Sliding Sequential CVAE with Time Variant Socially-aware Rethinking for Trajectory Prediction

Hao Zhou, Dongchun Ren, Xu Yang, Mingyu Fan, Hai Huang

Abstract—Pedestrian trajectory prediction is a key technology in many applications such as video surveillance, social robot navigation, and autonomous driving, and significant progress has been made in this research topic. However, there remain two limitations of previous studies. First, with the continuation of time, the prediction error at each time step increases significantly, causing the final displacement error to be impossible to ignore. Second, the prediction results of multiple pedestrians might be impractical in the prediction horizon, i.e., the predicted trajectories might collide with each other. To overcome these limitations, this work proposes a novel trajectory prediction method called CSR, which consists of a cascaded conditional variational autoencoder (CVAE) module and a socially-aware regression module. The cascaded CVAE module first estimates the future trajectories in a sequential pattern. Specifically, each CVAE concatenates the past trajectories and the predicted points so far as the input and predicts the location at the following time step. Then, the socially-aware regression module generates offsets from the estimated future trajectories to produce the socially compliant final predictions, which are more reasonable and accurate results than the estimated trajectories. Moreover, considering the large model parameters of the cascaded CVAE module, a slide CVAE module is further exploited to improve the model efficiency using one shared CVAE, in a slidable manner. Experiments results demonstrate that the proposed method exhibits improvements over state-of-the-art method on the Stanford Drone Dataset (SDD) and ETH/UCY of approximately 38.0% and 22.2%, respectively.

Index Terms—sliding sequence prediction, time variant socially-aware rethinking, trajectory prediction.

I. INTRODUCTION

The prediction of future pedestrian trajectories according to their past trajectories is a key problem for many applications such as autonomous driving [1]–[3], robotic systems [7], [4], and surveillance system [5], [6]. For example, in autonomous driving scenario, a accurate prediction of pedestrians trajectories is required before the vehicle can plan a safe and effective trajectory; otherwise, a car accident may occur.

A key challenge for pedestrian trajectory prediction is the high degree of uncertainty caused by two factors, i.e., the inherently multimodal attribute and the complex social interactions of pedestrians. In Figure 1, the uncertainties of one state-of-the-art method PECNet [7] are visualized using the average displacement error (ADE) between the ground truth and predictions. The visualization reveals that the degree of uncertainty increases over the prediction horizon. This is expected, as the movement patterns and social interactions change over time; thus, it is difficult to predict the long-term trajectories using the past trajectories. To cope with the uncertainty problem, numerous methods [7]–[9] have been proposed in recent years. The core notion of these methods is to first model the target prediction, and then to predict the future trajectory conditioned on the predicted target. Despite their competitive performance, the prediction of the target point, which has a tremendous influence on the final prediction, is inherently a difficult problem that has the maximum degree of uncertainty. For example, two different target points will lead to the generation of two completely different trajectories.

In Figure 1, the uncertainty can also be analyzed from another perspective, namely that the first prediction time step has the minimum degree of uncertainty; thus, it can be easily captured by the past trajectory and has the best prediction...
accuracy. This is also expected because the first prediction time step has a movement pattern similar to that of the past trajectories. Accordingly, in this paper, interest has been placed in the development of a model that can sequentially predict the following time step using the updated past trajectory as the input. Recurrent Neural Networks (RNN)-based sequential methods [10]–[14] are similar to the idea explored in this study. They use the hidden states of the predicted points before the prediction of the next point as the input. However, the hidden state feature of the decoder cannot explicitly encode the predicted points as the past trajectories.

This work presents a novel cascaded conditional variational autoencoder (CVAE) with socially-aware rethinking method, as illustrated in Figure 2. The proposed method consists of two modules, namely the cascaded CVAE module and the socially-aware regression module. In the cascaded CVAE module, unshared CVAEs are used to sequentially predict the future points of different time steps using the updated past trajectories as the input. The past trajectory after each prediction is updated by concatenating it with the newly predicted point at the current time. Via this strategy, the predicted points are treated as pseudo-labels to decrease the degree of prediction uncertainty of the next point, thereby improving the long-term trajectory prediction performance. The average displacement error of the proposed method is also visualized in Figure 1, from which it is evident that the proposed method is characterized by the decreases of the average displacement errors of long-term future points.

Moreover, the improvement is also partially attributed to the socially-aware regression module. The future trajectories predicted by the cascaded CVAE module may be socially unacceptable because the CVAE can not model interactions among pedestrians. Thus, the socially-aware regression module is proposed to refine the predicted trajectories via social interactions. This module first extracts features of the past trajectories and the predicted trajectories by two encoders. Then the two features are concatenated and fed to a social attention mechanism to extract the interaction feature. Finally, the decoder decodes the interaction feature to produce the offsets of the future trajectories, which are then used to refine the future trajectories predicted by the cascaded CVAE module.

