


THÖR-Magni: Comparative Analysis of Deep Learning Models for Role-conditioned Human Motion Prediction

Tiago Rodrigues de Almeida¹, Andrey Rudenko², Tim Schreiter¹, Yufei Zhu¹, Eduardo Gutierrez Maestro¹, Lucas Morillo-Mendez¹, Tomasz P. Kucner^{3,4}, Oscar Martinez Mozos¹, Martin Magnusson¹, Luigi Palmieri², Kai O. Arras², and Achim J. Lilienthal¹

Abstract

Autonomous systems, that need to operate in human environments and interact with the users, rely on understanding and anticipating human activity and motion. Among the many factors which influence human motion, semantic attributes, such as the roles and ongoing activities of the detected people, provide a powerful cue on their future motion, actions, and intentions. In this work we adapt several popular deep learning models for trajectory prediction with labels corresponding to the roles of the people. To this end we use the novel THÖR-Magni dataset, which captures human activity in industrial settings and includes the relevant semantic labels for people who navigate complex environments, interact with objects and robots, work alone and in groups. In qualitative and quantitative experiments we show that the role-conditioned LSTM, Transformer, GAN and VAE methods can effectively incorporate the semantic categories, better capture the underlying input distribution and therefore produce more accurate motion predictions in terms of Top-K ADE/FDE and log-likelihood metrics.

1. Introduction

Human motion understanding is a critical component in various domains, such as robotics, automated driving, au-

^{*1} Örebro University, Sweden {tiago.almeida, tim.schreiter, yufei.zhu, eduardo.gutierrez-maestro, lucas.morillo, oscar.mozos, martin.magnusson, achim.lilienthal}@oru.se

^{†2}Robert Bosch GmbH, Corporate Research, Stuttgart, Germany {andrey.rudenko, luigi.palmieri, kaioliver.arras}@de.bosch.com

^{‡3}Mobile Robotics Group, Department of Electrical Engineering and Automation, Aalto University, Finland tomasz.kucner@aalto.fi

^{§4}FCAI, Finnish Center for Artificial Intelligence, Finland

[¶]This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and by the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017274 (DARKO).

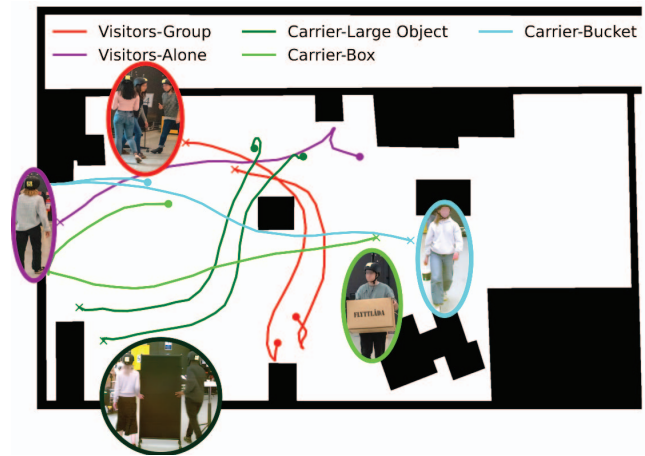


Figure 1: Example trajectories in the THÖR-Magni dataset. Participants undertake tasks according to their roles, tailored for industrial settings: visitors navigate individually and in groups between the various goals in the environment and workers carry boxes, buckets and large objects.

tonomous surveillance and security applications [22]. This is a challenging task since human motion is highly non-deterministic, uncertain, multimodal and influenced by various factors in the static and dynamic environment. To build more accurate models of human motion, modern methods attempt to move beyond the basic geometric input, such as position and velocity of the target agent, and incorporate more advanced cues, such as head orientations [9], full-body poses [27], emotions [16], and semantic attributes [14]. Among those, classes of agents are particularly useful in situations where diverse agents navigate in an interactive environment, such as an urban scene with mixed traffic. Defining and estimating classes of moving agents and making class-conditioned predictions is still an under-explored problem [4].

In the context of autonomous driving, integrating semantic attributes pertaining to road users (e.g., cars, trucks, pedestrians, or cyclists) yields safer planning and decision-making systems [10]. Similarly, in collaborative industrial

or healthcare settings, mobile robots are expected to possess the capability to interact with human counterparts. Henceforth, incorporating semantic attributes within these contexts can engender personalized human-robot interactions and facilitate social navigation [3, 16].

Training and evaluating advanced prediction methods require appropriate datasets, which contain the trajectories of moving agents and, as well as the relevant semantic information. Among the well-established outdoor datasets for pedestrian motion prediction, such as ETH/UCY [17, 13], Edinburgh [15] and the Stanford Drone Dataset (SDD) [19], only SDD incorporates a diverse environment with various classes of agents, such as pedestrians, bicyclists, skateboarders, carts, cars, and buses. In the indoor context, the THÖR [21] and the follow-up THÖR-Magni datasets [24] include novel attributes (activity-related roles) assigned to people, such as visitors, workers, and inspectors, specifically tailored for industrial and service robotics applications. These roles are chosen to reflect semantically meaningful tasks and schedules in an industrial environment¹, in which people navigate complex environments, interact with objects and robots, work alone and in groups. The type of activity people are engaged in according to their role provides a strong hint on the future motion, actions, and goals. In this paper, we aim to investigate the impact of considering role information on the accuracy of several popular classes of motion prediction algorithms.

