PECGAN: Endpoint Conditioned Trajectory Prediction via Generative Adversarial Network

1st Xiangyu Li
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
2019110959@mail.hfut.edu.cn

2st Yusheng Peng
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
wisionpeng@mail.hfut.edu.cn

3st Wenming Wu
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
wwming@hfut.edu.cn

4st Gaofeng Zhang
School of Software
Hefei University of Technology
Hefei, China
g.zhang@hfut.edu.cn

5st *Liping Zheng
School of Software
Hefei University of Technology
Hefei, China
zhenglp@hfut.edu.cn

2021 China Automation Congress (CAC) | 978-1-6654-2647-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAC53003.2021.9727437

***Abstract*—Pedestrian trajectory prediction is a key research topic in the field of computer vision and has been widely used in practical applications, such as robot navigation and autonomous driving. Previous studies predict the future trajectory by decoding the learned motion feature via a self-recurrent architecture, which leads to a significant prediction deviation of the endpoint. Therefore, we propose Predicted Endpoint Conditioned Generative Adversarial Network (PECGAN) to predict the future trajectory without significant endpoint deviations. In our model, endpoint prediction is the primary goal which is accomplished through a conditional variables autoencoder. The estimated endpoints, coupled with past trajectories are encoded as the motion feature, and refined by a social interaction module which adopts the self-attention mechanism for message passing. The refined motion features infer the intermediate trajectory more accurately. Experimental results demonstrate that PECGAN can generate a realistic and diverse set of trajectories that respect physical constraints. Our proposed model improves state-of-the-art performance on the Stanford Drone Dataset benchmark and the ETH-UCY benchmark.**

***Index Terms*—endpoint prediction, trajectory prediction, social interactions, conditional variational autoencoder, generative adversarial network.**

I. Introduction

Pedestrian trajectory prediction is a research hotspot in the field of computer vision, which has a wide range of applications in abnormal behavior detection [1], auto-driving of automobiles [2] and control of social robots [3], etc. Pedestrian travel is mainly motivated by destination and the social interactions. However, due to the uncertainty of the destination and the complex and subtle social interactions among pedestrians, predicting the future trajectory is full of challenges.

Most of the preceding works [4]–[9] focus on the prediction of a complete future trajectory based on the past trajectory and social interactions among pedestrians. However, due to the uncertainty of the destination, these models generate equal possible results in different directions. To address this, generating future trajectories according to estimated goals [9]–[12] has been proposed. Remarkably, Fang et al. [12] and Dendorfer et al. [11] utilize the past trajectories of all the pedestrians in the scene as well as the scene information to predict future endpoints, and then generate the future trajectory purposefully. Different from the above methods, Mangalam et al. [9] adopt a conditional variational autoencoder(CVAE) to estimate the endpoint, and utilize multi-layer perceptrons (MLPs) to interpolate the intermediate positions. This method does not make use of the scene information and also achieves good results. In this article, a two-stage motion prediction framework is proposed for predicting future trajectories without using scene information, and achieves the state-of-the-art results on the latest popular trajectory prediction datasets ETH [13], UCY [14] and Stanford Drone Dataset [15].

Inspired by [11] and [9], we propose Predicted Endpoint Conditioned Generative Adversarial Network (PECGAN). This is a novel two-stage method that makes the distribution of prediction targets more clear by adding constraints to Conditional variational autoencoder. The first stage estimates the endpoint of the future trajectory via a Variational Autoencoder model. The fused features of the predicted endpoints and the past trajectories are employed in trajectory generation module to infer the intermediate positions in the second stage.

To sum up, our principal contribution is triple: (1) A conditional generative adversarial network is proposed for multiple socially acceptable trajectory predictions. (2) We introduce a novel two-stage framework for the future trajectory prediction, in which the endpoint prediction is the primary objective, and the intermediate trajectory is conditionally inferred from the predicted endpoint. (3) The self-attention mechanism is employed for message passing to model social interactions, and it can be performed several times to model deep interactions.

