ForceFormer: Exploring Social Force and Transformer for Pedestrian Trajectory Prediction

Weicheng Zhang[1], Hao Cheng[2,*], Fatema T. Johora[3] and Monika Sester[1]

2023 IEEE Intelligent Vehicles Symposium (IV) | 979-8-3503-4691-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IV55152.2023.10186643

Abstract— **Predicting trajectories of pedestrians based on goal information in highly interactive scenes is a crucial step toward Intelligent Transportation Systems and Autonomous Driving. The challenges of this task come from two key sources: (1) complex social interactions in high pedestrian density scenarios and (2) limited utilization of goal information to effectively associate with past motion information. To address these difficulties, we integrate social forces into a Transformer-based stochastic generative model backbone and propose a new goal-based trajectory predictor called *ForceFormer*. Differentiating from most prior works that simply use the destination position as an input feature, we leverage the driving force from the destination to efficiently simulate the guidance of a target on a pedestrian. Additionally, repulsive forces are used as another input feature to describe the avoidance action among neighboring pedestrians. Extensive experiments show that our proposed method achieves on-par performance measured by distance errors with the state-of-the-art models but evidently decreases collisions, especially in dense pedestrian scenarios on widely used pedestrian datasets.**

I. INTRODUCTION

Accurate and plausible trajectory prediction in crowd scenarios for pedestrians plays a fundamental role in different applications, such as mobile robot navigation [1], Intelligent Transportation Systems and Intelligent Vehicles [2], and shared space safety [3]. Unlike vehicle movement governed by traffic rules, such as lane geometry, traffic lights and the headway direction, pedestrians may stop or turn at any time and interact more with neighbors, making their behavior highly stochastic.

In order to model pedestrian behavior, a variety of different methods have been applied. In rule-based models, the interactions among pedestrians, namely social interactions, are described as forces [4]. In data-driven models, attention mechanisms [5] and graph convolutional networks [6] are widely used to extract social interactions and obtain excellent results using supervised learning [7], [8], [9]. Goal information can also reduce the uncertainties of pedestrians' behavior. However, each of these methods has its own drawbacks. The rule-based models are relatively less robust and resilient in the face of complex scenarios. Data-driven models often achieve better performance but are data-dependent and less interpretable [10]. In goal-based models, goal information is often directly applied as an input or as the offset of the current position to the goal [11], [12], [13]. They are sub-optimal because the association between the goal and the current position is not established.

To this end, we propose a novel goal-based pedestrian trajectory prediction framework called *ForceFormer*. It takes as input not only the sequential motion information but also forces to train a Transformer-based backbone. Unlike the previous models that directly use last position information as an input feature parallel to other features describing motion dynamics, we apply the goal information to derive social forces so that the changes in velocity, position, and direction are better linked to the goal information.

In addition, we use the generative model AgentFormer [9] as the trajectory prediction backbone, which utilities the Transformer [5] network to learn social interactions in the temporal dimensions. Simultaneously, to estimate the temporary goal position for computing forces in the inference time, a goal-estimation module [14], [13] is applied. More specifically, history trajectories are concatenated with semantic scene information, and they are fed into a U-Net [15] structure to predict the potential goals. With this goal-estimation module, we can obtain reliable goal information and as well naturally take into account the constraints of environmental factors. Our major contributions are as follows:

- We propose a goal-based trajectory prediction framework **ForceFormer**. It imports more interpretable features, i. e., social forces, into a data-driven model to learn stochastic pedestrian behavior.
- Different variants of ForceFormer making use of goal information are studied, and we find two effective. Namely, a) **ForceFormer-Re** applies goal positions to derive repulsive forces, reinforcing the interactive information between the ego pedestrian and neighbors and decreasing the possibility of collisions; b) **ForceFormer-Dr** applies goals to derive driving force, enhancing destination guidance to predict the ego pedestrian's future trajectory.
- Extensive empirical studies are carried out on the widely used ETH/UCY [16], [17] pedestrian datasets. The experimental results show that ForceFormer performs on par with the state-of-the-art models measured by standard distance errors but it evidently decreases collisions, especially in dense pedestrian scenarios.

[1]Weicheng Zhang and Monika Sester are with the Institute of Cartography and Geoinformatics, Leibniz University Hannover, Appelstr. 9, 30167 Hannover, Germany weicheng.zhang@stud.uni-hannover.de, Monika.Sester@ikg.uni-hannover.de

[2]Hao Cheng is with the Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE Enschede, The Netherlands. Cheng is funded by MSCA European Postdoctoral Fellowships under the 101062870 — VeVuSafety project. h.cheng-2@utwente.nl

[2]Fatema T. Johora is with the Department of Informatics, Clausthal University of Technology, Julius-Albert-Str. 4, 38678 Clausthal-Zellerfeld, Germany fatema.tuj.johora@tu-clausthal.de

[*]Corresponding author

II. RELATED WORK

This section briefly reviews the works in sequence modeling, social interaction modeling, and goal-based models.

