Tra2Tra: Trajectory-to-Trajectory Prediction With a Global Social Spatial-Temporal Attentive Neural Network

Yi Xu, Dongchun Ren, Mingxia Li, Yuehai Chen, Mingyu Fan, and Huaxia Xia

***Abstract*—Accurate trajectory prediction plays a key role in robot navigation. It is beneficial for planning a collision-free and appropriate path for the autonomous robots, especially in crowded scenes. However, it is a particularly challenging task because there are complex and subtle interactions among pedestrians. There have been many studies focusing on how to model this spatial interactions but most of them neglected the temporal characteristic. Towards this end, we propose a novel Global Social Spatial-Temporal Attentive Neural Network for trajectory-to-trajectory prediction (Tra2Tra). In this model, we first extract features of spatial interactions with decentralization operation and attention mechanism, and then iteratively extract its temporal dependency through the Long Short-Term Memory network for obtaining the global spatial-temporal feature representation. We further aggregate this global spatial-temporal feature representation and velocity features into our encoder-decoder module for prediction. In order to make multimodality predictions, we introduce a random noise perturbation while decoding, which enhances the robustness and the generalization ability of our model. Experimental results demonstrate that our Tra2Tra model can achieve better performance than the state-of-the-art methods not only on two pedestrian-walking datasets, i.e. ETH and UCY, but also on three other complex trajectory datasets, i.e. Collisions, NGsim and Charges.**

***Index Terms*—Deep learning methods, human and humanoid motion analysis and synthesis, intention recognition, social HRI.**

I. Introduction

ROBOTICS technology becomes more and more common in our daily life, they have been found useful in applications such as households, offices and express delivery [1]–[3]. Almost all such robots have to be able to navigate in various and complex environments with moving pedestrians. It requires the robot to move from one place to another while interacting with surrounding pedestrians to avoid collision and increase efficiency. To achieve this, the robot needs to make accurate trajectory predictions of surrounding pedestrians and change its own path in real time.

Early works in robot navigation [4]–[6] focus on modeling individual pedestrian motion to make future trajectory predictions. However, such prediction models neglect the social interactions among pedestrians, which causes the unreliable prediction results for robot navigation [7]. Therefore, considering human-human interactions and make reliable future trajectory predictions are essential in robot navigation. It is a challenging problem because this kind of interactions is mainly driven by common sense and implicit social rules, which is difficult to model or quantify.

Recently, there have been many deep learning-based methods proposed for addressing this problem. In particular, Social-LSTM [8] method is proposed based on Recurrent Neural Networks (RNNs) architecture. A social pooling layer is designed in Social-LSTM for passing interaction information among pedestrian. Some researches follow this pattern [9]–[11] with different mechanisms to measure the interactions among pedestrians. Some papers applies Generative Adversarial Networks (GANs) in this task [12], [13] to generates multiple feasible trajectories.

In these methods, spatial features of interactions among pedestrians are extracted and calculated in pair-wise. We argue that this pair-wise calculation way is inadequate. Because it is at a local level, the collective influence of interactions is neglected. In addition to this limitation, temporal features of interactions are also neglected. Most methods [8], [9], [13] extract spatial feature of interactions at each time step while neglecting the temporal relevance among these spatial features from consecutive time steps.

Except for the features of social interactions, hidden information of the observed trajectories is also important for future trajectory prediction. In previous works [8], [10], only previous location information of pedestrians are aggregated for prediction. Model CF-LSTM [14] has considered the location information from previous two time steps. Model StarNet [15] applied a max-pooling layer for feature selection. However, these feature aggregation approaches are inefficient and may lose key information, especially in crowed scenes.

In this letter, we propose a Global Social Spatial-Temporal Attentive Neural Network for Trajectory-to-Trajectory Prediction (Tra2Tra) to overcome the above limitations. It includes two

Manuscript received October 14, 2020; accepted January 28, 2021. Date of publication February 4, 2021; date of current version March 2, 2021. This letter was recommended for publication by Associate Editor L. Natale and Editor M. Vincze upon evaluation of the reviewers' comments. This work was supported in part by the Beijing Nova Program under Grant Z201100006820046, in part by the National Natural Science Foundation of China under Grant 61772373, and in part by the Beijing Science, and Technology Project under Grant Z181100008918018. *(Yi Xu and Dongchun Ren are co-first authors.) (Corresponding author: Mingyu Fan.)*

Yi Xu, Dongchun Ren, and Huaxia Xia are with the Meituan, Beijing 100102, China (e-mail: xuyi14@meituan.com; rendongchun@meituan.com; xiahuaxia@meituan.com).

Mingxia Li and Yuehai Chen are with the Xi'an Jiaotong University, Xi'an 710049, China (e-mail: alicelmx@stu.xjtu.edu.cn; cyh0518@stu.xjtu.edu.cn).

Mingyu Fan is with the Wenzhou University, Wenzhou 325035, China (e-mail: fanmingyu@amss.ac.cn).

Digital Object Identifier 10.1109/LRA.2021.3057326

2377-3766 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.

main modules, spatial-temporal attention module and encoder-decoder module. The spatial-temporal attention module is designed to extract the global spatial-temporal feature representation of social interactions. At first, we extract spatial features by considering all the pedestrians at one time step instead of pairwise calculation. Then, we apply the attention mechanism [16] to measure the relative importance among these features. To capture the temporal features, we use Long Short-Term Memory (LSTM) Network [17] to extract temporal dependency of spatial features, which enables a compact spatial-temporal feature representation.