To that end, we introduce and evaluate several role-conditioned motion prediction methods, based on generative, recurrent and attention deep learning models, and compare them to the corresponding role-agnostic methods. As the baseline trajectory predictor [1], we assess a Long Short-Term Memory (LSTM)-based model (RED) and introduce its conditioned variant (cRED), where the assigned roles serve as additional input features to the decoder network. Similarly, we propose a Transformer-based [26] trajectory predictor (TFb) and its conditioned counterpart (cTFb). Furthermore, we study two categories of deep generative trajectory predictors: Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), which are able to generate multiple predictions. For these generative models, we propose their respective conditional counterparts, in which the role and the observed trajectory jointly contribute to the prediction. In our experiments, we compare all methods using the Top-K ADE/FDE scores, and additionally evaluate the deep generative models with negative conditional log-likelihood. We also qualitatively demonstrate the stronger conditioning signal impact on GAN-based models compared to RED variants. In summary, this work concludes that considering the roles improves the accuracy of trajectory prediction, biasing the dynamics to the motion patterns of the corresponding role.

¹<https://darko-project.eu/>

2. Related Work

2.1. Human Motion Datasets

Human motion is complex and diverse, influenced by a variety of factors such as goals, intentions, interactions and environment. Creating accurate and realistic models of human motion is a challenge, especially in the long-term perspective, considering the intentions, complex activity patterns, interactions with the static and dynamic environment elements. Addressing this challenge requires large and diverse human motion datasets that capture the richness and variability of natural human behavior in different scenarios and contexts. Many of the existing datasets [22] suffer from limitations such as uniform motion patterns [18], missing and incorrect detections [15], rough position estimations with bounding boxes, and imbalanced distribution of the agents' classes [19].

The recent THÖR dataset [21] is a high-quality and diverse human motion dataset that aims to address these issues by providing accurate motion capture data in crowded social spaces among the static obstacles and featuring a moving robot. The THÖR dataset also includes additional inputs such as maps of the environment and gaze directions of the participants. In the THÖR recording, participants navigate the industrial environment and perform tasks according to their role, such as inspector, visitor, utility and lab worker. Recording these semantic attributes of moving people is a key property, enabling the development of activity detection algorithms and class-conditioned motion prediction methods. However, as THÖR provides a limited amount of data (about 60 minutes of motion of the 9 participants), its applicability to training and generalizing the data-hungry learning-based approaches is reduced [20].

The follow-up THÖR-Magni dataset [24], which includes over 3.5 hours of motion with 30 participants, addresses this main limitation of THÖR. Furthermore, THÖR-Magni, in addition to basic navigation, includes diverse interaction scenarios with people working alone and in groups, interacting with the robot and transporting various objects (boxed, buckets and poster stands). Similarly to THÖR, the new recording features diverse roles for the participants, such as box carriers, bucket carriers, large object carriers, and visitors (navigating individually or in groups).

2.2. Class-conditioned Trajectory Predictors

In the context of trajectory prediction, semantic attributes of the moving agents can be any useful additional information that constraints the motion dynamics to the typical pattern of the underlying class, or biases the predictor towards the corresponding class-related motion distribution. The use of classes (or roles) is paramount in domains like Autonomous Driving [10] and Human-Robot Interaction [16], where the differentiating between classes has a

significant impact on the performance of the downstream decision-making systems.

To cope with Autonomous Driving safety requirements, Djuric et al. [4] propose MultiXNet, a framework that considers the heterogeneity of road agents to predict multi-modal trajectories. In the same line of research, Ivanovic et al. [10] present *HAICU*, a method that also leverages agents' class uncertainty in an end-to-end perception module for autonomous vehicles. Finally, in [8], the authors propose a contrastive learning approach that improves trajectory prediction by grouping together embeddings that belong to the same class (i.e. walking, looking, standing, and crossing).

In Human-Robot Interaction, Narayanan et al. [16] introduce a Transformer-based network that incorporates human intention and affect to enhance the efficacy of social robot navigation. Furthermore, T. Rodrigues de Almeida et al. [20] use THÖR to compare supervised and unsupervised deep conditional generative models for class-conditioned trajectory prediction. The authors demonstrate that the unsupervised class labeling enhances the accuracy of unconditional methods, while the supervised class labels from THÖR do not. This limitation may arise from various factors, such as imbalanced or limited data availability. Our objective in this study is to overcome these issues by gathering additional data with a more semantically enriched role assignment. Further, we compare conditional and unconditional trajectory predictors to prove the usefulness of the roles in THÖR-Magni.

3. Methods

3.1. The THÖR-Magni Dataset

In this work we aim to study diverse, natural, and goal-driven human motion in crowded social spaces with static obstacles and a moving robot. To this end, we use the novel THÖR-Magni [24] dataset, an extension of THÖR [21], which applies the THÖR data collection process with enriched and heterogeneous semantic attributes of the participants. THÖR-Magni is collected in a weakly-controlled laboratory environment using motion capture² to track the head position and orientation of every participant. THÖR-Magni includes five scenarios, each concentrating on various aspects of human motion, such as navigating in presence of static obstacles and a moving robot, interacting with the robot, and fulfilling various tasks according to the assigned roles. With the objective of investigating the impact of semantic attributes on human motion prediction, in this paper we concentrate on Scenarios 2 and 3, in which the roles and tasks assigned to the participants are featured prominently. These specific scenarios include 30 participants and entail a total motion duration of 1.5 hours.