Sequence Modeling. Essentially, motion trajectory is composed of positional information on time series. Therefore, converting trajectory prediction to sequence-to-sequence modeling is one of the most common approaches. In previous works, thanks to their powerful gating functions, Long Short-Term Memories (LSTMs) [18] have been widely applied to many pedestrian trajectory prediction tasks and achieved excellent results, especially in the temporal dimension [7], [6], [19], [20], [8]. In recent years, with the great achievements of the Transformer [5] network in the domain of Natural Language Process (NLP) [21], [22], Transformer-based models are also applied to trajectory forecasting. In contrast to LSTMs, Transformer networks have a better capability of modeling temporal dependencies in long sequences based on the self-attention mechanism [23], [24], [9]. In addition to the previously adopted deterministic approaches like LSTMs, an increasing number of deep generative models, such as conditional variational autoencoders (CVAEs) [25], [26] and generative adversarial networks (GANs) [27] are applied to trajectory forecasting. Rather than producing one single prediction, generative models learn the potential future trajectories as a distribution and generate multiple possible predictions from latent space. For example, Social GAN and Sophie [28], [29] are proposed for pedestrian multi-path trajectory prediction via jointly training a generator and a discriminator. Compared to GANs, CVAE models predict multiple plausible trajectories conditioned on the past trajectories and acquire better performance in recent works [30], [31], [32], [8], [33].

Social Interaction Modeling. Besides modeling individual trajectory sequences, establishing the influence of pedestrians on each other or from the environment has been a critical issue in pedestrian trajectory prediction. As groundbreaking work, Helbing et al. [4] leverage dynamic social forces to imitate the influence of the surroundings on pedestrians, e. g., a repulsive force for collision avoidance and an attractive force for social connection. The social force model has been effectively applied in various fields like robotics [34] and crowd analysis [35], [36]. Another pioneering data-driven work is Social LSTM [7]. It proposes a new structure, called the social pooling layer, to aggregate the interaction information from neighbors. With the development of graph neural networks (GNNs) [37], more recent works of deterministic models like [24], [38] resort to modeling a crowd as a graph and combining GNNs with attention mechanisms to learn spatial interactions. Other approaches like [32], [6] first encode features over social dimension at each independent time step. Then, these social features are fed into another temporal sequence model to summarize the social relations over time. Unlike these methods above, we import social forces at each time step to a Transformer-based backbone, facilitating the learning of social interactions among pedestrians.

Goal-based Model. Recently, goal-based models have become an effective way to improve prediction performance [11], [39]. Diverse goal information can provide more predictive possibilities to deterministic model [40], [13]. Moreover, pedestrians are motivated by their destinations. Therefore, high uncertainty behavior can be limited through the goal information [14]. In contrast to those models that directly use the goal information as an input feature, we use the goal information to calculate the social forces for each pedestrian [4], and the resulting forces are used as input to our prediction module. However, the goal information is not accessible in inference time. To circumvent this issue, we utilize a goal-estimation [14] module for estimating the goals in the inference time.

III. METHODOLOGY

A. Problem formulation

In the context of pedestrian trajectory prediction problems, a complete trajectory of a pedestrian can be divided into two parts, the observed and the future trajectories. The observed trajectory at time steps $t \leq 0$ is denoted as $X = (X^{-H}, X^{-H+1}, ...X^{0})$, which in total includes $H+1$ observed time steps; While, the future trajectory at time steps $t > 0$ is denoted as $Y = (Y^1, Y^2, ..., Y^T)$ over T future time steps. Similar to [9], we use the x- and y-coordinate and the velocity sequence in the 2D coordinate system to parameterize trajectories. In addition, the joint social sequences of all N pedestrians in the same scene at the same time step t are denoted as $X^t = (x_1^t, x_2^t, ..., x_N^t)$ for the observation and $Y^t = (y_1^t, y_2^t, ..., y_N^t)$ for the future trajectories. In our proposed generative model $p_\theta(Y|X, G, F)$, where θ are the model parameters, the task is to forecast future trajectories Y depending on not only observed trajectories X but also goal information G and social forces F.

Following [13], we use both the position of the last time step and the differences between every single position and the goal position to parameterize the goal information. It should be noted that we use the ground truth of the last position Y^T to derive the goal representation in the training phase, while we use the estimation from the goal-estimation module in Sec. III-B to derive the goal representation in the test phase. Additionally, two kinds of forces, i. e., driving force F_{Dr} and repulsive force F_{Re}, are calculated for each agent at every time step. They are also represented as sequences.

$$F = \begin{cases} F_{\mathrm{Dr}} = ({f_{\mathrm{Dr}}}_1^{-H}, ..., {f_{\mathrm{Dr}}}_N^{-H}, ..., {f_{\mathrm{Dr}}}_1^{T}, ..., {f_{\mathrm{Dr}}}_N^{T}), \\ F_{\mathrm{Re}} = ({f_{\mathrm{Re}}}_1^{-H}, ..., {f_{\mathrm{Re}}}_N^{-H}, ..., {f_{\mathrm{Re}}}_1^{T}, ..., {f_{\mathrm{Re}}}_N^{T}). \end{cases} \quad (1)$$

In order to explore different ways of incorporating the goal information, on the basis of the baseline model AgentFormer [9] that takes velocity and position sequences as input, we propose three variants of the additional goal information. As denoted in Figure 1, **ForceFormer-Goal** directly adds the additional goal sequence to the input. Alternatively, **ForceFormer-Dr** uses F_{Dr} as the additional conditional information. **ForceFormer-Re** uses both the goal sequence and the repulsive force F_{Re} sequence as the additional input.