Once we obtain the spatial-temporal feature representation, we design a novel feature aggregation layer to aggregate this spatial-temporal feature representation with useful local features of observed trajectories. Then, the aggregated feature is defined as the input of the encoder-decoder module. From the perspective of a robot, since there are multiple potential trajectories of pedestrians in the scene, it is necessary and beneficial to make multi-modality trajectory prediction. Therefore, we introduce a random noise perturbation while decoding to make multi-modality predictions.

Contributions are summarized as follows:

- A novel Tra2Tra model for pedestrian trajectory prediction, which can capture the compact spatial-temporal feature representation of social interactions at a global level is proposed.
- An efficient feature aggregation layer to aggregate global spatial-temporal feature representation of interactions with detailed local features of trajectories is included.
- State-of-the-art performance on five public datasets. Improve 7.7%/11.1% on ADE/FDE metrics for mono-modality prediction, and 28.6%/1.8% for multi-modality prediction on walking-pedestrian dataset ETH and UCY. Improve 4.7% on average RMSE on dataset Collisions, NGsim, and Charges.

The rest of this letter is structured as follows. In Section II, some related works are reviewed and discussed. Our proposed Tra2Tra model is detailed introduced in Section III. Experimental results on public datasets and detailed analysis are presented in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

A. Social Interactions in Trajectory Prediction

The most classic method is the social force model [18], it defined different interactions as different virtual forces upon the pedestrians. However, this definition cannot handle sophisticated scenes with a large number of pedestrians.

Some other early methods [19]–[21] constructed a two-dimensional grid graph of the scene and assigned costs of edges, which converts the social interaction modeling problem into an optimization problem. One major limitation of this kind of methods is that the performance is greatly influenced by how the cost is defined.

B. Deep Learning-Based Trajectory Prediction

With rapid development of neural networks, a large number of models based on deep learning are proposed. Social-LSTM [8] is one of the earliest models based on Long Short-term Memory Network for pedestrian trajectory prediction. In Social-LSTM model, a pooling layer is designed to connect each separate LSTM network representing a pedestrian. Later works such as Peeking into [9], SR-LSTM [10] followed this pattern, connecting each separate network with pooling layer for information delivery.

Based on the assumption that each pedestrian has multiple feasible trajectories, Social-GAN [13] was proposed using GANs [22] for multi-modality trajectory prediction. Sophie [23], another generative model which is similar with Social-GAN, used CNNs for scene feature extraction and designed two social and physical attention mechanisms for feature embedding. Similar with the architecture of Sophie, the CGNS [24] model used Gated Recurrent Units (GRUs) [25] instead of LSTMs. TF-based model [26] designed only-attention-based memory mechanisms for trajectory prediction.

C. Graph-Based Trajectory Prediction

Graph Neural Network (GNN) and its variants [27], [28] are powerful methods for data representation in the non-Euclidean space. Early work Social-Attention [29] utilized the S-RNN [30] architecture. Recent work Social-BiGAT [31] relies on graph attention networks [16] to model the social interactions. Social-STGCNN [32] also used GNNs for interaction modeling but directly from the beginning. Recur-Social [33] was proposed with a new insight of group-based social interaction modeling. Model Trajectron [34] predicts potential future trajectories of multiple agents simultaneously based on graph structure. In addition, VectorNet [35] presented a hierarchical GNN for vehicles behavior prediction.

III. Methodology

A. Problem Definition

Denote $o_i^t = (x_i^t, y_i^t) \in \mathbb{R}^2$ the coordinates of pedestrian i at time step t, $\Gamma_i^{obs} = \{o_i^1, \ldots, o_i^{obs}\}$ the single observed trajectory of pedestrian i from time step T_1 to T_{obs}. The trajectory prediction problem is defined as given all the observed trajectories of total N pedestrians $\Upsilon^{obs} = \{\Gamma_1^{obs}, \Gamma_2^{obs}, \ldots, \Gamma_N^{obs}\}$ in the scene, to jointly predict the future trajectories $\Upsilon^{pred} = \{\Gamma_1^{pred}, \Gamma_2^{pred}, \ldots, \Gamma_N^{pred}\}$ of all the pedestrians from time step T_{obs+1} to T_{pred}, where $\Gamma_i^{pred} = \{o_i^{obs+1}, \ldots, o_i^{pred}\}$ represents a predicted future trajectory of pedestrian i.

Formally, we need to learn the parameters W^* of a model $f(\cdot)$ to predict future trajectories of all pedestrians as follows:

$$\Upsilon^{pred} = f(\Upsilon^{obs}; W^*) \tag{1}$$

where W^* is the collection of all parameters used in the model.

B. Tra2Tra

The structure of our proposed Tra2Tra model is illustrated in Fig. 1, which consists of two main modules: the spatial-temporal attention module and the encoder-decoder module. First, the spatial-temporal attention module is designed to extract the compact spatial-temporal feature representation of complex social interactions in the scene. Then, this feature representation

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.