The remainder of this paper is organized as follows. Section II briefly reviews the related work of our method. Section III introduces the details of each module of PECGAN and

the regularization strategy. In Section IV, we illustrate the experiments and comparative analyses, with some conclusions presented in Section V.

II. Related Work

A. Trajectory Prediction

There are many methods to predict pedestrian trajectories based on deep-learning models. The Long Short Term Memory (LSTM) is one of the most well-known approaches for model trajectory prediction. Alahi et al. [5] propose a Social LSTM that predicts a trajectory based on a social pooling layer, which aggregates the neighbor pedestrians' hidden states. To generate multiple socially acceptable trajectories, Gupta et al. [6] firstly propose Social-GAN that predicts socially plausible future trajectories by training adversarially and encourage multiple predictions with a new variety loss. Recently, many GAN-based trajectory prediction methods [7], [11], [16] have made good progress. As another popular generative model, the conditional variational autoencoder (CVAE) [17], [18] is widely adopted in trajectory prediction [9], [19].

In this paper, GAN is used as a framework for predicting socially acceptable trajectories. The CVAE is viewed as a sub-module of the generator to generate the predicted endpoints.

B. Social Interaction Modeling

To model the complex and subtle social interactions among pedestrians, hand-crafted features are adopted to predict trajectories, such as Social Force [4] and Gaussian Processes [20]. Recently, with the development of deep-learning technology, data-driven approaches have proven highly successful in predicting pedestrian trajectories. In these methods, social interactions are modeled by the social pooling mechanism [5]–[7] and self-attention mechanism [8], [9], [12]. The social pool mechanism cannot model different impacts on different neighbor pedestrians. Self-attention mechanism can accurately acquire the respective weights for different pedestrians and has achieved great success in modeling social interaction.

Different from the traditional attention mechanism, self-attention mechanism has achieved excellent performance in modeling social interaction through a novel information transmission mechanism. Therefore we adopt a novel self-attention mechanism as a sub-module of the generator to model social interactions among pedestrians.

C. Endpoint-conditioned Prediction

The previous works are predicting the future path of the pedestrian according to the past trajectory, regardless of the pedestrians' destinations which play a key role in their motions in the scene. Rehder et al. [10] predict a probability map of each target pedestrian's endpoint based on the surrounding environment. However, Rehder et al. [10] predict possible destinations based on precomputed environment features, but ignore the pedestrians in the scene who are the key to predict short-term motions. Dendorfer et al. [11] tend to achieve the goal estimation first and estimate a set of plausible trajectories that route towards the estimated goals based on RNNs. However, the error accumulations exist in the recurrent-based prediction, which results in a large deviation of the endpoint of prediction. Different from the above method, Mangalam et al. [9] adopt a conditional variational autoencoder (CVAE) to estimate the endpoint, and utilize MLPs to interpolate the intermediate positions. However, Mangalam et al. [9] only optimize the distribution of samples by reducing the distance error between the predicted trajectory and the ground truth. This constraint is too weak to obtain a goal distribution explicitly. The generated endpoints still have a significant deviation from the endpoints of the ground truth.

To obtain a reasonable distribution of the predicted endpoints and generate multiple socially acceptable trajectories through the predicted endpoints, the CVAE is used as a sub-module of the generator to predict possible destinations in the proposed model.

III. Approach

A. Overview

In this section, we present the proposed PECGAN. As shown in Fig. 1, PECGAN has two components: the generator and discriminator. In the generator, there are two stages for trajectory prediction. In the first stage, the past trajectory along with the ground truth endpoint is used to train a CVAE for the endpoint of trajectory in the endpoint generation module. In the second stage, the trajectory generation module generates the intermediate positions by taking the endpoint of prediction into account. The discriminator module is composed of the feature encoder and classifier. The feature encoder captures the features from the intermediate trajectory and endpoint, and the classifier identifies whether the input trajectory is generated by the generator or not. The details of PECGAN are described in the following sections.

