A Set of Control Points Conditioned Pedestrian Trajectory Prediction

Inhwan Bae and Hae-Gon Jeon*

Gwangju Institute of Science and Technology (GIST)
inhwanbae@gm.gist.ac.kr and haegonj@gist.ac.kr

Abstract

Predicting the trajectories of pedestrians in crowded conditions is an important task for applications like autonomous navigation systems. Previous studies have tackled this problem using two strategies. They (1) infer all future steps recursively, or (2) predict the potential destinations of pedestrians at once and interpolate the intermediate steps to arrive there. However, these strategies often suffer from the accumulated errors of the recursive inference, or restrictive assumptions about social relations in the intermediate path. In this paper, we present a graph convolutional network-based trajectory prediction. Firstly, we propose a control point prediction that divides the future path into three sections and infers the intermediate destinations of pedestrians to reduce the accumulated error. To do this, we construct multi-relational weighted graphs to account for their physical and complex social relations. We then introduce a trajectory refinement step based on a spatio-temporal and multi-relational graph. By considering the social interactions between neighbors, better prediction results are achievable. In experiments, the proposed network achieves state-of-the-art performance on various real-world trajectory prediction benchmarks.

1 Introduction

Predicting the future trajectories of humans in crowds is an important task, especially for social robots, autonomous navigation, and surveillance systems. However, this task is challenging because such predictions require considering the desired destinations of each pedestrian, and the social norms of other moving agents, simultaneously.

Early works (Helbing and Molnar 1995; Pellegrini et al. 2009; Mehran, Oyama, and Shah 2009; Yamaguchi et al. 2011; Pellegrini, Ess, and Gool 2010) have attempted to capture social interactions using handcrafted Langevin equations, however, they often fail to model the complex social interactions that occur in crowded scenes. The recent development of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) combines with social pooling (Alahi et al. 2016; Gupta et al. 2018) and social attention (Vemula, Muelling, and Oh 2018), and has improved understanding of the social interactions among pedestrians. However, these approaches still suffer from severe errors in final destination, because they accumulate the errors inherent to problems in the recursive predictions.

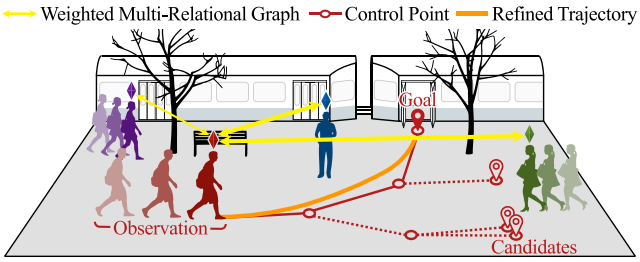


Figure 1: An illustration of Graph-TERN. Our Graph-TERN first constructs a multi-relational pedestrian graph (yellow) to capture social and temporal relationships, and then predicts a set of control points (red) to sample hypothetical destinations of pedestrians in scenes. After that, the proposed refinement module learns to estimate accurate final trajectories (orange).

To overcome this issue, two notable previous attempts have been made to infer the potential destinations of pedestrians in scenes. Works (Mohamed et al. 2020; Shi et al. 2021a; Bae and Jeon 2021) utilize several temporal CNNs that predict entire sequences in a single shot, which alleviates the accumulation of errors. In (Rehder et al. 2018; Deo and Trivedi 2020; Mangalam et al. 2020, 2021), the potential endpoints of the local trajectories are defined first, and their intermediate steps are then interpolated. Although these methods show promising performance improvements, issues remained. (1) Long-term predictions are performed without any consideration of events occurring in the intermediate steps, and (2) social interactions are not regarded in endpoint prediction.

In this paper, we propose a Graph-based pedestrian Trajectory Estimation and Refinement Network (Graph-TERN) using a set of control points that combines the advantages of both attempts. Our network consists of three parts: a control point prediction, trajectory refinement, and a multi-relational graph convolutional network (MRGCN). Firstly, we divide each pedestrian's future path into three sections and infer each stochastic goal, called control points. Each control point represents probabilistic decisions about the next future steps pedestrians will take. Mixture density networks (MDNs) are

*Corresponding author

used in this process, and stochastic endpoints are determined by connecting the three control points. With the benefit of the control point prediction, long-term predictions are feasible while respecting the social norms of interactions among people travelling in crowds or groups. Here, we introduce a Gaussian mixture model (GMM) pruning scheme, effectively cutting-off abnormal behaviors of pedestrians in Gaussian distributions. Secondly, we generate realistic future paths by introducing a refinement module that yields correction vector fields. We initially infer intermediate steps by linearly interpolating between the endpoints, and then refine the initial trajectory by adding the correction vector fields. Lastly, we design an MRGCN operator to take account of complex social interactions in both the control point prediction and the trajectory refinement. By effectively incorporating the three modules, our model achieves state-of-the-art results using a variety of public pedestrian trajectory prediction benchmarks.

2 Related Works

Context-aware Trajectory Prediction. Starting with handcrafted motion model (Helbing and Molnar 1995), there have been significant advances in the ability of methods to represent the movement of humans. Pioneering works (Alahi et al. 2016; Vemula, Muelling, and Oh 2018; Bartoli et al. 2018; Fernando et al. 2018; Zhang et al. 2019; Li, Ma, and Tomizuka 2019; Shi et al. 2020; Kothari, Sifringer, and Alahi 2021; Xu et al. 2022b) introduce schemes for human proximity relationships which have been used with social forces to model pedestrian interactions in crowds. Better prediction results are achievable using a fusion of visual features and coordinate information (Varshneya and Srinivasaraghavan 2017; Xue, Huynh, and Reynolds 2018; Manh and Alaghband 2018; Sadeghian et al. 2019; Liang et al. 2019; Kosaraju et al. 2019; Sun, Zhao, and He 2020; Dendorfer, Elflein, and Leal-Taixé 2021; Zhao et al. 2019; Tao, Jiang, and Duan 2020; Sun, Jiang, and Lu 2020; Shafiee, Padir, and Elhamifar 2021; Chai et al. 2019). Recently, the use of Gaussian distribution (Hug, Hübner, and Arens 2020; Hug et al. 2022; Xu, Yang, and Du 2020), generative adversarial networks (GANs) (Gupta et al. 2018; Sadeghian et al. 2019; Kosaraju et al. 2019; Li 2019; Dendorfer, Elflein, and Leal-Taixé 2021) and the Conditional Variational Auto-encoder (CVAE) (Lee et al. 2017; Ivanovic and Pavone 2019; Salzmann et al. 2020; Chen et al. 2021b; Yao et al. 2021; Xu et al. 2022a; Wang et al. 2022; Yue, Manocha, and Wang 2022; Xu, Hayet, and Karamouzas 2022; Wen, Wang, and Metaxas 2022) are proposed to infer socially-acceptable multiple trajectories.

More recently, graph neural network-based approaches (Liang et al. 2020; Li et al. 2020; Liang, Jiang, and Hauptmann 2020; Yu et al. 2020; Bae and Jeon 2021; Bae, Park, and Jeon 2022b,a; Xu et al. 2022c; Gu et al. 2022) have explicitly modeled human-human interactions and jointly predicted the trajectories of all agents, using graph-based representations. Among them, Graph Attention Networks (GATs) (Veličković et al. 2018) implicitly assign the social relations of pedestrians on graph nodes and edges (Huang et al. 2019; Kosaraju et al. 2019; Sun, Jiang, and Lu 2020; Shi et al. 2021a). Social-STGCNN (Mohamed et al. 2020) presents a Graph Convolutional Network (GCN)-based trajectory prediction, which effectively aggregates the spatial information of pedestrians. This method imposes physical constraints to capture the relative distance of humans in scenes. However, its performance is limited due to the single relations based on displacements among them.

