TPPO: A Novel Trajectory Predictor With Pseudo Oracle

Biao Yang, Member, IEEE, Caizhen He, Pin Wang, Member, IEEE, Ching-Yao Chan, Member, IEEE, Xiaofeng Liu, Senior Member, IEEE, and Yang Chen, Member, IEEE

Abstract—Forecasting pedestrian trajectories in dynamic scenes remains a critical problem in various applications, such as autonomous driving and socially aware robots. Such forecasting is challenging due to human-human and human-object interactions and future uncertainties caused by human randomness. Generative model-based methods handle future uncertainties by sampling a latent variable. However, few studies explore the generation of the latent variable. In this work, we propose the trajectory predictor with pseudo Oracle (TPPO), which is a generative model-based trajectory predictor. The first pseudo oracle is pedestrians’ moving directions, and the second one is the latent variable estimated from ground truth trajectories. A social attention module is used to aggregate neighbors’ interactions based on the correlation between pedestrians’ moving directions and future trajectories. This correlation is inspired by the fact that pedestrians’ future trajectories are often influenced by pedestrians in front. A latent variable predictor is proposed to estimate latent variable distributions from observed and ground-truth trajectories. Moreover, the gap between these two distributions is minimized during training. Therefore, the latent variable predictor can estimate the latent variable from observed trajectories to approximate that estimated from ground-truth trajectories. We compare the performance of TPPO with related methods on several public datasets. Results demonstrate that TPPO outperforms state-of-the-art methods with low average and final displacement errors. The ablation study shows that the prediction performance will not dramatically decrease as sampling times decline during tests.

Index Terms—Future uncertainty, generative adversarial network, latent variable predictor, social attention, trajectory prediction.

FORECASTING pedestrian trajectories in dynamic scenes remains a critical problem with various applications, such as autonomous driving [1], social behavior-aware robots [2], and intelligent transportation systems [3], [4]. For example, an autonomous driving car could plan a safer route in the crosswalks could be accurately predicted by the roadside equipment, such as cameras on buildings or roads. Fig. 1 shows a scenario of trajectory prediction. Pedestrians’ future trajectories labeled with different colored arrows are predicted based on past trajectories marked with different colored lines. A human observer can most likely forecast that the woman (blue) with a shoulder bag will turn right to avoid the still car. However, a robot without prior knowledge or solid training may not do the same thing since it is hard to understand human-human and human-object interactions. Several inherent human properties, including interpersonal, socially acceptable, and multimodal properties, also pose critical challenges for robots to perform accurate trajectory prediction.

Early works related to trajectory prediction always depend on simple motion models, such as constant velocity and acceleration models. However, trajectory prediction could not be performed in complex scenes. Later, researchers utilize several stochastic models (e.g., Gaussian mixture regression [5] or dynamic Bayesian networks (DBNs) [6]) to model complex motion patterns. However, such models need hand-crafted features, and thus are difficult to implement. Recent trends in deep learning, especially the recurrent neural network (RNN), provide a data-driven manner to understand complex motion patterns. Specifically, the long-short term memory (LSTM) network is used to encode pedestrians’ motion patterns, and their future trajectories are estimated by sampling from the hidden states [7]. Recently, a social generative adversarial network (SGAN) is proposed to generate socially acceptable and multimodal trajectories [8]. Except for the understanding of motion patterns, human-human interactions are modeled by a social pooling module based on the common-sense rules. Human-object interactions are captured by feeding the model with the scene-level and object-level information through object detection and semantic segmentation. A detailed review is discussed in the next section.

Although remarkable progress has been made over the past years, there remain two shortcomings that influence trajectory prediction performance. First, few works utilize...
the correlation between pedestrians’ head orientations and their future trajectories. Common sense is that a pedestrian’s future trajectory is always influenced by pedestrians in front. Utilizing such common sense is beneficial to model pedestrians’ social interactions. Second, few studies explore the latent variable, which plays a critical role in generating multimodal predictions. Most works use random Gaussian noise as a latent variable. However, such latent variable encompasses little knowledge about the scene or pedestrians and thus is not a proper choice to improve trajectory prediction.

In this work, we propose trajectory predictor with pseudo Oracle (TPPO), which adversarially trains a generator and discriminator. The generator contains an LSTM-based encoder-decoder and the discriminator has an LSTM-based encoder. As Hasan et al. [9] indicated, pedestrians’ head orientations can be used as an oracle for an improved trajectory prediction. We propose to use pedestrians’ moving directions as a pseudo oracle, which approximately reflects their head orientations. Then, we utilize the correlation between pedestrians’ moving directions and their future trajectories through an attention mechanism. Such a mechanism highlights the influences among correlated pedestrians and thus can better-model pedestrians’ social interactions. Besides, the latent variable distribution of ground truth trajectories can be regarded as another oracle. We propose a novel latent variable predictor, which estimates two latent variable distributions from observed and ground truth trajectories, respectively. Then, their Kullback–Leibler divergence (KL-divergence) is minimized during training. In the testing stage, the predictor estimates the latent variable distribution of observed trajectories. Such a distribution is similar to that of ground truth trajectories and thus can be regarded as another pseudo oracle. Pedestrians’ positions, velocities, and accelerations are extracted from their trajectories and then are fed into the latent variable predictor. Positions reflect the potential scene layout, whereas velocities and accelerations represent pedestrians’ motion patterns and radicalness. Notably, the proposed latent variable predictor learns knowledge from trajectories only, and thus increases little computational overhead.

Generally, our contributions are three-fold.

1) We propose a social attention pooling module that fully utilizes the correlation between pedestrians’ moving directions and future trajectories. The attention mechanism improves the modeled social interactions.

2) We propose a novel latent variable predictor that can estimate a knowledge-rich latent variable for an improved prediction performance. Such a latent variable is learned from trajectories and thus increase little computational overhead.

3) We embed the social attention pooling module and the latent variable predictor into a GAN framework to generate socially acceptable and multimodal outputs. Moreover, we achieve state-of-the-art performance on ETH [10] and UCY [11] datasets.