In the cascaded CVAE module, $\delta$ unshared CVAEs are used to predicted $\delta$ future points using incremental inputs. Thus, the proposed module has large model parameters and slow inference speed. Actually, the predictions of different future points fit the independent and identically distributed. Thus, one CVAE is enough to predict $\delta$ future points. Based on this observation, a more lightweighted slide CVAE module is introduced to adopt one shared CVAE to predict future points using the updated past trajectory of fixed length, in a slide manner, as illustrated in Figure 5c. Moreover, the proposed slide CVAE module is integrated with the socially-aware regression module resulting in an more efficiency end-to-end trajectory prediction method, called the S-CSR. The average displacement error of the S-CSR is also visualized in Figure 1, from which it is indicates that the slide CVAE module can further improve the prediction accuracy of the long-term future points.

The contributions are summarized as follows:

- A cascaded CVAE with socially-aware rethinking method is proposed for pedestrian trajectory prediction. The proposed method consists of the cascaded CVAE module or slide CVAE module and the socially-aware regression module.
- The proposed cascaded CVAE module uses unshared CVAEs to sequentially predict the future points using the updated past trajectories as the input.
- The proposed socially-aware regression module extracts interaction features to generate offsets to refine the future trajectories predicted by the cascaded CVAE module.
- The proposed slide CVAE module uses one shared CVAE to predict the future points using the updated past trajectory of fixed length, in a slideable manner.
- The results of extensive experiments demonstrate that the proposed CSR and S-CSR surpass state-of-the-art methods by a large margin on two benchmark datasets.

II. RELATED WORK

A. Direct Prediction Methods

The direct strategy involves the utilization of one model that is capable of predicting the entire future trajectory in a one-shot manner. Most trajectory prediction methods are characterized by direct prediction. For example, early works adopted the Bayesian network method [15], the Gaussian process regression method [16], and the kinematic method [17] to directly predict the future sequence. However, these method did not consider the interactions between agents, and usually failed in crowded scenarios.

Recently, with the development of deep learning, convolutional neural network (CNNs) and multilayer perceptron (MLPs) have been used in trajectory prediction. For example, GRIP [18] uses a temporal graph to model the interactions and a temporal CNN to predict the future sequences. VectorNet [19] utilizes self-attention to aggregate interactions and the MLP to predict future sequences. Social-STGCN [20] and AST-GNN [21] capture interactions using a graph attention (GAT) and predict future sequences using a temporal CNN.

Albert it is simple and efficient, the direct strategy uses only the past trajectory for prediction, but has no opportunity to extract useful information from the predicted future points. In contrast, the proposed method uses the predicted points to update the past trajectory, thereby improving the prediction accuracy of the following time step.

B. Recursive Prediction Methods

The recursive strategy involves the use of a one-step model multiple times, in which the predictions for the prior time steps are used for the prediction of the following time step. RNN models including long short-term memory (LSTM) [22] and gated recurrent unit (GRU) [23] have been widely adopted in the recursive prediction methods. For example, S-LSTM [10] uses LSTM to extract the past trajectories, and then aggregates the interactions of different pedestrians in the
Fig. 2: The overview of the proposed CSR model. First, the cascaded CVAE module estimates the future trajectory in a sequential manner. Specifically, each CVAE concatenates the past trajectory and the predicted points so far as the input and predicts the location at the following time step. Then, the socially-aware regression module extracts interaction features to refine the predicted trajectory.

hidden state using a social pooling mechanism. CF-LSTM [14] improves LSTM by employing a cascaded-feature that can simultaneously capture location and velocity information. S-GAN [11] extends S-LSTM by using GAN to generate multimodal predictions. SoPhie [12] and PIF [13] use a CNN to extract scene features and LSTM to extract motion features, and then adopt LSTM to predict scene-compliant trajectories. Trajectron++ [24] introduces a modular graph-structured recurrent model that can integrate agent dynamics and heterogeneous data into trajectory prediction. Although the hidden state of the predicted points has been adopted to assist in the prediction of the following time step, it does not explicitly model the predicted points as the past trajectories.

C. Target-conditioned Methods

The target-conditioned strategy first predicts the target points and then generates the future trajectories conditioned on the targets. For example, PECNet [7] first predicts distant trajectory endpoints using a CVAE and then generates multimodal predictions conditioned on the endpoints. TNT [8] improves PECNet by introducing a scoring and selection stage that ranks and selects multimodal predictions with likelihood scores. Y-net [9] further improves PECNet by iteratively predicting intermediate waypoints and trajectories. The performance of target-conditioned methods is easily affected by the predicted targets, however, target prediction is inherently a difficult problem with a high degree of uncertainty.

