Hierarchical Multi-Supervision Multi-Interaction Graph Attention Network for Multi-Camera Pedestrian Trajectory Prediction

Guoliang Zhao[1], Yuxun Zhou[2], Zhanbo Xu[1], Yadong Zhou[1], Jiang Wu[1]

[1]Institute of Automation Science and Engineering, Xian Jiaotong University, China
[2]Department of Electrical Engineering and Computer Sciences, UC Berkeley, United States
zgl934455716@stu.xjtu.edu.cn, yxzhou@berkeley.edu,
zbxu@sei.xjtu.edu.cn, {ydzhou, jiangwu}@xjtu.edu.cn

Abstract

Pedestrian trajectory prediction has become an essential underpinning in various human-centric applications including but not limited to autonomous vehicles, intelligent surveillance system and social robotics. Previous research endeavors mainly focus on single camera trajectory prediction (SCTP), while the problem of multi-camera trajectory prediction (MCTP) is often overly simplified into predicting presence in the next camera. This paper addresses MCTP from a more realistic yet challenging perspective, by redefining the task as a joint estimation of both future destination and possible trajectory. As such, two major efforts are devoted to facilitating related research and advancing modeling techniques. Firstly, we establish a comprehensive multi-camera Scenes Pedestrian Trajectory Dataset (mcScenes), which is collected from a real-world multi-camera space combined with thorough human interaction annotations and carefully designed evaluation metrics. Secondly, we propose a novel joint prediction framework, namely HM^{3}GAT, for the MCTP task by building a tailored network architecture. The core idea behind HM^{3}GAT is a fusion of topological and trajectory information that are mutually beneficial to the prediction of each task, achieved by deeply customized networks. The proposed framework is comprehensively evaluated on the mcScenes dataset with multiple ablation experiments. Status-of-the-art SCTP models are adopted as baselines to further validate the advantages of our method in terms of both information fusion and technical improvement. The mcScenes dataset, the HM^{3}GAT, and alternative models are made publicly available for interested readers.

Introduction

The information about human behavior, especially human trajectory, is of major importance for multiple human-centric application domains, including autonomous driving (Chai et al. 2019), intelligent surveillance system (Bastani, Marcenaro, and Regazzoni 2016) and social robotic navigation (Rhinehart, Kitani, and Vernaza 2018). Trajectory prediction, in a nutshell, aims to estimate socially acceptable trajectories in a near future, according to historical records in a time window from the past. Despite of an unprecedented development in this research direction, most works only consider pedestrian trajectory in a single scene from an aerial view, which is referred to as single camera trajectory prediction (SCTP). This simplification leads to a critical drawback: The location and the inter-connection (topological) information among different camera scenes are completely neglected. However in most public spaces, such as transportation hubs, shopping malls and subway stations, occupant spaces are usually composed of multi-camera networks, and pedestrian presence prediction in the next possible camera scene is much more valuable. With that, we re-define the *multi-camera trajectory prediction* (MCTP) task as follows: In a space with a multi-camera network, MCTP performs a comprehensive prediction about future trajectories, including both the next camera and the detailed trajectory in the current camera scene, based on topological information and historical knowledge available from the multi-camera network. A specific example is given in Fig. 1, where pedestrians move in a space with a multi-camera network composed of three cameras. The MCTP defined in this paper strives to learn a model that not only can predict a target pedestrian's future trajectory, but also can predict the next camera that the target pedestrian will likely reach. This combined approach would resolve practical needs in an unified framework, and as will be shown in this paper, would allow the fusion of multi-supervision information to benefit each sub-task.

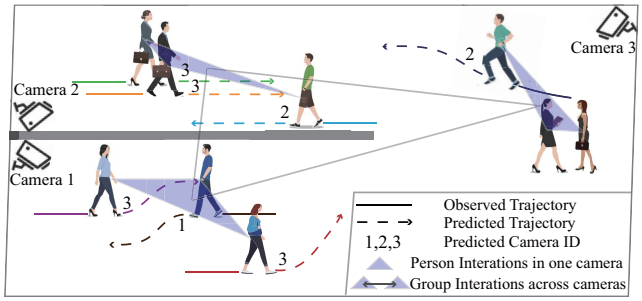

Figure 1: An illustration of the HM^{3}GAT. By leveraging the strong association between pedestrian trajectory and their potential intention, HM^{3}GAT achieves a joint prediction of both next camera and future trajectory.

The MCTP task defined here is confronted by two major challenges. Practically, most of the existing datasets are acquired for SCTP, such as ETH and UCY (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007), Although

the WNMF (Styles et al. 2020) dataset includes several cameras, it lacks the important pedestrian interaction annotations hence are not suitable for MCTP considered in this work. Technically, a joint modeling can be arduous as issues about representation learning and information sharing could arise when multiple layers of data sources are combined together. On the one hand, pedestrian interaction, randomness of motion and latent intent of pedestrian are already hard to describe with many entangled factors. On the other hand, it's intuitively helpful for MCTP to include multi-supervision information fusion, because knowledge about topological layout, pedestrians' mutual interaction, historical motion and latent intent are intrinsically entangled and carry intimations on each other.

