Agent Behaviour Prediction for Urban Autonomous Driving

Dmytro Zabolotnii[1]

[1]Institute of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu

Abstract

Behaviour prediction is a key aspect of urban autonomous driving and is required to ensure the safe and comfortable operation of emerging robo-taxis and self-driving assistants for private vehicles. Prediction of human motion is of particular interest due to the high concentration of pedestrians in urban environments, the unpredictability of their motion, and the high risk of injury in collisions. While the task of motion prediction is well established in Robotics and Autonomous Driving, the existing literature rarely delves into the practical aspects of implementing these solutions within existing autonomous driving software and evaluating them in real-world experimental conditions. Another common issue is the under-utilisation of available sensor data during the prediction process, with reliance instead on a heavily reduced input data modality. In this paper, we present the progress of our PhD study, focused on developing and evaluating real-time pedestrian motion prediction methods for implementation in modular autonomous driving stacks.

Keywords

Autonomous vehicle navigation, computer vision for transportation, datasets for human motion, human detection and tracking, intention recognition

1. Introduction

Autonomous driving vehicles are already part of day-to-day life, with Waymo spearheading deployment in cities across the United States, Japan [1] and Australia [2], promising safer, less accident-prone travel for passengers compared to human-driven vehicles [3]. Still, autonomous vehicles have to interact with other traffic agents, and in the context of dense urban areas, considering pedestrians' motion is critical [4]. The task is non-trivial, and a combined solution of behavioural science and complex engineering is needed to ensure the ego-vehicle's safe and appropriate responses to complex traffic scenarios involving variable traffic rules and a mix of separate types of other agents' intentions and their own interactions. At the same time, passenger safety, comfort, and needs also need to be considered, creating a balancing act with many variables.

It is also necessary to distinguish between the prediction of pedestrian motion and the prediction of other agents' motion - commonly vehicles and bicycles. The key difference is the much smaller effect of inertia, which, in the case of vehicles, strongly reduces the possible future trajectory directions [5]. Additionally, the pedestrians have a different level of awareness than other road agents [6], different social rules and understanding of them (in context to jaywalking acceptance [7]) and as such often uniquely approach different urban environments due to qualities that are not directly observable by the ego-vehicle. The theoretical autonomous vehicle can encounter a young runner jaywalking on a highway and a blind elderly person crossing a small street on the same day. It should efficiently assess a situation in real time and adjust its course of action accordingly.

Waymo historically followed the modular autonomous-driving approach [8]. In contrast to the end-to-end framework, every function in the autonomous driving framework is realised as a semi-independent module with its own inputs and outputs. In this format, prediction is a sub-task of the perception module and depends on the upstream detection and tracking modules. The detection module governs the classification of the input sensor data into recognisable objects - pedestrians, vehicles,

Baltic DB&IS 2026 Conference Forum and Doctoral Consortium, 28 June - 1 July 2026, Tartu, Estonia
dmytro.zabolotnii@ut.ee (D. Zabolotnii)
0000-0001-6524-9454 (D. Zabolotnii)

static structures, etc. The tracking module handles tracking and cross-referencing detected objects from separate sensor readings over time. Finally, prediction takes all available sensor data associated with a specific object over the time period t_{hist}, and uses it to predict the future object trajectory over the time period t_{fut}.

The specific focus on pedestrian motion prediction within the modular autonomous driving framework serves as the main research objective: *How can we accurately predict a pedestrian's future trajectory in a complex urban scenario in real time, with limited resources?*. An additional limitation of constrained resources is that all modules of the autonomous driving framework run on a single set of hardware, which is constrained by the vehicle's space and the budget allocated to internal hardware. An additional concern is backwards hardware compatibility - while the autonomous driving framework software can be upgraded continuously and remotely, the hardware in the deployed vehicle fleet is costly to upgrade and, in the case of privately owned vehicles, may be impossible without issuing a product recall that is both lengthy and legally complicated.

This paper introduces the research work for an ongoing PhD thesis. The paper is structured as follows: Section 2 introduces the primary research questions and their evolution throughout the thesis. Section 3 presents the research methods employed to answer the research questions. Section 4 goes over current results. Section 5 presents work in progress. Section 6 contains concluding remarks.

