF-map [10] models local motion patterns with Semi-Wrapped Gaussian Mixture Models (SWGMMs) on a discretized grid, effectively capturing multimodality in human flows. Online CLiFF-map [11] extends this with incremental updates as new observations arrive. STeF-map [12] incorporates temporal periodicity using frequency-domain models but discretizes motion orientations into histograms. However, prior MoDs rely on spatial discretization at a manually selected resolution, introducing information loss and preventing queries at arbitrary locations. Additionally, discretizing time into fixed intervals produces abrupt changes in the representation, even though human motion patterns evolve continuously throughout the day.

To address these challenges, we propose NeMo-map, a *continuous map of dynamics* using implicit neural representations. We learn a neural function that maps spatio-temporal coordinates to parameters of a local SWGMM distribution. This formulation allows the model to smoothly generalize across space and time, while maintaining multimodal motion representations. Unlike discrete MoDs, NeMo-map provides a compact, differentiable, and queryable model of human dynamics that can be integrated with gradient-based training pipelines.

Experiments on two real-world datasets show that NeMo-map achieves higher accuracy in motion representation while remaining computationally efficient. We further demonstrate that NeMo-map significantly improves long-term human trajectory prediction, validating its utility as a motion-pattern prior map for robot autonomy.

II. Method

A. Probabilistic Motion Modeling

A velocity $\mathbf{v}$ is defined by speed ($\rho \in \mathbb{R}^+$) and orientation ($\theta \in [0, 2\pi)$). We model human motion patterns with a Semi-Wrapped Gaussian Mixture Model (SWGMM) [13], which jointly captures circular-linear variables with correlations:

$$p(\mathbf{v} \mid \boldsymbol{\xi}) = \sum_{j=1}^{J} w_j \mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}(\mathbf{v}), \tag{1}$$

where each w_j is a mixing weight, $\boldsymbol{\mu}_j$ the component mean, and $\boldsymbol{\Sigma}_j$ the covariance. Each Semi-Wrapped Normal Distribution (SWND) is defined as $\mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}(\mathbf{v}) = \sum_{k\in\mathbb{Z}} \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\big([\rho,\theta]^\top + 2\pi[0,k]^\top\big)$, where k is a winding number (approximated with $k \in \{-1, 0, 1\}$). This formulation

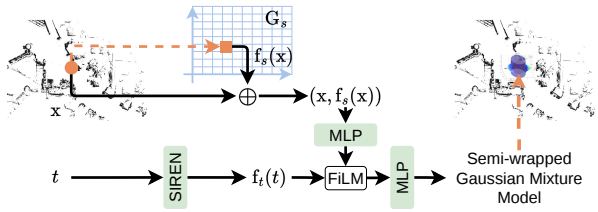

Fig. 1: Method overview. A spatio-temporal query $(\mathbf{x}, t)$ is mapped to SWGMM parameters. The spatial coordinate $\mathbf{x}$ retrieves features from a learnable grid $\mathbf{G}_s$ via bilinear interpolation, and the temporal coordinate t is encoded using a SIREN network. The spatial features, temporal encoding, and raw coordinates are fused via FiLM [14] and passed through an MLP that outputs the SWGMM parameters.

wraps the angular component periodically, yielding a flexible and interpretable probabilistic representation on a cylinder (Fig. 1), capable of capturing multimodality and correlations between orientation and speed at each location.