To overcome the challenges and solve the joint prediction problem raised by the redefined MCTP task, we first create a multi-camera pedestrian trajectory dataset, namely mcScenes, that allows to train and compare different models in a quantitative way. Then we establish a joint future destination and trajectory prediction framework, called hierarchical multi-supervision multi-interaction graph attention network (HM^{3}GAT), to capture the hierarchical, coupled data structure with a history motion encoder, a social interaction encoder, and a latent goal decoder. The final prediction is achieved through the fusion of topological information, pedestrian historical motion, hierarchical interaction and latent intent by a future trajectory decoder. To summarize the main contributions:

- We establish a carefully labeled multi-camera pedestrian trajectory dataset and evaluation metrics for MCTP, which would benefit the community for future research.
- A joint future destination and trajectory prediction framework is proposed for MCTP task, which effectively captures both the human social interaction across multi-camera scenes and the destination/topological information for better prediction.
- We establish a benchmark for MCTP and compare various methods on our multi-camera trajectory dataset.

Related Work

Single-camera Trajectory Prediction. Most previous works on pedestrian trajectory prediction focused on single-camera view, mainly based on two methods, i.e., RNNs and GNNs. Because pedestrian trajectory prediction is a seq-to-seq in nature, RNNs and their variants, e.g., LSTM, were also adopted for this problem. (Alahi et al. 2016) proposed a Social-LSTM model, which aggregates the hidden states of neighbor pedestrians on a grid by a social pooling layer. Extending the idea of Social-LSTM, Social-Attention (Vemula, Muelling, and Oh 2018) modeled pedestrian interactions as a spatio-temporal graph and adds attention mechanism to the social pooling layer. Group-LSTM (Bisagno, Zhang, and Conci 2018) and SR-LSTM (Zhang et al. 2019) reused the social pooling with different pooling mechanism. Social-GAN (Gupta et al. 2018) combined GAN with LSTM to generate multi-modal pedestrian trajectory.

With the pervasive success of graph neural networks (GNN) for modeling relations, a large body of research begins to use it to learn the social interaction graph representation. STGAT (Huang et al. 2019) is the first attempt to combine GAT (graph attention network) with LSTM in the context of modeling pedestrian motions. STGCNN (Mohamed et al. 2020) aggregated motion information using GCN (Graph Convolutional Network) on a spatio-temporal graph. DMRGCN (Bae and Jeon 2021) is also a GCN-based method, which can learn sophisticated social relations between pedestrians using multi-scale aggregation. Similarly, our HM^{3}GAT also employs graph embedding methods to capture pedestrian interactions. The difference is that our method divides pedestrian interactions into person interaction in single camera and group interaction across cameras.

Joint Prediction Framework and Multi-camera Trajectory Prediction. Until recently, only a few works have addressed Multi-Camera Trajectory Prediction. The first work (Styles et al. 2020) put forward a formulation of MCTP task with a simplified goal to predict the next camera given observed trajectory for several seconds in single camera. A complete MCTP task, on the other hand, should consist of both the next camera prediction and trajectory prediction. Other recent works, e.g., (Liang et al. 2019) and PECNet (Mangalam et al. 2020), attempted to combine the prediction of activity/endpoint and trajectory. However their modeling techniques were still traditional which might not be able to comprehensively capture the complex structure of MCTP data. To the best of our knowledge, this paper is the first one to handle both destination and trajectory prediction with a tailored and unified graph attention network.

Trajectory Datasets. Existing pedestrian trajectory datasets can be divided into three categories: single-camera trajectory datasets (Robicquet et al. 2016; Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007), multi-future trajectory datasets (Liang et al. 2020) and multi-camera trajectory datasets (Styles et al. 2020; Ristani et al. 2016). To overcome the limitation of previous datasets, we establish a new one with enriched pedestrian interactions. Note that this is the first real-world MTCP dataset with carefully labeled pedestrian interaction annotations in crowds.

The HM^{3}GAT Framework

In this section, we introduce a joint destination and trajectory prediction framework, namely HM^{3}GAT. It essentially consists of a history motion encoder, a social interaction encoder, a latent goal decoder and a future trajectory decoder. An overview is illustrated in Fig. 2.

Problem Definition

The multi-camera trajectory prediction problem involves a joint prediction of future destination and trajectory based on observed position sequences for all pedestrians across multi-camera network. We assume that there are M scenes involved in a multi-camera space. Given the topology of multi-camera network and a set of N pedestrians across the multi-camera network with their observed positions $tr_{obs}^{n,i}$, $n \in \{1, ..., N\}$, $i \in \{1, ..., M\}$ over a time period T_{obs}, we need to predict the next camera s_j and the future trajectory $tr_{pred}^{n,i}$ over a future time period T_{pred}. For each pedestrian n,