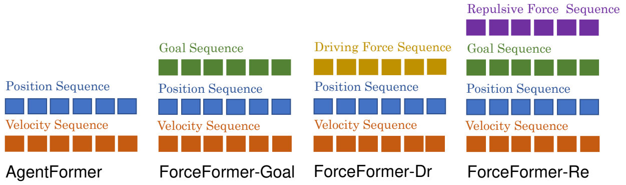

Fig. 1. Inputs of AgentFormer [9] and the proposed ForceFormer-Goal, ForceFormer-Dr, and ForceFormer-Re.

B. The Proposed Framework

Figure 2 depicts the overview of our proposed framework ForceFormer. It mainly consists of three modules: AgentFormer (X-Encoder, Y-Encoder, and Decoder) as the backbone, Goal-estimation module, and Social force model.

In the training process, the goal-estimation module and AgentFormer are trained separately. The goal information is supplied from ground truth Y^T, which is used for training the goal-estimation module and the calculation of social forces. The repulsive force and driving force are calculated by the position information, velocity information, and goal information. However, the ground truth Y^T is unavailable in the test phase. Hence, during the inference process, we sample K goal candidates for every trajectory from the goal-estimation module. Following the previous goal-conditioned trajectory prediction models [39], [11], [14], we evaluate all potential K goals against the ground truth and choose the one with the smallest $L2$ error as the estimated goal position in the test phase.

a) AgentFormer: The backbone prediction model is a CAVE-based model and establishes spatial and temporal relations using attention mechanisms. Based on the conditions of observed trajectory X, goal information G, and social forces F, the future trajectory distribution is modeled as $p_\theta(Y|X,G,F)$. The future trajectory distribution can be rewritten as

$$p_\theta(Y|X,G,F) = \int p_\theta(Y|Z,X,G,F)p_\theta(Z|X,G,F)dZ, \quad (2)$$

where $p_\theta(Z|X,G,F)$ is the conditional Gaussian prior, which is learned by X-Encoder. $p_\theta(Y|Z,X,G,F)$ is the conditional likelihood. Eq. (2) proposes a set of latent variables $Z = (z^{(1)},...,z^{(K)})$, reflecting the latent intent of pedestrian n to account for stochasticity and multi-modality in the pedestrian's future behavior.

The negative evidence lower bound $\mathscr{L}_{elbo}$ is used to address the intractable posterior $p_\theta(Z|Y,X,G,F)$. Concretely, the CVAE-based model is optimized using the loss function

$$\begin{aligned}\mathscr{L}_{elbo} = &-\mathbb{E}_{q_\phi(Z|Y,X,G,F)}[\log p_\theta(Y|Z,X,G,F)] \\ &+KL(q_\phi(Z|Y,X,G,F)||p_\theta(Z|X,G,F)),\end{aligned} \quad (3)$$

where $q_\phi(Z|Y,X,G,F)$ is the approximate posterior distribution parameterized by ϕ, which is learned by Y-Encoder. The first term in the above equation can be considered as the expected predicted probability of the future trajectory $p_\theta(Y|Z,X,G,F)$. The second term $KL(q_\phi(Z|Y,X,G,F)||p_\theta(Z|X,G,F))$ denotes the distribution difference between the prior and the approximate posterior, which both tend to be a standard normal distribution.

b) Social Forces: A modified social force model [4] that contains both driving and repulsive forces is applied in this work. The **driving force**, denoted by Eq. (4), describes the attractive effect related to the destination (goal position).

$$\vec{F}^0_\alpha = \frac{1}{\tau_\alpha}(v^0_\alpha \vec{e}_\alpha - \vec{v}_\alpha). \quad (4)$$

The value of the driving force depends on the deviation of the current velocity $\vec{v}_\alpha(t)$ from the desired velocity $v^0_\alpha(t) = v^0_\alpha \vec{e}_\alpha$. v^0_α is the speed at which a pedestrian would walk towards their destination in the desired direction $\vec{e}_\alpha$ if they were undisturbed. The relaxation time τ_α is a parameter that represents the expected time removing this deviation.