III. METHODOLOGY

Pedestrian trajectory prediction is a task to predict the future trajectories of pedestrians based on their past trajectories. First, it is assumed that videos of pedestrians walking are preprocessed to obtain the spatial coordinates of pedestrians. For example, \( p^i_t = (x^i_t, y^i_t) \) represents the coordinate of the \( i \)-th pedestrian at the \( t \)-th time step. Then, the coordinates of each pedestrian in the scene are classified into the past and future trajectories according to two default values, namely the observation horizon \( \tau \) and the prediction horizon \( \delta \). The definitions of the past trajectory \( X_i \) and the future trajectory \( Y_i \) are as follows:

\[
X_i = \left\{ (x^i_t, y^i_t) | \forall t \in \{1, 2, \dots, \tau\} \right\},
\]

(1)

\[
Y_i = \left\{ (x^i_t, y^i_t) | \forall t \in \{\tau + 1, \tau + 2, \dots, \tau + \delta\} \right\}.
\]

(2)

Finally, given the past trajectories \( \{X_i | \forall i \in \{1, 2, \dots, N\}\} \) of pedestrians in scene, the goal is to generate future predictions \( \{\hat{Y}_i | \forall i \in \{1, 2, \dots, N\}\} \) that minimize the average displacement error with the ground truth future trajectories \( \{Y_i | \forall i \in \{1, 2, \dots, N\}\} \), where \( N \) represents the number of pedestrians in a scene. The definition of \( \hat{Y}_i \) is as follows:

\[
\hat{Y}_i = \left\{ (x^i_t, y^i_t) | \forall t \in \{\tau + 1, \tau + 2, \dots, \tau + \delta\} \right\}.
\]

(3)

To generate reliable predictions, a pedestrian trajectory prediction method that consists of a cascaded CVAE module and a socially-aware regression module is proposed, as shown in Figure 2. Because the direct prediction of long-range future trajectory \( Y_i \) has a high degree of uncertainty, a cascaded CVAE module is first designed to predict the future points in a sequential manner using the updated past trajectories as the input. Then, considering that the CVAE is unable to model social interactions, a socially-aware regression module is proposed to further refine the predicted future trajectories using social interaction knowledge.
Fig. 3: The illustration of the cascaded CVAE unit. The CVAE uses the updated past trajectory and ground truth future point to train a model for multimodal future point prediction. The dots and circles respectively represent the predicted future points and the ground truth future points. Red dotted lines denote the layers utilized only during training.

A. Cascaded CVAE Module

The prediction of the future trajectories is inherently ambiguous and has a high degree of uncertainty. Thus, it is difficult to learn a deterministic function $f$ that directly maps the past trajectories $\{X_i | \forall i \in \{1, 2, \ldots, N\}\}$ to the future trajectories $\{Y_i | \forall i \in \{1, 2, \ldots, N\}\}$.

However, as described in the Introduction, the degrees of the uncertainty of the future points are increase with the continuation of time, and the first future point has the minimum degree of uncertainty. According to this phenomenon, the use of cascaded CVAEs is proposed to decompose the prediction of the future trajectories into the prediction of the future points using the updated past trajectories, in a sequential manner, as illustrated in Figures 2 and 5b.

The detailed structure of each cascaded CVAE is illustrated in Figure 3. The CVAE infers a distribution of future points based on the updated past trajectory. The training of the CVAE requires two inputs, namely the updated past trajectory $\hat{X}_t^i$ and the ground point $p_t^i = (x_t^i, y_t^i)$ of pedestrian $i$ at future time step $t$. The updated past trajectory consists of the past trajectory and the predicted future points before time step $t$ using the concatenation operation, such that:

$$\hat{X}_t^i = \text{Concat} (x_i, \{ (\hat{x}_t^i, \hat{y}_t^i) | \forall s \in \{\tau + 1, \tau + 2, \ldots, t - 1\}\}).$$

(4)

The updated past trajectory $\hat{X}_t^i$ and the ground truth future point $p_t^i$ are first encoded using two MLP encoders $E_{upast}$ and $E_{point}$, as follows:

$$f_{upast} = E_{upast}(\hat{X}_t^i),$$

(5)

$$f_{point} = E_{point}(p_t^i).$$

(6)

Then, the two feature representations $f_{upast}$ and $f_{point}$ are concatenated together and fed to the MLP encoder $E_{latent}$ to generate parameters $(\mu_t^i, \sigma_t^i)$ of the latent distribution, such that:

$$(\mu_t^i, \sigma_t^i) = E_{latent}(\text{Concat}(f_{upast}, f_{point})).$$

(7)

Finally, the latent variable $z_t^i$ is randomly sampled from distribution $\mathcal{N}(\mu_t^i, \sigma_t^i)$, and is concatenated with the feature $f_{opast}$ and yield prediction $\hat{p}_t^i = (\hat{x}_t^i, \hat{y}_t^i)$ using the MLP decoder $D_{latent}$, as follows:

$$\hat{p}_t^i = D_{latent}(\text{Concat}(z_t^i, f_{point})).$$

(8)

In the inference stage, because the ground truth point $p_t^i$ is unavailable, the latent variable $z_t^i$ is randomly sampled from a prior distribution $\mathcal{N}(0, I)$, and concatenated with the feature $f_{upast}$, and the prediction $\hat{p}_t^i$ is then generated using the trained decoder $D_{latent}$.