B. Learning Continuous Motion Fields

Previous MoDs (CLiFF-map, STeF-map) discretize the environment into grid cells and fit local models, leading to information loss and discontinuities. We instead learn a *continuous* neural implicit function Φ_θ that maps a spatio-temporal coordinate $(\mathbf{x}, t)$ to SWGMM parameters:

$$\Phi_\theta(\mathbf{x}, t) = \left\{ w_j(\mathbf{x}, t),\ \boldsymbol{\mu}_j(\mathbf{x}, t),\ \boldsymbol{\Sigma}_j(\mathbf{x}, t) \right\}_{j=1}^{J}. \quad (2)$$

a) Architecture.: The spatial coordinate $\mathbf{x}$ is used to bilinearly interpolate features $\mathbf{f}_s(\mathbf{x}) \in \mathbb{R}^{C_s}$ from a learnable grid $\mathbf{G}_s \in \mathbb{R}^{H \times W \times C_s}$, capturing local spatial variations in motion patterns while remaining continuous. The temporal coordinate t is encoded using SIREN [15], a sinusoidal representation network with periodic activations, producing $\mathbf{f}_t(t) \in \mathbb{R}^{C_t}$. This choice is motivated by the periodic nature of daily human motion patterns. The spatial features are processed by an MLP with ReLU activations, and the temporal encoding is fused via FiLM modulation [14]. The fused representation $\mathbf{z} = [\,\mathbf{x},\, t,\, \mathbf{f}_s(\mathbf{x}),\, \mathbf{f}_t(t)\,]$ is passed to a linear head producing $6J$ outputs that parameterize the SWGMM: for each of J components, the network predicts a mixing weight, a 2D mean (speed and orientation), two variances, and a correlation coefficient.

b) Training.: Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, t_i, \mathbf{v}_i)\}_{i=1}^{N}$ of spatio-temporal motion samples, where $\mathbf{x}_i \in \mathbb{R}^2$ is the spatial position, t_i is a temporal variable, and $\mathbf{v}_i = [\rho_i, \theta_i]^\top$ is the observed velocity, the model is trained by minimizing the negative log-likelihood:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log\, p\big(\mathbf{v}_i \mid \Phi_\theta(\mathbf{x}_i, t_i)\big). \quad (3)$$

At inference, querying Φ_θ at any arbitrary spatio-temporal coordinate yields the full SWGMM, providing a continuous probabilistic representation of motion dynamics without the boundary artifacts of grid-based methods.

Method	NLL↓	Reduction vs Ours	95% CI
Ours	**0.775 ± 2.052**	–	–
Online CLiFF	1.527 ± 4.156	+0.752	[0.749, 0.755]
CLiFF-map	1.964 ± 4.953	+1.189	[1.185, 1.192]
STeF-map	5.576 ± 9.314	+4.801	[4.794, 4.809]

TABLE I: Accuracy on ATC using average NLL (↓). Mean ± std.

Method	ETH	HOTEL	UNIV	ZARA
Ours	**-0.384**	**-0.838**	**0.404**	**-0.342**
CLiFF-map	0.112	0.701	0.518	0.068
Online CLiFF	0.086	1.241	0.577	0.186
STeF-map	2.315	3.349	10.932	2.784

TABLE II: Accuracy on ETH/UCY using NLL (↓).

III. Experiments

A. Setup

We evaluate on two real-world datasets. The ATC dataset [16] contains pedestrian trajectories in a large shopping mall ($\approx 900\,\mathrm{m}^2$), recorded over 92 days and downsampled to $1\,\mathrm{Hz}$. One day is used for training and three for testing. The ETH [17]/UCY [18] datasets contain outdoor pedestrian trajectories across four scenes (ETH, HOTEL, UNIV, ZARA). We compare against CLiFF-map [10], Online CLiFF-map [11], and STeF-map [12].

B. Motion Representation Accuracy

We evaluate MoD quality using negative log-likelihood (NLL) of motion observations, where lower values indicate more accurate representations.

Table I and Table II report the accuracy results for the ATC dataset and the ETH/UCY dataset, respectively. Our method achieves the lowest NLL, outperforming all baselines. Online CLiFF-map, CLiFF-map, and STeF-map exhibit higher NLLs, with paired t-tests showing $p < 0.01$ under the null hypothesis that baseline NLL is less than or equal to ours.