2. Research questions

The aim of the ongoing thesis is to develop an efficient pedestrian motion prediction solution that is deployable within existing/future modular autonomous driving frameworks. To achieve this, a thorough investigation of the existing agent-prediction literature across the robotics and autonomous driving fields was conducted. However, it quickly becomes apparent that while available research is vast, it is rarely evaluated in the field, but rather on static datasets. The crucial problem with this approach is the so-called dynamics gap [9] - in real life, the prediction module's results affect downstream modules, including the control that transfers to the motion of the ego-vehicle, and other agents react to it, whereas during static dataset evaluation, there is no such feedback. This immediately questions the transferability of the reported results to the solution performance when implemented in the autonomous driving framework (see Fig. 1) and, as such, creates the first research question:

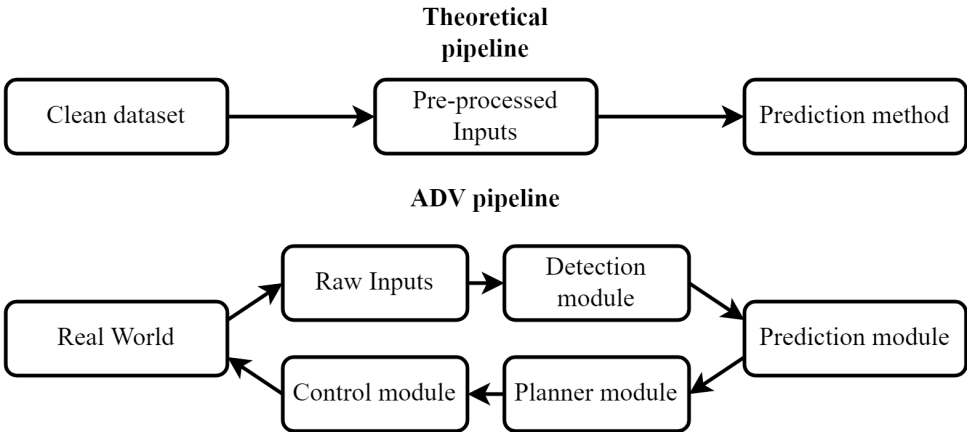

Figure 1: Architectural difference between implementation of state-of-the-art (SOTA) motion prediction methods and their potential implementation in a modular Autonomous Driving (AD) framework.

[RQ1]: Does the state-of-the-art motion prediction solution achieve comparable performance inside AD framework under time constraints?

In RQ1, our objective was to identify a top-of-the-line open-source motion prediction solution, adapt it to the existing open-source Autoware Mini framework [10], and evaluate it against other solutions and simple baselines. During the RQ1 research, it also became clear that existing solution rarely

fully incorporate all the information from the sensors, instead relying on simple representation of past pedestrian i trajectory encoded as the collection of the points over H previous states $X_i = x_i^1, x_i^2, ..., x_i^H$, while dropping potential meaningful data in the full LIDAR or camera data of the same pedestrian. This led to RQ2:

[RQ2]: How to incorporate camera and/or LIDAR input data into existing motion prediction solutions, and whether this leads to prediction quality improvement?

In RQ2, we validated the incorporation of the additional input data in the form of pedestrian gaze direction on the simpler knowledge-based method adapted because of RQ1 results, which allowed us to clearly see the impact of the additional input variable and gradually change its impact. However, the chosen approach still discarded available image information, reducing it to a single vector representing gaze direction.

[RQ3]: How to construct a state-of-the-art solution that incorporates as much of the additional input modalities as possible and validate it inside AD framework?

In RQ3, we combine best practices from SOTA prediction models and recent advances in image/LIDAR processing to create a single model that processes standard trajectory input and additional input modalities simultaneously, achieving performance comparable to SOTA models during validation within the Autoware Mini AD framework and on standard evaluation datasets. The solution formulated and validated during RQ3 constitutes the primary outcome of this thesis research and can potentially be commercially or freely adapted to additional AD frameworks, but this is outside the current research scope.

3. Research methods

In this section we outline research methods employed in answering every research question separately.