For each pedestrian $i$, there are $\delta$ future time steps. Thus, $\delta$ CVAEs with the same structure are sequentially operated to generate the predictions of $\delta$ future points $\hat{p}_{t+\delta}^i, \hat{p}_{t+2\delta}^i, \ldots, \hat{p}_{t+\delta}^i$. It should be noted that model parameters for the CVAE, including $E_{upast}$, $E_{point}$, $E_{latent}$ and $D_{latent}$, are not shared because the length of the updated past trajectory of each CVAE is different, and the CVAEs are used to predict future points at different future time steps.

B. Socially-aware Regression Module

The CVAE cannot model interactions among different pedestrians; thus, the future trajectories predicted by the cascaded CVAE module may rough and socially unacceptable. To cope with this problem, a socially-aware regression module, as illustrated in Figure 2, is proposed to estimate the offsets to refine the predicted future trajectories.

To generate effective offsets, two MLP encoders $E_{opast}$ and $E_{pfuture}$ are first used to extract features from the original past trajectory and the predicted future trajectory, as follows:

$$f_{opast} = E_{opast}(X_i),$$

(9)

$$f_{pfuture} = E_{pfuture}(Y_i).$$

(10)

Then, the two feature representations $f_{opast}$ and $f_{pfuture}$ are concatenated and passed into a social interaction layer $SE$ to produce an interaction feature, such that:

$$f_{st} = SE(\text{Concat}(f_{opast}, f_{pfuture})).$$

(11)
Finally, the extracted feature $f_{st}$ is fed to the MLP decoder $D_{offsets}$ to estimate the offsets of the predicted future trajectories:

$$\Delta \hat{Y}_i = D_{offsets}(f_{st}),$$

(12)

where $\Delta \hat{Y}_i = \{(\Delta \hat{x}_i^t, \Delta \hat{y}_i^t) | \forall t \in \{\tau + 1, \tau + 2, ..., \tau + \delta\}\}$. Using the estimated offsets, the predictions $\hat{Y}_i$ is refined by $Y_i + \Delta \hat{Y}_i$.

In this paper, the social non-local pooling layer proposed by [7] is adopted as the SE layer. This method integrates the social pooling [10] into the non-local attention mechanism [25], as illustrated in Figure 4. Specifically, three MLPs are first used to extract the query, key, and value features. Then, a matrix multiplication is applied between the query and key to generate a social interaction feature. Finally, a weighted sum between the masked attention map and the value feature is conducted to generate a social interaction feature.

C. Slide CVAE Module

Even though the updated past trajectory that contains the predicted future points before current time step is adopted by cascade CVAE module to assist in the predicting of the future point at the following time step. There exists two drawbacks of the proposed module. First, $\delta$ CVAEs are needed to predict $\delta$ future points, resulting in large model parameters. Second, with the continuation of time, the updated past trajectory is increase, and thus decrease the inference speed of CVAE. Actually, the predictions of different future points fit the independent and identically distributed. Thus, one CVAE is enough to predict $\delta$ future points. Based on this observation, different from the cascaded CVAE module, a slide CVAE module is proposed to decompose the prediction of the future trajectories into the prediction of the future points using the updated past trajectories of fixed length, in a slidable manner, as illustrated in Figure 5c. In this manner, the model parameters are decreased by using only one CVAE, and the inference speed is increased by using the updated past trajectory of fixed length.

The proposed slide CVAE has the same structure of each cascaded CVAE, except for the input. The slide CVAE module uses the updated trajectories of fixed length. Specifically, at the time step $t$, the input of slide CVAE is consists of the original past trajectory and the predicted points before current time step. To keep the fixed length of the updated past trajectory, the input $\tilde{X}_i^t$ of the slide CVAE module is defined as Equation 13, where $\alpha$ is the fixed length of the updated past trajectory.