Compared with STeF-maps, methods based on SWGMM, such as ours and CLiFF-map, offer two key advantages. They jointly model speed and orientation, whereas STeF-maps do not include speed information. In addition, SWGMMs represent orientation continuously rather than through a discretized 8-bin histogram as in STeF-map. These aspects lead to a more accurate representation of human motion and contribute to the improved performance.

Limitations of CLiFF-maps are from discretizing the environment into grid cells, with each cell storing a locally fitted SWGMM. This grid-based design limits spatial resolution and introduces discontinuities at cell boundaries in both space and time. In particular, dividing time into hourly intervals is a coarse approximation that can produce abrupt changes, since human motion patterns do not necessarily shift at exact hour boundaries. In contrast, our method models the MoD as a continuous neural implicit representation. This enables smooth generalization across space and time, supports queries at arbitrary spatio-temporal coordinates, and provides a compact representation that avoids the memory overhead of storing distributions for every grid cell.

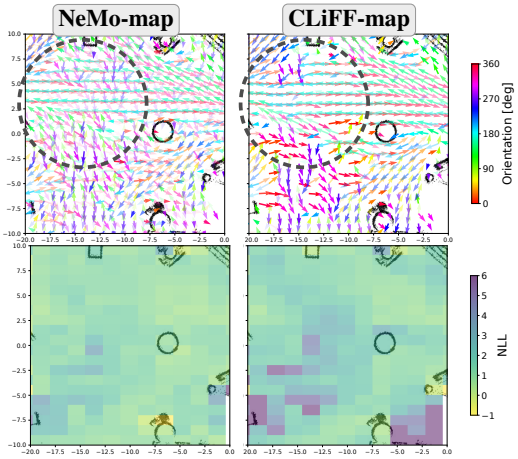

Fig. 2: NeMo-map (**left**) and CLiFF-map (**right**) in the central open area of the ATC mall. **Top**: velocity distributions where arrow color encodes orientation, arrow length encodes speed. The black circle highlights a region where horizontal and vertical flows intersect; NeMo-map preserves both modes while CLiFF-map breaks the vertical flow. **Bottom**: NLL heatmaps (lighter = lower NLL = more accurate).

C. Qualitative Results

a) Spatial multimodality: Fig. 2 compares NeMo-map with CLiFF-map on the central open area of the ATC shopping mall. In the region highlighted by a black circle, a dominant horizontal flow intersects with a vertical flow. NeMo-map preserves both motion modes, maintaining the full continuity of the crossing traffic. In contrast, CLiFF-map fails to capture this multimodality consistently, showing the vertical flow only at the top and bottom of the marked region while cutting off the middle section, resulting in an incomplete representation of the crossing motion. The NLL heatmaps (bottom row) confirm that NeMo-map achieves lower NLL values and more accurate representations across the entire area.

b) Temporal variation: A key advantage of continuous spatio-temporal modeling is the ability to capture smooth changes in motion patterns throughout the day. Fig. 3 shows ATC east corridor at different times. NeMo-map captures the temporal evolution: the dominant flow is directed leftward/upward in the morning (09:00), shifts around noon (12:00), and turns rightward/downward by evening (18:00). Importantly, NeMo-map not only captures the dominant corridor-following flow but also preserves the crossing flow perpendicular to the corridor at all times. In contrast, CLiFF-map, which trains separate models for each hour, produces abrupt transitions between time intervals and fails to represent the crossing flow consistently. The generated flow fields implicitly align with the environment's topology: pedestrian speeds decrease near resting benches, motion flows pass through exits, and flows follow the corridors, even though no explicit map was provided during training.

D. Computational Efficiency

We also compare the map building time as shown in Table III. For the baselines, the training time corresponds to convergence on all grid cells, while for our method it corresponds to the neural network training time.

Method	Train time (min)↓	Inference time (s)↓
Ours	19.26	1.363×10^{-6}
Online CLiFF	23.86	1.914×10^{-3}
CLiFF-map	1831	1.914×10^{-3}
STeF-map	0.82	5.665×10^{-5}

TABLE III: Training and inference times on the ATC dataset.

