

Multi-Vehicle Collaborative Learning for Trajectory Prediction With Spatio-Temporal Tensor Fusion

Yu Wang^{ID}, Shengjie Zhao^{ID}, Senior Member, IEEE, Rongqing Zhang^{ID}, Member, IEEE,
Xiang Cheng^{ID}, Senior Member, IEEE, and Liuqing Yang^{ID}, Fellow, IEEE

Abstract—Accurate behavior prediction of other vehicles in the surroundings is critical for intelligent transportation systems. Common practices to reason about the future trajectory are through their historical paths. However, the impact of traffic context is ignored, which means the beneficial environment information is deserted. Although a few methods are proposed to exploit the surrounding vehicle information, they simply model the influence according to spatial relations without considering the temporal information among them. In this paper, a novel multi-vehicle collaborative learning with spatio-temporal tensor fusion model for vehicle trajectory prediction is proposed, which introduces a novel auto-encoder social convolution mechanism and a fancy recurrent social mechanism to model spatial and temporal information among multiple vehicles, respectively. Furthermore, the generative adversarial network is incorporated into our framework to handle the inherent multi-modal characteristics of the agent motion behavior. Finally, we evaluate the proposed multi-vehicle collaborative learning model on NGSIM US-101 and I-80 benchmark datasets. Experimental results demonstrate that the proposed approach outperforms the state-of-the-art for vehicle trajectory prediction. Additionally, we also present qualitative analyses of the multi-modal vehicle trajectory generation and the impacts of surrounding vehicles on trajectory prediction under various circumstances.

Index Terms—Collaborative learning, spatio-temporal tensor fusion, vehicle trajectory prediction, generative adversarial networks.

Manuscript received November 22, 2019; revised April 23, 2020 and June 26, 2020; accepted July 6, 2020. Date of publication July 28, 2020; date of current version December 28, 2021. This work was supported in part by the National Key Research and Development Project under Grant 2019YFB2102300 and Grant 2019YFB2102301, in part by the National Natural Science Foundation of China under Grant 61936014, in part by the Scientific Research Project of Shanghai Science and Technology Committee under Grant 19511103302, and in part by the Fundamental Research Funds for the Central Universities. The Associate Editor for this article was S. Siri. (Corresponding author: Shengjie Zhao.)

Yu Wang is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China, and also with the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China (e-mail: wang_yu@tongji.edu.cn).

Shengjie Zhao is with the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China, and also with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: shengjiezhao@tongji.edu.cn).

Rongqing Zhang is with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: rongqingz@tongji.edu.cn).

Xiang Cheng is with the State Key Laboratory of Advanced Optical Communication Systems and Networks, Peking University, Beijing 100871, China (e-mail: xiangcheng@pku.edu.cn).

Liuqing Yang is with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80521 USA (e-mail: lqyang@engr.colostate.edu).

Digital Object Identifier 10.1109/TITS.2020.3009762

1558-0016 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See <https://www.ieee.org/publications/rights/index.html> for more information.

I. INTRODUCTION

ACCURATE and efficient trajectory prediction for intelligent vehicles is significant in sophisticated traffic scenarios where various vehicles and human crowds travel towards respective destinations with distinctive moving patterns [1], [2]. An intelligent vehicle is expected to be capable of taking actions proactively when encountering some emergencies, such as slowing down to enable surrounding intelligent vehicles to inlet and speeding up to switch lanes for overtaking. Consequently, intelligent vehicles are required to reason about accurate future trajectories of adjacent vehicles in order to conduct risk assessments of vehicle behaviors and further take appropriate actions.

Trajectory prediction is a fairly intractable problem since intelligent vehicles possess inherent multi-modality features, which means vehicles select actions randomly despite the same circumstance. For instance, it is possible for intelligent vehicles to carry out going straight or switching lanes in the same traffic scene. In addition, some trajectories are physically possible but socially unacceptable, e.g., intelligent vehicles are subjected to social norms like yielding right-of-way or keep a safe distance from others, and thus there exist possible multiple future trajectories that are socially acceptable. Moreover, the interactions among various vehicles in the scene potentially make a difference for the intelligent system to determine driving behaviors, and thus it is significant to take such interactions into consideration for the trajectory prediction.

Despite the ubiquity of these challenges in the field of trajectory prediction, extensive researches in computer vision have been done to address it in previous works. Traditional approaches for trajectory prediction mostly depend on hand-crafted features [3]–[7] or well-designed learning criteria [8], [9] in terms of a specific scene, rather than data-driven paradigms that can capture subtle and complicated interactive information among various heterogeneous agents. Recently, deep learning techniques manifest powerful capabilities in the field of computer vision [10]–[13] and natural language processing [14], [15]. Quantities of approaches based on deep neural networks have also been developed to address trajectory prediction problems [16]–[23] due to such great success.

Although a large number of existing researches have advanced in tackling with the above challenges, there still exist some defects that urgently need to be solved:

- 1) In terms of collaborative learning-based trajectory prediction, a large number of works have been developed using an agent-centric or spatial-centric method, and even with both combined, to model interactions on the basis of spatial relations among multiple agents up to now. However, they did not consider the temporal relations among agents.
- 2) To generate a socially plausible multi-modal future trajectory distribution while avoiding the effects of navigation style, most of the current researches are based on the assumption that future trajectory is subject to multivariate Gaussian distribution. However, this may result in a non-differentiable operation in an end-to-end neural network framework.

Responding to the above challenges, this paper proposes a novel method of multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion. In particular, a novel auto-encoder social convolution mechanism is proposed, which not only models the spatial relations among multiple vehicles but can also take full advantage of extra unlabeled data for model pre-training. Meanwhile, a fancy recurrent social mechanism is also provided to capture multi-vehicle temporal relations. Finally, we further fuse the spatial and temporal tensors derived from the above proposed mechanisms to encode subtle interaction clues among multiple vehicles.

Considering the superior ability of long short-term memory (LSTM) to cope with sequence models, we leverage LSTM encoder-decoder model for vehicle trajectory prediction in the proposed model. Additionally, generative adversarial networks (GANs) have made great progress in capturing intractable high dimensional probabilistic distribution. Inspired by that, we incorporate GANs into our framework to generate a multi-modal probability distribution of the future trajectory. Specially, we regard the encoder-decoder structure as the generator for trajectory generation while the encoder followed by a binary classification layer as the discriminator to distinct between positive and negative trajectory samples. Finally, we test and verify the proposed approach using quantitative and qualitative analysis on NGSIM US-101 and I-80 benchmark datasets. Experimental results demonstrate that our proposed approach outperforms the state-of-the-art methods, which further indicates that our algorithm is capable of predicting socially-acceptable vehicle trajectories.

In this paper, the contributions of our work are summarized as follows:

- 1) A novel multi-vehicle spatio-temporal tensor fusion framework that efficiently fuses vehicle spatio-temporal information in the scene for collaborative learning is proposed.
- 2) A novel auto-encoder social convolution mechanism that not only models the multi-vehicle spatial relations but can also utilize the unlabeled data pre-training model in a semi-supervised fashion is introduced.
- 3) A fancy recurrent social mechanism is introduced to delineate time relations among multiple vehicles. Particularly, to the best of our knowledge, our work is the first

to tackle multiple sequence relations on the temporal dimension.

- 4) Extensive experiments on driving trajectory prediction datasets are conducted using little training information without maneuver classes, and the experimental results indicate that our proposed approach outperforms the state-of-the-art methods comprehensively.

II. RELATED WORK

In this section, a brief review of existing researches on sequence prediction, multi-agent interaction, and multi-modal distribution learning is presented.

A. Sequence Prediction

To predict the vehicle trajectory accurately, approaches adopted in [8], [9] employed a well-designed objective function to predict the future trajectory in view of vehicle relative configurations. Despite that freeing from various scenarios and training sample constraints can make it transferred quickly and efficiently to new traffic scenes, it is still troublesome in terms of how to construct the corresponding objective function nicely.

Considerable progress has been made for sequence prediction models in numerous fields, including speech recognition [24]–[26], machine translation [15], [27], knowledge graph [28], caption generation [29]–[34]. These sequence models employ recurrent neural networks (RNNs) and a series of their derivatives, such as BiRNN [35], LSTM [36], GRU [37], and the attention mechanism [38]. These pioneering works indicate that RNNs are competent for tackling with sequence prediction issues. Motivated by these successful cases of RNNs and their variants in handling sequence prediction, Kuefler *et al.* [21] applied generative adversarial imitation learning to optimize policies of the gated recurrent unit for the generation of vehicle maneuver information. Lee *et al.* [19] predicted future multi-modal trajectories by means of fusing conditional variational auto-encoders with sequence-to-sequence structure, even though the interactions among multi-agent were not taken into consideration. Consequently, a recurrent sequence-to-sequence framework for trajectory prediction is also applied in our model.