D. Training the Model

To train the $\delta$ CVAEs in the cascaded CVAE module, two loss terms, namely the average point loss $L_{AP}$ and the Kullback-Leibler (KL) divergence loss $L_{KLD}$, are used. The definitions of $L_{AP}$ and $L_{KLD}$ are as follows:

$$L_{AP} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=\tau+1}^{\tau+\delta} \|p_i^t - \hat{p}_i^t\|^2,$$  

(14)

$$L_{KLD} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=\tau+1}^{\tau+\delta} D_{KL}(\mathcal{N}(\mu_i^t, \sigma_i^t) || \mathcal{N}(0, I)).$$  

(15)

The KL divergence loss is a regularization loss that measures the distance between the sampling distribution at the test stage and the sampling distribution of the latent variable learned at the training stage, and is used to train the CVAEs. The average point loss measures the displacement error between the predicted future points and the ground truth future points, and is used to train $E_{upath}$, $E_{point}$, $E_{latent}$, and $D_{latent}$. In addition, a regression loss $L_R$ is used to train the socially-aware regression module. The definition of $L_R$ is:

$$L_R = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=\tau+1}^{\tau+\delta} \|p_i^t - \hat{p}_i^t - \Delta \hat{p}_i^t\|,$$  

(16)

where $\Delta \hat{p}_i^t = (\Delta \hat{x}_i^t, \Delta \hat{y}_i^t)$. The regression loss measures the displacement between the predicted offsets and the real offsets, and is used to train $E_{past}$, $E_{future}$, and $D_{offsets}$. Finally, the total loss $L_T$ is defined as follows:

$$L_T = L_{KLD} + L_{AP} + L_R.$$  

(17)

This multi-task loss is used to train the entire model in an end-to-end manner.
The proposed CSR model was compared with the aforementioned baseline methods on the SDD and ETH/UCY datasets.

**Baselines:** The proposed method was compared with the following state-of-the-art baseline methods:

- **S-LSTM** [10]. Each pedestrian is modeled using an LSTM, and the interactions are extracted by pooling the hidden states among neighbors.
- **DESIRE** [33]. An inverse optimal control (IOC)-based trajectory planning method is used to refine the predicted trajectories.
- **S-GAN** [11]. This method improves S-LSTM by introducing a GAN to generate multimodal prediction results.
- **SoPhie** [12]. This method improves GAN-based methods by applying an attention to model social relationships and physical constraints.
- **P2TIRL** [27]. A maximum entropy inverse reinforcement learning strategy is introduced to learn a grid-based trajectory prediction method.
- **CF-VAE** [34]. A conditional normalizing flow-based prior is proposed to improve the VAE for the generation of effective predictions.
- **CGNS** [35]. Conditional variational divergence minimization and conditional latent space learning are used to predict the future trajectory.
- **CF-LSTM** [14]. A feature-cascaded method that can simultaneously capture location and velocity informations is proposed to improve LSTM.
- **Trajotron++** [24]. This method presents a modular graph-structured recurrent model that can interpret agent dynamics and heterogeneous data.
- **PECNet** [7]. This method first predicts distant trajectory endpoints and then generates the future predictions conditioned on endpoints.
- **TNT** [8]. This method improves PECNet using a scoring and selection stage that ranks and selects multimodal predictions with likelihood scores.
- **Y-net** [9]. This method improves PECNet and TNT by iteratively predicting intermediate waypoints and trajectories.

**C. Quantitative Analysis**

The proposed CSR model was compared with the aforementioned baseline methods on the SDD and ETH/UCY datasets.
| Methods   | S-GAN | CF-VAE | P2TIRL | SimAug | PECNet | Y-net* | CSR | S-CSR |
|-----------|-------|--------|--------|--------|--------|--------|-----|-------|
| ADE       | 27.23 | 12.60  | 12.58  | 10.27  | 9.96   | 7.85   | 4.87| 2.77  |
| FDE       | 41.44 | 22.30  | 22.07  | 19.71  | 15.88  | 11.85  | 6.32| 3.45  |

(a) Best of 20 samples

| Methods   | DESIRE | TNT   | PECNet | Y-net* | CSR | S-CSR |
|-----------|--------|-------|--------|--------|-----|-------|
| ADE       | 19.25  | 12.79 | 12.23  | 11.49  | 8.38| 6.02  |
| FDE       | 34.05  | 29.58 | 21.16  | 20.23  | 13.43| 8.74  |

(b) Best of 5 samples

TABLE II: The comparison of the proposed CSR and S-CSR models with several baseline methods and previous state-of-the-art methods (labeled by *) on the SDD dataset. (a) Results are tested using best of 5 samples. (b) Results are tested using best of 20 samples. The best results are highlighted in bold.