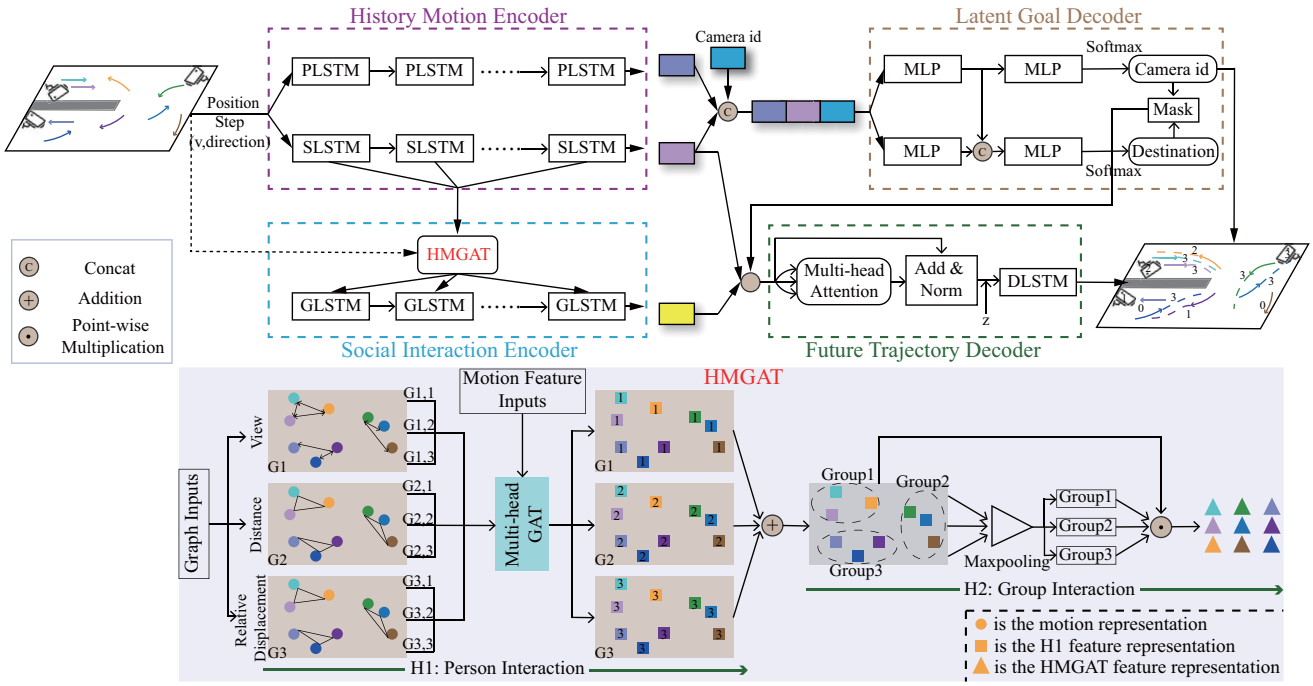

Figure 2: An overview of the HM^{3}GAT model for a joint future destination and trajectory prediction, elaborated in sections of diagram part. With 2-dimensional pixel coordinates of N pedestrians for T_{obs} frames, motion features are extracted by history motion encoder and the multi-interaction graphs are constructed for HMGAT to generate the social interaction features. Then pedestrian latent goals are decoded by HMC model. Finally, we use Transformer encoder to aggregate these three types of features, and extrapolate future trajectories with DLSTM.

we denote the corresponding trajectory by $tr_{obs}^{n,i} = \{p_t^{n,i} = (x_t^{n,i}, y_t^{n,i}) | \ t \in \{1, ..., T_{obs}\}\}$, where $p_t^{n,i} = (x_t^{n,i}, y_t^{n,i})$ are the pixel coordinates of pedestrian n at a specific time t in the ith camera scene. Similar to the predicted trajectory. We assume that the predicted coordinates $(\hat{x}_t^{n,i}, \hat{y}_t^{n,i})$ and s_j are random variables. Therefore, with the observed trajectory $tr_{1:T_{obs}}$ for all pedestrians across the multi-camera network, our goal is to predict the future camera s_j and the future trajectory $tr_{T_{obs}+1:T_{pred}}$.

History Motion Encoder

The generation of pedestrian trajectory is related to pedestrian motion state. Each pedestrian has his/her own motion pattern, including step length (representing direction and speed) and position. Extracting above information from observed trajectory is a key to the success of motion state representation. Based on several established works from computer vision (Alahi et al. 2016; Su et al. 2017), we propose to use SLSTM for step encoding and PLSTM for position encoding, which are two LSTMs that don't share weights.

We use PLSTM and SLSTM to encode pedestrian position information $p_{1:T_{obs}}^n = (x_{1:T_{obs}}^n, y_{1:T_{obs}}^n)$ and step information $\Delta p_{1:T_{obs}}^n = (\Delta x_{1:T_{obs}}^n, \Delta y_{1:T_{obs}}^n)$:

$$e_t^n = \phi^p(x_t^n, y_t^n; W_1) \tag{1}$$

$$v_t^n = \phi^s(\Delta x_t^n, \Delta y_t^n; W_2) \tag{2}$$

$$P_t^n = PLSTM(P_{t-1}^n, e_t^n; W_p) \tag{3}$$

$$S_t^n = SLSTM(S_{t-1}^n, v_t^n; W_s) \tag{4}$$

where $\phi^p(\cdot)$ and $\phi^s(\cdot)$ are embedding functions. W_1 and W_2 are the embedding weight. P_t^n and S_t^n are the hidden state of the PLSTM and SLSTM at t time step. W_p and W_s are the weight of PLSTM cell and SLSTM cell.

Social Interaction Encoder

Pedestrian trajectory is affected by the motion state of surrounding pedestrians, called social interaction. In the past few years, many methods were proposed to model social interaction, such as Social Force (Helbing and Molnar 1995), Social Pooling (Alahi et al. 2016; Zhang et al. 2019) and Social Graph Embedding (Huang et al. 2019; Mohamed et al. 2020). However, all existing methods are limited to single camera scene. To model the social interaction across the multi-camera network, this work proposes a novel Hierarchical Multi-interaction Graph Attention Network (HMGAT) to encode the social interaction from two levels:

- H1: encode the social interactions in single camera scene, referred to as the person interaction.
- H2: encode the social interactions of groups across multi-camera scenes, named group interaction.