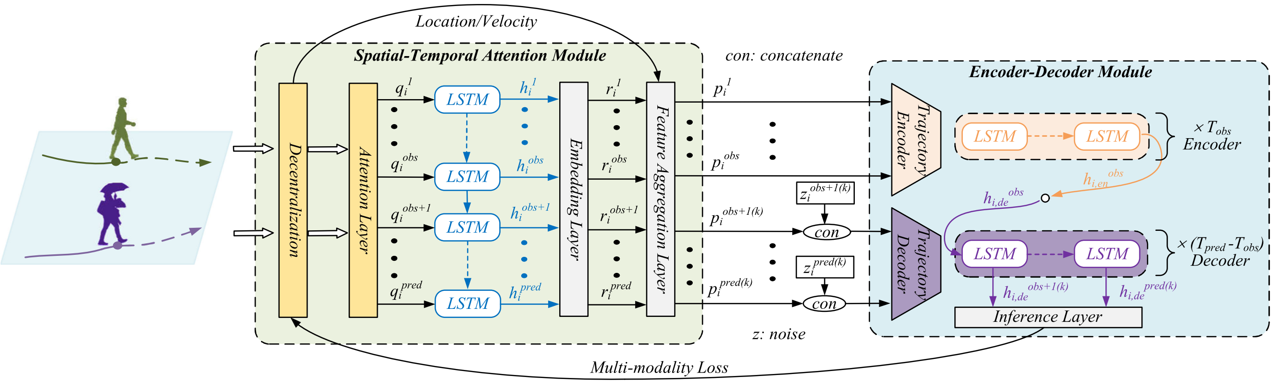


Fig. 1. The Tra2Tra model, including two main modules: The spatial-temporal attention module and the encoder-decoder module.

is aggregated with local trajectory features as the input of the encoder-decoder module. As argued above, we concatenate a random noise perturbation Z, sampled from a standard Gaussian distribution, to the input of the decoder for multi-modality predictions

1) Spatial-Temporal Attention Module: We start by passing the coordinates of all pedestrians through the decentralization layer, which is defined in (2).

$$o'^t_i = o^t_i - \frac{1}{N}\sum_{i=1}^{N} o_i^{obs} \tag{2}$$

Then, the decentralized coordinates o'^t_i are embedded into a vector $\bar{\boldsymbol{o}}_i^t \in \mathbb{R}^D$ as follows.

$$\bar{\boldsymbol{o}}_i^t = \phi\left(o'^t_i; W_o\right) \tag{3}$$

where ϕ is the ReLU non-linearity, $W_o \in \mathbb{R}^{2\times D}$ are embedding parameters and D is the dimension of feature space. The vector $\bar{\boldsymbol{o}}_i^t$ represents the spatial features of pedestrian i.

With the spatial feature representations all the pedestrians, we introduce an attention layer to calculate the relative importance of these spatial features and integrate them into a global spatial feature representation as follows.

We first compute the attention coefficients α_{ij}^t which measures the relative importance of different spatial features in (4).

$$\alpha_{ij}^t = \frac{\exp\left(\sigma\left(\overrightarrow{\mathbf{a}}^\top\left[\bar{\boldsymbol{o}}_i^t \oplus \bar{\boldsymbol{o}}_j^t\right]\right)\right)}{\sum_{j=1}^{N}\exp\left(\sigma\left(\overrightarrow{\mathbf{a}}^\top\left[\bar{\boldsymbol{o}}_i^t \oplus \bar{\boldsymbol{o}}_j^t\right]\right)\right)} \tag{4}$$

where $\overrightarrow{\mathbf{a}}^\top \in \mathbb{R}^{2D}$ is the weight of a single-layer feed-forward neural network, $\cdot^\top$ represents transposition and $\oplus$ is the concatenation operation, σ is the $\mathrm{LeakyReLU}$ non-linearity with negative input slope $\theta = 0.2$.

Once obtained, the normalized attention coefficients are used to compute the linear combination as the global spatial feature representation $\boldsymbol{q}_i^t$ in (5).

$$\boldsymbol{q}_i^t = \phi\left(\sum_{j=1}^{N}\alpha_{ij}^t \bar{\boldsymbol{o}}_j^t\right) \tag{5}$$

However, interactions change dynamically over time. Hence, we use a LSTM network to extract the temporal dependency between the spatial features, defined in (6).

$$\begin{cases} h_i^t = LSTM\left(h_i^{t-1}, \boldsymbol{q}_i^t; W_l\right) \\ \boldsymbol{r}_i^t = \phi\left(h_i^t; W_r\right) \\ t = 1, \ldots, T_{pred} \end{cases} \tag{6}$$

where $h_i^t \in \mathbb{R}^{d_1}$ is the hidden state of LSTM network (d_1 is the dimension of this LSTM network), $W_l \in \mathbb{R}^{D\times d_1}$ are the LSTM network parameters, $W_r \in \mathbb{R}^{d_1\times D}$ are the parameters of the embedding layer.

Note that $\boldsymbol{r}_i^t \in \mathbb{R}^D$ contains the information of not only the global spatial features but also the temporal features. Besides, this feature representation dynamically changes over time, which can describe the social interactions efficiently.

2) Feature Aggregation Layer: The global spatial-temporal feature representation $\boldsymbol{r}_i^t$ is defined as one input of our designed feature aggregation layer. Meanwhile, the relative location $\bar{o}_i^t \in \mathbb{R}^2$ and velocity $v_i^t \in \mathbb{R}^2$ of pedestrians are also defined as inputs. We define an integrated feature representation $\boldsymbol{p}_i^t$ as follows.

$$\boldsymbol{p}_i^t = \left[\boldsymbol{r}_i^t \odot \phi\left(\bar{o}_i^t; W_c\right)\right] \oplus v_i^t \oplus \|v_i^t\| \in \mathbb{R}^{D+3} \tag{7}$$

where $W_c \in \mathbb{R}^{2\times D}$ are embedding parameters, $\odot$ represents the element-wise multiplication.

As can be seen that $\boldsymbol{p}_i^t$ is aggregated from different terms. Interestingly, it tends to aid our model to learn detailed representation of trajectory features and achieves good performance in practice.