B. Multi-Agent Interactions

On the other hand, several researchers have noticed that social interactions among various agents are non-trivial for future trajectory prediction. This issue has been addressed utilizing either agent-centric that fuses multi-agent features via employing particular aggregation function, or spatial-centric that maintains the spatial relations among multiple agents, in a multi-agent collaborative manner. Social LSTM [16] proposed a novel social pooling mechanism, which performs max-pooling operation over feature vectors of adjacent agents whose spatial distances are less than a specifically set threshold. Social GAN [39] combined encoder-decoder structure that takes input as historical trajectories of agents with generative adversarial nets to reason about socially plausible

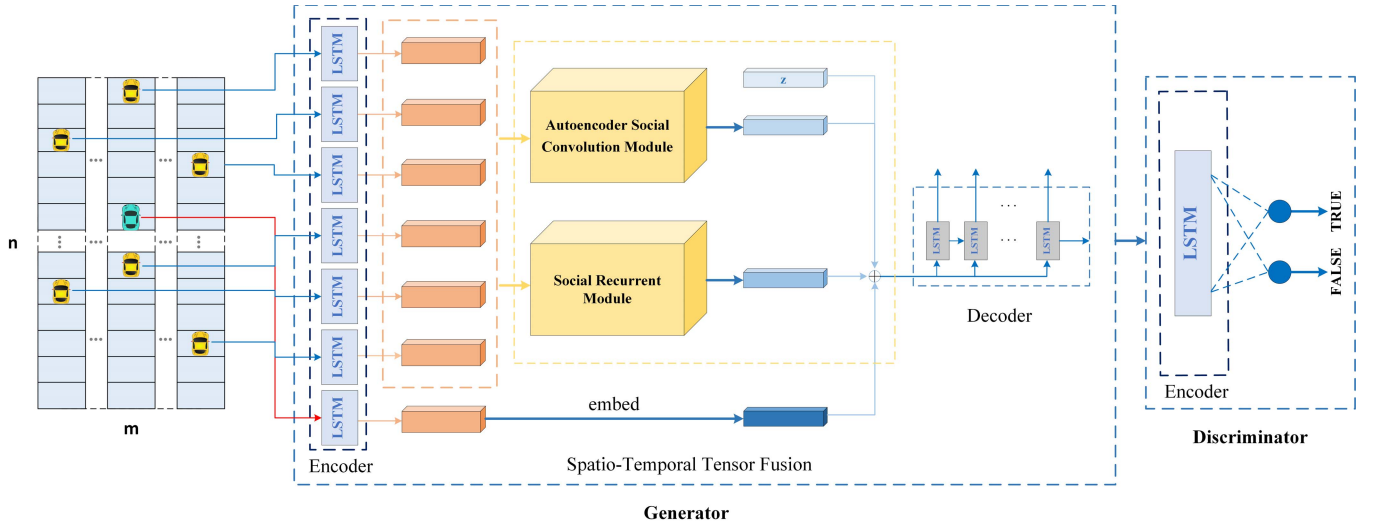


Fig. 1. Illustration of the TS-GAN model. The generative adversarial network is employed in our model for multi-modal distribution generation. The encoder in generator G shares parameters with the one in discriminator D . Two parallel fusion modules, namely auto-encoder social convolution module and recurrent social module, are developed to capture multi-vehicle interactions in spatial and temporal dimensions by manipulating LSTM recurrent encodings represented as orange cubes. Light blue cubes in Generator represent random noises, spatial encodings, temporal encodings, and embedded representations of the predicted vehicle from top to bottom, respectively.

future trajectories. For capturing interactions among multiple agents, a novel aggregation function was also proposed in Social GAN, which employs a multi-layer perceptron network followed by a max-pooling operation. The model was turned out to be more desirable in terms of the performance and overhead of time and space. Convolutional Social Pooling [17] not only modeled partially spatial structure of agents by using convolutional and max-pooling layers to fuse LSTM recurrent encodings of the adjacent vehicles' historical trajectories. It also utilized extra maneuver class information, which seems to be time and computational-consuming. Although these approaches can deal well with an uncertain number of agents in the scenario, they ignored individual uniqueness. As a consequence, the attention mechanism that focuses on the most critical part was introduced into Sophie [20] and Social Attention models [40] to address heterogeneity existing in diverse interactions [15], [38], [41]. ChauffeurNet [18] maintained spatial structure among multiple agents in the manner of operating on feature maps at first hand, which are described by bounding box areas instead of recurrent encodings. MATF [22] manipulated social interactions among agents and constraints from the scene context via convolutional fusion operation that retains the spatial structure of agents.

C. Multi-Modal Distribution Learning

The motions of agents inherently are subject to multi-modal distribution. Social LSTM [16] and Convolutional Social pooling [17] assumed that every timestamp of the future trajectory is subject to bivariate Gaussian distribution for the purpose of modeling possible motion selection within a plausible social range. Recently, generative models, e.g., variational autoencoders [42] and generative adversarial networks [43]–[46], have made significant advances in the aspect of complicated distribution generation. They learned distribution by maximizing the low bound of the likelihood function and

optimizing a two-player zero-sum game between a generator and a discriminator, respectively. In particular, there exist great advantages of GANs and their variants in capturing high dimensional data distribution while refraining from complicated probabilistic computation. Furthermore, it is a well-known fact that GANs have displayed remarkable performance in image synthesis [47]–[49], image-to-image translation [50]–[52], and super-resolution [53]–[55]. Accordingly, the conditional generative adversarial network (cGAN) is applied for multi-modal trajectory distribution generation in our model.

III. ARCHITECTURE

In this section, we propose a novel multi-vehicle collaborative learning model with spatio-temporal tensor fusion (TS-GAN) for vehicle trajectory prediction, and the proposed formulation achieves the goal by taking adjacent vehicles in the scene into account under an end-to-end framework. TS-GAN not only models spatio-temporal properties of the interactions among multiple vehicles but also efficiently estimates multi-modal trajectory distribution rather than simply minimizes ℓ_2 norm between the predicted value of future trajectory and the ground truth like most current methods.

A. Preliminaries

Our model takes as input historical trajectories of all vehicles in a scene denoted as $\mathbf{X} = (X_1, X_2, \dots, X_{n_0})$ and jointly reasons about the future trajectory of the predicted vehicle denoted as $\hat{Y}_{predict}$, where the corresponding coordinates of individual vehicle historical and predicted future trajectory are $X_i = (x_i^t, y_i^t)$ at each timestamp $t = t_1, t_2, \dots, t_{obs}$, and $\hat{Y}_{predict} = (\hat{x}_{predict}^t, \hat{y}_{predict}^t)$ at each timestamp $t = t_{obs}, t_{obs+1}, \dots, t_{pred}$, respectively. Additionally, the ground truth of the predicted vehicle is denoted as $Y_{predict}$, where the