The remainder of this article is organized as follows. Section II reviews related works. Section III describes the proposed method in detail. Section IV presents the experimental results. Section V provides the conclusion and discussion.

II. RELATED WORK

A. Trajectory Prediction Methods

Trajectory prediction is a modeling problem that attempts to understand pedestrians’ motion patterns by examining pedestrian time-series data. Early works often focus on predicting future trajectories with dynamics-based methods, including constant velocity and acceleration models [12]. However, a simple kinematic model is not suitable for long-term predictions. For long prediction horizons, flow-based methods [13], [14] are proposed to learn the directional flow from observed trajectories in the scene. Subsequently, trajectories are generated by recursively sampling the distribution of future motion derived from the learned directional flow. To cope with complicated scenarios, researchers have resorted to several learning-based methods, such as Gaussian mixture regression [5], Gaussian process [15], random tree searching [16], hidden Markov models [17], and DBNs [6]. However, it is nontrivial to handle high-dimensional data with these traditional methods.

With the rise of deep learning, RNN provides a data-driven manner of encoding pedestrians’ motion series. Prediction models have two categories, namely, deterministic and generative models. For deterministic models, the distribution of future trajectories is estimated from the hidden state encoded by an LSTM or a gated recurrent unit (GRU). Alahi et al. [7] presented Social-LSTM, which models pedestrian motions with a shared LSTM and then performs trajectory prediction through sampling. Zhang et al. [18] proposed SR-LSTM, which recursively refines the hidden states of LSTM, to recognize the critical current intention of neighbors. Refined hidden states guarantee improved prediction performance. Some works attempt to embed additional information. Xue et al. [19] and Syed and Morris [20] utilized three LSTMs to encode...
person, social, and scene scale information and then aggregate them for context-aware trajectory prediction. Ridel et al. [21] presented a joint representation of the scene and past trajectories by using Conv-LSTM and LSTM, respectively. Lisotto et al. [22] improved Social-LSTM by encompassing prior knowledge about the scene as a navigation map that embodies most frequently crossed areas. Moreover, the scene context is obtained through semantic segmentation to restrain motion for additional plausible paths. These methods improve the prediction performance of Social-LSTM. In contrast to LSTM-based trajectory predictions, several relevant studies attempted to the Transformer architecture [23], which introduces a well-designed attention mechanism to encode historical trajectories. Except for forecasting trajectories from the bird’s eye view [23], [24], these Transformer-based methods can predict pedestrians’ trajectories from the first perspective [25] and 3-D scenes [26], or deeply fused with scene semantic information [27], [28]. However, they suffer from future uncertainties, which pose a significant challenge in trajectory prediction.

Generative models handle future uncertainties by introducing an alterable latent variable. Lee et al. [29] presented DESIRE, which generates multimodal predictions with conditional variational auto-encoders. Later, Gupta et al. [8] proposed SGAN, which adversarially trains the generator and discriminator to produce socially acceptable trajectories. SGAN used the random Gaussian noise as the latent variable and thus generated diverse outputs. Zhu et al. [30] proposed StarNet, which is similar to SGAN except for the use of a query module. Amirian et al. [31] replaced the L2 loss used in SGAN with the information loss [32] to avoid mode collapse. Yang et al. [33] introduced multihop attention and memory network to conduct trajectory prediction. Unlike using random noise, they leveraged trajectory knowledge stored in the external memory network to forecast multimodal future trajectories. Recently, diffusion-based methods introduced motion indeterminacy diffusion, which gradually discards the indeterminacy to obtain the desired trajectory from ambiguous passable areas [34], [35].

Besides methods mentioned above, reinforcement learning-based trajectory predictors [36], [37], [38], [39], [40] exhibit a rising trend. Moreover, reinforcement learning-based methods always provide the optimal trajectory, whereas the true trajectory is suboptimal due to pedestrians’ randomness. However, estimating pedestrians’ destinations that are essential for reinforcement learning is difficult due to future uncertainties. Therefore, recent methods rely on goal forecasting for long-term trajectory prediction. For example, Chiara et al. [41] designed a scene-aware goal-estimation module to highlight that multimodality is a fundamental component of the prediction task. Chen et al. [42] proposed PPNet, the first goal-conditioned Transformer network that efficiently models the temporal features and maximizes the utilization of goal guidance. Lee et al. [43] introduced a multiscale goal-learning strategy for CVAE-based probabilistic models to realize environment-aware collision-free trajectory prediction. Similar to these state-of-the-art works, TPPO implicitly estimates goal information through the proposed latent variable predictor, achieving promising trajectory prediction performance.

B. Social Interaction Modeling

Great efforts in trajectory prediction are devoted to modeling pedestrians’ social interactions. These interactions can be defined by hand-crafted rules, such as social forces [44] and stationary crowds’ influences [45]. However, capturing complex interactions in the scene with these rules is difficult. Therefore, current works always learn social interactions in a data-driven manner. Social-LSTM [7] first proposed the social pooling layer to aggregate neighbors’ information in a particular grid. Yi et al. [45] presented a social-affinity map by replacing the grid with a bin. SGAN [8] removed the regional restraint and directly aggregated all neighbors’ information in the scene. To further capture pedestrians’ social interactions, Amirian et al. [31] replaced the relative displacement with the bearing angle, the Euclidean distance, and the closest distance between the target person and neighbors. These additional variables calculated from observed trajectories are more intuitional in modeling social interactions than relative displacements. They also used an attention mechanism, which is useful in highlighting key neighbors and features as proposed in SoPhie [46].