Similar aggregation procedure was employed in [15] for trajectory prediction and in [36] for point cloud semantic segmentation, which achieved promising performance. This aggregation way enables the model to obtain a compact feature representation without losing local information.

3) Encoder-Decoder Module: The function of the encoder-decoder module is to encode feature representation of observed trajectories into an implicit variable and decode future trajectories step by step from this implicit variable. In this case, due to the strong temporal correlation, we use LSTM networks both in the encoder and the decoder as feature extractor. The encoder is

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.

defined in (8).

$$\begin{cases} h_{i,en}^{t} = LSTM\left(h_{i,en}^{t-1}, \boldsymbol{p}_i^t; W_{en}\right) \\ t = 1, \ldots, T_{obs} \end{cases} \tag{8}$$

where $h_{i,en}^t \in \mathbb{R}^{d_{en}}$ is the hidden state (d_{en} is the dimension of this LSTM network), $W_{en} \in \mathbb{R}^{(D+3)\times d_{en}}$ are the LSTM network parameters.

Different from the encoder, we concatenate a random noise $\boldsymbol{z}_i^{t(k)}$, sampled from a standard Gaussian distribution, to the input of the decoder for multiple trajectory prediction at each time step. Assume we make total K predictions ($k = 1, 2, \ldots, K$), the decoder is defined in (9).

$$\begin{cases} h_{i,de}^{t(k)} = LSTM\left(h_{i,de}^{t-1(k)}, \left[\boldsymbol{p}_i^{t-1(k)} \oplus \boldsymbol{z}_i^{t-1(k)}\right]; W_{de}\right) \\ v_i^{t(k)} = \phi(h_{i,de}^{t(k)}; W_v) \\ h_{i,de}^{T_{obs}} = h_{i,en}^{T_{obs}} \\ t = T_{obs+1}, \ldots, T_{pred} \end{cases} \tag{9}$$

where $h_{i,de}^t \in \mathbb{R}^{d_{de}}$ is the hidden state (d_{de} is the dimension of this LSTM network), $W_{de} \in \mathbb{R}^{(D+3+z)\times d_{de}}$ are the LSTM network parameters (z is the dimension of the noise), $W_v \in \mathbb{R}^{d_{de}\times 2}$ is the weight matrix of the Inference Layer.

4) Multi-Modality Loss: Once we obtained the velocity predictions $v_i^{t(k)}$, we add it to the predicted coordinates $\hat{o}_i^{t-1}$ of last time step to obtain the predictions $\hat{o}_i^{t(k)}$.

The whole model is trained by minimizing the multi-modality loss L as follows. Different from normal L_2 loss, we introduce an exponential term related to time steps which increases over time.

$$L = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=T_{obs+1}}^{T_{pred}} \min_{k\in\{1,2,\ldots,K\}} \|\hat{o}_i^{t(k)} - \bar{o}_i^t\|^2 \cdot e^{\frac{t}{\lambda}} \tag{10}$$

where $T = T_{pred} - Tobs$ is the prediction length, $\bar{o}_i^t$ is the corresponding ground-truth and λ is a hyper parameter to control the value of the exponential term.

IV. Experiments

In this section, we demonstrate the experimental results on two public pedestrian-walking datasets: ETH [41] and UCY [42]. ETH contains two datasets named ETH and HOTEL, UCY contains three datasets named ZARA01, ZARA02, and UNIV. Trajectories are sampled every 0.4 seconds and we observe 8 frames (3.2 seconds) and predict the next 12 frames (4.8 seconds). In addition to the above two datasets, we also demonstrate the experimental results on three other more complex trajectory datasets, Collisions [43], NGsim [44], and Charges [45], illustrated in Section IV-F.

A. Evaluation Metrics

Two following metrics are used to evaluate the performance on ETH and UCY. Assume N is the total number of pedestrians, $\hat{o}_i^t = (\hat{x}_i^t, \hat{y}_i^t)$ represents the predicted coordinates of pedestrians and $\bar{o}_i^t$ represents the corresponding ground-truth.

TABLE I
PERFORMANCE STUDY OF λ. THE LOWER THE BETTER AND THE BEST IS IN BOLD

	Performance (ADE/FDE)					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
$\lambda = 5$	0.65/1.19	0.59/1.12	0.59/1.27	0.42/**0.88**	0.35/0.76	0.52/1.05
$\lambda = 10$	0.64/1.12	0.55/1.14	0.58/1.26	**0.41**/0.89	0.35/0.76	0.51/1.03
$\lambda = 20$	**0.56/1.06**	**0.52/0.91**	0.57/**1.20**	**0.41**/0.90	**0.33/0.74**	**0.48/0.96**
$\lambda = 30$	0.59/1.07	0.59/1.18	0.57/1.21	**0.41**/0.89	0.34/0.75	0.50/1.02
$\lambda = -$	0.61/1.14	**0.52**/1.11	**0.56**/1.21	**0.41**/0.90	0.34/0.75	0.49/1.02

- Average Displacement Error (ADE):

$$ADE = \frac{\sum_{i=1}^{N}\sum_{t=T_{obs+1}}^{T_{pred}} \|\hat{o}_i^t - \bar{o}_i^t\|_2}{N\left(T_{pred} - T_{obs}\right)} \tag{11}$$

- Final Displacement Error (FDE):

$$FDE = \frac{\sum_{i=1}^{N} \|\hat{o}_i^{T_{pred}} - \bar{o}_i^{T_{pred}}\|_2}{N} \tag{12}$$

B. Implementation Details and Training Setup

During training, we use the leave-one-out setting where we train and validate our model on four datasets and test on the remaining one. We set batch size to 64 and train the model for 200 epochs using Adam as optimizer. The initial learning rate is 0.0001 and change to 0.00 005 after 100 epochs. The dimension of all three LSTM networks is 64.