Regarding H1, the goal is to capture the person-to-person social interactions in single camera scene. The factors affecting the strength of pedestrian interactions mainly include view, distance and relative displacement. Hence we introduce a multi-relational social graph with three types of relations $R = \{view, distance, relative\_displacement\}$ (Li

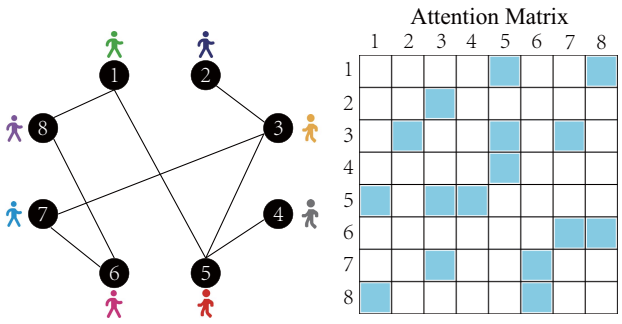

Figure 3: An illustration of self-attention for graph. The social interaction graph and corresponding attention matrix.

et al. 2019). As shown in Fig. 2, we assign different edges on multi-relational social graphs to represent the existing person-to-person interactions and place the output of history motion encoder as the nodes of multi-relational social graphs. Mathematically, our spatio-temporal social interaction graph can be described as a set of spatio social graph $G = \{G_t^{r,m} | r \in R, m \in \{1, ..., M\}, t \in \{1, ..., T_{obs}\}\}$, where $G_t^{r,m}$ is the rth relational social graph at t time step in m camera scene. Its adjacency matrix $A_t^{r,m} = \{a_{i,j,t}^{r,m} | r \in R, m \in \{1, ..., M\}, i, j \in \{1, ..., N\}, t \in \{1, ..., T_{obs}\}\}$ represents the physical relationships between pedestrians i and j. Its node representation $H_t^{r,m} = \{h_{n,t}^{r,m} | r \in R, m \in \{1, ..., M\}, n \in \{1, ..., N\}, t \in \{1, ..., T_{obs}\}\}$ represents the motion feature of pedestrian n. $A_t^{r,m}$ is normalized by:

$$\hat{A}_t^{r,m} = MinMaxScaler(A_t^{r,m}) \tag{5}$$

where $\hat{A}_t^{r,m} \in [0, 1]$.

As such, the node features of three types of social graphs can be updated as fellows:

$$\hat{H}_t^{r,m} = \sigma_1(A_t^{r,m} H_t^{r,m} W^r) \tag{6}$$

where $\sigma_1(\cdot)$ is a nonlinear activation function (such as ReLU), and W^r indicates a learnable weight matrix.

Since GAT (Busbridge et al. 2019) allows for aggregating information from neighbors by assigning different importance to different nodes, we use GAT as our information sharing mechanism. As shown in Fig. 3, the graph attention layer is introduced for single camera social graph, which enables a node to assign different importance to different nodes within a neighborhood and to aggregate features from them. The inputs of our Multi-head GAT are $\hat{H}_t^{r,m} = \{\hat{h}_{n,t}^{r,m} \in R^F | n \in \{1, ..., N\}\}$, where F is the feature dimension of each node. From the Multi-head GAT, the aggregated hidden state $\widetilde{H}_t^{r,m}$ can be obtained for all pedestrians of the rth social graph at t time step in m camera scene, which contains the spatial influence from other pedestrians in the same camera scene:

$$\widetilde{H}_t^{r,m} = MultiheadGAT(\hat{H}_t^{r,m}) \tag{7}$$

where MultiheadGAT is the success of self-attention mechanism in graph network (Veličković et al. 2017). Finally, the multi-relational aggregated motion features of all pedestrians at t time step in m camera scene, are given by:

$$\widetilde{H}_t^m = \sigma_2(\sum_{r \in R} \widetilde{H}_t^{r,m}; W) \tag{8}$$

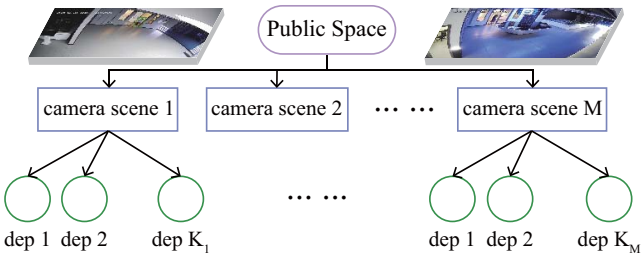

Figure 4: An illustration of the hierarchical structure between camera scenes and departures of camera scenes.

where $\sigma_2(\cdot)$ is a nonlinear activation function. W indicates a weight matrix. $\widetilde{H}_t^m = \{\widetilde{h}_{n,t}^m \in R^F | n \in \{1, ..., N\}\}$.

Next we discuss H2, which models the interactions between pedestrian groups in different camera scenes. Inspired by Starnet network (Zhu et al. 2019), the group motion feature of pedestrians in single camera scene m is given by:

$$\widetilde{g}_t^m = MaxPooling(\widetilde{h}_{1,t}^m, ..., \widetilde{h}_{N_m,t}^m) \tag{9}$$

then, we get the final interaction features of all pedestrians by the weighted point-wise multiplication between group motion features $\widetilde{g}_t^m$ and person motion feature $\widetilde{h}_{n,t}^m$ after H1:

$$\bar{h}_{n,t} = \sum_{m=1}^{M} (w_m \widetilde{g}_t^m \odot \widetilde{h}_{n,t}^m) \tag{10}$$

where w_m is the weight coefficient, representing the importance of different group motion features, and $\odot$ is the point-wise multiplication. $\bar{h}_{n,t}$ is the final output of HMGAT.