Except for the social pooling layer and its variants, graph neural network is another mainstream to model pedestrians’ social interactions [47], [48]. Such methods are especially suitable for handling heterogeneous agents [49] or crowd scenarios [50]. However, both methods neglect the correlation between pedestrian’s head orientation and trajectory prediction. Wang et al. [51] proposed a novel spatio-temporal interaction aware module to capture complex social interactions based on graph convolutional network (GCN). Chen and Wang [52] dynamically inferred the interacting relations and the interacting effects among agents, characterizing the interaction structure as an edge-featured graph. Chen et al. [53] leveraged Transformer-based networks to accurately extract interaction information both on the spatial and temporal dimensions. Yang and Pei [54] used GCN with a spatial transformer to capture the time-varying spatial dependencies. As Hasan et al. [9] reported, knowing the head orientation is beneficial to modeling social interactions. Specifically, future trajectories of the target person are influenced by pedestrians in front. Hasan et al. [55] proposed a novel trajectory prediction framework named MX-LSTM, which could predict future head orientations to increase the standard capabilities of the long-term trajectory forecasting approaches. Kooij et al. [56] established a context-based DBN, which incorporated the context cues (i.e., pedestrian head orientation, situation criticality, and distance to the curbside) as latent states for improved trajectory prediction performance. However, such an oracle cue is difficult to estimate from image data. In this work, we approximately take pedestrians’ moving directions as their head orientations and model their social interactions with a social attention pooling module.

C. Latent Variable Learning

Generative model-based prediction methods have gone mainstream due to their ability to handle future uncertainties. The latent variable used in the generator has a strong correlation with generated multimodal outputs. Several previous
works [8], [46] used random Gaussian noise as the latent variable and injected little knowledge about pedestrians or scenes into the generator. Amirian et al. [31] sampled part of the latent variable from a uniform distribution on the interval from 0 to 1. However, associating the trajectory or the scene with a specific latent variable remains difficult. Zhang et al. [57] presented a stochastic module to generate the latent variable based on pedestrians’ movements. The latent variable is sampled from the embedding outputs of a social graph network. Tang and Salakhutdinov [58] proposed a dynamic encoder to learn latent variables from multiple inputs, including trajectories and the environmental context. However, processing the visual context needs more computation than processing trajectory data alone.

In this work, we propose a novel predictor that only learns the latent variable from trajectory data. Specifically, we feed pedestrians’ positions, velocities, and accelerations into the predictor to understand the latent environmental context, pedestrians’ motion patterns and radicalness. Unlike References [57], [58] those only utilized observed data, we attempt to minimize the latent variable distribution gaps between observed and ground truth trajectories. Our inspiration comes from Reference [59], which focused on stochastic video generation with a learned prior.

### III. PROPOSED METHOD

In this work, we develop a trajectory predictor that can generate multiple future trajectories of all agents in a scene with high accuracy. Fig. 2 illustrates the system pipeline of TPPO.

#### A. Problem Definition

The trajectory prediction problem is a time-series analysis. For pedestrian $i$, we first denote the position $(x'_t, y'_t)$ at time step $t$ as $p'_t$. The goal of trajectory prediction is to estimate
the future trajectory $T_i = (p_i^{t+1}, \ldots, p_i^{t+T_{obs}})$, considering his motion history $H_i = (p_i^0, \ldots, p_i^t)$ and interactions with other pedestrians or objects. Then, the trajectory prediction problem is converted into training a parametric model that predicts future trajectory $T_i$ ($i=1, \ldots, n$), which can be formulated as follows:

$$\arg \max_{\Theta} P_b(T_1, \ldots, T_n | H_1, \ldots, H_n)$$

where $\Theta$ represents learnable parameters, and $n$ represents the number of pedestrians. Recently, the formulation mentioned above is always converted into a sequence-to-sequence prediction problem, which the data-driven RNN module can address.

### B. Generator and Discriminator

As reviewed above, generative model-based methods are always used for trajectory prediction due to their ability to generate multimodal outputs. Therefore, we design the structure of TPPO based on SGAN, which can produce multimodal and socially acceptable trajectories through adversarial training. We briefly introduce the generator and the discriminator as follows.

**Generator:** The generator consists of a shared encoder-decoder and a social attention pooling module used to model pedestrians’ social interactions. A linear layer is used to convert the relative displacements $[8], [31], [48]$ of pedestrian $i$ ($\Delta x_i = x_i^t - x_i^{t-1}, \Delta y_i = y_i^t - y_i^{t-1}$) into a fixed-length vector $e_i^t$. Then, the vector is fed into an LSTM-based encoder to generate the hidden state of pedestrian $i$ at time $t$ as follows:

$$e_i^t = \phi(\Delta x_i, \Delta y_i; W_{ee})$$

$$h_{ci}^t = \text{LSTM}(h_{ci}^{t-1}, e_i^t; W_{encoder})$$

where $\phi(\cdot)$ is a linear layer function with ReLU nonlinearity. $W_{ee}$ and $W_{encoder}$ are the learnable weights of $\phi(\cdot)$ and the encoder function LSTM($\cdot$), respectively.

Afterward, pedestrians’ interactions are aggregated by the social attention pooling module and latent variables are estimated by the latent variable predictor. We will discuss these two modules in detail later. Then, an LSTM-based decoder is used to estimate the future trajectories of pedestrian $i$ based on the concatenation of the hidden state $h_{ci}^t$, the pooling output $P_i^t$, and the estimated latent variable $z_i^t$. The decoding is formulated as follows:

$$h_{di}^t = \text{LSTM}(h_{di}^{t-1}, [h_{ci}^t, P_i^t, z_i^t]; W_{decoder})$$

$$(\hat{x}_i^t, \hat{y}_i^t) = \gamma(h_{di}^t)$$

where $\gamma(\cdot)$ is a linear layer function with ReLU nonlinearity to predict future trajectories $(\hat{x}_i^t, \hat{y}_i^t)$. $W_{decoder}$ represents the learnable weight of the decoder, which is shared by all pedestrians in the scene.

**Discriminator:** The discriminator is used to classify the predicted and ground truth trajectories as socially acceptable or not. Both trajectories are fed into an LSTM-based encoder to generate the embedding, and then a Softmax classifier is used to perform classification based on the embedding. Pedestrians’ interactions are not needed in the discriminator.