Endpoint Conditioned Approach. Endpoint conditioned trajectory prediction is a process that infers the hypothetical arrival points of pedestrians and then interpolates their paths like vehicle navigation systems. Rehder *et al*. (Rehder and Kloeden 2015; Rehder et al. 2018) propose a von-Mises distributed destination prediction using a particle filter with environment-based dynamics. TNT (Zhao et al. 2020) uses square lattice endpoints for initial target prediction. Goal-GAN (Dendorfer, Osep, and Leal-Taixe 2020) predicts the goal probability map on a scene image using CNN. P2TIRL (Deo and Trivedi 2020) introduces a grid-based goal planning with maximum entropy inverse reinforcement learning. Although the works show the potential of goal-based prediction with human-environment interactions, efforts to address both spatial and temporal aspects in a crowd remain insufficient. In very recent works, PECNet (Mangalam et al. 2020) proposed a CVAE-based endpoint conditioned trajectory prediction with social non-local pooling, and achieved state-of-the-art results. However, these approaches do not consider all of the pedestrians in scenes or agent interactions in the endpoint prediction time.

Trajectory Refinement. There are several works that refine initially predicted pedestrian trajectories. In (Lee et al. 2017), gated recurrent units and CVAE were used to predict the initial sampled trajectory and refine them by fusing the semantic context of scenes and the social interactions between agents. MANTRA (Marchetti et al. 2020, 2022) used a similar strategy with (Lee et al. 2017) on multi-modal initial trajectories sampled with novel memory augmented networks. These works have mainly focused on correcting the directions in the initial trajectory, which can result in severe final destination errors.

Compared to previous works, the proposed network adopts a multi-relational GCN structure to take complex agent interactions into account when considering various physical relations. Our trajectory refinement also corrects initial predictions well without any distortion of the destinations. What is unique in the proposed network is the use of control point prediction when predicting the potential destination of each pedestrian. The existing endpoint conditioned approaches directly predict a destination as a single hard constraint without any consideration of neighbors, and learn short-term destinations like waypoints. In contrast, our control points are used for endpoint sampling that reflects the agent interactions, not acting as waypoints.

3 Control Point Conditioned Prediction

Graph-TERN consists of two key components: (1) learning the probabilistic distribution for sampling endpoint candidates based on the control point; (2) yielding socially acceptable path prediction using a refinement module. Using an MRGCN framework, we develop a model that can success-

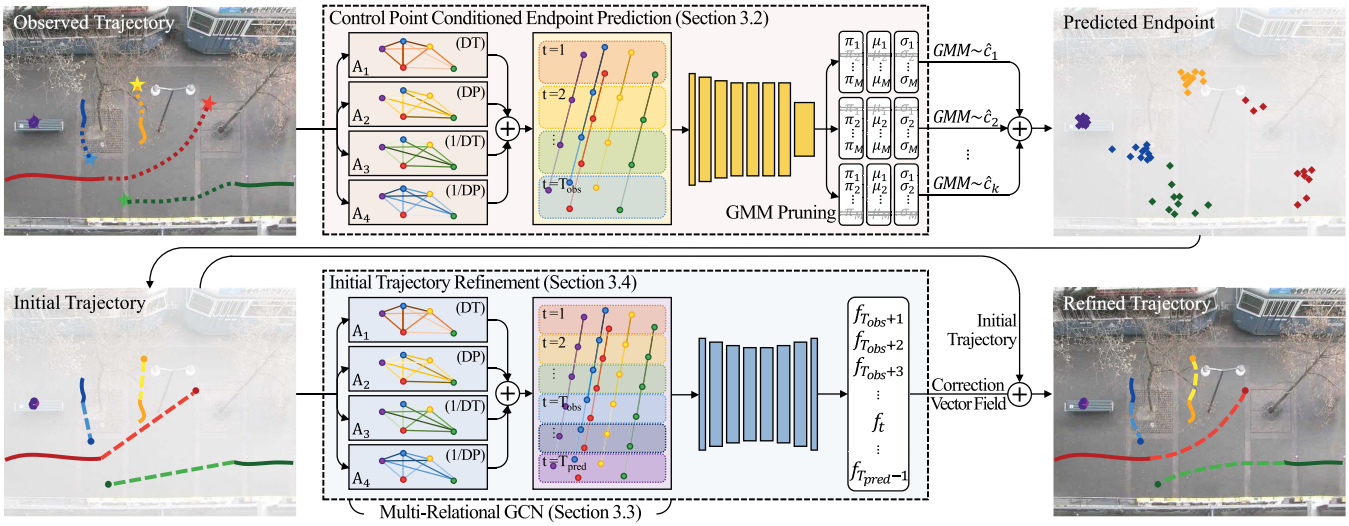


Figure 2: An overview of our Graph-TERN Architecture. First, a control point prediction module takes the observed trajectory $S_{1:T_{obs}}$, and then constructs a multi-relational pedestrian graph. With a spatio-temporal aggregation using GCN and CNN, we can predict hypothetical endpoints $\hat{E}$ through a summation of a set of randomly sampled control points $\hat{C}$. Second, a trajectory refinement modules takes the predicted endpoint $\hat{E}$ and the observed trajectory $S_{1:T_{obs}}$ to predict a correction vector field. A refined trajectory $\hat{S}_{T_{obs}+1:T_{pred}}$ is obtained by summing the initial trajectory with the correction vector field. DT and DP mean distance and displacement, respectively.

fully predict future trajectories while considering complex social relations. The overall framework is shown in Figure 2.

3.1 Preliminaries

Problem Definition. Pedestrian trajectory prediction attempts to determine future position sequences from observed position sequences for all agents in a scene. Suppose that there are N pedestrians in a scene at specific time t, and the corresponding positions of each pedestrian $n \in \{1, ..., N\}$ can be represented as $p_t^n = (x_t^n, y_t^n)$. The trajectory sequence from the first time frame to the observed time T_{obs} can be denoted as $S_{1:T_{obs}}^n = \{p_t^n \in \mathbb{R}^2 | t \in \mathbb{N}, 1 \leq t \leq T_{obs}\}$. The consecutive prediction time frames are represented as $S_{T_{obs}+1:T_{pred}}$.

An additional goal of this work is to estimate a set of potential destinations for each pedestrian, called endpoint E^n. The endpoint $\hat{E}$ can be predicted with the observed sequence $\hat{E}^n = \hat{p}_{T_{pred}}^n | S_{1:T_{obs}}$, then the future trajectories $\hat{S}_{T_{obs}+1:T_{pred}}$ can be inferred from the observed sequence $S_{1:T_{obs}}$ and the predicted endpoint $\hat{E}$.