	Ours	CLiFF	STeF	TUTR	MID
ADE↓	**5.08**	5.45	5.88	12.10	21.20
FDE↓	**10.97**	11.45	12.24	27.26	44.47

TABLE IV: Human motion prediction displacement error results on ATC at 60 s horizon. Lower is better.

Our method trains on a full day of data in under 20 minutes while achieving the highest accuracy, compared to CLiFF-map which requires over 30 hours. STeF-map trains fastest but at the cost of substantially lower accuracy due to its orientation discretization. These results highlight the practicality of continuous MoDs for real-time applications, combining both accuracy and efficiency.

E. Downstream Task: Long-Term Human Motion Prediction

We evaluate NeMo-map as a motion-pattern prior in the downstream task of long-term human motion prediction (LHMP). Accurate LHMP is essential for various applications, including optimized motion planning, advanced automated driving, and improved human-robot interaction. While short-term predictions can often rely on current state and immediate interactions, long-term predictions demand more comprehensive modeling of human motion patterns. Maps of Dynamics (MoDs) provide a way to encode such patterns and can act as a prior that guides motion prediction.

Table IV presents LHMP performance on the ATC dataset with the prediction horizon of 60 s. We compare against MoD-based predictors CLiFF-LHMP [7] and STeF-LHMP [12], a transformer-based baseline TUTR [19] and a diffusion-based baseline MID [20].

NeMo-map achieves the best performance among MoD-based predictors, with paired t-tests showing $p < 0.01$, under the null hypothesis that our method yields equal or higher mean error than the baseline. The improvements arise from its ability to provide smooth and accurate motion fields at arbitrary locations, capturing more precise local flow patterns than grid-based CLiFF-LHMP. Deep learning baselines (TUTR, MID) perform worse at long horizons, often producing physically infeasible trajectories that cross walls, whereas MoD-guided predictions implicitly respect the environment's topology.

Fig. 5 shows a qualitative prediction example. Predictions guided by NeMo-map follow the corridor structure realistically, while deep learning baselines produce trajectories that cross walls.

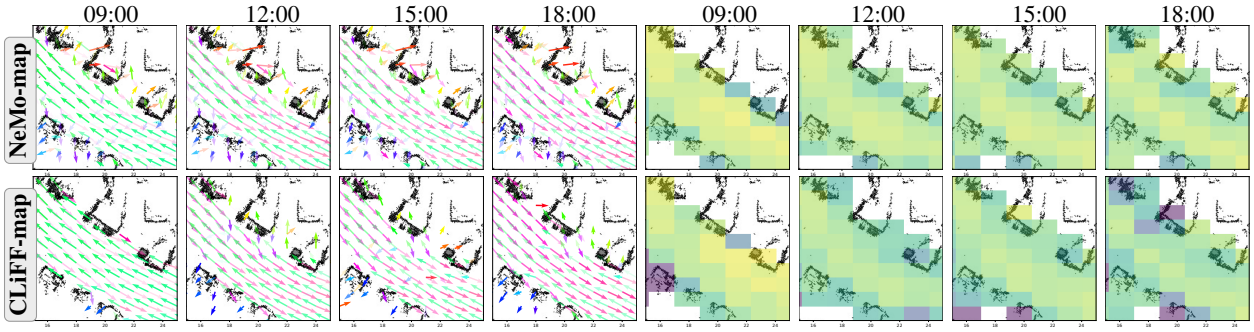

Fig. 3: NeMo-map and the baseline MoD in the east corridor, located on the right side of the ATC map (Fig. 5), at different times of the day. The **left** four columns show the generated velocity fields. Nemo-map captures the temporal variation of pedestrian flows. The **right** four columns show the corresponding NLL heatmaps. Compared with the baselines, NeMo-map preserves both the dominant corridor following flow and the crossing flow, leading to consistently lower NLL values.