To avoid collisions, pedestrians maintain a proper distance from other strangers. The **repulsive force**, denoted by Eq. 5, describes the avoidance phenomenon between the ego pedestrian α and other pedestrian β,

$$\vec{f}_{\alpha\beta}(\vec{r}_{\alpha\beta}) = -\nabla_{\vec{r}_{\alpha\beta}} V_{\alpha\beta}[b(\vec{r}_{\alpha\beta})]. \quad (5)$$

The repulsive potential $V_{\alpha\beta}(b)$ is a monotonic decreasing function related to b, which represents the semi-minor axis of an ellipse. Through b, the equipotential lines keep the form of an ellipse that is pointed to the direction of motion.

$$2b = \sqrt{(\left\|\vec{r}_{\alpha\beta}\right\| + \left\|\vec{r}_{\alpha\beta} - v_\beta \Delta t \vec{e}_\beta\right\|)^2 - (v_\beta \Delta t)^2}. \quad (6)$$

In addition to distance, the influence of viewpoint also needs to be considered when calculating repulsive forces. Thus a parameter w is introduced.

$$w(\vec{e},\vec{f}) = \begin{cases} 1 & \text{if } \vec{e}\cdot\vec{f} \geq \left\|\vec{f}\right\| cos\varepsilon, \\ c & \text{otherwise,} \end{cases} \quad (7)$$

where the effective angle of sight is 2ε, and c is a constant factor. Hence, the repulsive force, after taking the perspective factor into account, is constrained as $\vec{F}_{\alpha\beta} = w\vec{f}_{\alpha\beta}$.

Since the repulsive force is based on the premise that two pedestrians are strangers, they want to keep their distance from each other and avoid collisions. However, in reality, many pedestrians travel in pairs, such as classmates, relatives, and friends, who share a common destination. Therefore, it is not logical to consider their repulsive force within a group, which could cause significant errors, particularly in high-density scenes. So we adapt the DBSCAN method, a density-based spatial clustering of applications with noise [41], [42], for every time step. Group members are identified [43] if they are in the same cluster for more than σ time steps. Intra-group repulsive forces are then removed to avoid unnecessary repulsion between group members.

c) Goal-Estimation Module: The goal-estimation module predicts the goal information by modeling its multi-modality using a probability distribution, and the goal information is provided for the AgentFormer prediction module and the calculation of social forces in the test phase. We adopt the goal module proposed in [14], [13] for this purpose.

First, the past trajectories in a heat map form concatenate with the semantic map information. Semantic information

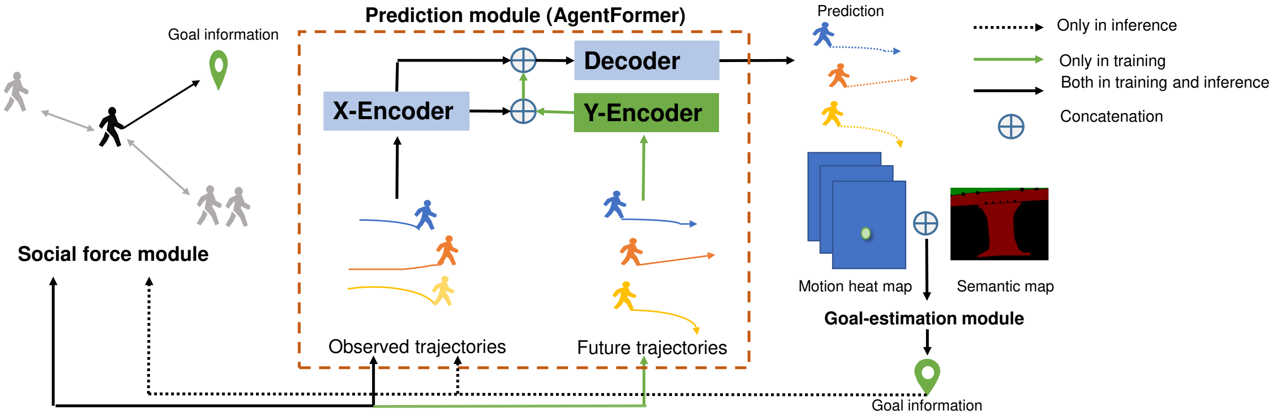

Fig. 2. An overview of the proposed framework ForceFormer.

is extracted from bird's-eye view RGB images to consider scene context, such as obstacles, pavement and terrain, using an off-the-shelf pre-trained semantic segmentation network [14]. The segmentation results in a tensor form $S \in \mathbb{R}^{W \times L \times C}$ containing C classes. W and L are the spatial sizes of the input image. The past trajectory $\{x_n^{-H}, x_n^{-H+1}, \ldots, x_n^{0}\}$ of agent n is mapped to the heat map representation $M \in \mathbb{R}^{W \times L \times (H+1)}$, which is a $2D$ Gaussian probability distribution map with mean p_i^t and variance $\sigma_S^2 I_2$. The trajectory-on-scene input tensor $M_s \in \mathbb{R}^{W \times L \times (C+H+1)}$ is created by concatenating the semantic tensor S with past motion history distribution maps M. The tensor is then fed to a U-Net [15] architecture consisting of encoder and decoder blocks. The final output is a spatial probability distribution of the final position.