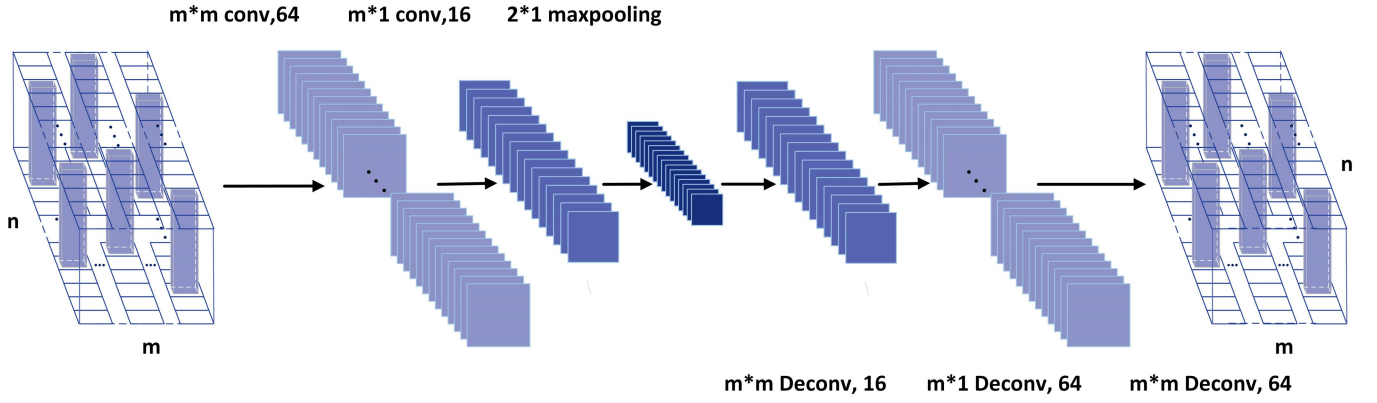


Fig. 2. Proposed auto-encoder social convolution module. We define a $n \times m$ grid on lanes as the social interaction scope of the predicted vehicle, and the historical trajectory encoding of each vehicle $e_i^{t_{obs}}$ at time-step t_{obs} is represented as a blue cube. They are placed in the sub-grid corresponding to the coordinates of time t_{obs} and finally make up a social tensor π , which retains the spatial relations of all vehicles within the scope. Afterwards, a convolutional auto-encoder architecture is further employed to get a robust representation of multi-vehicle historical trajectories.

corresponding coordinates are $Y_{predict}^t = (x_{predict}^t, y_{predict}^t)$ at each timestamp $t = t_{obs}+1, t_{obs}+2, \dots, t_{pred}$.

B. Generative Adversarial Networks

Generative Adversarial Networks (GANs) comprise a generator network and a discriminator network. Specially, the former takes random noises as input and then produces plausible samples, and the latter plays a role of distinguishing the actual samples from the false ones of the generator's output. In general, the generator and discriminator play a two-player zero-sum game by means of optimizing the following objective function:

$$\mathcal{L}_{GAN} = \min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]. \quad (1)$$

A conditional adversarial network (cGAN) is employed in our model to deal with the multi-modal phenomenon of future trajectories. In particular, the generator G , which is formulated as a sequence-to-sequence form, captures the future trajectory distribution of the predicted vehicle on the condition of joint encodings of all the vehicles' historical trajectories in the scene. Simultaneously, the discriminator D is employed to distinguish the actual trajectory from the false. It is worth noting that the generator G and the discriminator D in our model partly share the same architecture along with its parameters.

1) *Generator*: We first utilize LSTM cells to encode the historical trajectory of each vehicle in the scene as a fixed-length representation at time t :

$$e_i^t = LSTM(e_i^{t-1}, l_i^t; \theta) \quad (2)$$

where l_i^t is the location of the i th vehicle at time-step t and θ denote parameters of the LSTM cells. Notably, the parameters θ of LSTM cells are shared among all vehicles in the same scene. Nevertheless, vanilla LSTM cells are incapable of reasoning about subtle but significant interactions among multiple vehicles. To capture such interactions and further

fuse multi-vehicle information efficiently, we first propose an auto-encoder social convolution module (AESCM) to model spatial relations of multiple vehicles, as described in Section C. Furthermore, a recurrent social module (RSM) is proposed to fuse tensor information in terms of temporal relations, as described in Section D. After acquiring past trajectory encodings of all vehicles at time-step t_{obs} utilizing LSTM cells, AESCM and RSM modules are employed to obtain a spatial social tensor ν and a same length temporal social tensor τ , respectively. Besides, an embedding operation on the recurrent encoding of the predicted vehicle is conducted to generate an embedding encoding $h_{predict}$. Finally, a tensor fusion operation on the three encodings as mentioned above is conducted to form an informative representation, which is denoted as $f_{predict}$. In a nutshell, a spatio-temporal tensor fusion process is described as follows:

$$\begin{aligned} \nu &= AESCM(e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_v}^{t_{obs}}) \\ \tau &= RSM(e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_v}^{t_{obs}}) \\ h_{predict} &= \varphi(e_{predict}^{t_{obs}}) \\ f_{predict} &= (\nu + \tau) \oplus h_{predict} \end{aligned} \quad (3)$$

where φ denotes a multi-layer perceptron and the symbol \oplus is a concatenation operation. Specially, ν and τ are first fused by addition element-wise, and then concatenated with the encoding of the vehicle's past trajectory that is under prediction.

In our setting, cGAN is applied for capturing multi-modal trajectory distributions via combining a random noise z sampling from a standard Gaussian distribution with the fused informative tensor $f_{predict}$. Specially, the noise z and $f_{predict}$ are input to the decoder, and then a multi-layer perceptron is employed for coordinate prediction:

$$\begin{aligned} hd_i^t &= LSTM(hd_i^{t-1}, f_{predict} \oplus z) \\ (\hat{x}_i^t, \hat{y}_i^t) &= \varphi(hd_i^t) \end{aligned} \quad (4)$$

when we take diverse, randomly sampled noises, the model will generate possible plausible future trajectories.

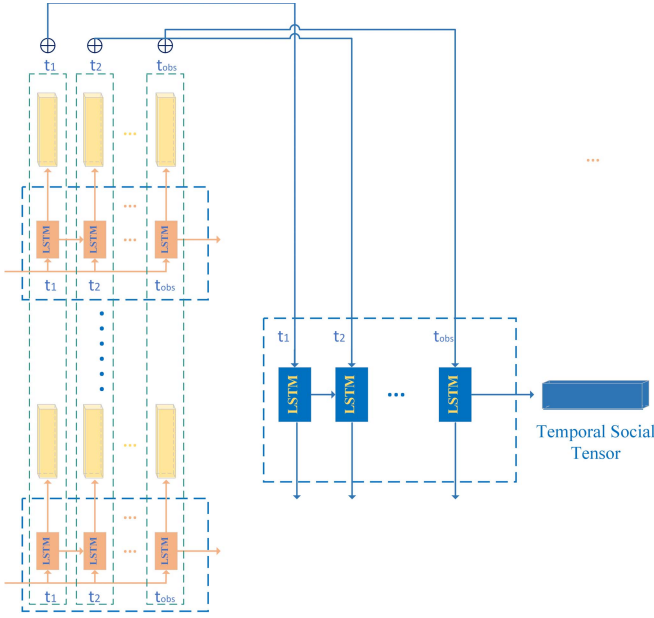


Fig. 3. Proposed recurrent social module. The Encoder is first employed to obtain the encoding vectors $\{[e_1^{t_1}, e_1^{t_2}, \dots, e_1^{t_{obs}}], \dots, [e_i^{t_1}, e_i^{t_2}, \dots, e_i^{t_{obs}}], \dots, [e_{n_v}^{t_1}, e_{n_v}^{t_2}, \dots, e_{n_v}^{t_{obs}}]\}$ of historical trajectories denoted as yellow cubes in the illustration for each adjacent vehicle within the grid scope at each time-step, then all vehicle encoding vectors of the same time-step are fused and as an input to another LSTM cell shown as the blue one at the corresponding time-step. Finally, we get a temporal social tensor τ , which fuses multiple vehicle information in the temporal dimension.

2) *Discriminator*: The discriminator D in our model is composed of an encoder which shares partly the structure as well as parameters with the one in generator G followed by a full connection layer for binary classification, which identifies its input $S_{real} = [X_i, Y_i]$ or $S_{fake} = [X_i, \hat{Y}_i]$ is true or false.

Furthermore, just like many pioneering approaches, we not only consider the adversarial loss in our work, but a ℓ_2 loss is also used to minimize the distance between the predicted trajectory and ground truth:

$$\mathcal{L}_{REC} = \sum_{t=obs+1}^{pred} \|\hat{Y}_i^t - Y_i^t\|_2^2. \quad (5)$$

C. Auto-Encoder Social Convolution Mechanism

For the purpose of tackling interactions among multiple vehicles, a novel auto-encoder social convolution mechanism is developed in our model, as illustrated in Fig. 2. Specially, the social interaction scope is defined by a $n \times m$ grid on lanes, where each column represents a longitudinal direction lane, and rows are split by the interval of 15 feet that is equivalent to an individual vehicle length approximatively. In our approach, the LSTM encoder is first utilized to acquire the encoding vectors of all surrounding vehicles' historical trajectories within the whole grid $\{e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_v}^{t_{obs}}\}$ at time-step t_{obs} . Afterwards, the encoding vectors $\{e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_v}^{t_{obs}}\}$ are placed in the corresponding grids to preserve spatial relations among vehicles, which results in forming a spatial tensor π . After that, two convolutional layers that possess the property of retaining local spatial structure, are employed to extract social

tensor v , which models multi-vehicle interactions at the aspect of spatial properties. Additionally, to raise the robustness of social encoding v , we reconstruct v into a tensor v_{rec} via two deconvolutional layers, so that v_{rec} approaches the original spatial tensor π as little information loss as possible. To this end, we minimize the following reconstruction loss:

$$\mathcal{L}_{AE} = \|\pi - v_{rec}\|_F^2. \quad (6)$$

It is evident that our proposed auto-encoder social convolution mechanism can take full advantage of some extra unlabeled data to optimize the above reconstruction loss, and thus gives rise to our whole model being able to work in a semi-supervised fashion.