²www.qualisys.com

In both scenarios, the participants fulfill two primary roles: visitors and industrial workers, while co-navigating with a mobile robot in the environment. Specifically, in Scenario 2, the robot serves as a static obstacle in the shared space, while in Scenario 3, the robot moves along with the participants, controlled by an operator. The participants are assigned to various tasks, including transporting different-sized objects between designated goal points. These tasks involve one participant transporting a small object (a *bucket*) between two specific points, another participant moving a medium-sized object (a *box*) between two distinct goal points in the room, and a two-person team collaboratively transporting a large object (a *poster stand*). One member of the two-person team responsible for moving the large object receives instructions over Discord (i.e., platform that enables voice calls), which facilitates the dynamic allocation of new goal points. Additionally, the remaining participants assume the roles of visitors who either move individually or in groups between pre-defined goals positions, assigned automatically through a system of cards (see [24] for more details). Therefore, importantly for this paper, in both scenarios we observe the emergence of five distinct agent roles: *Carrier-Bucket*, *Carrier-Box*, *Carrier-Large Object*, *Visitors-Alone*, and *Visitors-Group*.

The objective of **Scenario 2** is to capture the diverse motion patterns exhibited by the participants while undertaking their respective tasks (see left Fig. 2). In the scope of this study, we intend to leverage the inherent advantage derived from the fact that individuals performing the five different roles display dissimilar movement patterns.

Scenario 3 introduces a mobile robot as an active agent in the environment. The mobile robot is teleoperated and navigates the room, while the participants engage in the same tasks as in Scenario 2 (see right Fig. 2). The presence and behavior of the robot in Scenario 3 affects the participants' motion and interaction. Therefore, this scenario has two variations: 3A, where the robot moves as a regular differential drive robot, meaning that it can only move forward, backward in an arc trajectory; and 3B, where the robot moves in an omnidirectional way, meaning that it can also move sideways and diagonally.

We prepare the data from Scenario 2 and 3 as described below in Sec. 4.1. This data is used for training trajectory prediction methods, presented in the next section.

3.2. Role-conditioned Trajectory Prediction

In this work we show that information about the role of the agent can be used to improve the accuracy of trajectory prediction. To achieve this we are reformulating the standard mapping:

$$f : (X_i) \rightarrow Y_i, \quad (1)$$

where X_i denotes the sequence of observed states for trajectory i , and Y_i is the correspondent sequence of future states



Figure 2: Summary of the trajectories in Scenario 2 (left) and 3B (right) in the respective 4-minute recordings. Scenario 2 contains three static obstacles, while Scenario 3B has two. Each trajectory is color-coded according to the role of the participant. *Visitors-Group* and *Visitors-Alone* navigate freely between the various goal points in the environment. The *Carrier-Large Object* role involves a pair of individuals transporting a poster stand between designated goals. The *Carrier-Box* and *Carrier-Bucket* roles entail the transportation of a stack of objects (a box and bucket, respectively) between two fixed points. In Scenario 3B (right), trajectories are more dispersed across the width of the free space compared to Scenario 2 (left), due to the presence of the mobile robot in the scene.

to be predicted, to:

$$f : (X_i, r_i) \rightarrow Y_i, \quad (2)$$

where r_i encodes one of the recognized activity categories (*Carrier-Bucket*, *Carrier-Box*, *Carrier-Large Object*, *Visitors-Group*, and *Visitors-Alone*), assigned to the pair (X_i, Y_i) . Each observed state in X_i , denoted as s^o , is represented by the 2D Cartesian coordinates (x^o, y^o) and the corresponding velocity vector (v_x^o, v_y^o) . The primary aim of this study is to determine the function f that maps the pair (X_i, r_i) to a sequence of future velocity vectors, \hat{Y}_i . Then, we transform each estimated velocity vector $\hat{v} = (\hat{v}_x^f, \hat{v}_y^f)$ into the corresponding 2D position $\hat{p} = (\hat{x}^f, \hat{y}^f)$, which compose the predicted positions, \hat{S}_i . To this end, we utilize well-established trajectory prediction methods, namely RED [1], TFb [26], GAN, and VAE. Finally, we conduct a comparative analysis between these unconditional models and their respective conditional counterparts, namely conditional RED (cRED), conditional TFb (cTFb), conditional GAN (cGAN), and conditional VAE (cVAE).

3.2.1 Single Output Trajectory Predictors

In this section we study single output trajectory predictors, which produce one prediction per observed *tracklet*. In this setting, we explore two compelling methods in the research community for motion prediction: LSTM (i.e., RED [1]) and Transformer-based networks denoted by TFb.

Firstly, we evaluate RED and propose its conditional counterpart, cRED. The latter concatenates the features

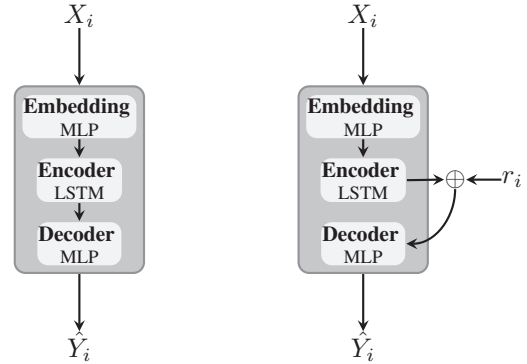


Figure 3: LSTM-based models: unconditional RED (left) and cRED (right).

from the encoder to the embeddings of the role and passes the summary vector to the Multilayer Perceptron (MLP) decoder. Fig. 3 depicts both networks. Lastly, we propose TFb (see Fig. 4), which leverages attention mechanisms enclosed in a Transformer-encoder network to encode the observed trajectory and decodes the memory vector as RED (via MLP). Similarly to cRED, cTFb concatenates the role label’s embeddings to the memory vector and processes the summary vector through the decoder.