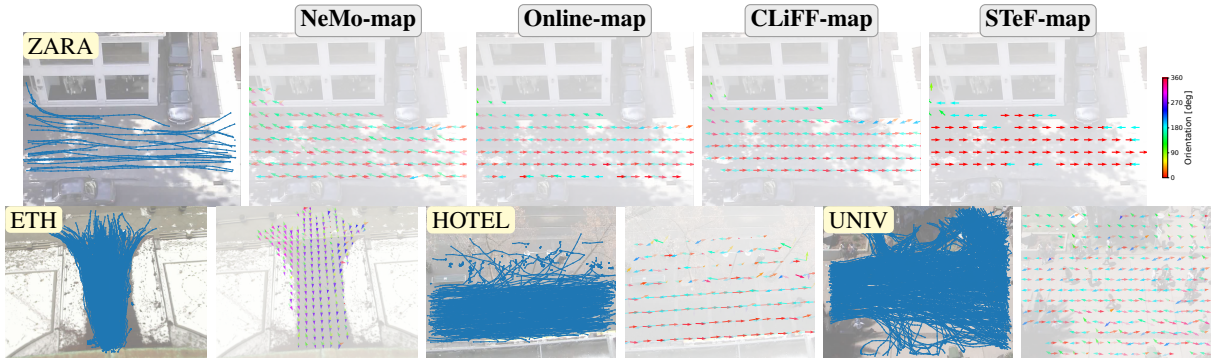

Fig. 4: **Top row**: NeMo-map and baseline MoDs in the ZARA scene at frame 600. Pedestrian trajectories are shown as reference. NeMo-map adapts to the changes from dominant horizontal flow to an emerging flow toward the upper-left entrance, while baseline MoDs remain aligned with the horizontal movement. **Bottom row**: NeMo-map results for the ETH, HOTEL and UNIV scenes with trajectories from the entire sequences. It captures multimodal human motion dynamics across different environments.

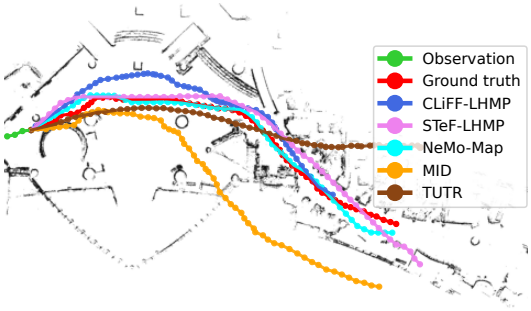

Fig. 5: Human motion prediction with a 60 s horizon. The **red** line represents the ground truth trajectory and the **green** line represents the observed trajectory. Prediction with NeMo-map achieves more accurate and realistic predictions than baseline methods. While the trajectories predicted by TUTR [19] and MID [20] are unfeasible (e.g., crossing the walls), predictions guided with MoDs follow human flow and implicitly respect the structure of the environment over long horizons.

IV. Conclusion

We introduced NeMo-map, a continuous spatio-temporal map of dynamics based on implicit neural representations. Unlike prior discretized methods, NeMo-map parametrizes a continuous neural function that outputs SWGMM parameters at arbitrary spatio-temporal coordinates, enabling smooth generalization across space and time with a compact representation. Experiments on the ATC and ETH/UCY datasets demonstrate that NeMo-map achieves substantially higher accuracy than existing MoD baselines while remaining computationally efficient. We further show that the learned motion fields serve as an effective prior for long-term human trajectory prediction, outperforming both MoD-based and deep learning baselines.

As a motion-pattern prior map, NeMo-map offers a complementary perspective to conventional geometric and semantic prior maps. While occupancy grids and floor plans encode *where* agents can move, NeMo-map encodes *how* they typically move, providing distributional information about expected human motion patterns that can guide robot navigation, motion planning, and prediction in human-populated environments.