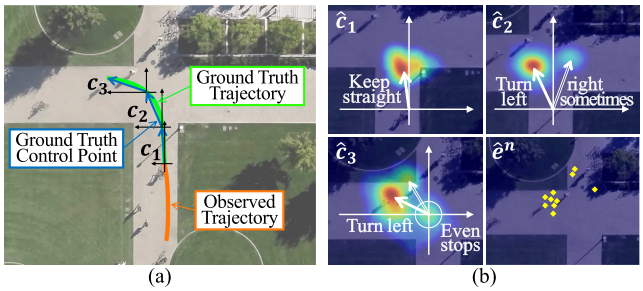
Graph Convolutional Network. In general, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents a set of nodes $\mathcal{V}$ and edges $\mathcal{E}$. In a pedestrian graph, the spatio-temporal graph $\mathcal{G}$ consists of a pedestrian node $\mathcal{V} = \{p_t^n | n, t \in \mathbb{N}, 1 \leq n \leq N, 1 \leq t \leq T\}$ and a set of spatial and temporal edges $\mathcal{E} = \mathcal{E}_t \cup \mathcal{E}_n$. The spatial edge $\mathcal{E}_t = \{a_t^{i,j} | i, j \in \mathbb{N}, 1 \leq i, j \leq N\}$ represents a spatial relation for each pedestrian at a specific time t, and the temporal edge $\mathcal{E}_n = \{a_n^{i,j} | i, j \in \mathbb{N}, 1 \leq i, j \leq T\}$ represents the temporal relation of each pedestrian n within an observed sequence. Node features are aggregated with both spatial and temporal dimensions using GCNs and CNNs (Yan, Xiong, and Lin 2018; Mohamed et al. 2020). With the node feature $H = \{h_t^n | n, t \in \mathbb{N}, 1 \leq n \leq N, 1 \leq t \leq T_{obs}\}$ and adjacency matrix $A = \{a_t^{i,j} | i, j, t \in \mathbb{N}, 1 \leq i, j \leq N, 1 \leq t \leq T_{obs}\}$, the GCN feature update rule is defined as $H' = \sigma(\hat{A}HW)$. Here, W and $\hat{A}$ indicate the learnable weight matrix and the normalized form with the formula $\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$, respectively. We denote the self-loop added adjacency matrix as $\tilde{A} = A + I$ and diagonal node degree matrix as $\tilde{D}$ from $\tilde{A}$.

3.2 Control Point Conditioned Endpoint Prediction

The methodology for sampling plausible endpoints and predicting socially acceptable paths has significantly improved for pedestrian trajectory prediction. The key to endpoint prediction is determining travelable roads. For this, it is essential to learn the complex social interactions between pedestrians in crowds, such as collision avoidance and group following. However, existing methods do not consider any factor that affects the endpoint prediction in the process. We often observe that pedestrians fail to arrive within the predicted time frame T_{pred}. For this reason, social interactions need to be considered in the endpoint prediction. In this work, we present a novel set of control point-based endpoint predictions to handle events that unexpectedly occur in the predicted time frames, and MRGCN to precisely capture social interactions.

Graph Control Point Prediction. The key idea of our model is to use multiple control points when inferring potential endpoints. In contrast to previous works, we incorporate our control point prediction into a GCN framework, and in this way, our model provides a socially compliant endpoint that considers intermediate social interactions.

First, we define a set of control points C based on a displacement in one section, which is equally divided into future sequences $S_{T_{obs}+1:T_{pred}}$ in Figure 3(a). The formula to obtain



(a) (b)

Figure 3: A set of control points prediction. (a) When a person turns left at the crossroad, three control points are defined based on the displacement. (b) Examples of predicted distributions for the control points and endpoint sampling.

C for a pedestrian n is as below:

$$C^n = \left\{ c_k^n = p^n_{T_{obs}+\tau\times k} - p^n_{T_{obs}+\tau\times(k-1)} \right\} \quad \text{for } \forall k \in \{1,...,K\}, \quad \tau = \frac{T_{pred}-T_{obs}}{K}, \tag{1}$$

where K is a user-defined parameter.

Next, we present a control point prediction module. We update feature maps for the observed input sequence using a multi-relational GCN, which will be explained in Section 3.3. We then use a multivariate GMM to sample the 2D displacements of a set of control points in a MDN, as shown in Figure 3(b). With the output of the module $z = \{z_k^\pi, \boldsymbol{z}_k^\mu, \boldsymbol{z}_k^\sigma | k \in \mathbb{N}, 1 \le k \le K\}$, we compute a mean $\hat{\boldsymbol{\mu}}_k = \{\boldsymbol{z}^\mu_{m,k} | m \in \mathbb{N}, 1 \le m \le M\}$, a standard deviation $\hat{\boldsymbol{\sigma}}_k = \{\exp(\boldsymbol{z}^\sigma_{m,k}) | m \in \mathbb{N}, 1 \le m \le M\}$, and a mixing coefficient $\hat{\pi}_k = \{\exp(z^\pi_{m,k}) / \sum_{i=1}^{M} \exp(z^\pi_{i,k}) | m \in \mathbb{N}, 1 \le m \le M\}$ for the multivariate GMM with $M = 8$ as illustrated in Figure 2. Here, the mean and standard deviation have x, y-axis values of $\hat{\boldsymbol{\mu}} = (\hat{\mu}_x, \hat{\mu}_y)$ and $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_x, \hat{\sigma}_y)$, respectively.

GMM Pruning. Public pedestrian trajectory datasets contain abnormal behaviors of agents. Since statistical models need to allocate a portion of its capacity to ensure the abnormal cases, it is left with relatively less capacity for generating realistic paths. In a previous GAN-based approach (Dendorfer, Elflein, and Leal-Taixé 2021), stacking multiple generators with individual parameters exclusively predicts path samples on disconnected manifolds. This multi-generator structure reduces unrealistic path generations in the test phase.

We observe that the Gaussian distribution for abnormal behaviors is formed to have large standard deviations and low mixing coefficients. These abnormal cases are considered as out-of-distribution samples drawn far away from the training distribution statistically and lead to performance drops. In a previous work (Mangalam et al. 2020), a truncation trick is used to restrict the distribution of samples. However, in the GMM-based prediction model, it limits the distribution of reasonable control points as well. To address this issue, we devise a GMM pruning which cuts off a lower half of the bivariate Gaussian based on predicted mixing coefficients as:

$$M^* = \left\lfloor \frac{M}{2} \right\rfloor, \quad z^* = \underset{z' \subset z, |z'| = M^*}{\operatorname{argmax}} \sum_{z'^\pi \subset z'} z'^\pi, \tag{2}$$

where M^* is the number of the selected mixture models. Through the GMM pruning, potential control points can be assigned to effectively feasible areas without any increase in the number of learnable parameters.

Endpoint Sampling. The final endpoint $\hat{e}$ is determined by using the set of control points $\hat{C}$, which is sampled through the probabilistic process in Figure 3(b). While existing works infer the final endpoint at once, our probabilistic model allows them to be determined by combining social interactions computed from each intermediate point. Because each control point represents a relative displacement from a previous point, the absolute coordinates of a final endpoint can be determined by adding all the control points to the last coordinates of the observed sequence, as below:

$$\hat{e}^n = p^n_{T_{obs}} + \sum_{k=1}^{K} \hat{c}^n_k. \tag{3}$$

Following the previous study (Gupta et al. 2018), we sample the $L = 20$ endpoints $\hat{E}^n = \{\hat{e}^n_l | l, n \in \mathbb{N}, 1 \le l \le L, 1 \le n \le N\}$ which represent multi-modality, and feed them into a trajectory refinement module in Section 3.4.

3.3 Multi-Relational Pedestrian Graph

Our model uses a GCN to manipulate a multi-relational graph. The selection between GAT (Huang et al. 2019; Kosaraju et al. 2019; Yu et al. 2020; Shi et al. 2021a) and GCN (Mohamed et al. 2020; Shi et al. 2022; Li et al. 2021) is an open issue in modeling social relations. While the GCN has the advantage of imposing physical constraints, conventional GCN-based models use a single relation edge, which makes capturing social relations limited. Due to this reason, the GCN-based approaches have gained less interest than those of the GAT-based approaches whose multi-head attention allows it. In this work, we fully take advantage of the GCN framework by overcoming the limitation through a multi-relational-based kernel function to produce each relational adjacent matrix.