| Scenes | S-LSTM | S-GAN | Sophie | CGNS | PECNet | Trajectron++ | Y-net* | CSR | S-CSR |
|--------|--------|-------|--------|------|--------|--------------|--------|-----|-------|
| ETH    | 0.70   | 1.40  | 1.00   | 1.52 | 0.70   | 1.43         | 0.62   | 1.40| 0.54  |
| ADE    | 0.37   | 1.75  | 0.72   | 1.61 | 0.76   | 1.67         | 0.90   | 0.93| 0.18  |
| FDE    | 0.52   | 2.24  | 0.96   | 1.22 | 0.38   | 0.48         | 0.12   | 0.12| 0.24  |
| ADE-FDE| 0.50   | 2.18  | 0.87   | 1.10 | 0.39   | 0.44         | 0.11   | 0.11| 0.20  |
| FDE    | 0.20   | 0.60  | 0.23   | 0.39 | 0.35   | 0.20         | 0.24   | 0.24| 0.24  |
| ADE    | 0.35   | 0.60  | 0.13   | 0.17 | 0.36   | 0.44         | 0.27   | 0.27| 0.13  |
| FDE    | 0.18   | 0.06  | 0.12   | 0.27 | 0.24   | 0.14         | 0.25   | 0.25| 0.05  |
| CSR    | 0.51   | 0.15  | 0.39   | 0.60 | 0.49   | 0.19         | 0.41   | 0.41| 0.05  |
| S-CSR  | 0.54   | 0.18  | 0.41   | 0.60 | 0.39   | 0.18         | 0.25   | 0.25| 0.10  |

TABLE III: The comparison of the proposed CSR and S-CSR models with several baseline methods and previous state-of-the-art methods (labeled by *) on the ETH/UCY dataset. Results are tested using best of 20 samples. The best results are highlighted in bold.

| Modules | BL | CCM | SCM | SRM |
|---------|----|-----|-----|-----|
| EADE    | 0.48 | 0.94 | 0.19 | 0.35|
| FDE     | 0.33 | 0.57 | 0.08 | 0.10|
| ADE     | 0.30 | 0.61 | 0.22 | 0.42|
| FDE     | 0.28 | 0.36 | 0.09 | 0.10|
| ADE     | 0.16 | 0.32 | 0.27 | 0.53|
| FDE     | 0.07 | 0.09 | 0.17 | 0.24|
| ADE     | 0.05 | 0.09 | 0.14 | 0.23|
| FDE     | 0.05 | 0.08 | 0.10 | 0.16|

TABLE IV: Ablation study of different modules of the proposed method on the ETH/UCY dataset. BL represents the introduced baseline method. CCM denotes the proposed cascaded CV AE module. SCM denotes the proposed slide CV AE module. SRM denotes the proposed socially-aware regression module.

| Modules | Performance | Targets | ADE | FDE |
|---------|-------------|--------|-----|-----|
| BL      |             |        | 11.56| 15.71|
| ✓       |             |        | 6.12 | 7.87 |
| ✓       |             |        | 4.87 | 6.32 |
| ✓       |             |        | 2.77 | 3.45 |

TABLE V: Ablation study of different modules of the proposed method on the SDD dataset. BL represents the introduced baseline method. CCM denotes the proposed cascaded CV AE module. SCM denotes the proposed slide CV AE module. SRM denotes the proposed socially-aware regression module.

via the ADE and FDE metrics.

1) Comparison with VAE-based Methods: The VAE is widely used to generate the latent variables of trajectory distributions, and many VAE-based prediction methods have been proposed. Unlike these methods, the proposed method uses multiple VAEs to predict the future points, in a cascaded manner. The proposed CSR was compared with several VAE-based methods, including DESIRE, CF-VAE, and CGNS. As reported in Table II, on the SDD, CSR respectively surpassed CF-VAE and DESIRE by 61.3%/71.7% and 56.5%/60.6% in terms of the ADE/FDE metrics. Moreover, as reported in Table III, CSR outperformed CGNS by 71.4%/76.3% on the ETH/UCY. This comparison validates that the proposed CSR, which utilizes VAEs in a cascaded manner, performs better than VAE-based methods.

2) Comparison with RNN-based Methods: The proposed CSR predicts the future trajectory in a sequential manner, which is similar to RNN-based methods. However, RNN based methods do not explicitly encode the predicted points as the past trajectories. To validate the updated past trajectory used in the proposed model, it was compared with several RNN-based methods, including DESIRE, CF-VAE, CGNS and Trajectron++. As shown in Table II, on the SDD, the proposed CSR exhibited respective improvements over DESIRE and CF-VAE of 56.5%/60.6% and 61.3%/71.7% in terms of the ADE/FDE metrics. Moreover, as shown in Table III, the proposed CSR exhibited respective improvements over CGNS and Trajectron++ of 71.4%/76.3% and 26.3%/52.1% on the ETH/UCY dataset. This comparison verifies that the proposed method, which explicitly encodes the predicted points as the past trajectories, achieves superior performance as compared to RNN-based methods.