B. Problem Formulation

In this paper, we focus on how to more accurately predict the pedestrian trajectory in crowded scenes. For each sample, we assume there are *N* pedestrians involved in the scene. Given observed trajectory $X_i = \{X_i^0, X_i^1, ..., X_i^{T_{obs}}\}$ of pedestrian *i*, where $\{X_i^t = (x_i^t, y_i^t)\}$ represents the ground truth position at time t, the problem requires predicting positions $\{\hat{Y}_i^{t'} | (\hat{x}_i^{t'}, \hat{y}_i^{t'}), t' = T_{obs}+1, T_{obs}+2, ..., T_{pred}\}$ of future T_{pred} key frames. The future positions of ground truth are denoted as $\{Y_i^{t'} | (x_i^{t'}, y_i^{t'}), t' = T_{obs}+1, T_{obs}+2, ..., T_{pred}\}$.

C. Generator

In the trajectory prediction task, GAN-based models such as social-GAN [6], NMMP [7] and goal-GAN [11]achieve the goal of effectively predicting the future trajectory. However, the importance of endpoint prediction is not taken into consideration in these methods. To achieve this, a novel GAN-based model (PECGAN) is proposed to predict the future trajectory. In the generator, the future trajectory is predicted in two steps: 1. Predicting the endpoint of the future trajectory; 2. Inferring the intermediate positions through the predicted endpoint.

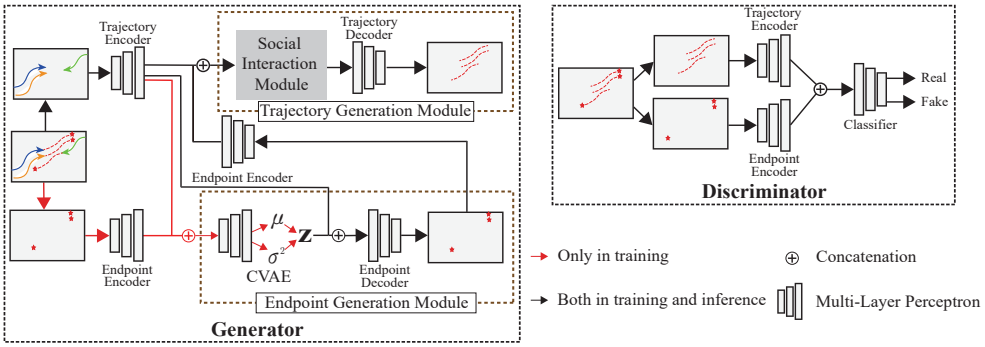

Fig. 1. Architecture of PECGAN. PECGAN has two components: the generator and the discriminator. The generator consists of an endpoint generation module and a trajectory generation module. In the generator, the endpoint is predicted through the endpoint generation module. The trajectory generation module is designed to predicts the intermediate positions. The discriminator network consists of multiple multi-layer perceptrons to distinguish between real and fake trajectories.

1) Feature Encoder: We propose two kinds of encoders, trajectory encoder and endpoint encoder. In order to capture the motion features of pedestrians, each trajectory's relative displacement vectors $(\Delta x_i^t, \Delta y_i^t)$ are fed into a trajectory encoder with multi-layer perceptrons. The obtained trajectory features are embedded into the endpoint generation module and the trajectory generation module to generate the endpoint and the intermediate trajectory respectively.

2) Endpoint Generation Module: In this section, a novel endpoint generation module is developed to predict the endpoint of the trajectory. As shown in Fig. 1, during the training process, the PECGAN fuses the motion features acquired by the trajectory encoder and the endpoint feature acquired by the endpoint encoder. The fused features are fed into the CVAE's latent encoder to encode the latent variable $z = \mathcal{N}(\mu, \sigma)$. Here, the latent codes can be sampled multiple times from the normal distribution to generate multiple socially acceptable endpoints.