Following the multi-relation-based GCN methods successfully adopted in other research areas such as action recognition and natural language processing (Marcheggiani and Titov 2017; Li et al. 2019; Shi et al. 2019), we construct a multi-relational weighted graph to predict pedestrian trajectories. Unlike the previous work (Mohamed et al. 2020) which has a graph considering only the relative distance between pedestrians, our multi-relational graph includes distance, displacement, and their inverse terms. We define its spatio-temporal feature update rule $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$ as below:

$$H' = \sigma\left[\text{CNN}\left(\sum_{r=1}^{R} \hat{A}_r H^T W_r \right)^T + H \right], \tag{4}$$

where R is the number of elements in a set of relations $\mathcal{R} = \{\textit{Distance}, \textit{Displacement}, 1/\textit{Distance}, 1/\textit{Displacement}\}$, $\hat{A}_r$ is the normalized term of $A_r = \{a^{i,j,r} \in \mathbb{R} | i, j, r \in \mathbb{N}, 1 \le i, j \le N, 1 \le r \le R\}$, and W_r is a learnable weight matrix. In the pedestrian graph, there are both spatial relations between pedestrians and temporal relations for the consecutive temporal sequences of each of them. Following prior studies on

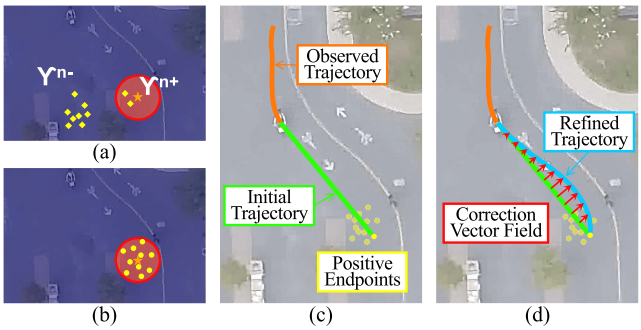


Figure 4: An example of trajectory refinement. (a) Endpoint candidates are classified into the positive and negative set. (b) Additionally, guided endpoints are randomly sampled. (c) The initial trajectory is predicted by linearly interpolating between the observed trajectory and the endpoints. (d) After that, the trajectory is refined by adding a correction vector field in the initial trajectory.

spatio-temporal data (Yan, Xiong, and Lin 2018; Mohamed et al. 2020), we perform both spatial and temporal aggregation on all nodes, with the MRGCN and CNN, respectively.

We observe that a model which only considers either the distance or the displacement sometimes suffers because it avoids either companions or persons who are walking behind their back. In contrast, our multi-relational graph deals with obstacles, stop and go motion, and group following in very challenging situations.

3.4 Initial Trajectory Refinement

Our refinement method proceeds with three steps: guided endpoint sampling, initial trajectory prediction, and graph trajectory refinement.

Guided Endpoint Sampling. To jointly train the control point prediction module and trajectory refinement module, we need to decouple the two modules. Since a predicted endpoint is not always close to a ground truth position, we define a rule to limit the predicted endpoints at training time.

We use not only a ground-truth endpoint, but also all the predicted endpoints close to the ground truth. Specially, as shown in Figure 4(a), we divide the predicted endpoints into a positive set Υ^+ and a negative set Υ^- as below:

$$\begin{aligned} \Upsilon^{n+} &= \{\hat{e}_l^n \,|\, \|\hat{e}_l^n - e^n\| \leq \Gamma\} \\ \Upsilon^{n-} &= \{\hat{e}_l^n \,|\, \|\hat{e}_l^n - e^n\| > \Gamma\} \\ \text{for} \quad \forall l \in \{1,...,L\}, \;\; \Gamma &= \frac{\|p_{T_{obs}}^n - p_1^n\|}{T_{obs} \times \gamma}, \end{aligned} \tag{5}$$

where γ is a scale indicator that adaptively adjusts the averaged displacement of each pedestrian. We only back-propagate gradients for the positive sets using a valid mask Ψ:

$$[\Psi]_{n,l} = \psi_l^n = \begin{cases} 1 & \text{if } \hat{e}_l^n \in \Upsilon^{n+}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

During the initial training phase, the number of positive sets might be extremely small because the endpoint candidates are not yet converged. To address this issue, we additionally sample the L positive endpoints within the range Γ of the ground-truth, called guided endpoints in Figure 4(b).

Initial Trajectory Prediction. The purpose of establishing an initial trajectory based on the guided endpoints is to make our trajectory refinement module tractable. The simplest way to do this is to connect these control points through linear interpolation. However, we observe that the use of a set of control points in the initial trajectory prediction acts as a hard constraint, even though it is helpful to infer accurate destinations of pedestrians, as will be demonstrated in Section 4.3.

Therefore, we first generate a single initial trajectory $\tilde{S}_{T_{obs}+1:T_{pred}}$ for one endpoint without any control point in Figure 4(c). To do this, we linearly interpolate between the last frame of observations $p_{T_{obs}}^n$ and the endpoint $\hat{e}_l^n$ as below:

$$\begin{aligned} \tilde{p}_{t,l}^n = & p_{T_{obs}}^n + \frac{\hat{e}_l^n - p_{T_{obs}}^n}{T_{pred} - T_{obs}} \times (t - T_{obs}) \\ & \tilde{S}_{T_{obs}+1:T_{pred},l} = \{\tilde{p}_{t,l}^n\} \\ \text{for} \quad & \forall t \in \{T_{obs}+1, ..., T_{pred}\}, \\ & \forall l \in \{1, ..., L\}, \;\; \forall n \in \{1, ..., N\}. \end{aligned} \tag{7}$$

Graph Trajectory Refinement. As a next step, we present a novel refinement module to yield an accurate trajectory from the observed trajectory $S_{1:T_{obs}}$ and the initial trajectory $\tilde{S}_{T_{obs}+1:T_{pred}}$. By concatenating the two trajectories along with the time axis, we can aggregate the social interactions for all time frames using the MRGCN. As shown in Figure 4(d), the correction vector field f_t^n is computed and the final refined trajectory can be obtained as below:

$$\begin{aligned} \hat{p}_{t,l}^n &= \tilde{p}_{t,l}^n + f_{t,l}^n \\ \hat{S}_{T_{obs}+1:T_{pred},l} &= \{\hat{p}_{t,l}^n\} \cup \hat{E} \\ \text{for} \quad & \forall t \in \{T_{obs}+1, ..., T_{pred}-1\}, \\ & \forall l \in \{1, ..., L\}, \;\; \forall n \in \{1, ..., N\}. \end{aligned} \tag{8}$$

Unlike existing methods (Mohamed et al. 2020; Mangalam et al. 2020; Shi et al. 2021a; Liu, Yan, and Alahi 2021; Chen et al. 2021a) which use social interactions based only on observations, our refinement module allows a more complex social relation because our MRGCN captures such relations even with the interpolated points and the endpoint.