We train single output networks with the Mean Squared Error (MSE) loss, given by:

$$L_T(S_i, \hat{S}_i) = \frac{1}{T_p} \sum_j^{T_p} \|p^j - \hat{p}^j\|_2, \quad (3)$$

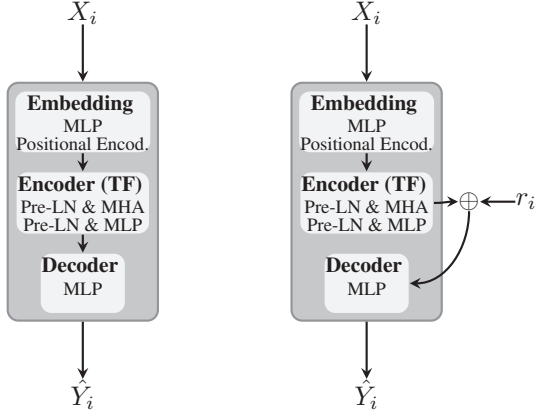


Figure 4: Transformer-based models: unconditional TFb (left) and cTFb (right). We first aggregate the input embeddings with the positional encoding employed in [6] and pass it to the encoder. The encoder includes two blocks: the initial block consists of pre-layer normalization (Pre-LN) succeeded by multi-head attention layers (MHA) with a skip connection, while the second block comprises Pre-LN followed by an MLP with a skip connection.

being $T_p = |S_i| = |\hat{S}_i|$ the prediction horizon, p^j the ground truth position at time step j , \hat{p}^j the predicted position at time step j , and S_i the ground truth sequence of positions.

3.2.2 GAN-based Trajectory Predictors

Following the network design choices of GAN-based encoder-decoder networks described in [12], we compare the GAN-based models depicted in Fig. 5. Both GAN-based methods rely on a generator (G) and a discriminator (D) network. The goal of the former is to generate samples that resemble the training data, whereas the latter aims to distinguish between real and generated samples. These two networks are trained together in an adversarial manner. Therefore, the generator is trained to fool the discriminator until both networks reach the Nash equilibrium, where each network can not decrease its loss without changing the other's parameters. In the trajectory prediction problem, the canonical GAN's generator aims to learn $f_G : (X_i, z_G) \rightarrow Y_i$, where z_G is a white noise vector while the discriminator's goal is to learn $f_D : W \rightarrow s$, where $W = Y_i \cup \hat{Y}_i$ and $s \in [0, 1]$ is a scalar value representing the likelihood that the input sample (from W) comes from the original set of samples, Y , rather than from the generator's space of outputs, \hat{Y} . In contrast, a cGAN adds extra information to both networks. Therefore, in this case, f_G and f_D take one more input, the role, r_i . The key insight is that by conditioning the GAN framework on the role, we control the generation process by forcing the networks to

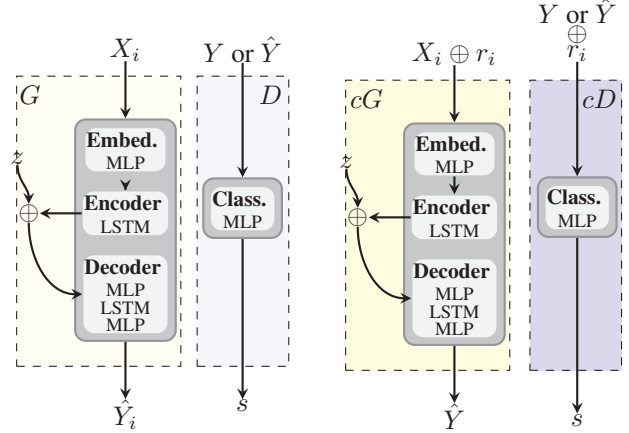


Figure 5: GAN-based models: unconditional GAN (left) and cGAN (right).

distinguish and generate the behaviors underlying each role in THÖR-Magni. We optimize GAN and cGAN discriminators with the binary cross entropy loss and the generator with a weighted sum given by:

$$L_G = \lambda L_T + (1 - \lambda) \left(\frac{1}{2} \mathbb{E}[(D(Y_i) - 1)^2] + \frac{1}{2} \mathbb{E}[D(\hat{Y}_i)^2] \right), \quad (4)$$

being L_T given by Eq. 3, and λ the weight applied to the MSE term for a single training example. To train the conditional variant (cGAN), we also pass the role class as an input to the generator and discriminator.