C. Ablation Study

In this section, we first analyze the relations between the hyper parameter λ and the model performance. Then we analyze two important components of our proposed model, including Temporal Dependency Extractor (6) and Feature Aggregation Layer (7). In Feature Aggregation Layer, it contains three terms: $\boldsymbol{r}_i^t \odot \phi(\bar{o}_i^t; W_c)$, v_i^t, and $\|v_i^t\|$. We named them as A, B, and C in Table II respectively.

Specifically, we set $K = 1$ without noise concatenation for comparison. Then the mono-modality loss is defined as follows:

$$L = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=T_{obs+1}}^{T_{pred}} \|\hat{o}_i^t - \bar{o}_i^t\|^2 \cdot e^{\frac{t}{\lambda}} \tag{13}$$

Table I shows the quantitative results of our proposed model with different λ and Table II shows the quantitative results of different model variants with different components.

1) Performance Study of λ*:* The normal L_2 Loss treats each frame equally, which is not applicable in this task. With accumulation error of each frame, the latter the frame, the more important it is. Beside, the final location of pedestrians is crucial in trajectory prediction. Thus, we introduce an exponential term related to time steps to control the proportion of each error in the whole loss. Table I demonstrates the results of our proposed model with different λ ('-' means directly using L_2 Loss). In general, by using $\lambda = 20$, the model can achieve the smallest average ADE and FDE. In particular, by using $\lambda = 20$, the FDE

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.

TABLE II
Performance (ADE/FDE) of Ablation Study. TD, A, B, C Represents Temporal Dependency Extractor, $\boldsymbol{r}_i^t \odot \phi(\bar{o}_i^t; W_c)$, v_i^t, and $\|V_i^t\|$, Respectively. the Lower the Better and the Best Is in Bold

Model ID	Components				Performance (ADE/FDE)					
	TD	A	B	C	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
1	✗	✓	✗	✗	0.97/1.56	1.30/2.24	1.42/2.56	1.31/2.39	1.21/2.19	1.24/2.19
2	✗	✓	✓	✗	0.66/1.15	0.67/1.52	0.61/1.30	0.47/1.02	0.36/0.78	0.55/1.15
3	✓	✓	✗	✗	0.61/1.08	0.78/1.46	0.98/1.81	0.58/1.15	0.51/0.99	0.69/1.30
4	✓	✓	✗	✓	0.63/1.20	0.70/1.20	0.88/1.71	0.48/0.98	0.42/0.83	0.62/1.20
5	✓	✓	✓	✗	0.61/**1.04**	**0.52**/0.92	**0.57**/1.23	**0.41**/0.91	**0.33/0.74**	0.49/0.97
6	✓	✓	✓	✓	**0.56**/1.06	**0.52/0.91**	**0.57/1.20**	**0.41/0.90**	**0.33/0.74**	**0.48/0.96**

metric is improved by 7.0% and 18.0% than using L_2 Loss respectively on relatively more crowded dataset ETH and HOTEL. However, if overemphasizing the latter frames by setting λ too small, it would have counterproductive performance according to the results in Table I. For instance, if $\lambda = 5$, the performance is worst.

2) Temporal Dependency Extractor: Model No.3 which has extracted temporal features of social interactions outperforms Model No.1 by 44.4%/40.6% on ADE/FDE, as the social interactions do have strong temporal features and comprehensive feature representation of social interactions is beneficial for accurate trajectory prediction.

Comparing Model No.2 with Model No.5, these two models both have term A and term B but Model No.5 has extracted temporal features of social interactions. Model No.5 outperforms Model No.3 by 10.9%/15.7% on ADE/FDE, which demonstrates the importance of temporal features of social interactions.

Model No.2 with term A and term B in Feature Aggregation Layer outperforms Model No.1 by 55.6%/47.5% on ADE/FDE. The improvement of this component is greater than that of temporal dependency extractor, one possible explanation is that we predict the future velocity to obtain the future coordinates in 9, without the velocity information as one input of the decoder, the prediction accuracy would be greatly affected.

Comparing from Model No.3 to Model No.6, these four models all have temporal dependency extractor but with different terms in Feature Aggregation Layer. We can see that: (1) Including all three terms leads to the best performance. (2) The term v_i^t plays important role in this component, primarily because the velocity information enables the network to be aware of the velocity (walking direction) pattern, which is also validated in [40] (3) Only aggregating the term $\|v_i^t\|$ (Model No.4) improves less performance than aggregating the term v_i^t (Model No.5), because the $\|v_i^t\|$ is a scalar without any direction information.

D. Quantitative Analysis

We compare our proposed Tra2Tra model with several recent works. For mono-modality trajectory prediction, baselines are: (1) Linear, (2) Vanilla-LSTM, (3) Social-LSTM [8], (4) TPNet [37], (5) TF-based [26], (6) SAGCN [38], (7) STGAT [39]. For multi-modality trajectory prediction, baselines are: (1) Social-GAN [13], (2) Trajectron [34], (3) TPNet [37], (4) Peeking into [9], (5) Social-BiGAT [31], (6) Recur-Social [33], (7) C-Velocity [40], (8) Social-STGCNN [32], (9) TF-based [26], (10) STGAT [39].

The quantitative results are shown in Table III. The number in the parenthesis denotes the number of predicted trajectories, i.e.,modality K. Note that all the methods are under the same dataset setting and evaluation methodology.