3.5 Implementation Details

Loss Function. We maximize an expectation to train the control point prediction module. We sum the probabilistic density functions of all the predicted control point distributions and pedestrians. The loss function Θ_w is defined as:

$$\Theta_w = \sum_{n=1}^{N} \sum_{k=1}^{K} -\log \left[\sum_{m=1}^{M} \hat{\pi}_{m,k}^n \frac{\exp\left(-\frac{(\boldsymbol{c}_k^n - \hat{\boldsymbol{\mu}}_{m,k}^n)^2}{2(\hat{\boldsymbol{\sigma}}_{m,k}^n)^2}\right)}{\sqrt{2\pi}\,\hat{\boldsymbol{\sigma}}_{m,k}^n} \right] \tag{9}$$

In addition, we minimize the trajectory refinement loss Θ_r. The loss is based on a mean square error (MSE) of an average displacement between a refined trajectory and a ground truth trajectory, and is formulated as below:

$$\Theta_r = \sum_{n=1}^{N} \sum_{l=1}^{2L} \sum_{t=T_{obs}+1}^{T_{pred}-1} \psi_l^n \left[\left(x_{t,l}^n - \hat{x}_{t,l}^n\right)^2 + \left(y_{t,l}^n - \hat{y}_{t,l}^n\right)^2 \right] \tag{10}$$

Model	Linear Regression	Social-LSTM	Social-GAN	SR-LSTM†	STGAT	IDL	RSBG	STAR†	Reciprocal Learning	Social-STGCNN	PECNet†
Year	-	2016	2018	2019	2019	2019	2020	2020	2020	2020	2020
ETH	1.33 / 2.94	1.09 / 2.35	0.87 / 1.62	1.01 / 1.93	0.65 / 1.12	0.59 / 1.30	0.80 / 1.53	0.78 / 1.47	0.69 / 1.24	0.64 / 1.11	0.65 / 1.13
HOTEL	0.39 / 0.72	0.79 / 1.76	0.67 / 1.37	0.35 / 0.72	0.35 / 0.66	0.46 / 0.83	0.33 / 0.64	0.33 / 0.83	0.43 / 0.87	0.49 / 0.85	0.22 / 0.38
UNIV	0.82 / 1.59	0.67 / 1.40	0.76 / 1.52	0.66 / 1.38	0.52 / 1.10	0.51 / 1.27	0.59 / 1.25	0.35 / 0.72	0.53 / 1.17	0.44 / 0.79	0.35 / 0.57
ZARA1	0.62 / 1.21	0.47 / 1.00	0.35 / 0.68	0.56 / 1.23	0.34 / 0.69	0.22 / 0.49	0.40 / 0.86	0.27 / 0.58	0.28 / 0.61	0.34 / 0.53	0.25 / 0.45
ZARA2	0.77 / 1.48	0.56 / 1.17	0.42 / 0.84	0.44 / 0.90	0.29 / 0.60	0.23 / 0.55	0.30 / 0.65	0.23 / 0.52	0.28 / 0.59	0.30 / 0.48	0.18 / 0.31
AVG	0.79 / 1.59	0.72 / 1.54	0.61 / 1.21	0.60 / 1.23	0.43 / 0.83	0.40 / 0.89	0.48 / 0.99	0.39 / 0.82	0.44 / 0.90	0.44 / 0.75	0.33 / 0.57
Model	Trajectron ++†	Causal-STGAT	NCE-STGCNN	TPNMS	Causal-STGCNN	SGCN	S-DPF	LBEBM†	DMRGCN	STT	**Graph-TERN**
Year	2020	2021	2021	2021	2021	2021	2021	2021	2021	2022	-
ETH	0.61 / 1.03	0.60 / 0.98	0.66 / 1.22	0.52 / 0.89	0.64 / 1.00	0.63 / 1.03	0.66 / 0.92	0.62 / 1.16	0.60 / 1.09	0.54 / 1.10	**0.42 / 0.58**
HOTEL	0.20 / 0.28	0.30 / 0.54	0.44 / 0.68	0.22 / 0.39	0.38 / 0.45	0.32 / 0.55	0.34 / 0.50	0.19 / 0.35	0.21 / 0.30	0.24 / 0.46	**0.14 / 0.23**
UNIV	0.30 / 0.55	0.52 / 1.10	0.47 / 0.88	0.55 / 1.13	0.49 / 0.81	0.37 / 0.70	0.50 / 0.69	0.37 / 0.67	0.35 / 0.63	0.57 / 1.15	**0.26 / 0.45**
ZARA1	0.24 / 0.41	0.32 / 0.64	0.33 / 0.52	0.35 / 0.70	0.34 / 0.53	0.29 / 0.53	0.34 / 0.59	0.23 / 0.43	0.29 / 0.47	0.45 / 0.94	**0.21 / 0.37**
ZARA2	0.18 / 0.32	0.28 / 0.58	0.29 / 0.48	0.27 / 0.56	0.32 / 0.49	0.25 / 0.45	0.32 / 0.45	0.19 / 0.36	0.25 / 0.41	0.36 / 0.77	**0.17 / 0.29**
AVG	0.31 / 0.52	0.40 / 0.77	0.44 / 0.76	0.38 / 0.73	0.43 / 0.66	0.37 / 0.65	0.43 / 0.63	0.32 / 0.59	0.34 / 0.58	0.43 / 0.88	**0.24 / 0.38**

Table 1: Comparison of our Graph-TERN with other state-of-the-art methods on ETH/UCY dataset (ADE/FDE, Unit: meter). The evaluation results are directly referred from (Shi et al. 2021a; Liu, Yan, and Alahi 2021; Chen et al. 2021a; Liang et al. 2021). The mark † means that the common data-loader in (Gupta et al. 2018) are used. Bold: Best, Underline: Second best.

Finally, the loss function Θ of the entire network is defined as a weighted sum of the control point loss and the refinement loss: $\Theta = \Theta_w + \lambda\Theta_r$, where λ is a scale factor between the control point prediction error and the trajectory refinement error, and is empirically set to $\lambda = 1$

Training Procedure. Our end-to-end network consists of one multi-relational GCN layer followed by eight CNN layers. DropEdge (Rong et al. 2020) with 0.8 rate is used for the GCN layer, and PReLU activation is used for all layers. Data augmentation schemes like random flip, rotation, and scaling are performed during the training phase. We train our model using a SGD optimizer with a batch size of 128 and learning rate of $1e-4$ for 512 epochs, which usually takes one day on a machine with an NVIDIA 2080Ti GPU.

4 Experiments

We conduct extensive experiments on various benchmark datasets. For strictly fair comparison with state-of-the-art models and our ablation study, we follow a standard evaluation protocol in (Gupta et al. 2018). In this experiments, we have tried our best to obtain the best results of competitive methods with the codes that are released publicly in a common pipeline, proposed by (Gupta et al. 2018).

4.1 Datasets

We evaluate our Graph-TERN using four real-world public datasets: ETH (Pellegrini et al. 2009), UCY (Lerner, Chrysanthou, and Lischinski 2007), Stanford Drone Dataset (SDD) (Robicquet et al. 2016), and Train Station dataset (Yi, Li, and Wang 2015). In ETH and UCY datasets, there are five different scenes (ETH, HOTEL, UNIV, ZARA1, and ZARA2) with various complex social interactions such as collision avoidance, group movement, and people stopping. SDD contains 20 scenes captured by a drone of top-down views around university environments. SDD consists of various objects exhibiting non-linear behaviors such as turning around and people stopping. The most crowded Train Station dataset contains up to 332 people simultaneously, and there are 10 entry/exit areas without any location information about them. We follow the standard evaluation strategy as used in (Alahi et al. 2016; Gupta et al. 2018; Kosaraju et al. 2019; Mohamed et al. 2020; Mangalam et al. 2020).