### C. Social Attention Pooling Module

The social pooling layer proposed in SGAN can aggregate pedestrians’ interactions in crowded scenes. As shown in Fig. 3, a multilayer perceptron (MLP) module is used to encode relative positions between target A and all other pedestrians. These embedding vectors are concatenated with each pedestrian’s hidden state. Another MLP is used to process the concatenated vectors and the results are pooled element-wise to compute target A’s pooling vector.

However, such a pooling layer only considers relative positions between target people and their neighbors. As common knowledge, pedestrians’ future trajectories are always influenced by people in front. As demonstrated in Fig. 3, target
A’s future trajectories are influenced by targets B and C, who are in A’s field of view (FoV). Target D does not interfere A’s trajectory decision even when he runs to A. Several previous works utilized pedestrians’ head orientations to infer the FoV [9]. However, precisely recognizing pedestrians’ head orientations from vision data is challenging. Thus, using such common knowledge for an improved trajectory prediction is difficult.

TPPO uses an attention mechanism to utilize the correlation between pedestrians’ head orientations and their trajectories. We approximately take pedestrians’ moving directions at the last observed time step as their head orientations. Then, the cosine values of all pedestrians’ bearing angles are calculated as follows:

$$\cos(B) = \begin{bmatrix} \cos(b_{11}) \cdots \cos(b_{1n}) \\ \vdots \ \ \vdots \\ \cos(b_{m1}) \cdots \cos(b_{mn}) \end{bmatrix}$$

where $n$ is the number of pedestrians in a scene. $b_{ij}$ represents the bearing angle of agent $j$ from agent $i$, that is, the angle between the velocity of agent $i$ and the vector joining agents $i$ and $j$.

Afterward, the attentional weights are calculated based on the cosine values. We perform hard and soft attention operations to refine the second MLP’s outputs. Finally, outputs of the attention module are max-pooled to generate pedestrian A’s pooling vector $P_A$. The hard and soft attention operations are formulated as follows.

**Hard Attention:** As discussed above, one pedestrian’s influence on another decreases when the bearing angle increases. Therefore, hard attention weight is represented as a matrix $H_A$ of the same size of $\cos(B)$, and each element $h_{ij}$ in $H_A$ is set to 0 or 1 by thresholding. $h_{ij}$ is set to 1 if $\cos(b_{ij})$ is larger than an empirically threshold $-0.2$, otherwise is set to 0.

**Soft Attention:** Unlike hard attention that calculates attention weight by thresholding, soft attention adaptively calculates various pedestrian pairs’ correlations. Soft attention weight $S_A$ can be formulated as follows:

$$S_A = \delta(\varphi(\cos(B)))$$

where $\delta(\cdot)$ represents the sigmoid activation, and $\varphi(\cdot)$ represents the $1 \times 1$ convolution.

**D. Latent Variable Predictor**

The latent variable predictor is used to estimate an alterable latent variable for generating multimodal outputs. Specifically, we attempt to train a well-learned predictor, which can estimate similar latent variable distributions from observed and ground truth trajectories, respectively. Afterward, the predictor can estimate knowledge-rich latent variables from observed trajectories only in the testing stage. Inputs to the latent variable predictor are pedestrians’ positions, velocities, and accelerations. As shown in Fig. 2, the latent variable predictor consists of two feed-forward neural networks, which are formulated as follows:

$$\begin{align*}
(\mu^k_i, \sigma^k_i) &= \Psi(\check{I}^k_i; W_{LP}^k) \\
(\hat{\mu}^k_i, \hat{\sigma}^k_i) &= \hat{\Psi}(\hat{I}^k_i; \hat{W}_{LP}^k)
\end{align*}$$

where $\Psi(\cdot)$ and $\hat{\Psi}(\cdot)$ are the feed-forward neural networks with learnable weights $W_{LP}^k$ and $\hat{W}_{LP}^k$, respectively. $I^k_i$ and $\hat{I}^k_i$ are the $k^{th}$ kinds of inputs (positions, velocities, and accelerations) we extract from observed and ground-truth trajectories, respectively. $(\mu^k_i, \sigma^k_i)$ and $(\hat{\mu}^k_i, \hat{\sigma}^k_i)$ represent the latent variable distributions of the $k^{th}$ kind of input estimated by $\Psi(\cdot)$ and $\hat{\Psi}(\cdot)$, respectively. The final latent variable $z$ is generated by concatenating the latent variable distributions of three kinds of inputs and the random Gaussian noise. $(\mu^k, \sigma^k)$ and $(\hat{\mu}^k, \hat{\sigma}^k)$ are used in the training and testing stages, respectively. The random Gaussian noise is used to generate multimodal outputs as that used in SGAN. The latent variable estimated by the proposed predictor is expected to provide rich information for precise trajectory prediction while maintaining the ability to handle future uncertainties.

**E. Loss Function**

The loss function used in this work consists of three parts: 1) adversarial; 2) variety; and 3) latent variable distribution losses. The adversarial loss is used to generate socially acceptable trajectories by performing a two-player min-max game. Unlike traditional GAN, the latent variable $z$ used in this work is sampled from the ground truth trajectory at the training stage with the latent variable predictor. Thus, the adversarial loss is formulated as follows:

$$L_{adv} = \mathbb{E}_{x \sim \text{data}(\alpha)}[\log D(x)] + \mathbb{E}_{c \sim \mathcal{P}([\varphi_1]^{20})}[\log(1 - D(G(z))]$$

The variety loss is used to fit the best-predicted trajectory in L2 loss while maintaining multimodal outputs. We follow its implementation proposed in SGAN, and the variety loss is defined as follows:

$$L_{\text{variety}} = \min_m \| \hat{T}_m - T^m \|_2$$

where $\hat{T}_m$ and $T^m$ are ground truth and the $m$th predicted trajectories, respectively. $m$ is a hyper-parameter and is set to 20 according to SGAN.