D. Recurrent Social Mechanism

The auto-encoder social convolution mechanism only models multi-vehicle interactions based on the spatial characteristic but fails to take the temporal characteristic of multi-vehicle interactions into account. As a result, a fancy recurrent social mechanism is introduced into our model to fuse multi-vehicle information in the temporal dimension. To the best of our knowledge, our work pioneers a fusion scheme concerning how to model the temporal relations between multiple sequences. At first, we get encoding vectors of historical trajectories $\{e_i^{t_1}, e_i^{t_2}, \dots, e_i^{t_{obs}}\}$ for each adjacent vehicle i within the whole grid scope at each time-step t . Afterwards, the encoding information of all nearby vehicles at each time-step t is fused by means of addition element-wise on all encoding vectors of adjacent vehicles' historical trajectories, namely $[\sum_{i=1}^{n_v} e_i^{t_1}, \dots, \sum_{i=1}^{n_v} e_i^{t_{obs}}]$ where n_v is the number of adjacent vehicles within the grid scope. Finally, the sequence $[\sum_{i=1}^{n_v} e_i^{t_1}, \dots, \sum_{i=1}^{n_v} e_i^{t_{obs}}]$ is input into another LSTM cell to acquire a temporal social tensor τ as illustrated in Fig. 3. Notably, the proposed recurrent social mechanism conducts information fusion at each time-step over multi-vehicle driving trajectory sequences during the encoding phase, which takes into account the relations of multiple time series models at each time-step instead of an individual model. Besides, as we will observe in subsequent experiments, the introduced recurrent social mechanism improves trajectory prediction performance by a large margin, which also implicitly demonstrates the benefits of multi-sequence collaborative learning for the trajectory prediction task.

E. Spatio-Temporal Generative Adversarial Networks

The proposed TS-GAN framework is devoted to addressing the vehicle trajectory prediction issue by means of multi-vehicle spatio-temporal tensor fusion manner as illustrated in Fig. 1, which takes full advantage of surrounding vehicle information instead of only considering single-vehicle itself. The proposed model applies a LSTM-based encoder-decoder framework that is famous for excellent ability in the aspect of addressing sequence generation problem for trajectory prediction. On this basis, a novel auto-encoder social convolution mechanism and a fancy recurrent social approach are introduced to capture spatial information and subtle timing relations of multi-vehicle driving trajectories,

Algorithm 1 TS-GAN Training Algorithm**Input:** $n_e \leftarrow \text{epoch_numbers};$ $n_b \leftarrow \text{batch_size};$ $\lambda, \gamma \leftarrow \text{hyper-parameter};$ $X_{\text{predict}} \leftarrow \text{historical trajectories of the predicted vehicle};$ $(X_1, X_2, \dots, X_{n_b}) \leftarrow \text{historical trajectories of neighbours within the grid scope};$ $Y_{\text{predict}} \leftarrow \text{ground truth of the predicted vehicle's future trajectory}$

```

1 for  $i = 1; i \leq n_e$  do
2   for  $j = 1; j \leq n_b$  do
3     Take as input  $(X_1, X_2, \dots, X_{n_b})$  to the encoder for
       acquiring historical trajectory encodings
        $[(e_1^{t_1}, e_2^{t_1}, \dots, e_{n_b}^{t_1}), \dots, (e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_b}^{t_{obs}})]$ ;
4     Input  $X_{\text{predict}}$  to the encoder for acquiring the
       encoding of the predicted vehicle
        $[e_{\text{predict}}^{t_1}, \dots, e_{\text{predict}}^{t_{obs}}]$ ;
5     Input  $(e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_b}^{t_{obs}})$  to the Auto-encoder
       Social Convolution Module, and acquire spatial
       social tensor  $v$ ;
6     Input  $[(e_1^{t_1}, e_2^{t_1}, \dots, e_{n_b}^{t_1}), \dots, (e_1^{t_{obs}}, e_2^{t_{obs}}, \dots, e_{n_b}^{t_{obs}})]$ 
       to the Recurrent Social Module, and output
       temporal social tensor  $\tau$ ;
7     Take as input  $e_{\text{predict}}^{t_{obs}}$  to a multi-layer perceptron,
       and acquire historical trajectory encoding  $h_{\text{predict}}$ 
       of the predicted vehicle;
8     Get spatio-temporal tensor  $\zeta$  by adding  $v$  and  $\tau$ 
       element-wise at first, then perform concatenation
       operation on  $\zeta$  and  $h_{\text{predict}}$  to obtain fused tensor
        $f_{\text{predict}}$ ;
9     Sample a random noise  $z$ , then concatenate  $z$  and
        $f_{\text{predict}}$  as the input of the decoder;
10    The decoder predicts the possible future trajectory
        $\hat{Y}_{\text{predict}}$ ;
11    The discriminator  $D$  distinguishes the one which is
       made up of the historical and predicted future
       trajectories  $[X_i, \hat{Y}_i], \dots, [X_{n_b}, \hat{Y}_{n_b}]$  from the ground
       truth which is composed of historical and realistic
       future trajectories  $[X_i, Y_i], \dots, [X_{n_b}, Y_{n_b}]$ ;
12    Update the generator  $G$  by descending its gradient:
       
$$\nabla_{\theta_g} [\mathcal{L}_{GAN} + \lambda \mathcal{L}_{REC} + \gamma \mathcal{L}_{AE}]$$

       ;
13    Update the discriminator  $D$  by ascending its gradient:
       
$$\nabla_{\theta_d} [\mathcal{L}_{GAN}]$$


```

respectively. Given the inherent multi-modal characteristic of the driving behavior, a conditional generative adversarial network is skillfully integrated into our model for capturing multi-modal trajectory distribution. The entire model is able to be trained in an end-to-end fashion by optimizing the following objective function:

$$\mathcal{L}_{TOTAL} = \mathcal{L}_{GAN} + \lambda \mathcal{L}_{REC} + \gamma \mathcal{L}_{AE} \quad (7)$$

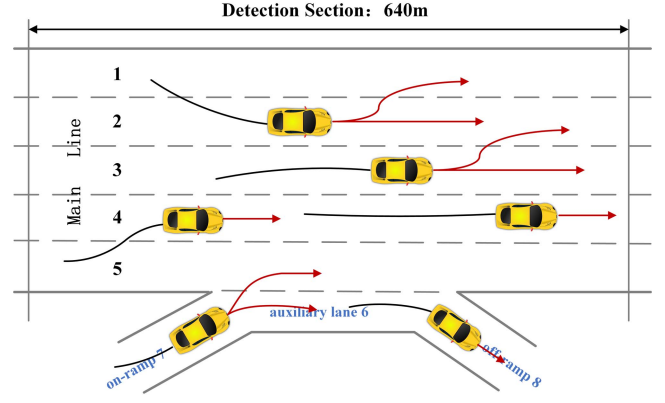


Fig. 4. Illustration of vehicle driving trajectories. Vehicles drive on the highway within a range of approximately 640 meters in length, including five mainline lanes, an on-ramp, an off-ramp, and an auxiliary lane. Black lines represent vehicles' historical trajectories, and red lines show the possible future trajectories within a reasonable range.

which combines adversarial loss, auto-encoder loss, and reconstruction loss as mentioned earlier, where λ and γ are hyper-parameters that weight the importance of different losses. Finally, the complete algorithm process is summarized as in Algorithm 1.