Finally, another LSTM is applied to explicitly incorporate the temporal correlations between interactions, named GLSTM (Huang et al. 2019):

$$G_t^n = GLSTM(G_{t-1}^n, \bar{h}_{n,t}; W_g) \tag{11}$$

where G_t^n is the hidden state of the GLSTM. W_g is the weight of GLSTM cell.

Latent Goal Decoder

Various works suggested that a better decoding of the future trajectory can be achieved by capturing the latent goal of target pedestrian. Note that the choice of destination with a pedestrian trajectory is often limited in a given scene. Moreover as shown in Fig. 4, the camera scenes and the exit-entrance pairs between cameras (called departure) exhibit a hierarchical structure. Consequently, we creatively propose a Hierarchical Multi-label Classification (HMC) (Vens et al. 2008) model to predict pedestrian latent intention, which consists of four MLPs and two Softmax layers in Fig. 2.

The inputs of HMC model are the outputs of history motion encoder: $P_{T_{obs}}^n$ and $S_{T_{obs}}^n$, and the one-hot encoder of current camera id $O_{T_{obs}}^n$. With these inputs, we can predict the next camera id and the departure of current camera scene for target pedestrian. The processing is as follows:

$$q_{l_1}^{n,1} = MLP(|||(P_{T_{obs}}^n, S_{T_{obs}}^n, O_{T_{obs}}^n); W_{l_1}^1) \tag{12}$$

$$q_{l_2}^{n,1} = MLP(|||(P_{T_{obs}}^n, S_{T_{obs}}^n, O_{T_{obs}}^n); W_{l_2}^1) \tag{13}$$

$$q_{l_1}^n = MLP(q_{l_1}^{n,1}; W_{l_1}^2) \tag{14}$$
$$q_{l_2}^n = MLP(q_{l_1}^{n,1} || q_{l_2}^{n,1}; W_{l_2}^2) \tag{15}$$
$$e_c^n = Softmax(q_{l_1}^n) \tag{16}$$
$$e_d^n = Softmax(q_{l_2}^n) \tag{17}$$

where $||$ represent concatenate. $W_{l_1}^1$, $W_{l_2}^1$, $W_{l_1}^2$ and $W_{l_2}^2$ are the weight matrix of four MLPs. $e_c^n \in R^{(M,1)}$ is the probability vector of the predicted next camera, and $e_d^n \in R^{(\sum_{i=1}^{M}(K_i),1)}$ is the probability vector of the predicted departure. Important prior knowledge and space outlines, such as the multi-camera network topology $A_c \in R^{(M,M)}$ and the camera-departure topology $A_d \in R^{M,\sum_{i=1}^{M}(K_i)}$, are usually accessible for a specific public area. Thus the masked probability vector of the predicted next camera and the masked probability vector of the predicted departure can be written as:

$$\hat{e}_c^n = O_{T_{obs}}^n A_c \odot e_c^n \tag{18}$$
$$\hat{e}_d^n = O_{T_{obs}}^n A_d \odot e_d^n \tag{19}$$

where $\odot$ is the point-wise multiplication.

Future Trajectory Decoder

At last, with the motion features, social interaction features and latent goal features, we can generate the predicted future trajectory. Firstly, the aforementioned three parts need to be combined to accomplish the information fusion of multi-supervision. To this end, three types of features are connected by the encoder part of Transformer network (Jaderberg et al. 2015) to achieve the effective fusion. More specifically, the feature alignment is done by:

$$E^n = \phi^l(\hat{e}_c^n, \hat{e}_d^n; W_l) \tag{20}$$

where $\phi^l(\cdot)$ is embedding function. W_l is the embedding weight. $E^n \in R^D$ indicates the latent goal feature. So the inputs of Transformer encoder reads:

$$H_F^n = Stack(S_{T_{obs}}^n, G_{T_{obs}}^n, E^n) \tag{21}$$

where the function of Stack is a feature stacking, $H_F^n \in R^{(3,D)}$. With that the fusion of features yields:

$$\hat{H}_F^n = MultiheadAttention(H_F^n) \tag{22}$$
$$\bar{H}_F^n = LayerNorm(\hat{H}_F^n + H_F^n) \tag{23}$$

Finally, we customize LSTM to decode future trajectory, namely DLSTM, with the following state vector:

$$D_{T_{obs}}^n = Flatten(\bar{H}_F^n) \| z \tag{24}$$

where the function of Flatten is flattening features. z is represents noise. The future trajectory is:

$$D_{T_{obs}+1}^n = DLSTM(D_{T_{obs}}^n, e_{T_{obs}}^n; W_d) \tag{25}$$
$$(\Delta x_{T_{obs}+1}^n, \Delta y_{T_{obs}+1}^n) = \delta(D_{T_{obs}+1}^n) \tag{26}$$

where W_d is the weight of DLSTM cell. δ is a linear layer. Similarly, we take the MSE of predicted positions and the CEL of predicted departures (compared with groundtruth) together as our loss functions. The proposed model is trained for 256 epochs with the Adam optimizer. We use a mini-batch size of 32. The initial learning rate is 0.01, and changed to 0.005 after 128 epochs. The training is performed on a NVIDIA TITAN V GPU. Each of the above components has been fine-tuned for optimal hyper-parameters and the readers are referred to the Github repo for more details.