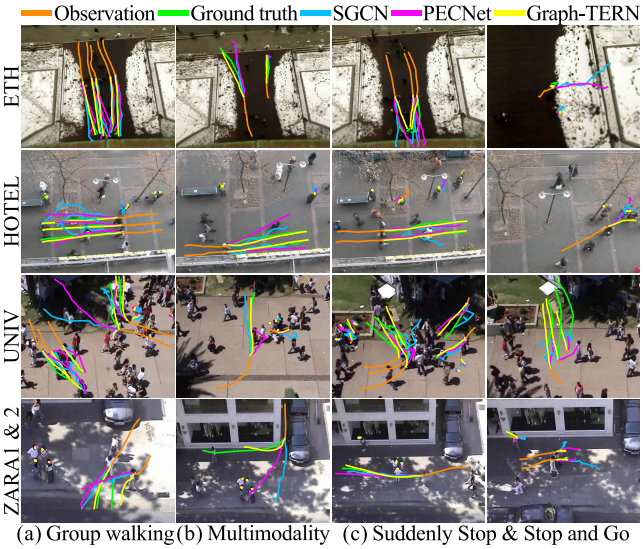


Figure 5: Visualization of prediction results. We compare Graph-TERN, SGCN and PECNet, whose results are reproduced with the pre-trained network. To aid visualization, trajectories with the best ADE on 20 samples are reported.

Evaluation Metrics. Following the standard evaluation strategy (Alahi et al. 2016; Gupta et al. 2018), all datasets are downsampled to 2.5 fps. The observation frames are $T_{obs} = 8$ (3.2s) and the prediction frames are $T_{pred} - T_{obs} = 12$ (4.8s). In our evaluations, we use common quantitative measures

Model	DESIRE	Social-GAN	STGAT	SoPhie	Social-STGCNN	EvolveGraph	P2TIRL	Trajectron++	PECNet	
Year	2017	2018	2019	2019	2020	2020	2020	2020	2020	
Samples	5	20	20	20	5	20	20	20	5	20
ADE	19.25	27.23	18.80	16.27	20.60	13.90	12.58	11.40	12.79	9.96
FDE	34.05	41.44	31.30	29.38	33.10	22.90	22.07	20.12	25.98	15.88
Model	SimAug	MG-GAN	DMRGCN		SGCN		LBEBM		**Graph-TERN**	
Year	2020	2021	2021		2021		2021		-	
Samples	20	20	5	20	5	20	5	20	5	20
ADE	10.27	13.60	17.57	14.31	14.62	11.67	13.58	9.03	**12.35**	**8.43**
FDE	19.71	25.80	32.24	24.78	26.17	19.10	26.57	15.97	**23.32**	**14.26**

Table 2: Comparison of our Graph-TERN with other state-of-the-art methods on SDD (Unit: pixel). We report evaluation results for both $L = 5$ and $L = 20$. The evaluation results are directly referred from (Mangalam et al. 2020; Li et al. 2020; Dendorfer, Elflein, and Leal-Taixé 2021; Shi et al. 2021a). Bold: Best, Underline: Second best.

Model	Social-GAN	STGAT	PECNet	Trajectron++
Year	2018	2019	2020	2020
ADE (meter)	0.39	0.40	0.46	0.33
ADE (pixel)	15.82	15.60	16.58	12.75
FDE (meter)	0.80	0.83	0.76	0.73
FDE (pixel)	32.50	31.90	28.38	24.23
Model	Social-STGCNN	DMRGCN	SGCN	**Graph-TERN**
Year	2020	2021	2021	-
ADE (meter)	0.37	0.37	0.30	**0.27**
ADE (pixel)	14.03	12.70	11.20	**9.52**
FDE (meter)	0.60	0.65	0.54	**0.46**
FDE (pixel)	22.87	21.01	20.66	**16.91**

Table 3: Comparison of our Graph-TERN with other state-of-the-art methods on Train Station dataset (Unit: meter, pixel). Evaluation results are reproduced with authors' provided source codes. Bold: Best, Underline: Second best.

to determine the accuracy of the pedestrian trajectory, the average displacement error metric (ADE), and the final displacement error metric (FDE). We select the best prediction from among $L = 20$ samples, following prior works.

4.2 Comparison with State-of-the-Art

We compare our Graph-TERN with the following state-of-the-art methods: Social-LSTM (Alahi et al. 2016), Social-GAN (Gupta et al. 2018), SR-LSTM (Zhang et al. 2019), STGAT (Huang et al. 2019), IDL (Li 2019), RSBG (Sun, Jiang, and Lu 2020), STAR (Yu et al. 2020), Reciprocal Learning (Sun, Zhao, and He 2020), Social-STGCNN (Mohamed et al. 2020), PECNet (Mangalam et al. 2020), Trajectron++ (Salzmann et al. 2020), Causal (Chen et al. 2021a), NCE (Liu, Yan, and Alahi 2021), TPNMS (Liang et al. 2021), SGCN (Shi et al. 2021a), S-DPF (Shi et al. 2021b), LBEBM (Pang et al. 2021), DMRGCN (Bae and Jeon 2021), STT (Monti et al. 2022), DESIRE (Lee et al. 2017), SoPhie (Sadeghian et al. 2019), EvolveGraph (Li et al. 2020), P2TIRL (Deo and Trivedi 2020), SimAug (Liang, Jiang, and Hauptmann 2020), and MG-GAN (Dendorfer, Elflein, and Leal-Taixé 2021).

The results on the ETH/UCY dataset are reported in Table 1, whose examples are displayed in Figure 5. Our Graph-TERN achieves the best ADE and FDE for all scenes. Interestingly, the performance gaps between Graph-TERN and the second best work are notable on ETH($\sim 37\%$), HOTEL($\sim$24%) and UNIV($\sim$16%). In ETH and HOTEL, there are many people who abruptly stop walking. Graph-TERN handles this case well by learning the probability of people stopping in the control point prediction as shown in Figure 3(b). For UNIV set with very crowded scenes, our multi-relational GCN synergizes well with the initial trajectory prediction and the refinement module by considering complex social relations. This infers accurate multimodal and stop-and-go predictions based on the well-estimated group movements in Figure 5.

In Table 2, our Graph-TERN also shows the best performance overall among all the comparison methods on SDD, whose examples are displayed in Figure 6. The performance gap between Graph-TERN and the second best method is large for L=5, indicating that Graph-TERN provides more accurate goal-directed paths with fewer samples. Remarkably,

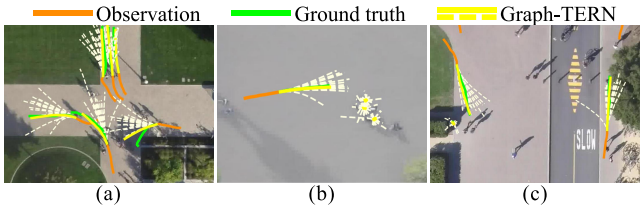


Figure 6: Visualization of multi-modal predictions with $L = 20$ samples on SDD. Our Graph-TERN predicts various plausible trajectories such as turning left, crossing the road, and collision avoidance.

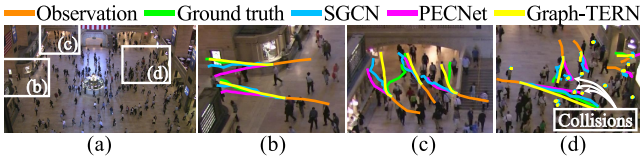


Figure 7: (a) An example of Train Station dataset. (b), (c) and (d) Enlarged images from (a) to visualize prediction results with the best ADE of 20 samples. We compare our Graph-TERN with the second-best and the third-best methods in the experiment on ETH/UCY dataset: PECNet and SGCN.