3.2.3 VAE-based Trajectory Predictors

A different type of deep generative models we evaluate is VAE-based methods (see Fig. 6). These models also contain an encoder-decoder branch to perform the prediction task [2]. The encoder's objective is to map each observed trajectory (X_i) to a feature vector, which is later concatenated to the latent variable z_V and reconstructed to produce the prediction (\hat{Y}_i). Additionally, VAE-based models have a recognition network that aims to learn $f_{q_\phi} : Y_i \rightarrow (\mu, \sigma^2)$. That is, the mapping from the ground truth to two probabilistic entities that define a lower dimensional latent space representation. During training, we employ the reparameterization trick to sample the latent variable z_V , which is generated by the recognition network. Additionally, we incorporate the Kullback-Liebler divergence to effectively regularize the learned distribution, aligning it with the prior standard normal Gaussian distribution. Thus, the variational loss function is as follows:

$$L_V = \lambda L_T - (1 - \lambda) \beta D_{KL}[q_\phi(z_V | Y_i) \| p(z_V | X_i)], \quad (5)$$

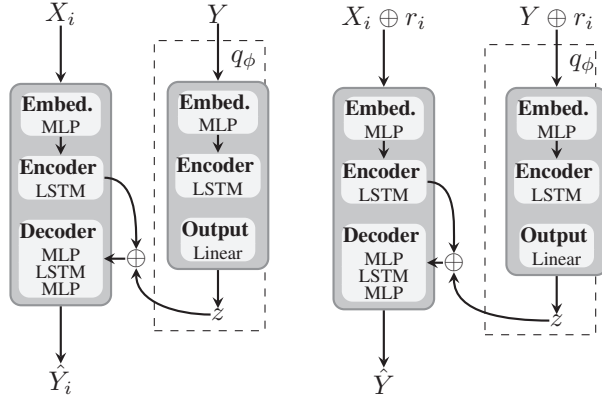


Figure 6: VAE-based models: unconditional VAE (left) and cVAE (right). The recognition networks (q_ϕ , enclosed with dashed border) are only available during training.

where β is the weight applied to the regularization loss. To train the conditional variant (cVAE), we also pass the role class as an input to the encoder-decoder and recognition networks.

4. Experiments and Results

In this section, we conduct a series of experiments to showcase the significant impact of the semantic attributes present in THÖR-Magni on the accuracy of trajectory prediction. We begin by outlining the experimental setup, including data preprocessing and analysis. We then describe the implementation details of these experiments, along with the evaluation metrics used to assess predictors performance. Finally, we present and analyze the results obtained in this investigation.

4.1. Dataset Preprocessing and Analysis

We first preprocess the raw trajectory data to form the training and validation sets for the trajectory predictors, separately for the THÖR-Magni Scenario 2, 3A and 3B, as described in Sec. 3.1. As the motion capture system relies on helmets equipped with reflective markers, in the end, we have multiple tracks (one per marker) per helmet. To minimize the number of discontinued trajectories, we identify the marker of each helmet with the highest tracking duration. Then, we single out these markers and preprocess the respective signal by applying the following steps: (1) linear interpolation of tracking discontinuities of 0.5 s at most; (2) resample the signal to 0.4 s; (3) smooth the signal with a moving average filter.

Following the existing trajectory prediction benchmark [12], we split each trajectory into tracks of 20-time steps (8 s). Table 1 shows a summary of the data comprising the number and percentage of 20-time steps *tracklets* and agents' velocity statistics per role in each scenario. A

Role	Magni-S2	Magni-S3A	Magni-S3B
<i>Carrier-Box</i>	223 (14.13%) $1.12 \frac{m}{s} \pm 0.21$	224 (13.74%) $1.15 \frac{m}{s} \pm 0.27$	220 (13.70%) $1.08 \frac{m}{s} \pm 0.26$
<i>Carrier-Bucket</i>	224 (14.20%) $1.21 \frac{m}{s} \pm 0.24$	226 (13.87%) $1.21 \frac{m}{s} \pm 0.20$	227 (14.13%) $1.13 \frac{m}{s} \pm 0.18$
<i>Carrier-Large Object</i>	394 (24.97%) $0.72 \frac{m}{s} \pm 0.27$	440 (26.99%) $0.68 \frac{m}{s} \pm 0.32$	405 (25.22%) $0.76 \frac{m}{s} \pm 0.36$
<i>Visitors-Alone</i>	452 (28.64%) $0.95 \frac{m}{s} \pm 0.20$	318 (19.51%) $0.92 \frac{m}{s} \pm 0.29$	322 (20.05%) $0.87 \frac{m}{s} \pm 0.32$
<i>Visitors-Group</i>	285 (18.06%) $0.92 \frac{m}{s} \pm 0.31$	422 (25.89%) $0.87 \frac{m}{s} \pm 0.26$	432 (26.90%) $0.84 \frac{m}{s} \pm 0.31$
Global	1578 (100%) $0.95 \frac{m}{s} \pm 0.51$	1630 (100%) $0.91 \frac{m}{s} \pm 0.48$	1606 (100%) $0.90 \frac{m}{s} \pm 0.48$

Table 1: Data summary per role in our experiments: number and ratio of 20-time steps tracklets, velocities average and standard deviation.

broad view of this table shows that the same role moved with similar velocity across the different scenarios. However, if we compare the roles within each scenario, we can observe that the *Carrier-Bucket* (small object) moved the fastest, followed by the *Carrier-Box*. On the opposite side of the spectrum, *Carrier-Large Object* was the slowest role. This is expected as transporting a small object implies less effort than moving a bigger object like a box. Moreover, a group of two people moved the poster stand (large object), which entails a team effort and, therefore, a slower pace. Finally, it is worth highlighting that, on average, *Visitors-Group* moved slower than *Visitors-Alone*.