The latent variable distribution loss is used to measure the latent variable distribution gaps between observed and ground truth trajectories. We introduce KL-divergence to calculate the loss, which is formulated as follows:

$$L_{LD} = D_{KL}(\text{\mu}_i^k, \sigma_i^k)\| \text{\hat{\mu}_i^k, \hat{\sigma}_i^k})$$

Afterward, the total loss is defined in a weighted manner as follows:

$$L_{\text{total}} = L_{adv} + \alpha \times L_{\text{variety}} + \beta \times L_{LD}$$

where $\alpha$ and $\beta$ are set to 1 and 10, respectively, by cross validation across benchmarking datasets.

**F. Implementation Details**

LSTMs with 32-dimensional hidden states are used in the encoder and the decoder. We use two three-layer MLPs ($2 \times 32 \times 32$ for the first one, and $64 \times 32 \times 16$ for the second one) in the social attention pooling module. The 16-D latent
variable $z$ contains three four-dimensional vectors that embedded from positions, velocities, and accelerations and another four-dimensional random Gaussian noise. We iteratively train the generator and discriminator with a batch size of 64 for 600 epochs using Adam [60], with the initial learning rate of 0.001. The training frequency of the discriminator is twice that of the generator for better convergence. The proposed model is built with a Pytorch framework and is trained with an NVIDIA GTX-1080 GPU.

IV. EXPERIMENTAL RESULTS

The proposed method is evaluated on two publicly available datasets, namely, ETH [10] and UCY [11]. Both of them consist of real-world pedestrian trajectories with rich human-human and human-object interaction scenarios. There are five sets of data, namely, ETH, HOTEL, UNIV, ZARA1, and ZARA2. In total, there are 1536 pedestrians in challenging scenarios, such as people crossing each other, group forming and dispersing, and collision avoidance. All the trajectories are converted to real-world coordinates, sampled every 0.4 s through the interpolation operation.

Similar to several prior works [8], [31], the proposed method is evaluated with two error metrics as follows.

1) **Average Displacement Error (ADE):** Average L2 distance between the ground-truth trajectory and the predicted trajectory over all predicted time steps.

2) **Final Displacement Error (FDE):** The Euclidean distance between the actual final destination and the predicted final destination at the last step of prediction.

We use the leave-one-out approach for evaluation [7]. Specifically, we train models on four sets and test them on the remaining set. The observed and predicted horizons are 8 (3.2 s) and 8/12 (3.2/4.8 s) time steps, respectively. We denote $T$ as the prediction horizon.

A. Quantitative Evaluations

**Comparisons With Baseline Methods:** We compare the proposed method against the following baselines.

1) **Linear:** A linear regressor is used to predict future trajectories by minimizing the least square error.

2) **LSTM:** An LSTM is used to embed the motion patterns of observed trajectories. Future trajectories are predicted based on the learned motion patterns.

3) **Social-LSTM** [7]: An improved LSTM-based trajectory prediction method by proposing a social pooling layer to aggregate hidden states of interested pedestrians. Future trajectories are predicted by decoding the concatenation of LSTM embedding and social pooling outputs.

4) **SGAN** [8]: An improved version of Social-LSTM by utilizing adversarial training to generate socially acceptable trajectories. Gaussian noises are used as latent variables to generate multimodal outputs in consideration of pedestrians’ future uncertainties. The model is trained using a variety loss with the hyper-parameter set to 20. During the test, 20 times are sampled from the generator, and the best prediction in L2 sense is used for quantitative evaluation.

Fig. 4 demonstrates the quantitative comparisons between TPPO, with either soft or hard attention modules, and the aforementioned baseline methods across five sets. It is obviously seen that the proposed TPPO outperforms all baseline methods in ADE and FDE for both prediction horizons. We owe the superiority of TPPO to the proposed social attention pooling module and the novel latent variable predictor. Furthermore, TPPO with hard attention module is slightly better than TPPO with soft attention module. Such a difference may indicate that an empirically defined threshold is more suitable to be used in the social attention pooling module since it is challenging to learn a well-trained soft attention function with limited trajectory data.

Table I presents the comparison between TPPO and state-of-the-art methods. Similar to results in Fig. 4, TPPO with the hard attention module outperforms others in both prediction horizons in UNIV, ZARA1, and ZARA2 sets. We also report the relative ADE and FDE gains of TPPO with the hard attention module compared with the STGAT, which is the best method among methods selected for comparison. As shown in the column “Rel. gain,” our method
Fig. 4. Quantitative results between the proposed method and baseline methods mentioned above across five sets. We report the ADE and FDE for $T = 8$ and $T = 12$ (8/12) in meters. Our method outperforms other baseline methods (low is preferred and is labeled with bold fonts).

Achieves appreciable gains except in ETH and HOTEL sets. Such satisfactory performance benefits from the proposed social attention pooling module and the novel latent variable predictor. Compared with traditional methods that capture pedestrians' social interactions in a data-driven manner, the social attention pooling module introduces obstacle avoidance.
TABLE I
Comparison Results With State-of-the-Art Methods Across All Datasets. We Report the ADE and FDE for $T = 12$ in Meters. Our Method Outperforms State-of-the-Art Methods in Univ, Zara1, and Zara2 Sets, and Is Specifically Good for Average ADE and FDE (Low Is Preferred and Is Labeled With Bold Fonts). “Rel. Gain” Shows the Relative ADE and FDE Gains of Our Best Model + Hard (Marked in Red) Compared with the Latest Method AC-VRNN (Marked in Blue)