K	ETH & UCY						SDD (meter)	Train Station
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG		
1	0.83	0.32	**0.50**	0.46	0.41	0.503	0.776	0.481
2	0.77	0.25	**0.50**	0.43	0.39	0.468	**0.771**	0.471
3	**0.71**	**0.23**	0.51	**0.39**	**0.33**	**0.433**	**0.771**	**0.462**
4	**0.71**	0.25	0.52	0.45	0.39	0.464	0.805	0.508
6	0.83	0.25	0.56	0.45	0.40	0.500	0.826	0.570
12	0.94	0.25	0.64	0.53	0.45	0.561	0.875	0.626

Table 4: Ablation study on the number of control points in FDE. Bold: Best, Underline: Second best.

	A Set of Control Points Prediction (FDE)			Initial Trajectory Refinement (ADE)		
	ETH & UCY	SSD (meter)	Train Station	ETH & UCY	SSD (meter)	Train Station
w/o GCN	0.472	0.834	0.553	0.276	0.477	0.284
GCN	0.476	0.795	0.522	0.275	0.443	0.274
w/o inv.	0.463	0.787	0.517	0.270	0.442	0.274
MRGCN	0.433	0.771	**0.462**	0.255	0.433	**0.265**
w/ Pruning	**0.385**	**0.701**	**0.462**	**0.240**	**0.414**	**0.265**

Table 5: Ablation study on the effectiveness of a MRGCN and GMM pruning. The initial trajectory refinement results are based on those of MRGCN in a set of control points prediction. Bold: Best, Underline: Second best.

Figure 6(a) includes a result that predicts all possible directions at the intersection, but in Figure 6(b,c), the path is not predicted where a collision is expected. This means that our set of control points successfully performs the disconnected traversable manifold prediction task using intermediate social interactions, compared to existing endpoint methodologies.

In Table 3, there is a significant performance gap between our Graph-TERN and others for the Train Station dataset. Interestingly, our control points-conditioned method captures the crowd interaction streaming to the exits in Figure 7. In contrast, the state-of-the-art methods have difficulty modeling, with no intermediate relation aggregation at the endpoint prediction.

4.3 Ablation Study

We conduct extensive ablation studies of all benchmark scenes in the metric system to examine the scene-agnostic effectiveness of each component of our Graph-TERN.

Number of Control Points. We examine our Graph-TERN performance with respect to the number of sections K. Table 4 shows that the averaged FDE is minimal around $K = 3$ and $K = 4$. We note that when using $K = 3$ and $K = 4$, we observe that Graph-TERN models a case of pedestrians who suddenly stop walking with a probabilistic distribution, especially in the ETH set. This leads to a moderate performance improvement. Because Graph-TERN with $K = 4$ shows the best performance with only ETH, we set to $K = 3$.

Social Interactions. We demonstrate the effectiveness of a multi-relational GCN. We compare three cases: without a GCN, with a single relational GCN (Mohamed et al. 2020) and with a multi-relational GCN, whose results are reported in Table 5. We apply the multi-relational GCN to a set of control points prediction and trajectory refinement. As expected, capturing complex social relations with the multi-relational GCN provides better performances in FDE for control point predictions and ADE for refinement. One interesting finding is that adding inverse terms to the multi-relational GCN achieves the best prediction results. Since different channels in the multi-relational GCN have independent spatial aggregation kernels, various types of spatial aggregations are available when the number of relations is increased by the inverse terms. Additionally, the GMM pruning is also helpful for accurate destination prediction by preventing out-of-distribution samples.

	ETH & UCY						SDD (meter)	Train Station
	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG		
Dir	0.48	0.26	0.30	0.26	0.23	0.306	0.462	0.272
Con	0.60	0.21	0.38	0.29	0.26	0.348	0.493	0.297
Lin	**0.42**	**0.14**	**0.26**	**0.21**	**0.17**	**0.240**	**0.414**	**0.265**
w/o GE	0.45	0.14	0.26	0.21	0.17	0.246	0.426	0.266

Table 6: Ablation study on initialization methods for the refinement module. Dir, Con, Lin, and GE denote a direct inference of intermediate points without any refinement, a connection between control points, a connection between a last observed frame and a predicted endpoint, and a guided endpoint sampling, respectively. ADE results are reported. Bold: Best, Underline: Second best.

Initial Trajectory. As discussed in Section 3.4, the trajectory refinement requires an initialization step. This can be provided in the form of an initial trajectory. The initial trajectories are obtained from (1) connections between control points and the predicted endpoint and (2) a connection between the last observed frame and a predicted endpoint. We report the evaluation results according to these initializations as well as the direct inference of intermediate points. Table 6 shows that the connection between the last observed frame and a predicted endpoint provides the best initialization. In other words, Graph-TERN appears to learn best when it finds the relationships between graph nodes starting with an unbiased initialization. In contrast, the connection between control points and a predicted endpoint acts as a hard constraint in our refinement modules. Additionally, the guided endpoints sampling increases the robustness of the trajectory refinement module with the slight performance gain.

5 Conclusion

We present a set of control points prediction and a refinement network for pedestrian trajectory prediction. The control point prediction allows the accurate computation of the final destinations of pedestrians and the refinement provides a socially acceptable trajectory. By incorporating a multi-relational GCN, our model achieves state-of-the-art results by modeling complex social interactions in real-world scenes.

Directions exist for improving Graph-TERN. One is to integrate scene semantics into the graph nodes of our multi-relational GCN. Another direction is to employ contextual geometry knowledge in the control point prediction.

Acknowledgements

This work is in part supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST), No.2021-0-02068, Artificial Intelligence Innovation Hub), Vehicles AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea (NIPA) funded by the Ministry of Science and ICT (No.S1602-20-1001) and the National Research Foundation of Korea (NRF) (No.2020R1C1C1012635) grant funded by the Korea government (MSIT).

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bae, I.; and Jeon, H.-G. 2021. Disentangled Multi-Relational Graph Convolutional Network for Pedestrian Trajectory Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Bae, I.; Park, J.-H.; and Jeon, H.-G. 2022a. Learning Pedestrian Group Representations for Multi-modal Trajectory Prediction. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Bae, I.; Park, J.-H.; and Jeon, H.-G. 2022b. Non-Probability Sampling Network for Stochastic Human Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bartoli, F.; Lisanti, G.; Ballan, L.; and Del Bimbo, A. 2018. Context-aware trajectory prediction. In *Proceedings of International Conference on Pattern Recognition (ICPR)*.

Chai, Y.; Sapp, B.; Bansal, M.; and Anguelov, D. 2019. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning (CoRL)*.

Chen, G.; Li, J.; Lu, J.; and Zhou, J. 2021a. Human Trajectory Prediction via Counterfactual Analysis. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Chen, G.; Li, J.; Zhou, N.; Ren, L.; and Lu, J. 2021b. Personalized Trajectory Prediction via Distribution Discrimination. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Dendorfer, P.; Elflein, S.; and Leal-Taixé, L. 2021. MG-GAN: A Multi-Generator Model Preventing Out-of-Distribution Samples in Pedestrian Trajectory Prediction. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Dendorfer, P.; Osep, A.; and Leal-Taixe, L. 2020. Goal-GAN: Multimodal Trajectory Prediction Based on Goal Position Estimation. In *Proceedings of Asian Conference on Computer Vision (ACCV)*.

Deo, N.; and Trivedi, M. M. 2020. Trajectory Forecasts in Unknown Environments Conditioned on Grid-Based Plans. *arXiv preprint arXiv:2001.00735*.

Fernando, T.; Denman, S.; Sridharan, S.; and Fookes, C. 2018. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural Networks*, 108: 466–478.

Gu, T.; Chen, G.; Li, J.; Lin, C.; Rao, Y.; Zhou, J.; and Lu, J. 2022. Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Helbing, D.; and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E*, 51(5): 4282.