F. Implementation Details

The model is trained using Adam optimizer [56] in an end-to-end fashion with learning rate 0.001 and betas (0.9, 0.999). The hidden state size of the encoder and decoder is 64-dimensional and 128-dimensional, respectively. And the random noise z is a 10-dimensional vector sampling from a standard Gaussian distribution. The spatial social tensor has the same size as the temporal social tensor, while the multi-layer perceptron that embeds the encoding of the predicted vehicle outputs an embedded vector whose size is 32. In our networks, the ReLU activation function is always applied for convolutional and deconvolutional layers, as well as the sigmoid function for fully connected layers. During the training process, batch size is set as 64 and the number of epochs is set as 50. In addition, the hyper-parameter λ is set as 5 and γ is set as 3 in our experiments.

IV. EXPERIMENTS

A. Datasets

In this section, we conduct experiments on NGSIM US-101 [57] and I-100 [58] public benchmark datasets. Both datasets record realistic vehicle driving trajectories on the highway by sampling at 10Hz cross a 45-minute time span within a range of approximately 640 meters in length, as illustrated in Fig. 4. It is noteworthy that these data contain abundant interaction scenarios. Diverse traffic conditions, such as mild, moderate and congested, are taken into account in both datasets. The datasets provide vehicle driving trajectory coordinates. The same dataset partition manner and coordinate system settings are employed as in [17], leading to 5922867 training data, 859769 validate data, and 1505756 test data. Each data item

TABLE I
QUANTITATIVE EXPERIMENT EVALUATION ON NGSIM DATASET USING ROOT SQUARED ERROR IN METERS AS THE EVALUATION CRITERIA

Prediction horizon(s)	CV	C-VGMM +VIM [9]	LSTM	S-LSTM [16]	CS-LSTM [17]	GAIL-GRU [21]	MATF-Multi-Agent [22]	MATF-GAN [22]	TS-GAN (ours)
1	0.73	0.66	0.68	0.65	0.61	0.69	0.67	0.66	0.60
2	1.78	1.56	1.65	1.31	1.27	1.51	1.51	1.34	1.24
3	3.13	2.75	2.91	2.16	2.09	2.55	2.51	2.08	1.95
4	4.78	4.24	4.46	3.25	3.10	3.65	3.71	2.97	2.78
5	6.68	5.99	6.27	4.55	4.37	4.71	5.12	4.13	3.72

TABLE II
QUANTITATIVE EXPERIMENT ANALYSIS ABOUT DIFFERENT GRID SIZES USING ROOT SQUARED ERROR IN METERS AS THE EVALUATION CRITERIA

Grid Size	1s	2s	3s	4s	5s	Grid Size	1s	2s	3s	4s	5s
$3 \times 13_grid$	0.62	1.26	2.04	3.04	4.28	$5 \times 13_grid$	0.60	1.26	2.07	3.02	4.23
$3 \times 17_grid$	0.62	1.24	2.02	2.98	4.15	$5 \times 17_grid$	0.62	1.25	2.02	2.97	4.14
$3 \times 21_grid$	0.60	1.23	1.98	2.86	3.93	$5 \times 21_grid$	0.62	1.28	2.06	3.00	4.12
$3 \times 25_grid$	0.62	1.26	1.99	2.82	3.82	$5 \times 25_grid$	0.61	1.27	2.05	2.97	4.06
$3 \times 29_grid$	0.61	1.25	1.98	2.81	3.79	$5 \times 29_grid$	0.61	1.28	2.04	2.94	3.96
$3 \times 33_grid$	0.62	1.24	1.97	2.81	3.74	$5 \times 33_grid$	0.63	1.28	2.06	2.95	3.97
$3 \times 37_grid$	0.60	1.24	1.95	2.78	3.72	$5 \times 37_grid$	0.63	1.29	2.04	2.93	3.94
$3 \times 41_grid$	0.61	1.25	1.96	2.80	3.76	$5 \times 41_grid$	0.63	1.28	2.04	2.92	3.90

tackles with 8s horizon driving trajectories, where 3s of them are historical trajectories, and 5s are predictions.

B. Baselines

In our experiments, comparisons between our model with the following baselines are conducted:

CV: A typical constant velocity Kalman filter is employed for trajectory prediction.

LSTM: A vanilla LSTM-based sequence-to-sequence model is used for trajectory prediction with the same network setting as our model for a fair comparison.

C-VGMM+VIM: The model manipulates multi-agent interactions via combining Markov random field and maneuver-based variational Gaussian mixture model. However, this method exists the disadvantage of complex probability calculations for trajectory prediction.

GAIL-GRU: The model applies generative adversarial imitation learning for optimizing recurrent policies. However, the GAIL-GRU takes advantage of the ground truth of surrounding vehicles' future trajectories. We use it as a baseline on account of the same datasets being used in both types of research.

S-LSTM: The S-LSTM model, as described in [16], is considered since the recurrent social model takes multi-agent in the scene into account and introduces a pooling mechanism based on local spatial distance modeling interactions among multiple vehicles.

CS-LSTM: The convolution social pooling model is the closest to our work, which aligns surrounding agents using convolution operations for maintaining the spatial relations. However, CS-LSTM doesn't consider the temporal properties of multiple sequence models. Remarkably, the CS-LSTM model takes advantage of maneuver class information during the training process, but better performance is achieved in our model using less training information without maneuver classes.

MATF-GAN: The multi-agent tensor fusion network handles social interactions among various numbers and types of agents and constraints from the scene context that achieves state-of-the-art level. However, the MATF-GAN utilizes an extra global bird's-eye view of the scene, which may be inaccessible in many scenarios. Particularly, MATF-Multi-Agent removes the generative adversarial network from the entire MATF-GAN.

C. Quantitative Evaluation

As in [17], this paper uses root mean squared error (rmse) between the predicted results and ground truth regarding each timestamp t within the horizon of prediction over the test set as the algorithm performance measure: $rmse(t) = \sqrt{\frac{1}{n_t} \sum_{i=1}^{n_t} [(\hat{x}_i^t - x_i^t)^2 + (\hat{y}_i^t - y_i^t)^2]}$ where n_t denotes the number of test set data. The quantitative experimental

TABLE III

COMPLEXITY AND RUNNING SPEED COMPARISON BETWEEN DIFFERENT MODELS. MODEL COMPLEXITY W.R.T VEHICLES NUMBER n IN A GIVEN SCENARIO IS PRESENTED. COMPARED WITH THE BASELINE, THE PERCENTAGE OF ALGORITHM ACCURACY IMPROVED, THE PERCENTAGE OF MODEL TIME-CONSUMPTION INCREASED, AND THE RATIO OF THE TWO ARE DISPLAYED. ALL METHODS ARE BENCHMARKED ON NVIDIA GeForce RTX 2080 Ti Graphics Cards

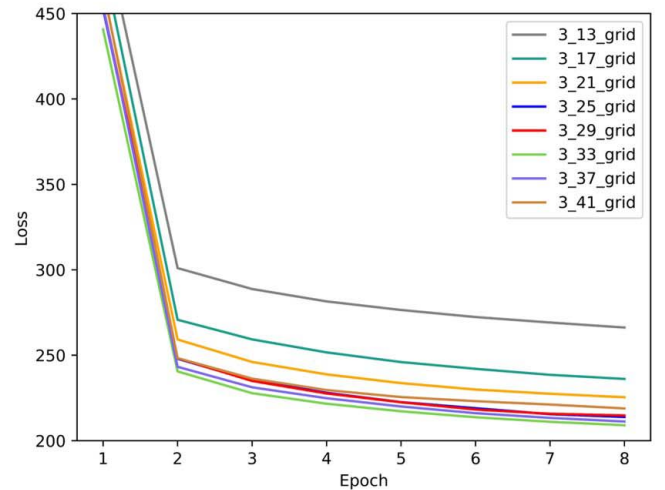
		LSTM (baseline)	S-LSTM	CS-LSTM	GAIL-GRU	MATF-GAN	TS-GAN
Complexity		$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$
Accuracy(m)	MSE	6.27	4.55	4.37	4.71	4.13	3.72
	Accuracy-Up	0%	27.4%	30.3%	24.9%	34.1%	40.7%
Speed(ms)	running time	0.11	0.28	0.31	0.27	0.39	0.34
	Speed-Up	0%	155%	182%	145%	255%	209%
Accuracy-Up/Speed-Up		0	0.18	0.17	0.17	0.14	0.19

evaluations on NGSIM US-101 and I-80 datasets are represented in Table. I. The results demonstrate that our proposed approach for trajectory prediction comprehensively outperforms these baselines, where most of them achieve state-of-the-art level. An interesting phenomenon, as we can observe in the comparisons, represents that our proposed model appears to obtain better performance in long-term trajectory prediction relative to short-term ones. We attribute this desirable fruit to the spatiotemporal property fusion of multi-vehicle driving trajectories, which captures the subtle and sophisticated interactions among multiple vehicles.