| Metric | Dataset | SR-LSTM | Sophie | S-Ways | Social-BiGAT | CGNS | STGAT | AC-VRNN | TPPO + Soft | TPPO + Hard | Rel. gain |
|--------|---------|---------|--------|--------|--------------|------|-------|---------|------------|------------|-----------|
| ADE    | ETH     | 0.63    | 0.70   | 0.39   | 0.69         | 0.62 | 0.65  | 0.61    | 0.79       | 0.75       | -18.67%   |
|        | HOTEL   | 0.37    | 0.76   | 0.39   | 0.49         | 0.70 | 0.35  | 0.30    | 0.37       | 0.36       | -16.67%   |
|        | Univ    | 0.51    | 0.54   | 0.55   | 0.55         | 0.48 | 0.52  | 0.58    | 0.48       | 0.39       | +48.72%   |
|        | Zara1   | 0.41    | 0.30   | 0.44   | 0.30         | 0.32 | 0.34  | 0.34    | 0.22       | 0.22       | +54.55%   |
|        | Zara2   | 0.32    | 0.38   | 0.51   | 0.36         | 0.35 | 0.29  | 0.28    | 0.18       | 0.23       | +21.74%   |
| AVG    | ETH     | 0.45    | 0.54   | 0.46   | 0.48         | 0.49 | 0.43  | 0.42    | 0.41       | 0.39       | +7.69%    |
| FDE    | HOTEL   | 1.25    | 1.43   | 0.64   | 1.29         | 1.40 | 1.12  | 1.09    | 1.42       | 1.27       | -14.17%   |
|        | Univ    | 1.10    | 1.24   | 1.31   | 1.32         | 1.22 | 1.10  | 1.22    | 0.94       | 0.74       | +64.86%   |
|        | Zara1   | 0.90    | 0.63   | 0.64   | 0.62         | 0.59 | 0.69  | 0.68    | 0.39       | 0.37       | +83.78%   |
|        | Zara2   | 0.70    | 0.78   | 0.92   | 0.75         | 0.71 | 0.60  | 0.59    | 0.31       | 0.45       | +31.11%   |
| AVG    | ETH     | 0.94    | 1.15   | 0.83   | 1.00         | 0.97 | 0.83  | 0.83    | 0.75       | 0.71       | +16.90%   |

TABLE II
Ablation Study. We Report the ADE and FDE for $T = 8$ and $T = 12$ (8/12) in Meters Across Five Sets. SLVP/MLVP Represent Single/Multiple Inputs Latent Variable Predictor, Respectively. Soft/Hard Represent Soft and Hard Attention Mechanisms, Respectively (Low Is Preferred and Is Labeled With Bold Fonts)

| Metric | Dataset | SGAN         | + SLVP  | + MLVP  | + Soft   | + Hard    | + MLVP + Soft | + MLVP + Hard |
|--------|---------|--------------|---------|---------|----------|-----------|---------------|---------------|
| ADE    | ETH     | 0.60/0.87    | 0.71/0.94 | 0.64/0.67 | 0.73/0.84 | 0.67/0.86 | 0.64/0.79    | 0.51/0.75     |
|        | HOTEL   | 0.52/0.67    | 0.42/0.65 | 0.29/0.44 | 0.36/0.49 | 0.50/0.51 | 0.28/0.37    | 0.22/0.36     |
|        | Univ    | 0.44/0.76    | 0.40/0.70 | 0.27/0.45 | 0.38/0.63 | 0.39/0.62 | 0.28/0.48    | 0.25/0.39     |
|        | Zara1   | 0.22/0.35    | 0.23/0.40 | 0.17/0.24 | 0.21/0.32 | 0.21/0.34 | 0.14/0.22    | 0.13/0.22     |
|        | Zara2   | 0.29/0.42    | 0.21/0.34 | 0.13/0.21 | 0.20/0.29 | 0.20/0.29 | 0.12/0.18    | 0.11/0.23     |
| AVG    | ETH     | 0.41/0.61    | 0.39/0.61 | 0.30/0.40 | 0.38/0.51 | 0.39/0.52 | 0.29/0.41    | 0.24/0.39     |
| FDE    | HOTEL   | 1.19/1.62    | 1.35/1.96 | 1.12/1.13 | 1.37/1.67 | 1.28/1.71 | 1.12/1.42    | 0.82/1.27     |
|        | Univ    | 1.02/1.37    | 0.81/1.40 | 0.58/0.91 | 0.68/0.98 | 0.96/1.08 | 0.46/0.70    | 0.35/0.70     |
|        | Zara1   | 0.84/1.52    | 0.83/1.49 | 0.51/0.69 | 0.79/1.34 | 0.81/1.34 | 0.54/0.94    | 0.46/0.74     |
|        | Zara2   | 0.43/0.68    | 0.46/0.82 | 0.30/0.42 | 0.41/0.64 | 0.40/0.68 | 0.24/0.39    | 0.21/0.37     |
| AVG    | ETH     | 0.58/0.84    | 0.43/0.73 | 0.22/0.36 | 0.40/0.60 | 0.40/0.61 | 0.20/0.31    | 0.18/0.45     |
|        | HOTEL   | 0.81/1.21    | 0.78/1.28 | 0.55/0.70 | 0.73/1.05 | 0.77/1.08 | 0.51/0.75    | 0.40/0.71     |

experiences into the data-driven learning process to better-capture social interactions. Furthermore, the latent variable predictor first predicts pedestrians’ future intentions from observed trajectories, then forecasts accurate future trajectories based on predicted intentions. Compared with the proposed method, S-Ways performs well in ETH and HOTEL sets by designing an attention-based module that focuses on the bearing angle, the Euclidean distance, and the distance of the closest approach. Such a module can capture human-human and human-object interactions in complex scenes that contain lots of obstacles. However, S-Ways performs poorly in ZARA1 and ZARA2 sets. AC-VRNN achieves the suboptimal average ADE and FDE because of the usage of the graph-based attention mechanism. It reveals that the graph model is good at capturing social interactions, which are significant for accurate trajectory prediction. The online attentive hidden state refinement strategy used in AC-VRNN makes it advantageous over STGAT and Social-BiGAT that also use graph attention networks. Compared with the state-of-the-art methods, our method achieves the best performance in average ADE and FDE values. Moreover, the hard attention module performs better than the soft attention module.