Huang, Y.; Bi, H.; Li, Z.; Mao, T.; and Wang, Z. 2019. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Hug, R.; Becker, S.; Hübner, W.; Arens, M.; and Beyerer, J. 2022. Bézier Curve Gaussian Processes. *arXiv preprint arXiv:2205.01754*.

Hug, R.; Hübner, W.; and Arens, M. 2020. Introducing probabilistic bézier curves for n-step sequence prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ivanovic, B.; and Pavone, M. 2019. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Kothari, P.; Sifringer, B.; and Alahi, A. 2021. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, N.; Choi, W.; Vernaza, P.; Choy, C. B.; Torr, P. H. S.; and Chandraker, M. 2017. DESIRE: Distant Future Prediction in Dynamic Scenes With Interacting Agents. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. *Computer Graphics Forum*, 26(3): 655–664.

Li, J.; Ma, H.; and Tomizuka, M. 2019. Conditional Generative Neural System for Probabilistic Trajectory Prediction. *Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Li, J.; Yang, F.; Tomizuka, M.; and Choi, C. 2020. EvolveGraph: Multi-Agent Trajectory Prediction with Dynamic Relational Reasoning. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-structural graph convolutional networks for

skeleton-based action recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, S.; Zhou, Y.; Yi, J.; and Gall, J. 2021. Spatial-Temporal Consistency Network for Low-Latency Trajectory Forecasting. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Li, Y. 2019. Which Way Are You Going? Imitative Decision Learning for Path Forecasting in Dynamic Scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, J.; Jiang, L.; and Hauptmann, A. 2020. SimAug: Learning Robust Representations from Simulation for Trajectory Prediction. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Liang, J.; Jiang, L.; Murphy, K.; Yu, T.; and Hauptmann, A. 2020. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, J.; Jiang, L.; Niebles, J. C.; Hauptmann, A. G.; and Fei-Fei, L. 2019. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, R.; Li, Y.; Li, X.; Tang, Y.; Zhou, J.; and Zou, W. 2021. Temporal Pyramid Network for Pedestrian Trajectory Prediction with Multi-Supervision. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Liu, Y.; Yan, Q.; and Alahi, A. 2021. Social NCE: Contrastive Learning of Socially-aware Motion Representations. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mangalam, K.; An, Y.; Girase, H.; and Malik, J. 2021. From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mangalam, K.; Girase, H.; Agarwal, S.; Lee, K.-H.; Adeli, E.; Malik, J.; and Gaidon, A. 2020. It is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Manh, H.; and Alaghband, G. 2018. Scene-lstm: A model for human trajectory prediction. *arXiv preprint arXiv:1808.04018*.

Marcheggiani, D.; and Titov, I. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marchetti, F.; Becattini, F.; Seidenari, L.; and Bimbo, A. D. 2020. MANTRA: Memory Augmented Networks for Multiple Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marchetti, F.; Becattini, F.; Seidenari, L.; and Del Bimbo, A. 2022. SMEMO: Social Memory for Trajectory Forecasting. *arXiv preprint arXiv:2203.12446*.

Mehran, R.; Oyama, A.; and Shah, M. 2009. Abnormal crowd behavior detection using social force model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Monti, A.; Porrello, A.; Calderara, S.; Coscia, P.; Ballan, L.; and Cucchiara, R. 2022. How Many Observations Are Enough? Knowledge Distillation for Trajectory Forecasting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pang, B.; Zhao, T.; Xie, X.; and Wu, Y. N. 2021. Trajectory Prediction with Latent Belief Energy-Based Model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pellegrini, S.; Ess, A.; and Gool, L. V. 2010. Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Rehder, E.; and Kloeden, H. 2015. Goal-Directed Pedestrian Prediction. In *Proceedings of International Conference on Computer Vision Workshop (ICCVW)*.

Rehder, E.; Wirth, F.; Lauer, M.; and Stiller, C. 2018. Pedestrian Prediction by Planning Using Deep Neural Networks. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Rong, Y.; Huang, W.; Xu, T.; and Huang, J. 2020. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations (ICLR)*.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Shafiee, N.; Padir, T.; and Elhamifar, E. 2021. Introvert: Human Trajectory Prediction via Conditional 3D Attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zheng, F.; Zheng, N.; and Hua, G. 2022. Social Interpretable Tree for Pedestrian

Trajectory Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zhou, M.; Niu, Z.; and Hua, G. 2021a. SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, X.; Shao, X.; Fan, Z.; Jiang, R.; Zhang, H.; Guo, Z.; Wu, G.; Yuan, W.; and Shibasaki, R. 2020. Multimodal interaction-aware trajectory prediction in crowded space. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Shi, X.; Shao, X.; Wu, G.; Zhang, H.; Guo, Z.; Jiang, R.; and Shibasaki, R. 2021b. Social-DPF: Socially Acceptable Distribution Prediction of Futures. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Sun, H.; Zhao, Z.; and He, Z. 2020. Reciprocal learning networks for human trajectory prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, J.; Jiang, Q.; and Lu, C. 2020. Recursive Social Behavior Graph for Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tao, C.; Jiang, Q.; and Duan, L. 2020. Dynamic and Static Context-aware LSTM for Multi-agent Motion Prediction. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Varshneya, D.; and Srinivasaraghavan, G. 2017. Human Trajectory Prediction using Spatially aware Deep Attention Models. *arXiv preprint arXiv:1705.09436*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.

Vemula, A.; Muelling, K.; and Oh, J. 2018. Social attention: Modeling attention in human crowds. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*.

Wang, C.; Wang, Y.; Xu, M.; and Crandall, D. J. 2022. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters (RA-L)*.

Wen, S.; Wang, H.; and Metaxas, D. 2022. Social ODE: Multi-agent Trajectory Forecasting with Neural Ordinary Differential Equations. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Xu, C.; Li, M.; Ni, Z.; Zhang, Y.; and Chen, S. 2022a. GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction with Relational Reasoning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, C.; Mao, W.; Zhang, W.; and Chen, S. 2022b. Remember Intentions: Retrospective-Memory-based Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, P.; Hayet, J.-B.; and Karamouzas, I. 2022. SocialVAE: Human Trajectory Prediction Using Timewise Latents. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Xu, Y.; Wang, L.; Wang, Y.; and Fu, Y. 2022c. Adaptive Trajectory Prediction via Transferable GNN. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, Y.; Yang, J.; and Du, S. 2020. CF-LSTM: Cascaded feature-based long short-term networks for predicting pedestrian trajectory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Xue, H.; Huynh, D. Q.; and Reynolds, M. 2018. SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Yamaguchi, K.; Berg, A. C.; Ortiz, L. E.; and Berg, T. L. 2011. Who are you with and where are you going? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Yao, Y.; Atkins, E.; Johnson-Roberson, M.; Vasudevan, R.; and Du, X. 2021. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters (RA-L)*.

Yi, S.; Li, H.; and Wang, X. 2015. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, C.; Ma, X.; Ren, J.; Zhao, H.; and Yi, S. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Yue, J.; Manocha, D.; and Wang, H. 2022. Human trajectory prediction via neural social physics. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, H.; Gao, J.; Lan, T.; Sun, C.; Sapp, B.; Varadarajan, B.; Shen, Y.; Shen, Y.; Chai, Y.; Schmid, C.; Li, C.; and Anguelov, D. 2020. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning (CoRL)*.

Zhao, T.; Xu, Y.; Monfort, M.; Choi, W.; Baker, C.; Zhao, Y.; Wang, Y.; and Wu, Y. N. 2019. Multi-Agent Tensor Fusion for Contextual Trajectory Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.