To encode the latent variable $z = \mathcal{N}(\mu, \sigma)$, a novel CVAE is proposed to encourage the generation of the multiple possible endpoints. The endpoint $(x_i^{t'}, y_i^{t'})$ of the pedestrian trajectory is regarded as a known condition of the VAE. we sample the latent code z from $\mathcal{N}(\mu, \sigma)$, and the fused features of motion feature and latent code z are fed into the endpoint decoder to obtain the predicted endpoint $\hat{Y}_i^{T_{pred}}$. However, it is impossible to take the ground-truth endpoint as input data during the inference process. Therefore, the latent code z will be sampled from $\mathcal{N}(0, 1)$. Here, the latent codes are sampled multiple times from the standard normal distribution to generate multiple socially acceptable endpoints. The predicted endpoint will be served as the input data of the Trajectory Generation Module to predict the intermediate trajectory.

3) Trajectory Generation Module: We obtain the endpoint $\hat{Y}_i^{T_{pred}}$ of the trajectory through the Endpoint Prediction Module, the next goal is interpolating the intermediate trajectory $\{\hat{Y}_i^{t'} | (\hat{x}_i^{t'}, \hat{y}_i^{t'}), t' = T_{obs} + 1, T_{obs} + 2, ..., T_{pred} - 1\}$. To generate the trajectories of all pedestrians, the latent interactions among pedestrians are modeled through a social interaction module in the trajectory generation module. The structure of the trajectory generation module is provided by Fig. 2. To better interpolate intermediate trajectories, we fuse the

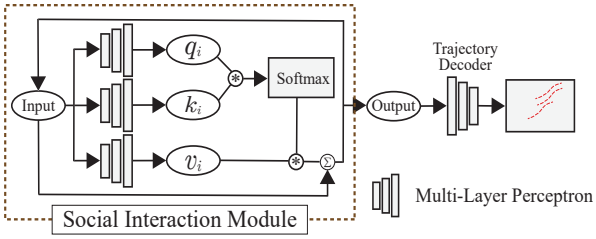

Fig. 2. Trajectory Generation Module. The social interaction features are generated by the social interaction module, and the intermediate positions are generated through the trajectory decoder.

features from the endpoints of future trajectories and the past trajectories. The fused features h_i are regarded as latent motion features to model the social interactions among pedestrians through the self-attention mechanism. For each feature h_i, we acquire its corresponding query vector as $q_i = f_Q(h_i)$, key vector as $k_i = f_K(h_i)$, and value vector $v_i = f_V(h_i)$, where $f_Q(\cdot)$, $f_K(\cdot)$ and $f_V(\cdot)$ represent MLPs. The self-attention mechanism is formalized as follows:

$$u_{i,j} = \frac{q_i \times k_i}{\sqrt{d_k}} \tag{1}$$

$$a_{i,j} = \frac{\exp u_{i,j}}{\sum_k \exp u_{i,k}} \tag{2}$$

$$h_i = \sum_k a_{i,k} \cdot v_k \tag{3}$$

where the d_k is the dimension of the query vector. The output h_i serves as input to the prediction decoder P_{future} to obtain the intermediate positions from the last observed position to the endpoint. The transmission of the message is performed several times for modeling the deep interactions among humans.

D. Discriminator

To get socially acceptable trajectories, the Generative Adversarial Network (GAN) is proposed to train the generators in PECGAN. The discriminator learns to distinguish the actual samples from the false ones, while the generator learns to generate the future trajectory.

The structure of the discriminator is given in Fig. 1. The discriminator of PECGAN includes the feature encoders and the classifier. As mentioned above, the generator of PECGAN has two components: the endpoint generation module and the trajectory generation module. The gradient optimization will also be performed on the two modules. Similar to the generator of PECGAN, the feature encoders of the discriminator consist of the trajectory encoder and the endpoint encoder. First the endpoints from ground truth or the generator are fed into the endpoint encoder to capture the endpoint features h_e. Concurrently, the intermediate positions before endpoint are fed into the trajectory encoder to capture motion features h_f. Accordingly, a multi-layer perceptron is applied as the classifier to obtain a classification score, such that:

$$h_e = E_{end}(Y_i^{T_{pred}}) \tag{4}$$

$$h_f = E_{future}(Y_i) \tag{5}$$

$$p_i = D_{cls}(h_e \oplus h_f) \tag{6}$$

where E_{end}, E_{future} and D_{cls} is a multi-layer perceptron with ReLU non-linearity, $\oplus$ is the concatenate operator.