[RQ1]: Does the state-of-the-art motion prediction solution achieve comparable performance inside AD framework under time constraints?

To answer RQ1, we began by conducting and drafting a systematic literature review of existing motion prediction methods. However, upon realising that the reported results are not directly applicable to our research due to a dynamics gap, we shifted our research to evaluating several selected SOTA methods. Following existing literature reviews [11, 12, 4, 13, 14] and cross-referencing with recently published novel methods, we have chosen 5 SOTA machine-learning-based methods with open-source code and available pre-trained weights on the standard dataset. We adapted their code to run as a separate module of the Autoware Mini AD framework running in ROS. Using the .bag recordings from previous trips made by the Lexus vehicle running Autoware Mini on the streets of Tartu, Estonia, we evaluated the performance of the methods across standard metrics.

[RQ2]: How to incorporate camera and/or LIDAR input data into existing motion prediction solutions, and whether this leads to prediction quality improvement?

To answer RQ2, we first researched other applications of the data derived from camera/LIDAR. The main candidates are pedestrian gaze, used for predicting pedestrian intention to cross the road [15], general pedestrian motion prediction [16], and pedestrian skeleton, which is widely used for 3D-space pedestrian motion prediction [17]. However, as 3D space pedestrian motion prediction usually requires higher-resolution/sensor-size readings than are available from AD vehicle sensors, the gaze was selected. To incorporate gaze direction, we extended a pedestrian motion predictor built upon the standard Reciprocal Velocity Obstacle (RVO) model [18]. RVO is a knowledge-based model that relies on a constructed motion model rather than a learned one, unlike the models evaluated in RQ1. Similarly, we adapted the RVO model within the Autoware Mini framework and evaluated its performance relative to the baseline and RQ1 models.

[RQ3]: How to construct a state-of-the-art solution that incorporates as much of the additional input modalities as possible and validate it inside AD framework?

To answer RQ3, we build upon lessons learned from RQ1 and RQ2. We choose the DINOv3 model [19] and finetune it using the LoRA approach for a learnable way to incorporate image features not

limited to pedestrian gaze or skeleton. While knowledge-based methods have their own advantages in interpretability and customisation, they tend to perform poorly on standard datasets that machine-learning-based models learn well. As a result, the main skeleton for the prediction solution is the recent transformer-based deep learning architecture. The work is ongoing with more details available in Section 5. Final validation of the constructed model will be performed both on multiple static datasets and inside Autoware Mini framework against all methods adapted before.

4. Current results

The first 18 months of the PhD thesis were focused on narrowing research direction, formulating RQ1, and following research, adaptation engineering, evaluation, and writing the published paper [20]. Following RQ1 formulation, 5 existing SOTA pedestrian prediction methods were selected for further adaptation; details are available in table 1.

PECNet [21], SGNet [22], and MUSE [23] belong to the CVAE (Conditional Variational Autoencoders) model architecture family that encodes input (in this case, past pedestrian trajectories) to the latent space and samples future trajectories from it. SGNet encodes only the predicted agent's past trajectory at runtime, whereas PECNet encodes the predicted agent's past trajectory and the past trajectories of nearby pedestrians, modelling social interactions. MUSE additionally encodes the semantic pixel representation of the map, where each pixel has a binary value indicating whether the underlying area is drivable or non-drivable.

GATraj [24] instead uses a vector-level map representation and agents' locations to create graph nodes, with edges encoding spatial and interaction relationships. While the original paper supports map input, as we lack a compatible vector map representation of our area, a simplified version was used, where only agent locations and their past trajectories are processed. Nevertheless, the simplified GATRaj model showed the best results during testing among all selected methods in the published paper [25]. LED [26] represented at the time cutting-edge development in the field of diffusion models, which denoise future trajectory samples step-by-step and found a great success in testing in the sports-focused datasets like the NBA dataset [27]. However, while the method claimed performance improvements, our tests have shown the method to be too slow for real-time processing within an autonomous driving framework. Finally, we chose the Constant Velocity Method as our baseline. This simple physics-based method assumes that the agent goes in the same direction with the same velocity vector as its last sensor reading.