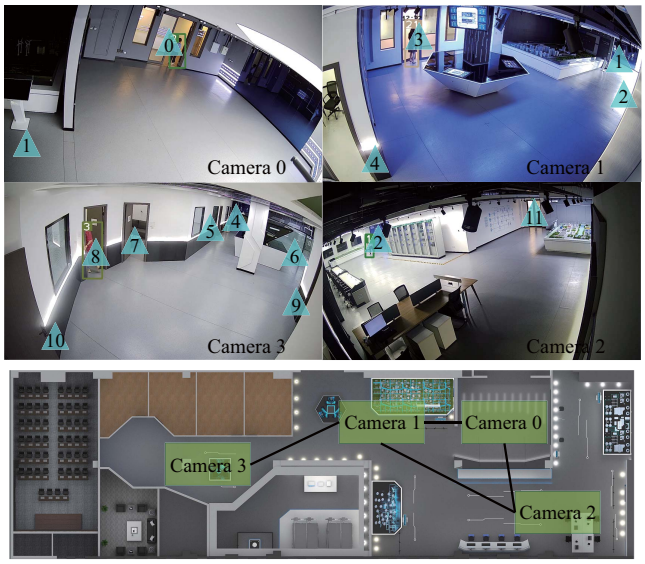

Figure 5: Visualization of camera scenes and camera network topology. Where, these triangles represent the departures in each camera scene.

The Pedestrian Trajectory Dataset: mcScenes

In this section, we introduce our multi-camera scenes pedestrian trajectory dataset annotated by human, called mcScenes, for multi-camera trajectory prediction evaluation.

Existing datasets. Since the trajectory prediction problem was proposed, many trajectory datasets have been established, such as: (1) Single camera trajectory datasets: SDD (Robicquet et al. 2016), ETH/UCY (Pellegrini et al. 2009; Lerner, Chrysanthou, and Lischinski 2007) and VIRAT/ActEV (Awad et al. 2018; Oh et al. 2011); (2) Single camera multi-future trajectory datasets: Forking Paths (Liang et al. 2020); (3) Multi-camera trajectory datasets: WNMF (Styles et al. 2020).

mcScenes Overview. As shown in Fig. 5, Our dataset is constructed from a laboratory space with a surveillance camera angle of view, which consists of $M = 4$ camera scenes and 11 departures in the whole multi-camera network. We collected surveillance video data, and use video processing tools to extract frames from original videos (5 $frame/s$). Then we use Deepsort (Veeramani, Raymond, and Chanda 2018), that is a pedestrian tracking model, to extract pedestrian trajectory roughly. Finally we hired some volunteers to annotate the trajectories. The dataset contains these fields: {frame, id, x, y, interaction_category, camera_id, next_camera_id, departure_id}. In total, mcScenes contains 263 pedestrian trajectories with multiple human interactions and 8843 frames. Due to the page limit, details about the data collection and processing are deferred to the supplement material for interested readers.

Experiments

In this section, we evaluate various methods for multi-camera trajectory prediction on our mcScenes dataset.

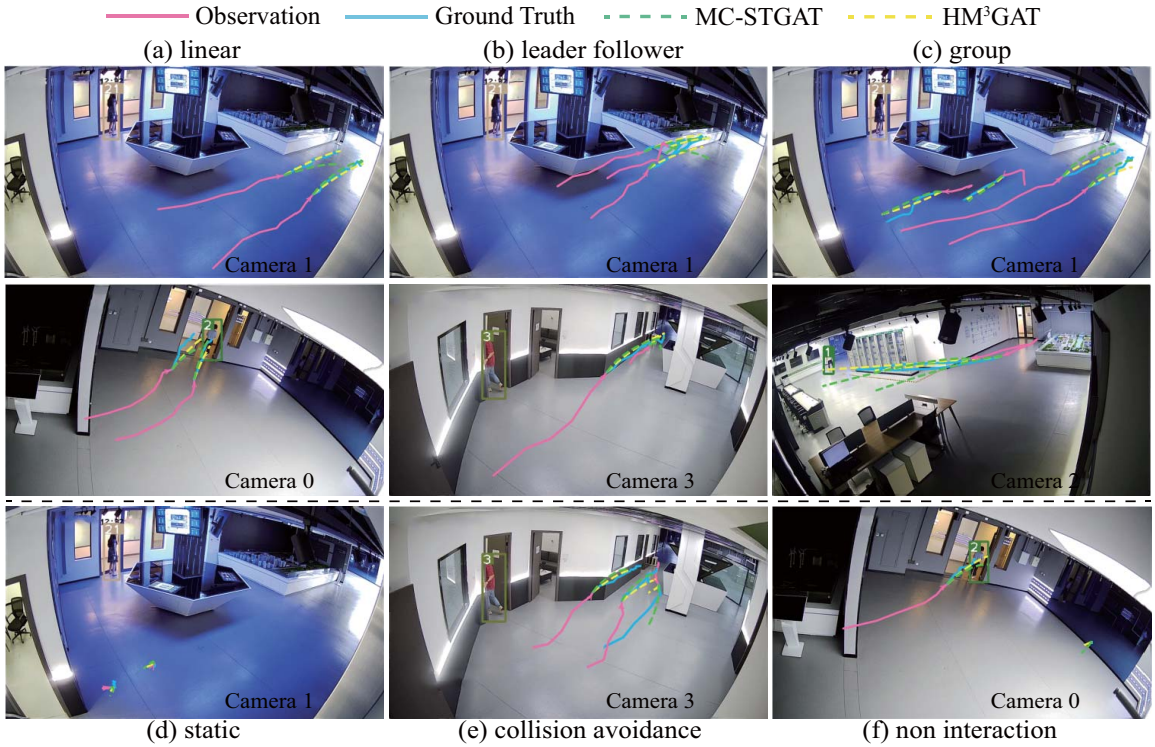

Figure 6: Qualitative analysis. According to the social interaction categories (a∼f), we compare HM^{3}GAT with MC-STGAT in visualization. MC-STGAT is retrained by merging with our proposed joint prediction framework. See text for details.