E. Losses

To train our PECGAN, two loss terms are designed: the generator loss L_G for the generator fooling the discriminator, and the discriminator loss L_D for the discriminator correctly classifying the generator. Both loss terms are as follows:

$$L_G = D_{kl}(\mathcal{N}(\mu,\sigma)\|\mathcal{N}(0,1)) + \|Y_i - \hat{Y}_i\|^2 + \|Y_i^{t_{pred}} - \hat{Y}_i^{t_{pred}}\|^2 \tag{7}$$

$$L_D = \log D(X_i, Y_i) + \log(1 - D(X_i, \hat{Y}_i)) \tag{8}$$

where G represents the generator and D represents the discriminator, the Kullback-Leibler divergence D_{kl} pushes the approximated posterior $\mathcal{N}(\mu,\sigma)$ to the prior distribution $\mathcal{N}(0,1)$.

IV. Experiments

In this section, we assess our approach using three publicly available pedestrian trajectory datasets: ETH Dataset [13], UCY Dataset [14] and Stanford Drone Dataset [15]. Our model is compared with recent state-of-the-art baselines on these datasets.

A. Implementation Details

Our experiments on the three datasets are carried out under the same hardware environment, the processor is AMD Ryzen 7 3700X 8-Core Processor, and the graphics card is NVIDIA 2080Ti GPU. All sub-networks in our proposed model are MLPs with ReLU non-linearity. The dimensions of MLPs of each of the subnetworks are list in Table. I. The adam optimizer is used to iteratively train the generator and the discriminator. For the ETH-UCY dataset, the batch size is set to 256, the learning rates of generator and discriminator are set to 0.0002 and 0.0008, respectively, and it takes about 500 epochs for our network to converge. For the SDD dataset, the batch size is set to 512, the learning rates of generator and discriminator are set to 0.0001 and 0.0004, respectively, and it takes about 800 epochs for our network to converge. The social interaction module is performed 1 and 3 times on ETH-UCY and SDD, respectively.

TABLE I
THE DIMENSIONS OF MLPS OF EACH OF THE SUBNETWORKS USED IN THE MODULE.

		Network architecture
Generator	E_{past}	$16 \to 512 \to 256 \to 16$
	E_{end}	$2 \to 8 \to 16 \to 16$
	E_{latent}	$32 \to 8 \to 50 \to 32$
	D_{latent}	$32 \to 1024 \to 64 \to 128$
	f_Q, f_K	$32 \to 512 \to 64 \to 128$
	f_V	$32 \to 512 \to 64 \to 32$
	P_{future}	$32 \to 1024 \to 512 \to 1024 \to 22$
Discriminator	E_{future}	$22 \to 512 \to 256 \to 16$
	E_{end}	$2 \to 8 \to 16 \to 16$
	D_{cls}	$32 \to 64 \to 32 \to 1$

1) Evaluation Metrics: Similar to prior works [6], [21], the proposed method is evaluated with two types of metrics as follows:

1. *Average Displacement error(ADE)*: the Mean Square Error(MSE) between the predicted trajectory and the ground-truth trajectory of all predicted time steps.
2. *Final Displacement error(FDE)*: the Mean Square Error(MSE) between the predicted trajectory and the ground-truth trajectory at the last predicted time step.