Over all, our proposed Tra2Tra model outperforms all previous models both on mono-modality and multi-modality trajectory prediction.

As for mono-modality trajectory prediction, the previous state-of-the-art model on ADE metric is SAGCN with an error of 0.52 and our Tra2Tra model has an error of 0.48 on ADE metric, which improves by 7.7%. The previous state-of-the-art model on FDE metric is TPNet with an error of 1.08 and our Tra2Tra model has an error of 0.96 on FDE metric, which improves by 11.1%. In particular, the improvement on most crowded dataset ETH is significant, improving by 37.8%/45.9% on ADE/FDE metrics.

As for multi-modality trajectory prediction, the previous state-of-the-art model on ADE metric is C-Velocity with an error of 0.28 and our Tra2Tra model has an error of 0.20 on ADE metric, which improves by 28.6%. The previous state-of-the-art model on FDE metric is TF-based with an error of 0.55 and our Tra2Tra model has an error of 0.54, which slightly improves by 1.8%.

Comparing our two different Tra2Tra models for mono-modality and multi-modality trajectory prediction. In general, our Tra2Tra model for multi-modality prediction is better than that for mono-modality prediction with an improvement of 58.3%/43.8%. The qualitative results explains that the introduction of noise can enhance the overall generalization ability of the model. Interestingly, for the FDE metric, errors of dataset ETH are very close (1.06 and 1.02). It will be our future work to figure out the inherent difference of social interactions between crowded and sparse datesets.

E. Qualitative Analysis

Fig. 2 shows some prediction examples of our Tra2Tra model. In general, we can see that our Tra2Tra model could handle social interactions among pedestrians. These predicted trajectories can reflect their movements accurately and have no collision.

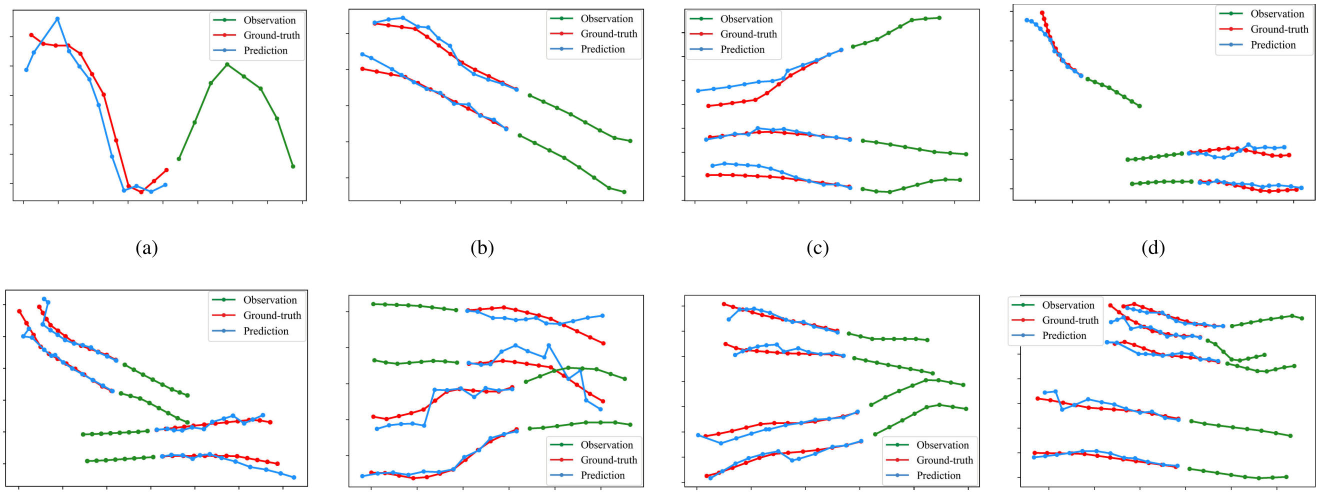
Fig. 2(a) is a simple example of one pedestrian walking, the predicted trajectory is consistent with the ground-truth. Fig. 2(b), 2(c) and 2(d) are examples where pedestrians have

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.

TABLE III

QUANTITATIVE RESULTS (ADE/FDE) OF BASELINES COMPARED TO OUR PROPOSED MODEL. METHODS IN THE ABOVE TABLE ARE MONO-MODALITY TRAJECTORY PREDICTION MODELS ($K = 1$), METHODS IN THE BELOW TABLE ARE MULTI-MODALITY TRAJECTORY PREDICTION MODELS ($K = 100$ FOR MODEL [34], $K = 20$ FOR THE REST). THE LOWER THE BETTER AND THE BEST IS IN BOLD