Limitations and future work. The proposed method has limitations in handling sharp spatial or temporal discontinuities, such as temporary barriers or sudden event-driven changes. Adapting to such changes requires additional observations, which is a limitation shared by existing MoD approaches. Future work will extend this formulation with online update mechanisms to adapt continuously to evolving crowd behaviors, and to better handle sharp, event-driven discontinuities in motion patterns, further bridging the gap toward long-term real-world deployment.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard. "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age". In: *Trans. Rob.* 32.6 (Dec. 2016), pp. 1309–1332. ISSN: 1552-3098.

[2] 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.

[3] T. P. Kucner et al. "Survey of maps of dynamics for mobile robots". In: *Int. J. of Robotics Research* (2023).

[4] Y. Zhu, S.-M. Yang, M. Magnusson, and A. Wang. "HiCrowd: Hierarchical Crowd Flow Alignment for Dense Human Environments". In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2026.

[5] L. Palmieri, T. P. Kucner, M. Magnusson, A. J. Lilienthal, and K. O. Arras. "Kinodynamic motion planning on Gaussian mixture fields". In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2017, pp. 6176–6181.

[6] C. S. Swaminathan, T. P. Kucner, M. Magnusson, L. Palmieri, S. Molina, A. Mannucci, F. Pecora, and A. J. Lilienthal. "Benchmarking the utility of maps of dynamics for human-aware motion planning". In: *Frontiers in Robotics and AI* 9 (2022).

[7] Y. Zhu, A. Rudenko, T. P. Kucner, L. Palmieri, K. O. Arras, A. J. Lilienthal, and M. Magnusson. "CLiFF-LHMP: Using Spatial Dynamics Patterns for Long-Term Human Motion Prediction". In: *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*. 2023.

[8] Y. Zhu, H. Fan, A. Rudenko, M. Magnusson, E. Schaffernicht, and A. J. Lilienthal. "LaCE-LHMP: Airflow Modelling-Inspired Long-Term Human Motion Prediction By Enhancing Laminar Characteristics in Human Flow". In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2024.

[9] Y. Zhu, A. Rudenko, T. P. Kucner, A. J. Lilienthal, and M. Magnusson. "Long-Term Human Motion Prediction Using Spatio-Temporal Maps of Dynamics". In: *IEEE Robotics and Automation Letters* 10.11 (2025), pp. 12229–12236.

[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] Y. Zhu, A. Rudenko, L. Palmieri, L. Heuer, A. J. Lilienthal, and M. Magnusson. "Fast Online Learning of CLiFF-maps in Changing Environments". In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2025.

[12] S. Molina, G. Cielniak, and T. Duckett. "Robotic Exploration for Learning Human Motion Patterns". In: *IEEE Trans. on Robotics and Automation (TRO)* (2022).

[13] A. Roy, S. K. Parui, and U. Roy. "SWGMM: a semi-wrapped Gaussian mixture model for clustering of circular-linear data". In: *Pattern Anal. Appl.* 19.3 (2016), pp. 631–645.

[14] E. Perez, F. Strub, H. Vries, V. Dumoulin, and A. C. Courville. "FiLM: Visual Reasoning with a General Conditioning Layer". In: *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*. 2018.

[15] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. "Implicit Neural Representations with Periodic Activation Functions". In: *Advances in Neural Inf. Proc. Syst. (NeurIPS)*. 2020.

[16] 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.

[17] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. "You'll never walk alone: Modeling social behavior for multi-target tracking". In: *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*. 2009, pp. 261–268.

[18] A. Lerner, Y. Chrysanthou, and D. Lischinski. "Crowds by example". In: *Computer Graphics Forum*. Vol. 26. 2007, pp. 655–664.

[19] L. Shi, L. Wang, S. Zhou, and G. Hua. "Trajectory Unified Transformer for Pedestrian Trajectory Prediction". In: *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*. 2023, pp. 9675–9684.

[20] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu. "Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion". In: *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*. 2022.