2) Baselines: We compare PECGAN against several published baselines including previous state-of-the-art methods. S-LSTM [5] combines LSTM with a social pooling layer, which aggregates the hidden states of the neighbor pedestrians. S-GAN [6] leverages a GAN framework using pooling to model social interactions. SoPhie [16] predicts trajectories compliant to social and physical constraints based on S-GAN. Goal-GAN [11] proposes a two-stage trajectory prediction model, which first estimates the target and proposes a set of credible trajectories to the estimated target based on the recurrent neural network (CNN). TPNet [12] first generates potential future trajectories as proposals, then classifies and refines the proposal using a DNN-based model. NMMP [7] models the interactions and learn representations for direct interactions among actors via the neural motion message passing. Transformer-TF [22] is a novel trajectory prediction architecture via transformer. PECNet [9] predict multiple trajectories via CVAE model.

B. Quantitative Results

In this part, we compare our model with the above baselines on the ADE and FDE metrics on the ETH-UCY and Stanford Drone Dataset.

ETH-UCY: Table. II presents the quantitative results of our model on the ETH-UCY dataset. In the baselines, S-LSTM, S-GAN, SoPhie, NMMP and Transformer-TF all directly predict a complete trajectory based on the observation trajectory. However, Goal-GAN, TPNet and PECNet perform trajectory prediction based on endpoint targets. Compared with above baselines, our proposed PECGAN not only adopts a novel self-attention mechanism, but also assumes an adversarial strategy

TABLE II
QUANTITATIVE RESULTS OF THE SEVERAL RECENTLY PUBLISHED BASELINES AND OUR METHOD ON ETH-UCY DATASET.

Methods	Performance (ADE/FDE)					
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
S-LSTM [5]	1.09/2.35	0.79/1.73	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
S-GAN [6]	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.78	0.61/1.21
SoPhie [16]	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
Goal-GAN [11]	0.59/1.18	0.19/0.35	0.60/1.19	0.43/0.87	0.32/0.65	0.43/0.85
TPNet [12]	0.84/1.73	0.24/0.46	0.42/0.94	0.33/0.75	0.26/0.60	0.42/0.90
NMMP [7]	0.61/1.08	0.33/0.63	0.52/1.11	0.32/0.66	0.29/0.61	0.41/0.82
Transformer-TF [22]	0.61/1.12	**0.18**/0.30	0.35/0.65	**0.22**/0.38	0.17/0.32	0.31/0.55
PECNet [9]	**0.54/0.87**	**0.18**/0.24	0.35/0.60	**0.22**/0.39	0.17/0.30	0.29/0.48
Ours	0.57/0.90	**0.18/0.20**	**0.27/0.42**	0.23/**0.37**	**0.16/0.28**	**0.28/0.43**

TABLE III
QUANTITATIVE RESULTS OF THE BASELINES AND OUR MODEL ON STANFORD DRONE DATASET.

Methods	S-GAN	SoPhie	Goal-GAN	PECNet	Ours
K	20	20	20	20	20
ADE	27.33	16.27	12.20	9.96	**9.72**
FDE	41.44	29.38	22.10	15.88	**15.47**

to train the generator. Consequently, PECGAN has a better metric performance against baselines. As shown in Table. II, PECGAN pushes the state-of-the-art on FDE metric by 11.6%. Meanwhile, PECGAN pushes state-of-the-art performance on ADE metric by 3.6%. We obtain the best ADE metrics in HOTEL, UNIV & ZARA2 and the best FDE metric in HOTEL, UNIV, ZARA1 & ZARA2.

Stanford Drone Dataset: As shown in Table. III, we compare PECGAN on the Stanford Drone Dataset with other baselines on ADE and FDE metrics. We find that our proposed method is 2.5%/2.7% and 25.5%/42.9% higher than the second & third methods on ADE/FDE metrics respectively. This result clearly demonstrates that PECGAN is effective in predicting pedestrian trajectories.