The methods were adapted for the Autoware Mini AD framework that is running on a Lexus RX450h vehicle deployed on the streets of Tartu, Estonia. The vehicle is equipped with Ouster OS1-128 and Velodyne VLP-32C lidars as the primary sensors for object detection, and two Mako frontal cameras for auxiliary tasks. All sensor data recorded during regular autonomous and manual trips is stored inside .bag file format. This raw sensor data collection serves as the primary dataset. During evaluation, the raw sensor data is replayed according to recorded timestamps and processed through the full framework, running all detection, tracking, prediction, and planning modules (see Fig. 2. This allows for evaluation

Table 1
Overview of chosen pedestrian motion prediction methods evaluated in RQ1

Model	Year	Architecture	Input	Cross-agent interaction consideration	Output Modality
PECNet [21]	2020	CVAE	Trajectory	Yes	Stochastic
SGNet [22]	2022	CVAE	Trajectory	No	Stochastic
GATraj [24]	2023	GNN + Attention	Trajectory	Yes	Probabilistic
MUSE [23]	2022	CVAE	Trajectory+Map	No	Probabilistic
LED [26]	2023	Diffusion Model	Trajectory	Yes	Probabilistic
CVM	-	Physics-based	Final Velocity	No	Deterministic

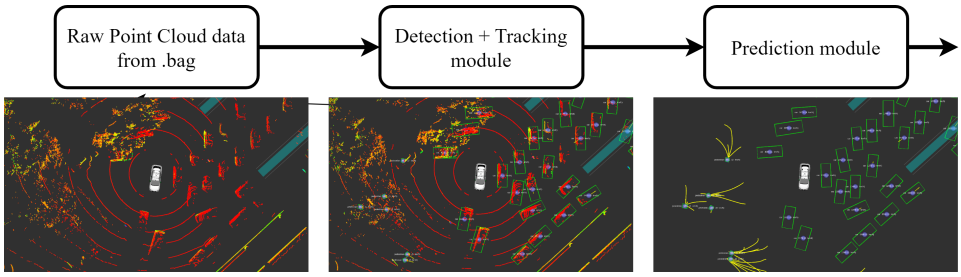

Figure 2: Data flow inside Autoware Mini framework [20] Detection module extracts the shapes of the objects from raw point cloud, and classifies them to pedestrian/vehicle/other objects. Afterwards, selected prediction model outputs candidate trajectories, represented here as yellow curves.

under conditions close to those in real life. Unfortunately, there is no reactivity to the ego-vehicle's motion during the data reply; the upstream half of the full framework is working correctly. Experiments to achieve full reactivity were attempted, resulting in highly non-deterministic evaluations with limited replicability.

During the following 12 months, work on formulating and researching RQ2, subsequent experiments, and publication [28] was performed. As previously established, pedestrian gaze direction was established as one auxiliary input that can be derived from a pedestrian image crop. We employed the Real-time 6DoF Full-Range Markerless Head Pose Estimation method [29] and converted the head pose heading in 3D space to a gaze direction vector in a Birds-Eye-View 2D vector. Next, we proposed FOV-RVO, an extension of the well-researched RVO model. In the RVO paradigm, for each agent, the range of velocities that will lead to collision with other agents is calculated, and the minimal viable velocity deviation that will not lead to collision is predicted. Our contribution enabled us to incorporate both gaze direction and driveable area map data (encoded in arbitrary polygons) as additional constraints on the range of potential velocities for the RVO model. As in RQ1, we adapted FOV-RVO to serve as a module in Autoware Mini and evaluated it on the .bag dataset vs the baseline and selected methods from RQ1. The results are available in Table 2. These results are valid for the contribution related to RQ2; RQ1 contribution results differ due to differences in the dataset and the upstream detection and tracking module, and are available in the linked publication [20].