3) Comparison with Target-conditioned Methods: Target-conditioned trajectory prediction methods have recently become popular. These methods first predict the target point and
Inference time

Parameters

SRM

Performance

Modules | SCM | CCM | Parameters | Inference time

| ✓ | ✓ | ✓ | 15.88 M | 0.32 S/B

| ✓ | ✓ | ✓ | 2.24 M | 0.30 S/B

Fig. 6: Visualization of the trajectories predicted by the cascaded CVAE module with or without the socially-aware regression module. The visualized trajectories are best predictions sampled from 20 trials. The white dots, yellow dots, green dots, and red dots respectively represent the past trajectories, ground truth future trajectories, predictions without using the socially-aware regression module, and predictions using the socially-aware regression module.

TABLE VI: Analysis of the model efficiency of the cascaded CVAE module and the slide CVAE module on the SDD dataset. CCM denotes the proposed cascaded CVAE module. SCM denotes the proposed slide CVAE module. SRM denotes the proposed socially-aware regression module. "S" represents million and "S/B" represents seconds per batch. All experiments are evaluated on an NVIDIA 1080TI GPU.

1) Cascaded CVAE Module: The proposed cascaded CVAE module predict future trajectory in a cascaded manner, in which each CVAE sequentially predict a future point using a updated past trajectory that consists of the original past trajectory and the predicted future points before the current time step. To prove the effect of the cascaded CVAE module, it is compared with a baseline method which \( \delta \) CVAEs that using the original past trajectory as the input are respectively used to predict \( \delta \) future points. As reported in Tables IV and V, the cascaded CVAE module outperforms the baseline by 37.0%/54.7% on the ETH/UCY and 47.1%/49.9% on the SDD, in terms of the ADE/FDE metrics. This demonstrates that the cascaded CVAE module can extract useful information from the predicted future points, and thus improves the prediction performance of the future trajectory.

2) Socially-aware Regression Module: The future trajectory predicted by the cascaded CVAE module may be impractical because the CVAE cannot model interactions among pedestrians. Thus, the socially-aware regression module was introduced to refine the final predictions via social interaction. To prove its effectiveness, the model are respectively trained with and without it. As presented in Tables IV and V, the socially-aware regression module can further boost the performance of the cascaded CVAE module by 17.6%/4.2% on the ETH/UCY and 20.4%/19.7% on the SDD, in terms of the ADE/FDE metrics, which indicates that this module can extract interaction features from the predicted trajectories to refine the final predictions. Besides, the predictions of the model trained with or without the socially-aware regression module are visualized in Figure 6. Figures 6a, 6b and 6c reveal that, compared to the less coordinated trajectories predicted by the cascade CVAE module, more socially compliant trajectories are refined by the socially-aware regression module. Figure 6d demonstrates that the collision problem in trajectory prediction can be improved by the socially-aware regression module using interactions between pedestrians.

3) Slide CVAE Module: The cascaded CVAE module is effectiveness in using the predicted future points to assist in trajectory prediction. But its efficiency is directly affected by two factors, namely the multiple unshared CVAEs and the incremental inputs. To increase the inference speed and decrease the model parameters, a slide CVAE module is proposed to further improve the cascaded CVAE module. As shown in the 6-th and 7-th columns of Tables IV and V, the slide CVAE module improves the cascaded CVAE module

D. Ablation Study

This experiment is designed to validate the effectiveness of the proposed cascaded CVAE module, socially-aware regression module and slide CVAE module. To validate those modules, a baseline method that consists of \( \delta \) CVAEs is introduced for comparison. In the baseline method, \( \delta \) future points are respectively predicted by \( \delta \) CVAEs that using the original past trajectory as the input, as illustrated in Figure 5a.
Fig. 7: Qualitative results of CSR in SDD dataset. The visualized trajectories are best predictions sampled from 20 trials. The white dots, yellow dots, green dots, and red dots respectively represent the past trajectories, ground truth future trajectories, predictions of PECNet, and predictions of the proposed CSR.

E. Qualitative Analysis

Qualitative results of CSR and PECNet on the SDD are visualized in Figure 7. Figures 7a and 7b present predictions of linear trajectories. The two methods can both produce trajectories that very close to the ground truth. It is expected because the prediction of linear trajectories is easy. Figure 7c and 7d show predictions of crooked trajectories. In this case, the predictions of CSR are more accurate than that of the PECNet. These results demonstrate that the proposed CSR has better performance in predicting nonlinear trajectories.

Figures 7e and 7f respectively visualize two easy interaction samples, i.e. turning together and straight going together. Both PECNet and CSR can generate trajectories that are close to the ground truth because the interactions between pedestrians are simple. Figure 7g and 7h visualize predictions of two complex interacions such as crossing and collision avoidance. The predictions of CSR are socially compliant and more reasonable than PECNet. These results demonstrate that the proposed CSR can effectively model the complex interactions between pedestrians.