C. Qualitative Results

In order to visually compare PECGAN with the state-of-the-art PECNet, we visualize the predicted trajectories of both methods in Fig. 3. Fig. 3(a) shows the trajectory of a single pedestrian. For this case in the third row of Fig. 3(a), during the movement, the sudden change of direction of the target pedestrian increases the risk of traffic accidents and also brings challenges to the prediction. In this scenario, our method takes the occurrence of this situation into account, and predicts the corresponding trajectory. Fig. 3(b) shows the trajectories of pedestrians with the same heading direction. From the third row of Fig. 3(b), one can observe that the target pedestrian often needs to avoid obstacles during the movement, which also increases the difficulty of prediction. In this scenario, our model will generate the continuous trajectories in the process of avoiding pedestrians standing still, which are more similar to the ground truth than the PECNet's predicted trajectories. Fig. 3(c) shows the trajectories of pedestrians with the opposite direction of travel. In this scenario, the social interaction among pedestrians will be more intricate and more difficult. Compared with PECNet, PECGAN still predict the future trajectories which are closer to the ground truth.

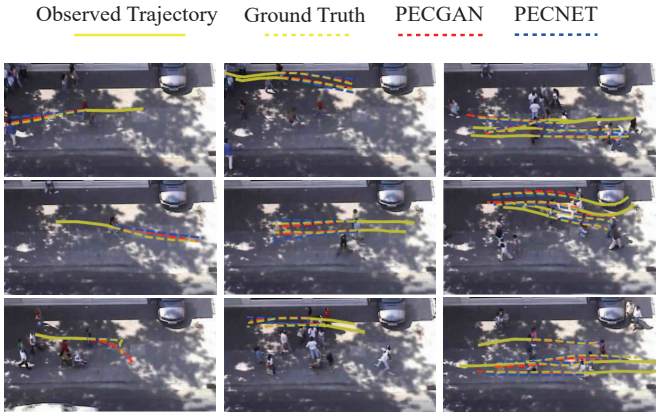

Fig. 3. Trajectory visualization of ZARA2 scenario in UCY dataset. (a) Single pedestrian; (b) Pedestrians in the same direction; (c) Pedestrians in the opposite direction. The solid line represents the observed trajectory (3.2s), and the dotted line represents the future trajectory (4.8s). (Yellow: ground truth; red: PECNet; blue: PECGAN).

D. Ablation Study

To capture the impacts of pedestrian intent and social interaction on target pedestrians, a two-stage strategy is proposed to generate the future trajectory in the generator of PECGAN. The first stage generates the predicted endpoint according to a CVAE and the intermediate positions are generated by the self-attention mechanism of the second stage. To demonstrate the effectiveness of the CVAE and the self-attention mechanism, the influence of each approach on the experimental results needs to be further verified. In this section, two sets of ablation experiments are carried out on SDD, W/O-CVAE experiment and W/O-SELF-ATTENTION experiment. In the W/O-CVAE experiment, predicted endpoints are generated without CVAE. During the training phase, the potential variable z is directly sampled from $\mathcal{N}(0,1)$ and fed into the endpoint decoder by fusing with motion features to generate future endpoints. In the W/O-SELF-ATTENTION experiment, intermediate positions are generated without the self-attention mechanism. During the training phase, the fused features of the past trajectory feature and the endpoint feature are considered to be the input into the prediction decoder to obtain the intermediate positions.

As shown in Table IV, the results of both groups of ablation experiments were worse than those of PECGAN, which proved the effectiveness of both approaches. Furthermore, according to the experimental deviation, one can see that the CVAE has a greater influence on the experimental results, which also demonstrates the necessity of the two-stage strategy. In the two-step strategy, we first generate the predicted endpoints of pedestrians, and then generate the intermediate coordinates purposefully.