After the U-Net model predicts a probability distribution of plausible final positions at the time step T, estimated goals are sampled from the probability maps. Because multiple samples are required, the Test-Time-Sampling-Trick (TTST) proposed in [14] is used. This technique involves initially sampling 10,000 possible goals and clustering them into 20 output modalities using the K-means algorithm.

IV. EXPERIMENTS

A. Dataset

The proposed framework is evaluated on ETH [16] and UCY [17], which have been widely used as the benchmark for pedestrian trajectory prediction. The datasets contain five different subsets as listed in Table I. A valid trajectory denotes a single pedestrian's track information in 20 consecutive frames captured at 2.5 Hz. These twenty-time steps are divided into two parts – the first eight time steps (3.2 s) are observed trajectories X, based on which twelve future time steps (4.8 s) as future trajectories Y are predicted. The position of the goal is located at the twentieth time step. It can be seen that the density of pedestrians varies across the subsets. The density in a scene largely influences the prediction results, especially in this work, because the calculation of social forces is closely related to crowd density.

TABLE I
THE NUMBER OF FRAMES AND VALID TRAJECTORIES IN THE ETH/UCY [16], [17] DATASETS.

	ETH	Hotel	Univ	zara01	zara02
Frame	1142	1788	947	883	1033
Valid trajectory	364	1197	24334	2356	5910

B. State-of-the-art models and baseline

We compare our proposed method, ForceFormer, with the following models. AgentFormer [9] is the baseline model without using any goal information. Sophoie [29] proposes a GAN-based model that combines trajectory information with context information. Trajectron++ [32] is a CVAE-based model that maintains top performance on the ETH/UCY benchmark. STAR [24] proposes a Temporal Transformer and a Spatial Transformer to model spatial-temporal information for pedestrian trajectory prediction. Moreover, ForceFormer is compared with a bunch of goal-based models. Namely, PECNet [11] is a goal-conditioning model for short-term trajectory prediction. Goal-GAN [12] integrates goal information in a GAN-based model for trajectory prediction. Heading [39] proposes a goal retrieval module that provides goal information for trajectory prediction. Y-net [14] combines scene information with goals and waypoints for trajectory prediction. Goal-SAR [13] proposes an attention-based recurrent network combined with the same goal-estimation module as ForceFormer.

C. Evaluation Metrics and Protocol

Three metrics are used to evaluate the proposed model. First, two standard error metrics are applied to measure the trajectory prediction performance of ForceFormer and compare it fairly with the previous models. These two distance errors are average displacement error ADE_K and final displacement error FDE_K of K trajectory samples of each agent compared to the corresponding ground truth.

$$ADE_K = \frac{1}{T} \min_{k=1}^{K} \sum_{t=1}^{T} \left\| \hat{y}_n^{t,(k)} - y_n^t \right\|^2, \tag{8}$$

$$FDE_K = \min_{k=1}^{K} \left\| \hat{y}_n^{T,(k)} - y_n^T \right\|^2 . \tag{9}$$

In addition, the number of collisions as another metric is leveraged to verify the social forces applied to ForceFormer.

$$NC = \sum_{m,n=1}^{N} \sum_{t=1}^{T} (\left\| \hat{y}_n^t - \hat{y}_m^t \right\|) < \gamma, m \neq n, \tag{10}$$

where m and n are different pedestrians in the same scene at time step t. The threshold γ is set for determining whether a collision occurs. In this paper, $\gamma = 0.1m$. All the metrics are computed with $K = 20$ samples. The calculation of NC is based on trajectories with the best ADE. Following prior works [44], [9], [32], [11], we adopt the leave-one-out strategy for the evaluation.

D. Implementation Details

For calculating social forces, we adopt 200° as the effective angle 2ε. The factor c for the field of view is 0.5. The threshold of minimum frames in the same cluster for grouping σ is four. Also, we consider that the desired direction cannot be calculated when the pedestrian position overlaps with the goal position, so we set the social forces at these positions to zero. For the AgentFormer backbone, we use all the same settings as in Yuan et al. [9]. But we only train the CVAE model using Adam optimizer [45] for 50 epochs, shorter than the original paper. For the goal-estimation module, in addition to the same settings as in Chiara et al., [13], we add a goal-specific MSE loss function $\mathscr{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \left\| y_i^T - \hat{y}_i^T \right\|^2$ and a hyper-parameter $\lambda = 1e^6$ to balance the original BCE loss function. All our models are trained on Google Colab with a single Tesla P100 GPU.