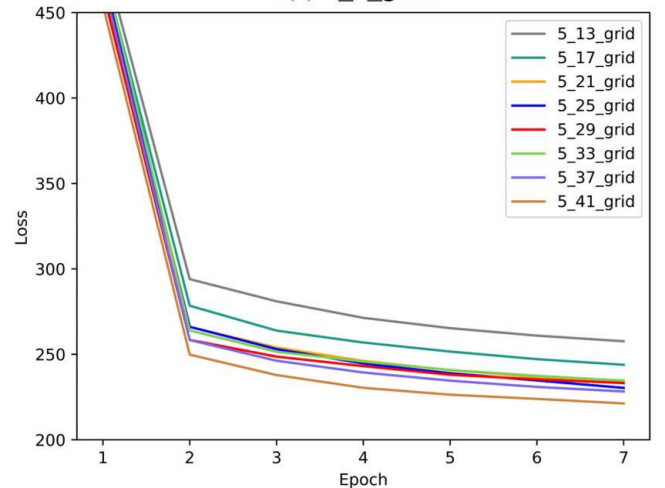
Experiments for varying number of surrounding vehicles are investigated with various grid sizes, which increases along with the lateral and longitudinal directions, respectively. First of all, while keeping the lateral scope unchanged, we range the longitudinal scope from 13 grids to 41 grids at the interval of 4 grids. Likewise, we set the lateral scope as three grids and five grids respectively while keeping the longitudinal scope unchanged. Quantitative experimental evaluations are represented in Table II. Generally, the performance improves as the grid size increases and finally converges after the grid size reaches a threshold. The possible cause is that remote vehicles have almost no effect on the current prediction. Besides, experimental results indicate that the grid size increment in the longitudinal direction has a significant impact on predicted accuracies, while in the lateral direction seems to be worse instead. The phenomenon in the lateral direction, in our opinions, is reasonable because the vehicle under the prediction is mainly affected by vehicles in adjacent lanes, and other lanes have little influence on it.

To further investigate the effect of various grid sizes on the prediction performance, training loss curves of the 3_x and 5_x grid sizes are depicted in Fig. 5. As the grid size increases along the longitudinal direction, the corresponding model converges much faster and reaches a lower loss in the end, which benefits from more effective information provided.

Furthermore, real-time is critical for trajectory prediction in real-world scenarios, and thus the computational complexity and running speed of different models are also compared and analyzed, as represented in Table III. We choose the LSTM model as the baseline and compare several typical methods,



(a) 3_x_grid



(b) 5_x_grid

Fig. 5. Loss descent curves using various grid sizes. (a) 3_x grids are employed in the proposed model, where x is the size in the longitudinal direction. (b) 5_x grids are employed in the proposed model, where x is the size in the longitudinal direction.

all of which model the interaction between multiple agents. Firstly, we theoretically analyze the complexity of different

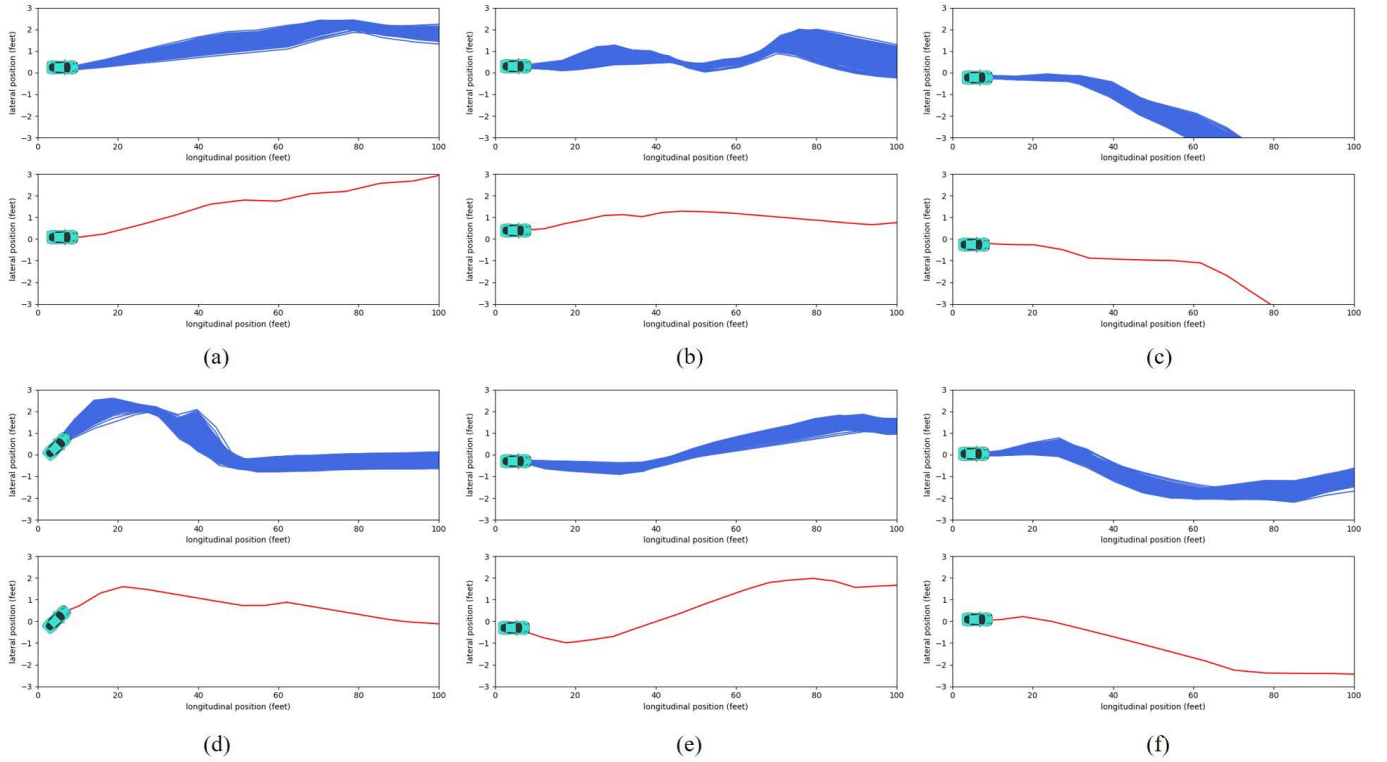


Fig. 6. The predicted trajectory distribution within the 5s horizon is visualized via sampling the distribution represented by the generator G 500 times. The blue lines represent the predicted trajectory distribution, and the red lines represent the corresponding ground truth. Apparently, the predicted distribution effectively covers the realistic future trajectory while capturing the inherent multimodal characteristics of driving behaviors.

approaches and find that they are all proportional to the number of vehicles in the environment. Compared with baseline, we calculate the percentage of algorithm accuracy improved, the percentage of model time-consumption increased, and the ratio of them. Although our model consumes more running time, we also achieved higher accuracy, and experimental results show that our model achieves the optimal trade-off between accuracy and speed.

D. Ablation Study

In order to verify the amazing benefits derived from the developed auto-encoder social convolution and recurrent social mechanisms, some quantitative ablation investigations are conducted through some experiments designed to control the special module affecting the model performance. Hence, the corresponding fusion modules are preserved or removed from the entire TS-GAN model, and the quantitative experiment results are shown in Table. IV. It is worth noting that the VS-GAN only considers single vehicle and ignores interactions among multiple vehicles in the scenario, but T-GAN, S-GAN, and TS-GAN all consider multi-vehicle interactions in a manner of collaborative learning for trajectory prediction:

VS-GAN: The vanilla sequence-to-sequence model with generative networks for multi-modal trajectory prediction and the VS-GAN model only tackles with a single-vehicle for reasoning about the future trajectory without considering any interaction information among multiple vehicles. To this end, neither auto-encoder social convolution module nor recurrent

social module is utilized in the VS-GAN. As we have observed in Table. IV, the VS-GAN model achieves poor performance on all predicted horizons.

T-GAN: Remove the auto-encoder social convolution module from our TS-GAN model and only consider the multi-vehicle interactions in the temporal dimension via the recurrent social module. We can discover that the performance of T-GAN is significantly superior to VS-GAN, which benefits from multi-vehicle collaborative learning in the aspect of temporal information fusion.