TABLE IV
ABLATION STUDY ON PECGAN. W/O MEANS THAT THE METHOD IS NOT USED IN THE EXPERIMENT, W MEANS THAT THE METHOD IS USED IN THE EXPERIMENT

Methods	CVAE		SELF-ATTENTION	
	w/o	w	w/o	w
ADE	18.52	**9.72**	9.99	**9.72**
FDE	34.89	**15.47**	16.13	**15.47**

V. CONCLUSION AND DISCUSSION

In this paper, we propose Predicted Endpoint Conditioned Generative Adversarial Network(PECGAN) for multiple trajectory prediction. A two-stage strategy is adopted to estimate future trajectory from the past trajectory. In the first stage, we utilize the endpoint of ground truth as the condition to estimate the endpoint via a CVAE model. In the second stage, the estimated endpoint features are combined with past trajectory features, and are refined by social interactions modeling via self-attention mechanism. After that, the refined motion features infer the intermediate trajectory more accurately. An adversarial strategy is adopted to obtain the predicted endpoints distribution during training generator. Compared with the baselines, PECGAN achieves the state-of-the-art performance under ADE and FDE metrics on both Stanford Drone Dataset and ETH-UCY datasets.

The proposed PECGAN model has achieved better results in predicting future trajectories. However, the scene information is not taken into account during inferring future targets. In the future, we will try to extract scene information through CNN and combine it with motion features to estimate the destinations of pedestrians.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61972128), the Natural Science Foundation of Anhui Province (Grant No.1808085MF176), the Fundamental Research Funds for the Central Universities of China (PA2021KCPY0050).

REFERENCES

[1] A. S. Öğrencı, "Anomaly detection in walking trajectory," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*. Izmir: IEEE, 2018, pp. 1–4.

[2] M.-j. Lee and Y.-g. Ha, "Autonomous driving control using end-to-end deep learning," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. Busan: IEEE, 2020, pp. 470–473.

[3] K. Wu, W. Han, M. Abolfazli Esfahani, and S. Yuan, "Learn to navigate autonomously through deep reinforcement learning," *IEEE Transactions on Industrial Electronics*, pp. 1–11, 2021.

[4] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, pp. 4282–4286, 05 1998.

[5] A. Alahi, K. Goel, V. Ramanathan, and et.al., "Social LSTM: human trajectory prediction in crowded spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016, pp. 961–971.

[6] A. Gupta, J. Johnson, L. Fei-Fei, and et.al., "Social GAN: socially acceptable trajectories with generative adversarial networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City: IEEE, 2018, pp. 2255–2264.

[7] Y. Hu, S. Chen, Y. Zhang, and et.al., "Collaborative Motion Prediction via Neural Motion Message Passing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 6318–6327.

[8] C. Yu, X. Ma, J. Ren, and et.al., "Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction," in *Computer Vision – ECCV 2020*, vol. 2. Cham: Springer International Publishing, 2020, pp. 507–523.

[9] K. Mangalam, H. Girase, S. Agarwal, and et.al., "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 759–776.

[10] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian prediction by planning using deep neural networks," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5903–5908, 2017.

[11] P. Dendorfer, A. Ošep, and L. Leal-Taixé, "Goal-GAN: Multimodal Trajectory Prediction Based on Goal Position Estimation," in *Computer Vision – ACCV 2020*, vol. 12623 LNCS, pp. 405–420.

[12] L. Fang, Q. Jiang, J. Shi, and B. Zhou, "Tpnet: Trajectory proposal network for motion prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 6796–6805.

[13] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 452–465.

[14] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[15] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 549–565.

[16] A. Sadeghian, V. Kosaraju, A. Sadeghian, and et.al., "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, 2019, pp. 1349–1358.

[17] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in Neural Information Processing Systems*, vol. 4, pp. 3581–3589, 2014.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014, pp. 1–9.

[19] K. D. Katyal, G. D. Hager, and C.-M. Huang, "Intent-aware pedestrian prediction for adaptive crowd navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3277–3283.

[20] M. K. C. Tay and C. Laugier, "Modelling smooth paths using gaussian processes," in *Field and Service Robotics: Results of the 6th International Conference*. Berlin, Heidelberg: Springer, 2008, pp. 381–390.

[21] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 12 077–12 086.

[22] F. Giuliari, I. Hasan, M. Cristani, and et.al., "Transformer Networks for Trajectory Forecasting," in *2020 25th International Conference on Pattern Recognition*. Milan: IEEE, 2020, pp. 10 335–10 342.