We have performed evaluation experiments for a variable number of candidate future trajectories, K, using standard minimal Average Displacement Error/minimal Final Displacement Error/Miss Rate metrics. While machine-learning-based models like GATraj outperform our FOV-RVO model when they are allowed to output a higher amount of trajectories (FOV-RVO outputs only 1 due to the architecture), they do not outperform with equal K, and high K values negatively affect downstream modules in the AD framework - planning and control. We also evaluate the FOV-RVO+ model, which combines FOV-RVO prediction with the best prediction from the GATraj model, and combined $K = 2$ trajectories outperform all machine-learning-based models with much higher K. We also evaluate the non-Drivable Area Coverage metric, which determines whether predicted candidate trajectories fall into areas where pedestrians are allowed to walk (sidewalks, crosswalks), and the high performance of FOV-RVO confirms

Table 2
Quantitative results with varying amount of candidate trajectories K, **bold** is the best entry and underlined is the second best entry

Model	FOV-RVO	FOV-RVO+	PECNet			SGNet			GATraj		
K	1	2	1	5	10	1	5	10	1	5	10
.bag dataset											
minADE (m) ↓	0.973	0.867	1.497	1.357	1.311	1.541	0.982	0.883	1.716	0.979	**0.828**
minFDE (m) ↓	1.783	1.566	2.791	2.436	2.346	2.885	1.844	1.680	3.321	1.691	**1.371**
MR_2 ↓	0.322	0.278	0.488	0.436	0.421	0.624	0.348	0.310	0.622	0.267	**0.193**
nonDAC ↑	**0.976**	0.944	0.964	0.964	0.963	0.934	0.954	0.949	0.922	0.919	0.915

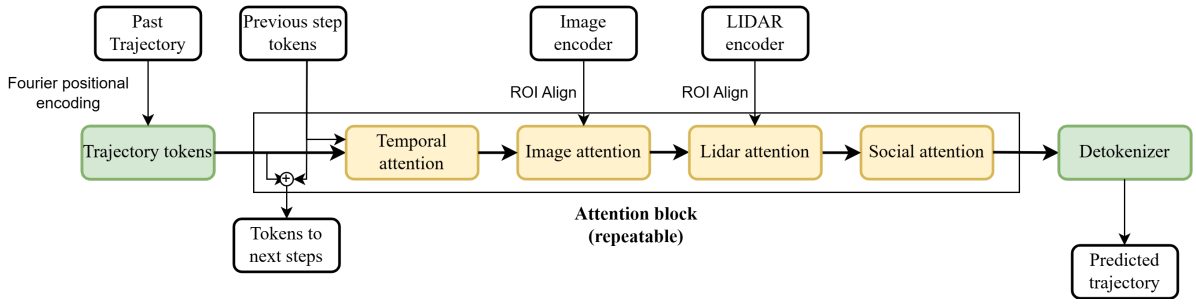

Figure 3: In-progress work on the decoder-only pedestrian motion prediction model.

that it correctly incorporates surrounding map information.

5. Ongoing work

Ongoing work focuses on answering RQ3 and developing a foundational pedestrian motion prediction solution. Building on the latest advances in transformer-based prediction models [30], we formulate a decoder-only model that can process an arbitrary number of additional input modalities with extra adaptation (see Fig. 3). The proposed solution offers a novel approach to current motion prediction methods in the field, which usually focus on processing only past trajectory history as their main input or, at times, incorporate pedestrian skeleton or lidar input in limited contexts, such as restricting the application field to indoor environments.

The model takes the standard past trajectory (relative coordinates, velocity vector, heading) as its main input and converts them into per-timestep tokens using Fourier positional encoding. Then it iterates over $t_{hist} + t_{fut}$ and auto-regressively predicts the next step of the trajectory while processing both historical and future steps. The encoded tokens are processed through a variable number of attention blocks, with each block containing multiple attention modes with different mechanisms:

1. Temporal attention: at step t_s, the model incorporates all tokens from timesteps $t_0..t_{s-1}$ and attends to them. This means that during prediction of the future trajectory, the model attends to the previous tokens it itself generated, rather than those pre-processed from a known past trajectory.
2. Image attention: at step t_s, the model finds the closest agents in a defined radius and creates attention links between trajectory tokens and the closest agents corresponding to encoded image tokens. Images are processed with a pre-trained image feature encoder such as DINOv3 [19], then processed with ROI Align to limit the number of image feature tokens per agent. Additionally, each trajectory token attends only to images extracted and processed at the same or the closest time stamp.
3. LIDAR attention: similarly to image attention. This is work-in-progress for now, but since common datasets for prediction include both image and LIDAR data for agents, this is a potential near-future direction.
4. Social attention: similarly to image attention, the model establishes attention between the modified trajectory token and the closest agents' trajectory tokens. Crucially, during multi-modal output (that generates multiple candidate trajectories), the model produces separate tokens for every mode, and social attention works only on the tokens from the same mode, leading to increased variance of decoded trajectories
5. The proposed model can support other input modalities by adding separate attention modes; for example, there are proposals in the literature to include map data, which is a potential future direction limited by available datasets.

The proposed model was engineered and trained/evaluated on Nuscenes [31] and JRDB [32] datasets; the initial results show very limited improvement from the addition of image features. As the model

reuses existing image encoders, the current hypothesis is that a different model or light fine-tuning of the encoder is necessary. Alternatively, a larger dataset such as Argoverse v2 [33] can be used to prevent the model from overfitting to a rich image representation with limited training data.

6. Concluding remarks

In this paper, we present progress of ongoing PhD thesis work on pedestrian motion prediction in urban scenarios for autonomous driving applications. We have introduced three major research questions and demonstrated progress and existing results for the first two. We have described the ongoing work for RQ3 and our general approach and research methods for tackling the problem. During the final validation steps of the thesis, the proposed solution will be tested not only on the standard evaluation datasets, as per the field standard, but also within the Autonomous Driving Framework.

The final model will be available open-source, both as pre-trained .onnx weights that can use trajectory history, camera images and lidar scans as the inputs, as well as real-time example integration in the Autoware Mini framework. The distribution of such a model provides a novel contribution to the field, where the usual published models use the trajectory as the only input.

The PhD thesis manuscript, built upon the developed method, will provide more comprehensive answers to all established Research Questions, frame the evolution of the research, discuss the discovered pitfalls, and outline future research directions.

Acknowledgments

This PhD thesis is supervised by Prof. Naveed Muhammad at the Institute of Computer Science, University of Tartu, Estonia and co-supervised by Prof. Yar Muhammad at the Department of Computer Science at the University of Hertfordshire, U.K.

This work was supported in part by European Social Fund through the "ICT Programme" Measure and in part by Bolt Technologies through the Collaboration Project under Grant LLTAT21278.

Declaration on Generative AI

During the preparation of this work, the author used Grammarly in order to: Paraphrase and reword, Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] K. Nagata, Waymo-backed robotaxis quietly ply the streets of Tokyo as tests continue, The Japan Times (2026). URL: https://www.japantimes.co.jp/business/2026/01/25/companies/japan-robotaxis/, accessed: 2026-04-06.

[2] D. McCowen, Waymo gears up for Aus debut, news.com.au (2026). URL: https://www.news.com.au/technology/motoring/motoring-news/waymo-gears-up-for-aus-debut/news-story/c695331f9cc10faeed52eb435749eb90, accessed: 2026-04-06.

[3] L. Di Lillo, T. Gode, X. Zhou, M. Atzei, R. Chen, T. Victor, Comparative safety performance of autonomous-and human drivers: A real-world case study of the waymo driver, Heliyon 10 (2024).

[4] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, K. O. Arras, Human motion trajectory prediction: A survey, The International Journal of Robotics Research 39 (2020) 895–935.

[5] M. Gulzar, Y. Muhammad, N. Muhammad, A survey on motion prediction of pedestrians and vehicles for autonomous driving, IEEE Access (2021).

[6] J. F. P. Kooij, N. Schneider, F. Flohr, D. M. Gavrila, Context-based pedestrian path prediction, in: European Conference on Computer Vision, Springer, 2014, pp. 618–633.