E. Results

In table II, we compare our approaches with current state-of-the-art methods. First, our proposed methods ForceFormer-Dr and ForceFormer-Re achieve better performance in all the subsets compared to the baseline model AgentFormer, e. g., on average, ForceFormer-Dr reduces FDE by 26% and ForceFormer-Re reduces ADE by 17%. In addition, when comparing to models that also use goal information, our methods perform on par with the previous best method Y-net. In particular, when we compare the results on each subset, we can find that our models achieve better performance than Y-net on the other four subsets, except for ETH. Moreover, when comparing the results in the high-density scenes, i. e., on Univ and Zara2, ForceFormer-Dr decreases FDE by 22% and 23%, respectively, than Y-net. The improvements indicate better performance of our method on final position predictions in high-density scenes.

TABLE II

QUANTITATIVE PERFORMANCES OF THE STATE-OF-THE-ART MODELS AND OUR PROPOSED MODELS ON THE ETH/UCY DATASETS.

Method	$ADE_k/FDE_k(m)\,K = 20$ Samples					
Datasets	ETH	Hotel	Univ	Zara1	Zara2	Average
Sophie [29]	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
STAR [24]	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
Trajectron++ [32]	0.67/1.18	0.18/0.28	0.30/0.54	0.25/0.41	0.18/0.32	0.32/0.55
AgentFormer [9]	0.45/0.75	0.14/0.22	0.25/0.45	0.18/0.30	0.14/0.24	0.23/0.39
PECNet [11]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
Goal-GAN [12]	0.59/1.18	0.19/0.35	0.60/1.19	0.43/0.87	0.32/0.65	0.43/0.85
Heading [39]	0.37/0.65	0.11/0.15	**0.20**/0.44	0.15/0.31	**0.12**/0.26	0.19/0.36
Y-net [14]	**0.28/0.33**	0.10/**0.14**	0.24/0.41	0.17/0.27	0.13/0.22	**0.18/0.27**
Goal-SAR [13]	**0.28**/0.39	0.12/0.17	0.25/0.43	0.17/0.26	0.15/0.22	0.19/0.29
ForceFormer-Dr	0.43/0.58	0.12/0.16	0.21/**0.32**	**0.14/0.20**	**0.12/0.17**	0.20/0.29
ForceFormer-Re	0.36/0.52	**0.09/0.14**	0.21/0.42	0.15/0.22	**0.12**/0.20	0.19/0.30

[*] The results of Trajectron++ and Heading are updated according to the implementation issue 53 [46] and sampling trick [47]. The underlined methods use goal information.

TABLE III

PERFORMANCE DIFFERENCES FOR THE METHODS USING GOAL INFORMATION MEASURED BY FDE.

	$FDE_k(m)K = 20$ Samples								
Method	Goal	F_{Dr}	F_{Re}	ETH	Hotel	Univ	Zara1	Zara2	Average
AgentFormer [9]				0.75	0.22	0.45	0.30	0.24	0.39
ForceFormer-Goal	✓			0.55	0.17	0.49	0.30	0.28	0.36
ForceFormer-Dr		✓		0.58	0.16	**0.32**	**0.20**	**0.17**	**0.29**
ForceFormer-Re	✓		✓	0.52	**0.14**	0.42	0.22	0.20	0.30

F. Ablation study

The variants of our proposed model making use of the goal information are compared against the FDE values in Table III and the number of collisions in Table IV. First, it can be seen clearly that, compared to the baseline model AgentFormer, all the variants making use of the additional goal information achieve smaller average FDE. Except ForceFormer-Goal, ForceFormer-Dr and ForceFormer-Re have evidently smaller numbers of collisions. Among the three variants of ForceFormer, ForceFormer-Goal, in general, performs worse than the other models in terms of FDE and the number of collisions across the subsets. This indicates that directly utilizing the goal sequences may not as effective as the social forces. With closer observation, we can see that ForceFormer-Dr achieves the smallest FDE in high-density pedestrian scenes like Univ and Zara02. In contrast, ForceFormer-Re has the smallest total collisions, with a 19.8% reduction compared to the baseline model AgentFormer.

TABLE IV

PERFORMANCE DIFFERENCES FOR THE METHODS USING GOAL INFORMATION MEASURED BY COLLISION NUMBERS.

	Collision number $CN_K, K = 20$ **samples**								
Method	Goal	F_{Dr}	F_{Re}	ETH	Hotel	Univ	Zara1	Zara2	Sum
AgentFormer [9]				0	2	655	4	22	683
ForceFormer-Goal	✓			**0**	**0**	672	**3**	22	697
ForceFormer-Dr		✓		**0**	1	556	**3**	28	588
ForceFormer-Re	✓		✓	**0**	1	**529**	5	**13**	**548**

G. Qualitative results

Figure 3 shows the qualitative results predicted by AgentFormer (left column), ForceFormer-Dr (middle column), and ForceFormer-Re (right column), respectively. From the upper row, we can see that, compared to AgentFormer, ForceFormer-Dr benefits from the goal information and driving force to predict trajectories around corners or turns. Although ForceFormer-Re predicts less accurate curving