B. Ablation Study

We conduct an ablation study to demonstrate the effects of various modules used in TPPO. Concretely, SGAN is used as the baseline method because the proposed TPPO is an improved version of SGAN by introducing the attention mechanism and the latent variable predictor. Soft and Hard represent soft and hard attention modules, respectively. The MLVP module represents the latent variable predictor with multiple inputs. For comparison, we propose the SLVP module, which uses the velocity only to estimate the latent variable. As presented in Table II, the SLVP module does not improve the prediction performance of SGAN. However, the MLVP module drastically decreases the ADE and FDE errors in most sets compared with SGAN. Such findings confirm the
effectiveness of multiple inputs in latent variable prediction. Furthermore, the improved results verify that the proposed latent variable predictor could accurately predict pedestrians’ future trajectories by leveraging their predicted intentions. The latent variable predictor is our significant contribution compared with other trajectory prediction approaches. Besides, we embed hard and soft attention modules into SGAN to evaluate the social attention pooling layer’s effectiveness. For SGAN, the soft attention module is slightly better than the hard attention module, whereas an opposite conclusion is drawn when both attention modules are used in TPPO. The possible reason for such slight nonconformity is the differences in network structures. Specifically, TPPO makes a tradeoff between the learning of the soft attention module and latent variable predictor, whereas SGAN only focuses on the soft attention module during training.

C. Evaluations of Different Sampling Times

Despite the necessity of generating multimodal outputs, a good trajectory predictor should estimate precise future trajectories with few shots. One advantage of TPPO is its ability to forecast precise trajectories with few shots. We perform a comparison between SGAN and TPPO with the MLVP module only by gradually decreasing sampling times. Fig. 5 illustrates the comparison results of ADE and FDE between TPPO and SGAN when using different sampling times across five sets. Except for the ADE and FDE for $T = 12$ in the HOTEL set, TPPO significantly outperforms SGAN, especially with few sampling times. We can observe that the ADE and FDE values of TPPO gently increase while sampling times are decreasing. However, those of SGAN drastically increase when the number of samples is close to 1. The difference indicates the power of TPPO in predicting precise future trajectories with few shots.

D. Qualitative Evaluations

TPPO performs precise trajectory prediction based on the latent variable predictor, which estimates a knowledge-rich latent variable from the trajectory data only. Fig. 6 demonstrates the trajectory prediction results by using SGAN, SR-LSTM, Sophie, and TPPO with different modules in different datasets. Specifically, we show the prediction results of TPPO with the MLVP module only, with MLVP and soft attention modules, and with MLVP and hard attention modules. Each subfigure represents a scenario with multiple pedestrians. For each pedestrian, the predicted trajectory is the best one with the lowest-ADE value among the 20 samples generated by each method. Generally, all methods can predict future trajectories with high accuracy at most of the time. However, TPPO performs better than the selected state-of-the-art methods with predicted future trajectories closer to the ground truth. Moreover, as illustrated in the fourth scenario of Fig. 6(c), TPPO can handle a sudden motion change, which is a challenging issue in trajectory prediction. Therefore, the information learned from positions, velocities, and accelerations is useful for accurate trajectory prediction in most cases. However, such information may mislead TPPO to generate wrong results in rare cases, as shown in the third scenario of Fig. 6(b). We will investigate the robustness of the latent variable predictor in our future work.
V. CONCLUSION AND DISCUSSION

In this work, we propose TPPO, which forecasts future trajectories with two pseudo oracles. One pseudo oracle is pedestrians’ moving directions, which are used to approximate pedestrians’ head orientations. A social attention pooling module utilizes this pseudo oracle to improve trajectory prediction performance. Another pseudo oracle is the latent variable distribution estimated from ground truth trajectories. We propose a novel latent variable predictor, which estimates the latent variable distribution from observed trajectories. Such a distribution is similar to that from ground truth. The random Gaussian noise is injected into the estimated latent variable to handle future uncertainties. Evaluations are performed in two commonly used metrics, namely, ADE and FDE, across two benchmarking datasets. Comparisons with state-of-the-art approaches indicate the effectiveness of the proposed latent variable predictor. Ablation studies reveal the necessity of learning information from multiple inputs and the superiority of accurate trajectory prediction with few sampling times. Besides, the proposed method only learns knowledge from trajectories and thus increases little computing overhead. Our future work focuses on how to control the latent variable for an improved prediction performance and forecast pedestrians’ future trajectories from the first perspective. Moreover, the proposed method will be deployed on a self-navigating robot, which can make more informed decisions on selecting an optimal path after sensing agents’ future trajectories.
Fig. 7. Density maps of the predicted trajectories by using SGAN and TPPO with different modules, including (a) SGAN, (b) + Soft, (c) + Hard, (d) + MLVP, (e) + MLVP + Soft, and (f) + MLVP + Hard. The first, second, third, and fourth rows are typical scenarios randomly selected from the ETH, HOTEL, ZARA1, and ZARA2 sets, respectively. The density maps are generated by sampling 300 times from the learned generators. The red stars represent the true future trajectories, and different colors indicate the density distributions of different pedestrians (best viewed in color and zoom-in).

REFERENCES

[1] B. Yang, F. Fan, R. Ni, H. Wang, A. Jafaripournimchahi, and H. Hu, “A multi-task learning network with a collision-aware graph transformer for traffic-agents trajectory prediction,” IEEE Trans. Intell. Transp. Syst., early access, Jan. 2, 2024, doi: 10.1109/TITS.2023.3345296.

[2] B. Yang, G. Yan, P. Wang, C. Chan, X. Song, and Y. Chen, “A novel graph-based trajectory predictor with pseudo-oracle,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 7064–7078, Dec. 2022.

[3] B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, “Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 6, pp. 5338–5349, Jun. 2021.

[4] B. Yang, Z. Wei, H. Hu, R. Wang, C. Yang, and R. Ni, “DP-CIAN: A novel dual-channel pedestrian crossing intention anticipation network,” IEEE Trans. Intell. Transp. Syst., early access, Nov. 23, 2023, doi: 10.1109/TITS.2023.3333328.