4.2. Implementation Details

As in the prior art [12], from the 20-time steps tracklets, we observe 8 (3.2 s) and estimate the next 12 (4.8 s). To evaluate the unconditional and conditional methods, we conducted k-fold cross-validation within each Scenario (we set k to 10). That is, we train with k-1 folds and leave the remaining for validation. We used the same hyperparameters (i.e., optimizer, batch size, type of networks and activation functions) in all deep learning-based models. Also, we stopped the training when no improvement was observed after 20 epochs for a maximum of 100 epochs per fold. Finally, to study the generative process of GAN and VAE-based models, we used the *k-variety loss* proposed in [7].

For metrics reporting, we show the average and standard deviations for the k-fold cross validation. To compare the trajectory predictors, we use the *Top-K Average and Final Displacement Errors* (Top-K ADE and Top-K FDE, in meters), as in [23, 11]. ADE measures the root mean squared error between the ground truth track and the closest prediction (out of K samples). FDE measures the

Euclidean distance between the last ground truth position and the closest predicted counterpart (out of K samples). Furthermore, following [2], we use the negative conditional log-likelihood metric (CLL) to compare generative models. Finally, besides the models described in Section 3.2, we also report the metrics obtained by the Constant Velocity Model (CVM) [25].

4.3. Results

This section presents the quantitative results and analysis of the trajectory predictors applied in Scenarios 2, 3A and 3B. Throughout the quantitative experiments, bold scores indicate the superior performance of conditional models compared to their canonical counterparts (e.g., cGAN versus GAN). Table 2 reports the Top-1 ADE/FDE values for the three datasets. A comprehensive examination of the table shows that deep learning models consistently outperform the CVM. However, in addition to considering the observed velocity vector (v_x^o, v_y^o) , deep learning models also take into account the spatial layout information provided by the observed positions (x^o, y^o) , which contributes to their improved performance. Moreover, even when solely using the velocity vector as inputs (which renders the models spatial layout agnostic), simple models such as RED, TFb, cTFb, cGAN, VAE, and cVAE still outperform CVM. While CVM may serve as a relevant baseline in ETH/UCY benchmarks [17, 13], the THÖR-Magni dataset presents significantly more challenging and complex data, where simplistic baselines like CVM face substantial limitations.

In the three datasets, cRED and cTFb outperform the other baselines (underlined results), furthermore, all conditional models attain better or similar scores to the unconditional counterparts. Specifically, in Magni-S2, cGAN provides a 8.2% and 10.6% of ADE and FDE improvements, respectively. Also, in these datasets, the improvement given by the role is plainer on GAN-based models than VAE-based models. In fact, in Magni-S3B, surprisingly, VAE outperforms cVAE, which might be because VAE-based models learn a rather flexible latent representation. Consecutively, the condition may not include any extra information to lead to lower ADE/FDE scores.

Table 3 shows the Top-3 ADE/FDE and CLL scores obtained by deep generative models. Analogously to Top-1 ADE/FDE, in general, Top-3 ADE/FDE is also positively affected by conditioning roles. Further, CLL is lower in conditioned models, which means these models fit better the input data distribution than the respective counterparts.

In Fig. 7, we present qualitative results from RED, cRED, GAN, and cGAN applied to two test samples extracted from Magni-S3A. These outcomes are consistent with the quantitative evaluation metrics, ADE/FDE, where the conditional setting induces a greater improvement in GAN-based predictors compared to LSTM-based

Models	Scores	Magni-S2	Magni-S3A	Magni-S3B
CVM	ADE	1.19 ± 0.05	1.18 ± 0.08	1.17 ± 0.05
	FDE	2.58 ± 0.11	2.54 ± 0.20	2.52 ± 0.15
RED	ADE	0.71 ± 0.05	0.70 ± 0.03	0.73 ± 0.04
	FDE	1.42 ± 0.09	1.41 ± 0.07	1.48 ± 0.08
cRED	ADE	0.69 ± 0.05	0.68 ± 0.03	0.72 ± 0.04
	FDE	1.35 ± 0.08	1.35 ± 0.06	1.45 ± 0.07
TFb	ADE	0.72 ± 0.05	0.72 ± 0.03	0.75 ± 0.04
	FDE	1.42 ± 0.09	1.43 ± 0.08	1.50 ± 0.08
cTFb	ADE	0.68 ± 0.06	0.69 ± 0.03	0.73 ± 0.05
	FDE	1.32 ± 0.10	1.37 ± 0.07	1.47 ± 0.07
GAN	ADE	0.97 ± 0.12	0.93 ± 0.10	0.99 ± 0.12
	FDE	1.98 ± 0.24	1.90 ± 0.20	2.03 ± 0.22
cGAN	ADE	0.89 ± 0.08	0.84 ± 0.06	0.89 ± 0.06
	FDE	1.77 ± 0.15	1.69 ± 0.15	1.77 ± 0.12
VAE	ADE	0.82 ± 0.05	0.82 ± 0.07	0.84 ± 0.05
	FDE	1.65 ± 0.08	1.66 ± 0.16	1.68 ± 0.10
cVAE	ADE	0.82 ± 0.05	0.81 ± 0.04	0.86 ± 0.05
	FDE	1.62 ± 0.11	1.58 ± 0.06	1.72 ± 0.08

Table 2: Top-1 ADE/FDE (\downarrow). Bold values indicate the superior performance of conditional models compared to their canonical counterparts.