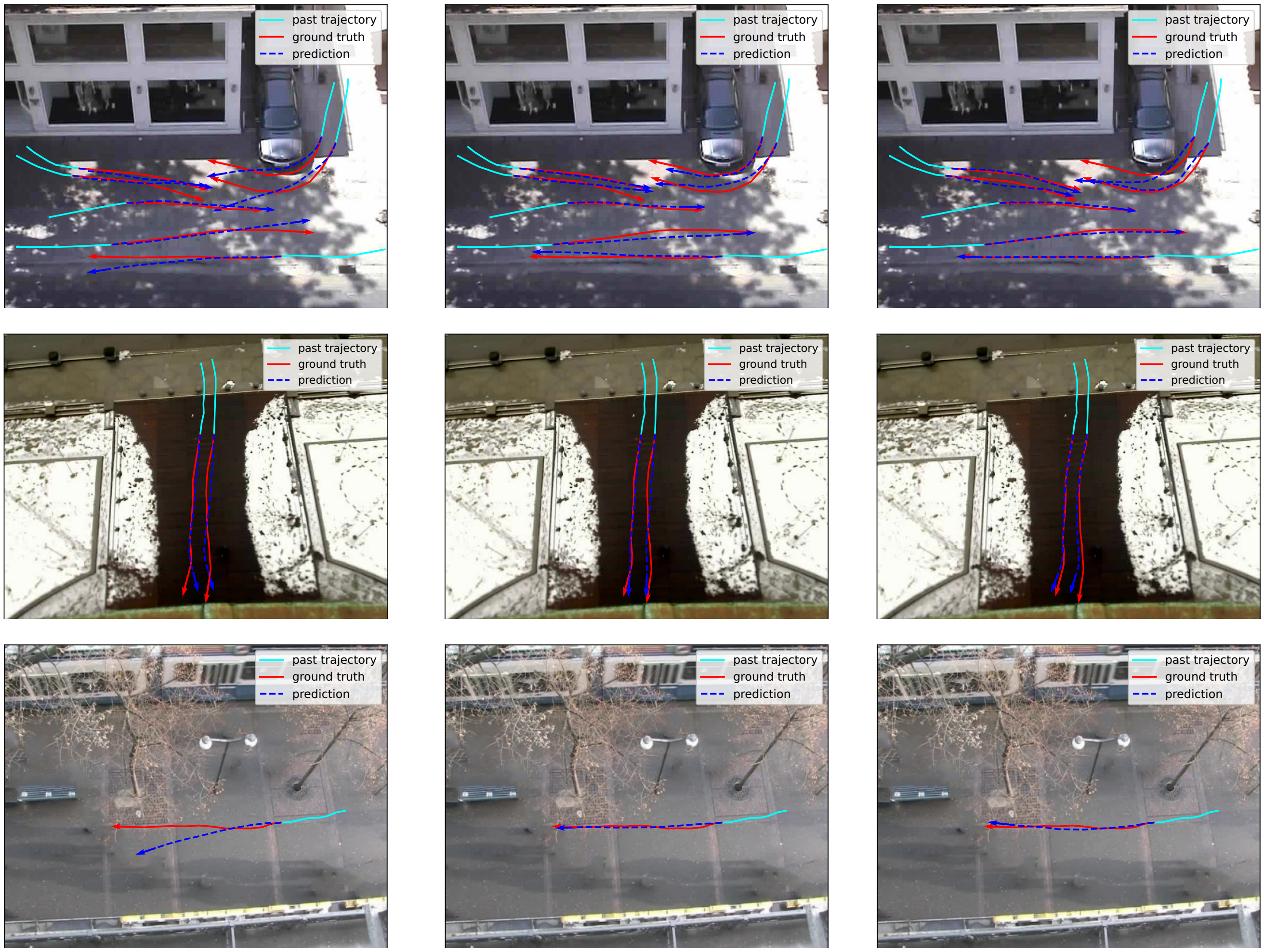

Fig. 3. Prediction results generated by AgentFormer (left column), ForceFormer-Dr (middle column), and ForceFormer-Re (right column). Each row represents a different scene.

trajectories, its prediction for other trajectories is closer to the corresponding ground truth. In the middle row for a scenario with two pedestrians walking in parallel, both ForceFormer-Dr and ForceFormer-Re predict more accurate final positions as the pedestrians make a left turn. In contrast, AgentFormer does not explore the goal information from the goal-estimation module and predicts walking in the middle of the road. A more visible scenario of predicting the final position can be seen in the bottom row. The prediction from AgentFormer largely deviates from the ground truth trajectory, while the predictions from ForceFormer-Dr and ForceFormer-Re are well aligned with the ground truth trajectory.

Limitations. Despite the enhanced performance brought by the social forces and the goal-estimation module, several limitations of the proposed model need to be noted. The collisions have been reduced, but the predictions from ForceFormer are not totally collision-free. One possible reason might be that in the social force module, we do not consider the interactions and forces within groups, which also may cause collisions. In future work, we will build a more comprehensive social force module and apply it to better simulate interactions among group members. Moreover, the overall performance of ForceFormer, especially the calculation of social forces, relies on the reliability of the goal-estimation module. Sub-optimal performance of this module can lead to compound errors in the final prediction. On the other hand, if we can access the ground truth goal information, we can quickly turn our model into a motion planning model.