V. Conclusion

In this work, a novel cascaded CV AE with socially-aware rethinking (CSR) method is proposed for pedestrian trajectory prediction. It consists of a cascaded CV AE module and a socially-aware regression module. The cascaded CV AE module predicts the future trajectories in a sequential manner using the updated past trajectory. And the socially-aware regression module extracts interactions features from both the past trajectories and the predicted trajectories to refine the final predictions. Besides, a slide CV AE module is proposed to further improve the model efficiency and performance of the cascade CV AE module using one shared CV AE, in a slidable manner. Extensive experiments on the SDD and ETH/UCY datasets demonstrate the effectiveness of the proposed methods. Further more, the results indicate that the proposed CSR and S-CSR outperforms other state-of-the-art methods by a large margin and achieves a new state-of-the-art performance.

REFERENCES

[1] J. Hong, B. Sapp, and J. Philbin, “Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[2] P. Raksincharoensak, T. Hasegawa, and M. Nagai, “Motion planning and control of autonomous driving intelligence system based on risk potential optimization framework,” International Journal of Automotive Engineering, vol. 7, no. AVEC14, pp. 53–60, 2016.

[3] A. Vemula, K. Muelling, and J. Oh, “Social attention: Modeling attention in human crowds,” in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 4601–4607.
Fig. 8: Qualitative results of CSR and S-CSR in ETH/UCY dataset. The visualized trajectories are best predictions sampled from 20 trials. The red lines, yellow lines, green dashed lines, and blue dashed lines respectively represent the past trajectories, ground truth future trajectories, predictions of CSR, and predictions of S-CSR.
[4] Y. Luo, P. Cai, A. Bera, D. Hsu, W. S. Lee, and D. Manocha, “Porca: Modeling and planning for autonomous driving among many pedestrians,” IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3418–3425, 2018.

[5] H. Xue, D. Huynh, and M. Reynolds, “Location-velocity attention for pedestrian trajectory prediction,” in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Jan 2019, pp. 2038–2047.

[6] M. Lu, J. A. Stork, G. D. Tipaldi, and K. O. Arras, “People tracking with human motion predictions from social forces,” in 2010 IEEE International Conference on Robotics and Automation. IEEE, 2010, pp. 464–469.

[7] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, “It is not the journey but the destination: Endpoint conditioned trajectory prediction,” in European Conference on Computer Vision. Springer, 2020, pp. 759–776.

[8] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, S. Chmid et al., “Tat: Target-driven trajectory prediction,” arXiv preprint arXiv:2008.08294, 2020.

[9] K. Mangalam, Y. An, H. Girase, and J. Malik, “From goals, waypoints & paths to long term human trajectory forecasting,” arXiv preprint arXiv:2012.01526, 2020.

[10] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social lstm: Human trajectory prediction in crowded spaces,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 961–971.

[11] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, “Social gan: Socially acceptable trajectories with generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2255–2264.

[12] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, “Sophie: An attentive gated rnn for predicting paths compliant to social and physical constraints,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1349–1358.

[13] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, “Peeking into the future: Predicting future pedestrian activities and locations in videos,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5725–5734.

[14] Y. Xu, J. Yang, and S. Du, “Cl-lstm: Cascaded feature-based long short-term networks for predicting pedestrian trajectory,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12541–12548.

[15] S. LeFevre, C. Laugier, and J. Ibañez-Guzmán, “Exploiting map information for driver intention estimation at road intersections,” in 2011 ieee intelligent vehicles symposium (iv). IEEE, 2011, pp. 583–588.

[16] C. Ramussen and C. Williams, “Gaussian processes for machine learning (adaptive computation and machine learning),” 2006.

[17] R. MLBoto-Moreo and M. A. Zamora-Izquierdo, “Imm-based lane-change prediction in highways with low-cost gps/sins,” IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 1, pp. 180–185, 2009.

[18] X. Li, X. Ying, and M. C. Chuah, “Grip: Graph-based interaction-aware trajectory prediction,” in 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 3960–3966.

[19] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, “Vocnet: Encoding hd maps and agent dynamics from vectorized representation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 525–11 533.

[20] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, “Social-stgcn: A social spatio-temporal graph convolutional neural network for human trajectory prediction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14424–14432.

[21] H. Zhou, D. Ren, H. Xia, M. Fan, X. Yang, and H. Huang, “A-st-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction,” Neurocomputing, vol. 445, pp. 298–308, 2021.

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

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in NIPS 2014 Workshop on Deep Learning, December 2014, 2014.

[24] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectron++: Dynamically-featural trajectory forecasting with heterogeneous data,” in Computer Vision–ECCV 2020: 16th European Conference. Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer, 2020, pp. 683–700.