Models	Score	Magni-S2	Magni-S3A	Magni-S3B
GAN	ADE	0.68 ± 0.08	0.67 ± 0.04	0.67 ± 0.05
	FDE	1.35 ± 0.15	1.34 ± 0.07	1.36 ± 0.10
	CLL	4.99 ± 0.20	5.24 ± 0.39	5.12 ± 0.33
cGAN	ADE	0.62 ± 0.06	0.62 ± 0.04	0.64 ± 0.04
	FDE	1.23 ± 0.11	1.21 ± 0.10	1.25 ± 0.05
	CLL	4.68 ± 0.28	4.68 ± 0.25	4.92 ± 0.23
VAE	ADE	0.60 ± 0.05	0.62 ± 0.03	0.64 ± 0.04
	FDE	1.19 ± 0.07	1.21 ± 0.08	1.24 ± 0.07
	CLL	4.64 ± 0.27	4.71 ± 0.28	4.95 ± 0.28
cVAE	ADE	0.60 ± 0.04	0.58 ± 0.06	0.63 ± 0.05
	FDE	1.17 ± 0.07	1.15 ± 0.04	1.23 ± 0.08
	CLL	4.51 ± 0.25	4.45 ± 0.28	4.75 ± 0.29

Table 3: Top-3 ADE/FDE (\downarrow) and CLL (\downarrow) for generative models. Bold values indicate the superior performance of conditional models compared to their canonical counterparts.

approaches. This improvement gap between cRED and cGAN with respect to their canonical counterparts might be because GANs rely on a latent space while LSTM-based methods do not (deterministic models). This latent space attempts to learn the various modes of the input underlying distribution, while deterministic approaches require more discernible cues in the input features to be effective. When conditioning a GAN, the latent space is influenced

by the condition, which helps in structuring and guiding the learned latent representations. On the other hand, the latent space in unconditional GANs exhibit more erratic behavior due to the lack of conditioning information. Finally, cGAN demonstrates an ability to learn the intrinsic data distribution pertaining to each role by capturing its various modes, resulting in diverse yet accurate predictions.

In summary, the results outlined in this section reveal that the THÖR-Magni dataset does not adhere to the constant velocity profile. Instead, deep learning-based models demonstrate superior performance over the CVM. The CLL scores substantiate the efficacy of conditioning trajectory predictors, while the ADE and FDE scores demonstrate that incorporating the role information generally enhances the accuracy of trajectory predictors. Based on these findings, it is evident that the role information provided in the three scenarios from THÖR-Magni dataset constitutes a valuable feature for the investigation of novel role-conditioned trajectory predictors. These results highlight the importance of considering the role attribute for trajectory prediction.

5. Conclusion

In this paper we exploit the role assignment in THÖR-Magni dataset in role-conditioned DL-based trajectory prediction methods. We conduct a comprehensive evaluation using four distinct approaches: deterministic baselines utilizing LSTM and Transformer-encoder and two deep generative methodologies employing GANs and VAEs. Our findings demonstrate that the conditional methods consistently outperform or achieve comparable performance to the unconditional counterparts in the trajectory prediction task. Specifically, the conditional GAN exhibits a significant margin of improvement over the unconditional GAN across all prediction metrics (Top-K ADE/FDE). Additionally, the performance of conditional generative models, including cGAN and cVAE, surpass their role-agnostic counterparts (GAN and VAE) in data fitting metrics (CLL). These results emphasize the efficacy of incorporating role information provided in THÖR-Magni dataset to enhance the prediction and fitting capabilities of trajectory estimators and highlight the potential advantages of role-conditioned deep learning-based approaches in trajectory prediction tasks.

In future work, we intend to expand upon the current manual role assignment to automatic techniques. In doing so, we seek to explore more efficient and scalable methods for assigning roles to individuals in human trajectory datasets. Moreover, as we transition towards automated role assignment, the THÖR-Magni dataset can serve as a valuable benchmark and a ground truth reference to evaluate the performance of the developed methods. The dataset’s existing manual role annotations will enable us to rigorously assess the accuracy and reliability of the automatic role assignment techniques, facilitating their vali-

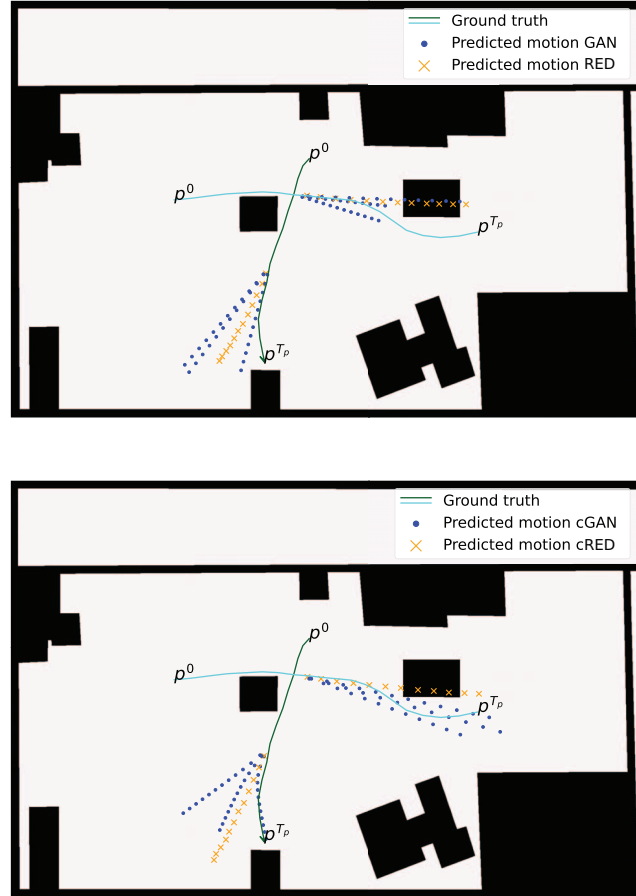


Figure 7: Trajectory predictions of unconditional methods (top) and conditional counterparts (bottom) for two test samples in Magni-S3A. The **cyan** sample that goes from the left to right corresponds to a *Carrier-Bucket* while the **green** one is from a *Carrier-Large Object*. While the discrepancy between RED and cRED predictions is minimal, a substantial difference is observed between GAN and cGAN predictions. cGAN effectively utilizes role conditioning to offer diverse yet more accurate trajectory predictions.