[7] L. Demerath, D. Levinger, The social qualities of being on foot: A theoretical analysis of pedestrian activity, community, and culture, City & Community 2 (2003) 217–237.
[8] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE Access 8 (2020) 58443–58469. doi:10.1109/ACCESS.2020.2983149.
[9] H. Wu, T. Phong, C. Yu, P. Cai, S. Zheng, D. Hsu, What truly matters in trajectory prediction for autonomous driving?, arXiv preprint arXiv:2306.15136 (2023).
[10] T. Matiisen, Ut-adl/autoware_mini: Autoware mini is a minimalistic python-based autonomy software., 2023. URL: https://github.com/UT-ADL/autoware_mini/.
[11] P. A. Lasota, T. Fong, J. A. Shah, et al., A survey of methods for safe human-robot interaction, Now Publishers, 2017.
[12] N. Brouwer, H. Kloeden, C. Stiller, Comparison and evaluation of pedestrian motion models for vehicle safety systems, in: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2016, pp. 2207–2212.
[13] E. Schuetz, F. B. Flohr, A review of trajectory prediction methods for the vulnerable road user, Robotics 13 (2023) 1.
[14] Z. Fu, K. Jiang, C. Xie, Y. Xu, J. Huang, D. Yang, Summary and reflections on pedestrian trajectory prediction in the field of autonomous driving, IEEE Transactions on Intelligent Vehicles (2024).
[15] K. M. Abughalieh, S. G. Alawneh, Predicting pedestrian intention to cross the road, IEEE Access 8 (2020) 72558–72569.
[16] Y. Su, J. Du, Y. Li, X. Li, R. Liang, Z. Hua, J. Zhou, Trajectory forecasting based on prior-aware directed graph convolutional neural network, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 16773–16785.
[17] N. Nilavadi, A. Rudenko, T. Linder, Uptor: Unified 3d human pose dynamics and trajectory prediction for human-robot interaction, in: 2025 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2025, pp. 13927–13933.
[18] J. Van den Berg, M. Lin, D. Manocha, Reciprocal velocity obstacles for real-time multi-agent navigation, in: 2008 IEEE international conference on robotics and automation, Ieee, 2008, pp. 1928–1935.
[19] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al., Dinov3, arXiv preprint arXiv:2508.10104 (2025).
[20] D. Zabolotnii, Y. Muhammad, N. Muhammad, Pedestrian motion prediction evaluation for urban autonomous driving, in: 2025 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, 2025, pp. 1–7.
[21] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, A. Gaidon, It is not the journey but the destination: Endpoint conditioned trajectory prediction, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 759–776.
[22] C. Wang, Y. Wang, M. Xu, D. J. Crandall, Stepwise goal-driven networks for trajectory prediction, IEEE Robotics and Automation Letters 7 (2022) 2716–2723.
[23] M. Lee, S. S. Sohn, S. Moon, S. Yoon, M. Kapadia, V. Pavlovic, Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2221–2230.
[24] H. Cheng, M. Liu, L. Chen, H. Broszio, M. Sester, M. Y. Yang, Gatraj: A graph-and attention-based multi-agent trajectory prediction model, ISPRS Journal of Photogrammetry and Remote Sensing 205 (2023) 163–175.
[25] D. Zabolotnii, Y. Muhammad, N. Muhammad, Pedestrian motion prediction evaluation for urban autonomous driving, arXiv preprint arXiv:2410.16864 (2024).
[26] W. Mao, C. Xu, Q. Zhu, S. Chen, Y. Wang, Leapfrog diffusion model for stochastic trajectory prediction, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5517–5526.
[27] R. A. Yeh, A. G. Schwing, J. Huang, K. Murphy, Diverse generation for multi-agent sports games,

in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4610–4619.

[28] D. Zabolotnii, Y. Muhammad, N. Muhammad, Fov-rvo: Velocity obstacle-based pedestrian motion predictor, in: 2025 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, 2025, pp. 35–42.

[29] R. Algabri, H. Shin, S. Lee, Real-time 6dof full-range markerless head pose estimation, Expert Systems with Applications 239 (2024) 122293.

[30] M. Knoche, D. de Geus, B. Leibe, Donut: A decoder-only model for trajectory prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 28903–28912.

[31] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631.

[32] R. Martin-Martin, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, S. Savarese, Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments, IEEE transactions on pattern analysis and machine intelligence 45 (2021) 6748–6765.

[33] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al., Argoverse: 3d tracking and forecasting with rich maps, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8748–8757.