V. Conclusion

This paper proposes a new goal-based trajectory predictor called ForceFormer that incorporates social forces into a Transformer-based generative model backbone. A U-Net-based goal-estimation module is adopted to predict the goals of pedestrians' trajectories. Additional to the position and velocity information, we derive the driving force from the estimated goal to efficiently simulate the guidance of a target on a pedestrian. Also, repulsive forces are used to help the model learn collision avoidance among neighboring pedestrians. ForceFormer achieves performance on par with the state-of-art models and better performance in the high-density scenarios on widely used pedestrian datasets.

REFERENCES

[1] Y. Luo, P. Cai, A. Bera, D. Hsu, W. S. Lee, and D. Manocha, "Porca: Modeling and planning for autonomous driving among many pedestrians," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3418–3425, 2018.

[2] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese, "Knowledge transfer for scene-specific motion prediction," in *Proceedings of ECCV*. Springer, 2016, pp. 697–713.

[3] Y. Li, H. Cheng, Z. Zeng, H. Liu, and M. Sester, "Autonomous vehicles drive into shared spaces: ehmi design concept focusing on vulnerable road users," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 1729–1736.

[4] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proceedings of CVPR*, 2019, pp. 6272–6281.

[7] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of CVPR*, 2016, pp. 961–971.

[8] H. Cheng, W. Liao, M. Y. Yang, B. Rosenhahn, and M. Sester, "Amenet: Attentive maps encoder network for trajectory prediction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 253–266, 2021.

[9] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proceedings of CVPR*, 2021, pp. 9813–9823.

[10] H. Cheng, F. T. Johora, M. Sester, and J. P. Müller, "Trajectory modelling in shared spaces: Expert-based vs. deep learning approach?" in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. Springer, 2021, pp. 13–27.

[11] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Proceedings of ECCV*. Springer, 2020, pp. 759–776.

[12] P. Dendorfer, A. Osep, and L. Leal-Taixé, "Goal-gan: Multimodal trajectory prediction based on goal position estimation," in *Proceedings of ACCV*, 2020.

[13] L. F. Chiara, P. Coscia, S. Das, S. Calderara, R. Cucchiara, and L. Ballan, "Goal-driven self-attentive recurrent networks for trajectory prediction," in *Proceedings of CVPR*, 2022, pp. 2518–2527.

[14] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *Proceedings of CVPR*, 2021, pp. 15 233–15 242.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proceedings of CVPR*. IEEE, 2009, pp. 261–268.

[17] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *Proceedings of CVPR*, 2019, pp. 12 085–12 094.

[20] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Proceedings of CVPR*, 2018, pp. 5275–5284.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[23] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 10 335–10 342.

[24] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Proceedings of ECCV*. Springer, 2020, pp. 507–523.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[26] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[28] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of CVPR*, 2018, pp. 2255–2264.

[29] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of CVPR*, 2019, pp. 1349–1358.

[30] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proceedings of CVPR*, 2019, pp. 2375–2384.

[31] C. Tang and R. R. Salakhutdinov, "Multiple futures prediction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[32] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Proceedings of ECCV*. Springer, 2020, pp. 683–700.

[33] G. Chen, J. Li, N. Zhou, L. Ren, and J. Lu, "Personalized trajectory prediction via distribution discrimination," in *Proceedings of ICCV*, October 2021, pp. 15 580–15 589.

[34] G. Ferrer, A. Garrell, and A. Sanfeliu, "Robot companion: A social-force based approach with human awareness-navigation in crowded environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1688–1694.

[35] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of CVPR*. IEEE, 2009, pp. 935–942.

[36] F. T. Johora, D. Yang, J. P. Müller, and Ü. Özgüner, "On the generalizability of motion models for road users in heterogeneous shared traffic spaces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 23 084–23 098, 2022.

[37] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[38] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[39] H. Zhao and R. P. Wildes, "Where are you heading? dynamic trajectory prediction with expert goal examples," in *Proceedings of ICCV*, October 2021, pp. 7629–7638.

[40] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid *et al.*, "Tnt: Target-driven trajectory prediction," *arXiv preprint arXiv:2008.08294*, 2020.

[41] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[42] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[43] H. Cheng, Y. Li, and M. Sester, "Pedestrian group detection in shared space," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1707–1714.

[44] H. Cheng, M. Liu, L. Chen, H. Broszio, M. Sester, and M. Y. Yang, "Gatraj: A graph-and attention-based multi-agent trajectory prediction model," *arXiv preprint arXiv:2209.07857*, 2022.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," https://github.com/StanfordASL/Trajectron-plus-plus/issues/53, 2020.

[47] H. Zhao and R. P. Wildes, "Where are you heading? dynamic trajectory prediction with expert goal examples," https://github.com/JoeHEZHAO/expert_traj, 2021.