S-GAN: Remove the recurrent social module from our proposed TS-GAN model and only take into account the multi-vehicle interactions in spatial dimension via the auto-encoder social convolution mechanism. The performance of S-GAN is also superior to the scheme where using single-vehicle and ignoring multi-vehicle interactions to reason about future trajectory.

TS-GAN: The proposed TS-GAN captures multi-vehicle spatio-temporal interaction information via introducing not only an auto-encoder social convolution mechanism but also a recurrent social module. The performance of TS-GAN is nearly prominent surpass all other ablation models for different grid sizes as represented in Table. IV, which demonstrates that the proposed both auto-encoder social convolution and recurrent social mechanisms play crucial roles in efficiently capturing subtle and sophisticated interactions among multiple vehicles. Additionally, we find that, dramatically, the pioneering work proposed by us concerning modeling multi-agent

TABLE IV

ABLATION ANALYSES ON ALGORITHM STRUCTURE FOR VARIOUS GRID SIZES USING ROOT SQUARED ERROR IN METERS AS THE EVALUATION CRITERIA

Grid Size	Model	1s	2s	3s	4s	5s	Grid Size	Model	1s	2s	3s	4s	5s
3_13_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11	3_17_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11
	T-GAN	0.61	1.27	2.07	3.09	4.34		T-GAN	0.62	1.25	2.01	2.94	4.09
	S-GAN	0.63	1.28	2.09	3.11	4.35		S-GAN	0.62	1.24	1.99	2.92	4.05
	TS-GAN	0.62	1.26	2.04	3.04	4.28		TS-GAN	0.62	1.24	2.02	2.98	4.15
3_21_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11	3_25_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11
	T-GAN	0.61	1.30	2.09	2.98	4.06		T-GAN	0.62	1.27	2.02	2.92	3.96
	S-GAN	0.63	1.28	2.01	2.90	3.98		S-GAN	0.63	1.26	2.02	2.91	3.96
	TS-GAN	0.60	1.23	1.98	2.86	3.93		TS-GAN	0.62	1.26	1.99	2.82	3.82
3_29_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11	3_33_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11
	T-GAN	0.65	1.29	2.03	2.89	3.84		T-GAN	0.62	1.26	1.99	2.82	3.79
	S-GAN	0.64	1.26	1.99	2.84	3.83		S-GAN	0.64	1.32	2.07	2.95	3.95
	TS-GAN	0.61	1.25	1.98	2.81	3.79		TS-GAN	0.62	1.24	1.97	2.81	3.74
3_37_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11	3_41_grid	VS-GAN	0.67	1.60	2.82	4.34	6.11
	T-GAN	0.64	1.29	2.04	2.89	3.86		T-GAN	0.62	1.27	2.00	2.83	3.80
	S-GAN	0.63	1.27	2.02	2.86	3.83		S-GAN	0.65	1.31	2.08	2.96	3.99
	TS-GAN	0.60	1.24	1.95	2.78	3.72		TS-GAN	0.61	1.25	1.96	2.80	3.76

TABLE V

TIME-CONSUMPTION ANALYSIS OF FOUR ABLATION MODELS FOR DIFFERENT TIME WINDOWS IN MILLISECONDS DURING THE TEST PHASE

Model	1s	2s	3s	4s	5s
VS-GAN	0.241	0.245	0.269	0.273	0.279
T-GAN	0.268	0.270	0.279	0.284	0.288
S-GAN	0.304	0.318	0.321	0.326	0.328
TS-GAN	0.317	0.323	0.326	0.329	0.335

interactions based on time characteristics seems to be better than space characteristic-based.

An average run-time comparison of four ablation models during the test phase has also been investigated using NVIDIA GeForce RTX 2080 Ti Graphics Cards, and the results are represented in Table V. In general, the average time-consumption of prediction in milliseconds for all four ablations, which can be used for real-time vehicle trajectory prediction. Although TS-GAN is slightly more complicated than the other three models, for different window sizes (e.g., 1s, 2s, 3s, 4s, 5s), the run-time hardly increases, but the performance is significantly improved.

E. Qualitative Analysis

Our proposed TS-GAN outperforms state-of-the-art approaches, as presented in Table. I. In this section, some qualitative analyses concerning the predicted trajectories of

the proposed model are presented. At first, a qualitative experiment about multi-modal characteristics of future trajectory is illustrated in Fig. 6. It describes predicted trajectory distributions of our proposed model via sampling 500 times from the distribution represented by the generator G , as well as the corresponding ground truth. Predictions by the proposed algorithm are desirable in terms of covering the ground truth. Particularly, the proposed model generates a series of socially-acceptable future trajectories and successfully captures the inherent multi-modal characteristics of vehicle motion, which indicates that generative adversarial networks introduced in our model are competent for generating multi-modal trajectory distribution. Additionally, as we saw in the illustration, the proposed model can appropriately handle the complex motions during the process of driving. For instance, the model enables one to recognize fast steering as represented in (d), (e), and (f), as well as some subtle changes shown in (a), (b), and (c).

Since the interactions among multiple vehicles play a critical role in the trajectory prediction task, some qualitative analyses are also conducted with regard to how the surrounding vehicles affect future trajectories, and the visualization of experimental results is presented in Fig. 7. The illustration depicts surrounding vehicles' historical trajectories and complete trajectories of predicted vehicles. Still, our model efficiently reasons about the future trajectory while considering the impact of other vehicles in the scenario, which gives rise to effective avoidance of a vehicle collision. As in the depiction presented, vehicles receive driving information from

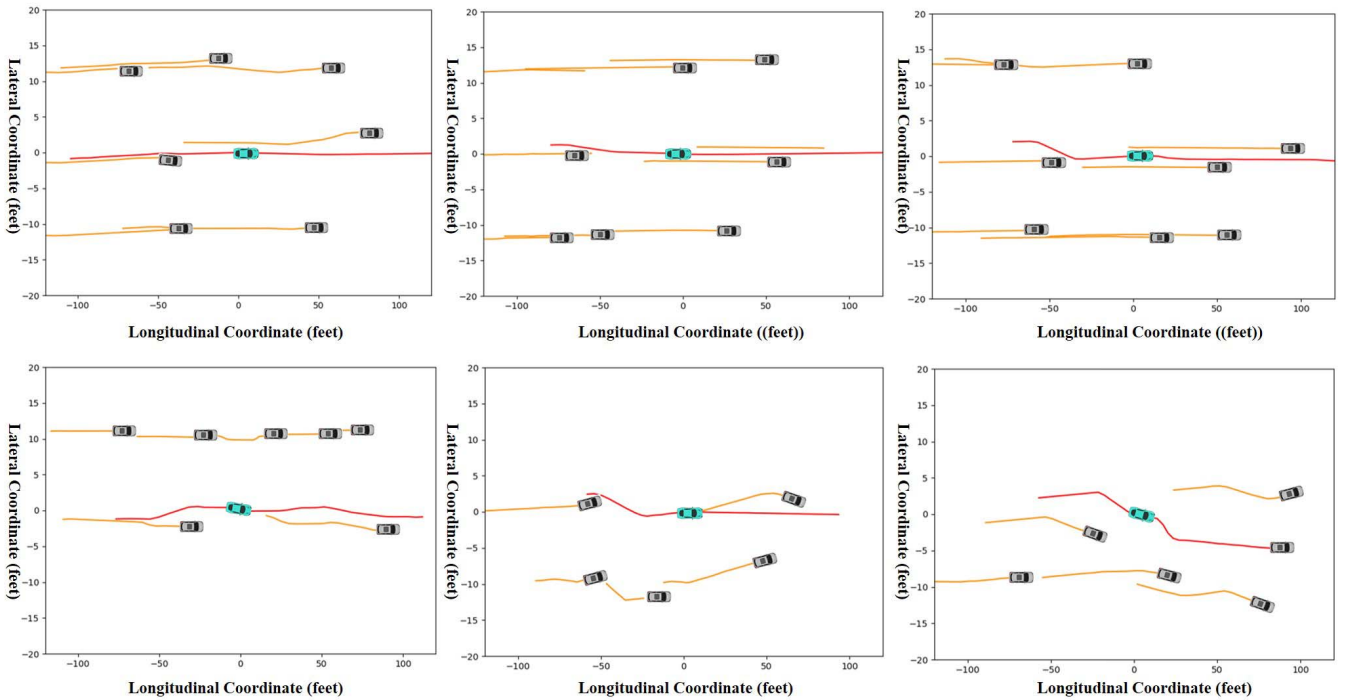


Fig. 7. **Multi-Vehicle Interactions:** The illustration shows impacts of surrounding vehicles on the predicted vehicle. Orange lines represent historical trajectories of surrounding vehicles in the scenario, and red lines represent the complete trajectories of predicted vehicles, each of which is composed of 3s past and 5s future trajectories.

adjacent ones in time. As a result, the reasonable planning of future trajectories is inferred from our model, which implicitly demonstrates that the subtle and sophisticated interactions among multiple vehicles are captured efficiently by the proposed auto-encoder social convolution and recurrent social mechanism.