[5] J. Li, W. Zhan, and M. Tomizuka, “Generic vehicle tracking framework capable of handling occlusions based on modified mixture particle filter,” in Proc. IEEE Intell. Veh. Symp. (IV), 2018, pp. 936–942.

[6] D. Kasper et al., “Object-oriented Bayesian networks for detection of lane change maneuvers,” IEEE Trans. Intell. Transp. Syst. Mag., vol. 4, no. 3, pp. 19–31, Aug. 2012.

[7] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 961–971.

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, “Social GAN: Socially acceptable trajectories with generative adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2255–2264.

[9] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, M. Cristani, and F. Galasso, “‘seeing is believing’: Pedestrian trajectory forecasting using visual frustums of attention,” in Proc. Winter Conf. Appl. Comput. Vis. (WACV), 2018, pp. 1178–1185.

[10] S. Pellegrini, A. Ess, and L. Van Gool, “Improving data association by joint modeling of pedestrian trajectories and groupings,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 452–465.

[11] L. Leal-Taixé, M. Fzeni, A. Kaznetsova, B. Rosenhahn, and S. Savarese, “Learning an image-based motion context for multiple people tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 3542–3549.

[12] S. Zernetsch, S. Kohnen, M. Goldhammer, K. Doll, and B. Sick, “Trajectory prediction of cyclists using a physical model and an artificial neural network,” in Proc. IEEE Intell. Veh. Symp. (IV), 2016, pp. 833–838.

[13] W. Zhi, R. Senanayake, L. Ott, and F. Ramos, “Spatiotemporal learning of directional uncertainty in urban environments with kernel recurrent mixture density networks,” IEEE Robot. Autom. Lett., vol. 4, no. 4, pp. 4306–4313, Oct. 2019.

[14] S. Molina, G. Cielniak, T. Krajnik, and T. Duckett, “Modelling and predicting rhythmic flow patterns in dynamic environments,” in Proc. Annu. Conf. Towards Auton. Robot. Syst., 2018, pp. 135–146.

[15] C. Laugier et al., “Probabilistic analysis of dynamic scenes and collision risks assessment to improve driving safety,” IEEE Intell. Transp. Syst. Mag., vol. 3, no. 4, pp. 4–19, Oct. 2011.

[16] G. Aoude, J. Joseph, N. Roy, and J. How, “Mobile agent trajectory prediction using Bayesian nonparametric reachability trees,” in Proc. Infotech@Aerospace, 2011, pp. 29–31.

[17] W. Wang, J. Xi, and D. Zhao, “Learning and inferring a driver’s braking action in car-following scenarios,” IEEE Trans. Veh. Technol., vol. 67, no. 5, pp. 3887–3899, May 2018.

[18] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, “SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 12085–12094.

[19] H. Xue, D. Q. Huynh, and M. Reynolds, “SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2018, pp. 1186–1194.

[20] A. Syed and B. T. Morris, “SSEG-LSTM: Semantic scene segmentation for trajectory prediction,” in Proc. IEEE Intell. Veh. Symp. (IV), 2019, pp. 2504–2509.

[21] D. Ridel, N. Deo, D. Wolf, and M. Trivedi, “Scene compliant trajectory forecast with agent-centric spatio-temporal grids,” IEEE Robot. Autom. Lett., vol. 5, no. 2, pp. 2816–2823, Apr. 2020.
Caizhen He was born in Yangzhou, China, in 1996. He received the B.S. degree in engineering from the Changshu Institute of Technology, Suzhou, China, in 2018, where he is currently pursuing the master’s degree in computer science.

Pin Wang (Member, IEEE) received the Ph.D. degree in transportation engineering from Tongji University, Shanghai, China, in 2016.

She was a Postdoctoral Fellow with California PATH, University of California, Richmond, CA, USA, from 2016 to 2018, where she is currently a Researcher and a Team Leader. Her current research is focused on deep learning algorithms and applications for autonomous driving, including driving behavior learning, trajectory planning, control policy learning, and pedestrian intention prediction. She also collaborates with industries on projects, such as intelligent traffic control and advanced vehicular technology assessment.

Xiaofeng Liu (Senior Member, IEEE) received the B.S. degree in electronics engineering and the M.S. degree in computer science from the Taiyuan University of Technology, Taiyuan, China, in 1997 and 1999, respectively, and the Ph.D. degree in biomedical engineering from Xi’an Jiaotong University, Xi’an, China, in 2006.

From 2013 to 2014, he was a Visiting Scholar with the University College London, London, U.K. Since 2011 he has been a Full Professor with Hohai University, Nanjing, China, where he is also the Leader of the Cognition and Robotics Laboratory and the Vice Director of the Jiangsu Key Laboratory of Special Robotic Technologies. He has over 16 grants as a PI and over 12 grants as a Researcher, including the National High-Tech Research and Development Program (863) and the National Basic Research Program (973). He has been granted 15 patents and authored 20 accredited journal papers. His current research interests focus on human robot interactions, social robotics, bioinspired navigation, and neural engineering.

Ching-Yao Chan (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of California, Richmond, CA, USA, in 1988.

He has three decades of research experience in a broad range of automotive and transportation systems, and currently is the Program Leader with California PATH and the Co-Director of Berkeley DeepDrive Consortium (bdd.berkeley.edu). He is currently a Leading Research in several topics, including driving behavior learning, pedestrian-vehicle interaction, sensor fusion for driving policy adaptation, and supervisory control in automated driving systems, human factors in automated driving.

Yang Chen (Member, IEEE) received the B.S. degree in electronic information engineering, the M.S. degree in signal and information processing, and the Ph.D. degree in underwater acoustic engineering from Harbin Engineering University, Harbin, China, in 2005, 2008, and 2010, respectively.

He is currently an Associate Professor with the School of Information Science and Engineering, Changzhou University, Changzhou, China. His research interests include underwater array signal processing and acoustic signal processing.