Benchmark and Evaluation Metrics

Although no off-the-shelf method is previously available for MCTP, we establish several non-trivial baselines and compare them with our HM^{3}GAT model on the mcScenes dataset, to reveal the insight behind HM^{3}GAT and to justify the designed architecture. The first alternative is called ***LSTM+FC****, which can be viewed as a segregated block of our proposed framework. A comparison with ***LSTM+FC**** will show that the proposed framework can achieve effective information fusion and can reveal latent goals that would otherwise be undetectable using individual data source. The second class of alternatives are constructed by replacing the social interaction encoder part of our framework with existing modules in literature, so as to verify the technical advancement of the proposed HM^{3}GAT. More specifically, the following classical methods are included. ***Linear*** (Alahi et al. 2016): a linear regressor that predicts the next coordinates based on previous points. ***Social-LSTM*** (Alahi et al. 2016): a LSTM model with social pooling layer. ***STGAT*** (Huang et al. 2019): a graph attention model for spatio-temporal social graph. ***Social-STGCNN*** (Mohamed et al. 2020): a graph embedding model, GCN, for spatio-temporal social graph.

For the sake of fairness, random re-sampling is performed with a ratio of 8:1:1 to divide the training, validation and test set. Same as prior works, the number of observed time steps is 8 (3.2 sec) of each person and the upcoming trajectory is 12 (4.8 sec). Three metrics are considered to evaluate the performance of different models: *minADE*, *minFDE* and *maxACC* (the proportion of correctly predicted samples in the total number of samples for the next camera prediction).

	mcScenes		
	minADE_{20}	minFDE_{20}	maxACC_{20}
LSTM+FC*	0.066	0.127	0.641
MC-LSTM	0.058	0.113	0.747
MC-S-LSTM	0.050	**0.095**	0.786
MC-STGAT	0.056	0.108	0.774
MC-STGCNN	0.054	0.101	0.783
HM^{3}GAT	**0.049**	**0.095**	**0.789**

Table 1: Comparison of different methods on the mcScenes dataset. Methods are marked with $*$ (LSTM+FC*), which indicates not adopting our proposed joint prediction framework. The pixel coordinate (x, y), used to calculate the ADE and FDE, is normalized by the resolution ($[856, 480]$).

Quantitative Analysis

Table 1 reports the results of our HM^{3}GAT and other baseline methods on the mcScenes dataset in terms of the evaluation metrics. It appears that:

- All baseline methods except for LSTM+FC* are adaptable to our proposed joint future destination and trajectory prediction framework. Methods using the proposed framework all yield acceptable estimation as far as the three evaluation metrics are concerned, e.g., ADE (0.49∼0.58), FDE (0.095∼0.113) and ACC (average

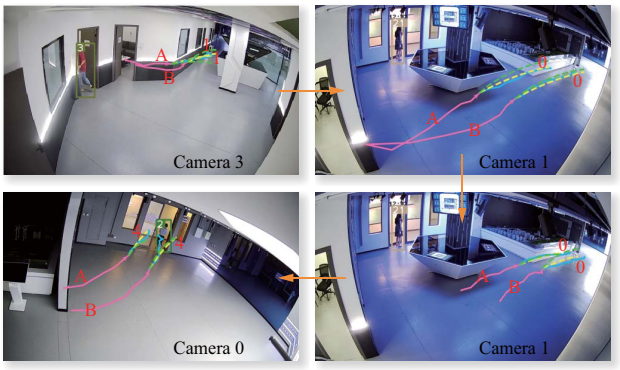

Figure 7: Example of a group of complete trajectories across the multi-camera network.

Variants	Components				AVG
	M&T	SIE		LGD	
		H	MI		(ADE/FDE)
MC-STGAT	✓	×	×	✓	0.056/0.108
HGAT	✓	✓	×	✓	0.054/0.100
HM^{3}GAT	✓	✓	✓	✓	**0.049/0.095**
HMGAT	✓	✓	✓	×	0.050/0.095

Table 2: Ablation experiments. M&T, SIE and LGD respectively denote motion encoder, trajectory decoder, social interaction encoder and latent goal decoder. SIE contains H (hierarchical interaction) and MI (multiple interaction).

0.776). By contrary, LSTM+FC*, not adopting the joint prediction framework, performs the worst with ADE (0.066), FDE (0.127) and ACC (0.641). This justifies our designed information fusion scheme for the MCTP task, and implies that the knowledge about destination would benefit the estimation of trajectory and vice versa.

- Compared with these baseline approaches, HM^{3}GAT achieves remarkable performance gains. our method HM^{3}GAT outperforms all the previous approaches in terms of the average ADE (0.049) and FDE (0.095). This demonstrates that our HM^{3}GAT with a hierarchical multi-interaction graph attention network to model the social interaction between all pedestrians across the multi-camera network indeed helps for MCTP task.

Qualitative Analysis

In this subsection, we provide examples to show intuitively how our HM^{3}GAT successfully captures complex motion behaviors of pedestrians. The prediction results are qualitatively compared between MC-STGAT (Huang et al. 2019) and HM^{3}GAT. It's evident that our model performs better in most instances, especially when the pedestrians' motion exhibit nonlinear patterns.