V. CONCLUSION

In this paper, a novel multi-vehicle collaborative learning framework for trajectory prediction with a spatio-temporal tensor fusion mechanism is proposed. To tackle interactions among multiple vehicles, a novel auto-encoder social convolution module and a recurrent social mechanism are developed, and they can fuse multi-agent information via manipulating on the spatial and temporal structure, respectively, in the manner of tensor fusion. Particularly, to the best of our knowledge, we are the first to propose an approach to model the relations of multiple time sequences. At the same time, for the purpose of handling the inherent multi-modal characteristics of driving motion, a generative adversarial network is also integrated into our model for generating possible socially-acceptable trajectories. Finally, quantitative evaluations on standard datasets indicated that our model outperforms the state-of-the-art methods. Qualitative analyses about the predicted trajectory distribution and the influence of surrounding vehicles in the scenario are also represented in this paper.

REFERENCES

- [1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.
- [2] J. Huang and H.-S. Tan, "Error analysis and performance evaluation of a future-trajectory-based cooperative collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 175–180, Mar. 2009.
- [3] M. Bahram, C. Hubmann, A. Lawitzky, M. Aeberhard, and D. Wollherr, "A combined model- and learning-based framework for interaction-aware maneuver prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1538–1550, Jun. 2016.
- [4] W. Choi and S. Savarese, "Understanding collective activities of people from videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1242–1257, Jun. 2014.
- [5] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.
- [6] G. Antonini, M. Bierlaire, and M. Weber, "Discrete choice models of pedestrian walking behavior," *Transp. Res. Part B, Methodol.*, vol. 40, no. 8, pp. 667–687, Sep. 2006.
- [7] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1345–1352.
- [8] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 201–214.
- [9] N. Deo, A. Rangesh, and M. M. Trivedi, "How would surround vehicles move? A unified framework for maneuver classification and motion prediction," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 2, pp. 129–140, Jun. 2018.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
- [11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–12.

- [15] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [16] 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. (CVPR)*, Jun. 2016, pp. 961–971.
- [17] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1468–1476.
- [18] M. Bansal, A. Krizhevsky, and A. Ogale, "ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst," in *Proc. Robot., Sci. Syst. 15th*, vol. 15, Jun. 2019, pp. 1–10.
- [19] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 336–345.
- [20] 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. (CVPR)*, Jun. 2019, pp. 1349–1358.
- [21] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer, "Imitating driver behavior with generative adversarial networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 204–211.
- [22] T. Zhao *et al.*, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12126–12134.
- [23] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-aware trajectory prediction," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1941–1946.
- [24] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.
- [25] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2980–2988.
- [26] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," 2014, *arXiv:1412.1602*. [Online]. Available: <http://arxiv.org/abs/1412.1602>
- [27] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
- [28] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.
- [29] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
- [31] A. Karpathy, A. Joulin, and F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.
- [32] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.
- [33] T. Shu, S. Todorovic, and S.-C. Zhu, "CERN: Confidence-energy recurrent network for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5523–5531.
- [34] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. S. Kweon, "AttentionNet: Aggregating weak directions for accurate object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2659–2667.
- [35] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
- [36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [37] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: <http://arxiv.org/abs/1412.3555>
- [38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [39] 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.*, Jun. 2018, pp. 2255–2264.
- [40] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–7.
- [41] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.
- [42] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.
- [43] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [44] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: <http://arxiv.org/abs/1411.1784>
- [45] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
- [46] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–16.
- [47] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.
- [48] E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.
- [49] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5077–5086.
- [50] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [51] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
- [52] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 702–716.
- [53] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
- [54] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3760–3768.
- [55] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–333.
- [56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [57] J. Colyar and J. Halkias, "U.S. Highway 101 dataset," Federal Highway Admin. (FHWA), Washington, DC, USA, Tech. Rep. FHWA-HRT-07-030, 2007.
- [58] J. Colyar and J. Halkias, "U.S. Highway i-80 dataset," Federal Highway Admin. (FHWA), Washington, DC, USA, Tech. Rep. FHWA-HRT-07-030, 2007.



Yu Wang received the B.S. degree from Fuzhou University, China, in 2017. He is currently pursuing the Ph.D. degree with Tongji University, Shanghai, China. His research interests include computer vision, cooperative learning, cross-modal data matching, intelligent transportation systems, and generative models.



Shengjie Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the M.S. degree in electrical and computer engineering from the China Aerospace Institute, Beijing, China, in 1991, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2004. He is currently the Dean of the College of Software Engineering and a Professor with the College of Software Engineering and the College of

Electronics and Information Engineering, Tongji University, Shanghai, China. In previous postings, he conducted research at Lucent Technologies, Whippany, NJ, USA, and the China Aerospace Science and Industry Corporation, Beijing. His research interests include artificial intelligence, big data, wireless communications, image processing, and signal processing. He is a fellow of the Thousand Talents Program of China.



Rongqing Zhang (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) from Peking University, Beijing, China, in 2009 and 2014, respectively. From 2014 to 2018, he worked as a Post-Doctoral Research Fellow with Colorado State University, CO, USA. Since 2019, he has been an Associate Professor with Tongji University, Shanghai, China. He was also an International Presidential Fellow of Colorado State University in 2017. He has authored or coauthored two books, two book chapters, and over 90 papers in refereed journals and conference

proceedings. His current research interests include Internet of vehicles (IoV), physical layer security, and autonomous driving. He was a recipient of the Academic Award for Excellent Doctoral Students, Ministry of Education of China and a co-recipient of the First-Class Natural Science Award, Ministry of Education of China. He received the best paper awards at the IEEE ITST'12, ICC'16, GLOBECOM'18, and ICC'19. He is currently serving as an Associate Editor of *IET Communications* and *Complexity*.



Xiang Cheng (Senior Member, IEEE) received the Ph.D. degree from Heriot-Watt University and the University of Edinburgh, Edinburgh, U.K., in 2009. He is currently a Professor with Peking University. His general research interests include the areas of channel modeling and mobile communications, on which he has published more than 200 journals and conference papers, five books, and holds seven patents. He was a recipient of the IEEE Asia-Pacific (AP) Outstanding Young Researcher Award in 2015, a Co-Recipient of the 2016 IEEE JSAC

Best Paper Award: Leonard G. Abraham Prize, the NSFC Outstanding Young Investigator Award, and the First Rank and Second-Rank Awards in Natural Science, Ministry of Education, China. He received the best paper awards at the IEEE ITST'12, ICC'13, ITSC'14, ICC'16, ICNC'17, and GLOBECOM'18. He also received the Postgraduate Research Thesis Prize from the University of Edinburgh. He served as the symposium leading-chair, co-chair, and a member of the Technical Program Committee for several international conferences. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and *Journal of Communications and Information Networks*, and a Distinguished Lecturer of the IEEE.



Liuqing Yang (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Minnesota, Minneapolis, in 2004. She is currently a Professor with Colorado State University. Her general interests include signal processing with applications to communications, networking, and power systems, on which she has published more than 310 journals and conference papers, four book chapters, and five books. She was a recipient of the ONR Young Investigator Program (YIP) Award in 2007, the NSF Faculty

Early Career Development (CAREER) Award in 2009, and the Best Paper Award at the IEEE ICUWB'06, ICC'13, ITSC'14, Globecom'14, ICC'16, WCSP'16, Globecom'18, ICCS'18, and ICC'19. She served as the program chair, track/symposium or TPC chair for many conferences. She is the Editor-in-Chief of *IET Communications*, and has served as an Associate/Senior Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the IEEE INTELLIGENT SYSTEMS, and *PHYCOM: Physical Communication*.