ation and comparison. By leveraging automatic role assignment techniques and validating them against the natural well-established role annotations in THÖR-Magni, we anticipate enhancing the dataset utility and contributing to the advancement of trajectory prediction and crowd analysis research. Further, we can augment the present study along a subset of JRDB-Act [5] available classes such as standing, walking, cycling, scootering, skating, or running. In this dataset, the labeling granularity is finer as it involves per frame activities rather than per trajectory. This characteristic presents an opportunity to delve deeper into the research of role-conditioned trajectory prediction along an innovative temporal dimension.

References

- [1] Stefan Becker, Ronny Hug, Wolfgang Hübner, and Michael Arens. Red: A simple but effective baseline predictor for the trajnet benchmark. In *ECCV Workshops*, 2018. 2, 4
- [2] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2018. 5, 7
- [3] Konstantinos Charalampous, Ioannis Kostavelis, and Antonios Gasteratos. Recent trends in social aware robot navigation: A survey. *Robotics and Autonomous Systems*, 93:85–104, 2017. 2
- [4] Nemanja Djuric, Henggang Cui, Zhaoen Su, Shangxuan Wu, Huahua Wang, Fang-Chieh Chou, Luisa San Martin, Song Feng, Rui Hu, Yang Xu, Alyssa Dayan, Sidney Zhang, Brian C. Becker, Gregory P. Meyer, Carlos Vallespi-Gonzalez, and Carl K. Wellington. Multixnet: Multiclass multistage multimodal motion prediction. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 435–442, 2021. 1, 3
- [5] M. Ehsanpour, F. Saleh, S. Savarese, I. Reid, and H. Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20951–20960, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 8
- [6] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso. Transformer networks for trajectory forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society. 5
- [7] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 6
- [8] Marah Halawa, Olaf Hellwich, and Pia Bideau. Action-based contrastive learning for trajectory prediction. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 143–159, Cham, 2022. Springer Nature Switzerland. 3
- [9] Blake Holman, Abrar Anwar, Akash Singh, Mauricio Tec, Justin Hart, and Peter Stone. Watch where you're going! gaze and head orientation as predictors for social robot navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3553–3559, 2021. 1
- [10] Boris Ivanovic, Kuan-Hui Lee, Pavel Tokmakov, Blake Wulfe, Rowan McIlister, Adrien Gaidon, and Marco Pavone. Heterogeneous-agent trajectory forecasting incorporating class uncertainty. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12196–12203, 2022. 1, 2, 3
- [11] Parth Kothari and Alexandre Alahi. Safety-compliant generative adversarial networks for human trajectory forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):4251–4261, 2023. 6
- [12] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2021. 5, 6
- [13] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3542–3549, 2014. 2, 7
- [14] W. Ma, D. Huang, N. Lee, and K. M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4636–4644, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 1
- [15] Barbara Majecka. Statistical models of pedestrian behaviour in the forum. *Master's thesis, School of Informatics, University of Edinburgh*, 2009. 2
- [16] Venkatraman Narayanan, Bala Murali Manoghar, Rama Prashanth RV, and Aniket Bera. Ewaret: Emotion-aware pedestrian intent prediction and adaptive spatial profile fusion for social robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7569–7575, 2023. 1, 2, 3
- [17] S. Pellegrini, A. Ess, K. Schindler, and L. 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, 2009. 2, 7
- [18] S Pellegrini, A Ess, K Schindler, and L 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, 2009. 2
- [19] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9912, pages 549–565. Springer International Publishing, 2016. 2
- [20] Tiago Rodrigues de Almeida, Eduardo Gutierrez Maestro, and Oscar Martinez Mozos. Context-free self-conditioned gan for trajectory forecasting. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1218–1223, 2022. 2, 3
- [21] Andrey Rudenko, Tomasz P Kucner, Chittaranjan S Swaminathan, Ravi T Chadalavada, Kai O Arras, and Achim J Lilienthal. Thör: Human-robot navigation data collection and accurate motion trajectories dataset. *IEEE Robotics and Automation Letters*, 5(2):676–682, 2020. 2, 3
- [22] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020. 1, 2
- [23] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, edi-

- tors, *Computer Vision – ECCV 2020*, pages 683–700, Cham, 2020. Springer International Publishing. 6
- [24] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. The magni human motion dataset: Accurate, complex, multi-modal, natural, semantically-rich and contextualized. *arXiv preprint arXiv:2208.14925*, 2022. 2, 3
- [25] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020. 7
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 2, 4
- [27] Jianqi Zhong, Hao Sun, Wenming Cao, and Zhihai He. Pedestrian motion trajectory prediction with stereo-based 3d deep pose estimation and trajectory learning. *IEEE Access*, 8:23480–23486, 2020. 1