Method	Performance (ADE/FDE)					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Linear [8]	1.33/2.94	0.39/0.72	0.82/1.59	0.62/1.21	0.79/1.59	0.79/1.59
Vanilla-LSTM [8]	1.09/2.41	0.86/1.91	0.61/1.31	**0.41/0.88**	0.52/1.11	0.70/1.52
Social-LSTM [8]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
TPNet [37]	1.00/2.01	**0.31/0.58**	0.55/1.15	0.46/0.99	0.33/0.72	0.71/1.08
TF-based [26]	1.03/2.10	0.36/0.71	0.53/1.32	0.44/1.00	0.34/0.76	0.54/1.17
SAGCN [38]	0.90/1.96	0.41/0.83	0.57/1.19	**0.41/0.89**	0.32/0.70	0.52/1.11
STGAT [39]	0.88/1.66	0.56/1.15	**0.52/1.13**	**0.41**/0.91	**0.31/0.68**	0.54/1.11
Tra2Tra (Ours)	**0.56/1.06**	0.52/0.91	0.57/1.20	**0.41**/0.90	0.33/0.74	**0.48/0.96**
Social-GAN [13]	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
Trajectron [34]	**0.37/0.72**	0.20/0.35	0.48/0.99	0.32/0.62	0.34/0.66	0.34/0.67
TPNet [37]	0.84/1.73	0.24/0.46	0.42/0.94	0.33/0.75	0.26/0.60	0.42/0.90
Peeking into [9]	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00
Social-BiGAT [31]	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00
Recur-Social [33]	0.80/1.53	0.33/0.64	0.59/1.25	0.40/0.86	0.30/0.65	0.48/0.99
C-Velocity [40]	0.43/0.80	0.19/0.35	0.34/0.71	0.24/0.48	0.21/0.45	0.28/0.56
Social-STGCNN [32]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
TF-based [26]	0.61/1.12	**0.18/0.30**	0.35/0.65	0.22/0.38	0.17/**0.32**	0.31/0.55
STGAT [39]	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
Tra2Tra (Ours)	**0.37**/1.02	**0.18**/0.37	**0.23/0.63**	**0.10/0.31**	**0.12**/0.36	**0.20/0.54**



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Visualization results of some different prediction scenarios. Green line represents the observation of the trajectory, blue line represents the prediction, and red line represents the corresponding ground-truth.

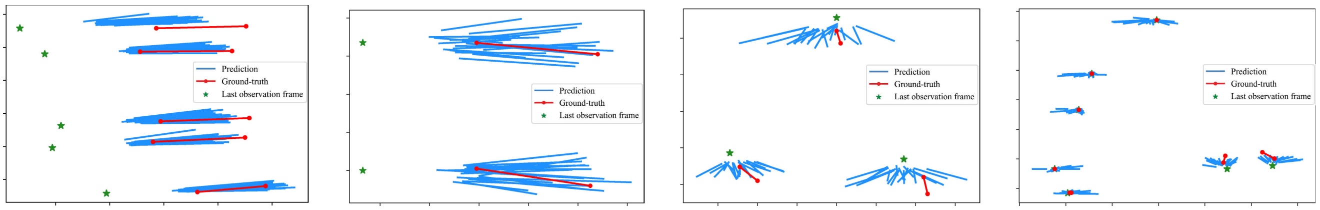
the same direction and the error is very small. Fig. 2(e) and 2(f) are two similar scenarios where pedestrians have the opposite directions. Due to our comprehensive feature representation of social interactions, there is no collision. Fig. 2(g) and 2(h) are two scenarios where pedestrians moving in two small groups and the prediction trajectories are reliable.

Fig. 3 shows some examples of multi-modality trajectory prediction. Fig. 3(a) and 3(b) are two scenarios of pedestrians parallel moving, most of the predicted results are close to the ground-truth and the predicted trend of their movement is accurate. Fig. 3(c) and 3(d) are two complex scenarios because the distance between two time steps is very short (five pedestrians in Fig. 3(d) stand still) which leads to larger error. We will focus on how to predict this kind of anomaly in the future work.

F. Additional Experiments and Analysis

Besides frequently-used ETH and UCY datasets which only contain pedestrian walking scenes, we also demonstrate the experimental results on three other more complex trajectory datasets: Collisions [43], NGsim [44], and Charges [45]. Collision dataset contains 9500 scenes of synthetic physics data with balls moving on a friction-less 2D plane, fixed circular landmarks and boundary walls. NGsim dataset contains 3500

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c) (d)

Fig. 3. Visualization results of multi-modality trajectory prediction scenarios with two time steps. Green star marker represents the last frame of observation, red line represents the ground-truth and blue line represents the prediction. Note that for each red line, there are 20 blue lines, namely 20 trajectory predictions.

TABLE IV
QUANTITATIVE RESULTS (RMSE) OF BASELINES COMPARED TO OUR PROPOSED MODEL. THE NUMBER OF PREDICTED TRAJECTORIES IS 8. THE LOWER THE BETTER AND THE BEST IS IN BOLD

Method	Performance (RMSE)			
	Collisions	NGsim	Charges	AVG
Vanilla-LSTM [8]	0.245	5.972	0.533	2.250
Social-LSMT [8]	0.211	6.453	0.485	2.383
NRI [45]	0.254	7.491	0.557	2.767
Graph-Net [46]	0.234	5.901	0.508	2.214
Graph-SAGE [47]	0.238	5.582	0.522	2.114
Graph-AT [16]	0.237	6.100	0.524	2.287
FQA [43]	0.176	5.071	0.409	1.885
Tra2Tra (Ours)	**0.149**	**4.857**	**0.383**	**1.796**

scenes of freeway traffic data with fast moving vehicles. Charges dataset contains 3600 scenes of positive and negative charges moving under other charges electric fields and colliding with bounding walls.

Similar in [43], we use Root Mean Square Error (RMSE) between ground truth and predictions over all predicted time steps as the evaluation metrics. As for dataset Collisions and Charges, we observe 13 frames and make predictions of next 12 frames, as for dataset NGsim, we observe 8 frames and make predictions of next 12 frames.

Table IV shows quantitative results of our proposed model and other seven baselines: (1) Vanilla-LSTM, (2) Social-LSTM [8], (3) NRI [45], (4) Graph-Net [46], (5) Graph-SAGE [47], (6) Graph-AT [16], (7) FQA [43].