Fig. 6 shows diverse predictions from multiple social interaction situations, including linear, leader follower, group, static, collision avoidance and non interaction. Fig. 6(a)(b)(c) show three different situations (linear, leader follower and group). By comparison, we observe that our model generates trajectories much closer to ground truth, especially for walking in the same direction in group and following others. Fig. 6(d)(f) show two different situations (static, non interaction). Both of two models can predict the future trajectories precisely. Fig. 6(e) shows the most difficult collision avoidance situation, in which one pedestrian and another two pedestrians are heading towards the opposite directions. The prediction trajectories of the single pedestrian present the avoidance intention away from another two pedestrians, but the final accuracy is deteriorated. This demonstrates that our model can learn the initial phase in collision avoidance, but fails to emulate the subsequent trajectory. One possible reason is that the samples of collision avoidance are still insufficient in our mcScenes dataset.

Fig. 7 shows a group of complete trajectories across the multi-camera network. From the pedestrians entering the multi-camera space to leaving the scene, our model accurately predicts the future trajectories and the next camera in each scene. At the same time, our model achieves the continuity of motion state between adjacent scenes, demonstrating the advantages and effectiveness of our model for MCTP.

Ablation Experiments

To further evaluate each component of our model systematically, we conduct a series of ablation experiments with the following variants of our model: *MC-STGAT*: modified model without hierarchical interaction or multi-interaction module; *HGAT*: modified model without multi-interaction module, but with hierarchical interaction and latent goal module; *HMGAT*: modified model without latent goal module, but including hierarchical interaction and multi-interaction module.

In Table 2, comparing HM^{3}GAT with HMGAT, the average ADE and FDE reduce from 0.050/0.095 to 0.049/0.095 with LGD module. It proves that the latent goal decoder module of our joint prediction framework provides effectively destination information for trajectory prediction. From the comparison of MC-STGAT, HGAT and HM^{3}GAT, we can see that the average ADE and FDE are reduced from 0.056/0.108 to 0.054/0.100 to 0.049/0.095, which indicates that our HM^{3}GAT can successfully capture the multiple interactions between pedestrians and the hierarchical interaction among different camera scenes.

Conclusion

In this paper, we redefined the multi-camera trajectory prediction problem and introduced the mcScenes dataset for MCTP. Our study is the first to combine the next camera prediction and the future trajectory prediction within a novel joint future destination and trajectory prediction framework. Besides, by introducing existing methods of single camera trajectory prediction into our framework, we provided a quantitative benchmark and evaluation methodology for multi-camera trajectory prediction. It's shown that our method achieves state-of-the-art performance on our proposed mcScenes dataset. We believe our dataset, together with the proposed models, will facilitate future research and uphold applications on multi-camera trajectory prediction.

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 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Awad, G.; Butt, A.; Curtis, K.; Lee, Y.; Fiscus, J.; Godil, A.; Joy, D.; Delgado, A.; Smeaton, A.; Graham, Y.; et al. 2018. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*.

Bae, I.; and Jeon, H.-G. 2021. Disentangled Multi-Relational Graph Convolutional Network for Pedestrian Trajectory Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 911–919.

Bastani, V.; Marcenaro, L.; and Regazzoni, C. S. 2016. Online nonparametric bayesian activity mining and analysis from surveillance video. *IEEE Transactions on Image Processing*, 25(5): 2089–2102.

Bisagno, N.; Zhang, B.; and Conci, N. 2018. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Busbridge, D.; Sherburn, D.; Cavallo, P.; and Hammerla, N. Y. 2019. Relational graph attention networks. *arXiv preprint arXiv:1904.05811*.

Chai, Y.; Sapp, B.; Bansal, M.; and Anguelov, D. 2019. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*.

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 the IEEE Conference on Computer Vision and Pattern Recognition*, 2255–2264.

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 the IEEE/CVF International Conference on Computer Vision*, 6272–6281.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28: 2017–2025.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer graphics forum*, volume 26, 655–664. Wiley Online Library.

Li, B.; Li, X.; Zhang, Z.; and Wu, F. 2019. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8561–8568.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10508–10518.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5725–5734.

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 *European Conference on Computer Vision*, 759–776. Springer.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14424–14432.

Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.-C.; Lee, J. T.; Mukherjee, S.; Aggarwal, J.; Lee, H.; Davis, L.; et al. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, 3153–3160. IEEE.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, 261–268. IEEE.

Rhinehart, N.; Kitani, K. M.; and Vernaza, P. 2018. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 772–788.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, 17–35. Springer.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, 549–565. Springer.

Styles, O.; Guha, T.; Sanchez, V.; and Kot, A. 2020. Multi-camera trajectory forecasting: Pedestrian trajectory prediction in a network of cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1016–1017.

Su, H.; Zhu, J.; Dong, Y.; and Zhang, B. 2017. Forecast the Plausible Paths in Crowd Scenes. In *IJCAI*, volume 1, 2.

Veeramani, B.; Raymond, J. W.; and Chanda, P. 2018. DeepSort: deep convolutional networks for sorting haploid maize seeds. *BMC bioinformatics*, 19(9): 1–9.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vemula, A.; Muelling, K.; and Oh, J. 2018. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, 4601–4607. IEEE.

Vens, C.; Struyf, J.; Schietgat, L.; Džeroski, S.; and Blockeel, H. 2008. Decision trees for hierarchical multi-label classification. *Machine learning*, 73(2): 185.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12085–12094.

Zhu, Y.; Qian, D.; Ren, D.; and Xia, H. 2019. Starnet: Pedestrian trajectory prediction using deep neural network in star topology. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8075–8080. IEEE.