In general, our proposed Tra2Tra model outperforms the previous state-of-the-art FQA model by 4.7% on average RMSE. In specific, our model improves by 15.3%, 4.2%, and 6.4% on dataset Collisions, NGsim, and Charges, respectively. These results validate the effectiveness of our proposed Tra2Tra model on trajectory prediction task.

V. CONCLUSIONS

In this letter, we propose a trajectory-to-trajectory (Tra2Tra) prediction model with a global social spatial-temporal attentive neural network. We introduce a spatial-temporal attention module to extract a compact spatial-temporal feature representation of social interactions, which considers all the pedestrians at a global level. In addition, we present an efficient feature aggregation layer to aggregate the global feature representation by features of the history trajectories, combining as a comprehensive feature representation. Subsequently, this feature representation is defined as the input of the Encoder-Decoder Module for trajectory-to-trajectory prediction. In order to make multi-modality predictions, we introduce a random noise perturbation while decoding, which also enhances the robustness and the generalization ability of our model. We demonstrate that our Tra2Tra model can achieve better performance than the state-of-the-art methods not only on two pedestrian-walking datasets but also on three other complex trajectory datasets.

REFERENCES

[1] A. Binch, G. P. Das, J. P. Fentanes, and M. Hanheide, “Context dependant iterative parameter optimisation for robust robot navigation,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3937–3943.

[2] J. Hooks *et al.*, “Alphred: A multi-modal operations quadruped robot for package delivery applications,” *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 5409–5416, Oct. 2020.

[3] Y. Hada, H. Gakuhari, K. Takase, and E. I. Hemeldan, “Delivery service robot using distributed acquisition, actuators and intelligence,” in Proc. IEEE/*RSJ Int. Conf. Intell. Robots Syst.*, 2004, vol. 3, pp. 2997–3002.

[4] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun, “Learning motion patterns of people for compliant robot motion,” *Int. J. Robot. Res.*, vol. 24, no. 1, pp. 31–48, 2005.

[5] F. Large, D. Vasquez, T. Fraichard, and C. Laugier, “Avoiding cars and pedestrians using velocity obstacles and motion prediction,” in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 375–379.

[6] S. Thompson, T. Horiuchi, and S. Kagami, “A probabilistic model of human motion and navigation intent for mobile robot path planning,” in *Proc. IEEE Int. Conf. Auton. Robots Agents*, 2009, pp. 663–668.

[7] P. Trautman and A. Krause, “Unfreezing the robot: Navigation in dense, interacting crowds,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 797–803.

[8] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 961–971.

[9] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, “Peeking into the future: Predicting future person activities and locations in videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5725–5734.

[10] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, “Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 120 85–12094.

[11] Z. Pei, X. Qi, Y. Zhang, M. Ma, and Y.-H. Yang, “Human trajectory prediction in crowded scene using social-affinity long short-term memory,” *Pattern Recognit.*, vol. 93, pp. 273–282, 2019.

[12] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds,” in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 314–330.

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

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.

[14] Y. Xu, J. Yang, and S. Du, "Cf-lstm: Cascaded feature-based long short-term networks for predicting pedestrian trajectory," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12541–12548.

[15] Y. Zhu, D. Qian, D. Ren, and H. Xia, "Starnet: Pedestrian trajectory prediction using deep neural network in star topology," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 8075–8080.

[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

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

[18] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, 1995, Art. no. 4282.

[19] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3302–3309.

[20] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu, "Learning and inferring "dark matter" and predicting human intents and trajectories in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1639–1652, Jul. 2018.

[21] S. Huang *et al.*, "Deep learning driven visual path prediction from a single image," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5892–5904, Dec. 2016.

[22] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[23] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1349–1358.

[24] J. Li, H. Ma, and M. Tomizuka, "Conditional generative neural system for probabilistic trajectory prediction," *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 6150–6156.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *NIPS 2014 Workshop Deep Learn.*, 2014.

[26] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *Proc. Int. Conf. Pattern Recognitionc*, 2020, *arXiv:2003.08111*.

[27] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[28] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[29] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 1–7.

[30] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5308–5317.

[31] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 137–146.

[32] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A. social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 424–14432.

[33] J. Sun, Q. Jiang, and C. Lu, "Recursive social behavior graph for trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 660–669.

[34] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2375–2384.

[35] J. Gao *et al.*, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11522–11530.

[36] Q. Hu *et al.*, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 108–11117.

[37] L. Fang, Q. Jiang, J. Shi, and B. Zhou, "Tpnet: Trajectory proposal network for motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6797–6806.

[38] Y. Sun, T. He, J. Hu, H. Huang, and B. Chen, "Socially-aware graph convolutional network for human trajectory prediction," in *IEEE Inf. Technol., Netw., Electron. Automat. Control Conf.*, 2019, pp. 325–333.

[39] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6272–6281.

[40] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1696–1703, Apr. 2020.

[41] S. Pellegrini, A. Ess, K. Schindler, and L. J. V. Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2009, pp. 261–268.

[42] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 655–664, 2010.

[43] N. Kamra, H. Zhu, D. Trivedi, M. Zhang, and Y. Liu, "Multi-agent trajectory prediction with fuzzy query attention," *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[44] B. Coifman and L. Li, "A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset," *Transp. Res. Part B: Methodol.*, vol. 105, pp. 362–377, 2017.

[45] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2688–2697.

[46] A. Tacchetti *et al.*, "Relational forward models for multi-agent learning," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJlEojAqFm

[47] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on October 08,2021 at 12:29:16 UTC from IEEE Xplore. Restrictions apply.