Trajectory Prediction With Heterogeneous Graph Neural Network

Abstract. Trajectory prediction with dense traffic is a challenging task. The heterogeneity caused by multi-type of road agents complicates the mutual and dynamic relationship between agents. Besides, scene context will affect the trajectory of the agent. To address the aforementioned challenges, we present a novel model named HTFNet. Specifically, we use heterogeneous graph network to model multi-type of agents in traffic. In order to handle varying influence between nodes, interactions between nodes are modeled by a heterogeneous transformer neural network, which use node and edge dependent parameters to distinguish heterogeneous attention over each edge. In addition, scene context and dynamic pattern are considered in the process of multi-model destinations prediction. Our proposed model HTFNet is validated on the Stanford Drone Dataset and the Intersection Drone Dataset. The results show that our model achieves a better performance on heterogeneous traffic dataset and produces more reasonable trajectories for different types of road agents.

Keywords: Trajectory forecast · Heterogeneous graph transformer · Multi-type agents.

1 Introduction

Predicting trajectory is an essential component for many applications. For example, autonomous driving needs accurate trajectory prediction to avoid collisions and ensure safety. When robots deliver goods or complete tasks in a complex environment, trajectory forecasting can help them take appropriate strategies to improve efficiency.

In a traffic scenario, there are multi-type road agents, such as cars, buses, pedestrians, bicycles. Most existing trajectory prediction works focus on one type of road agents such as pedestrian [1,2,3] or vehicle [4,5]. These methods ignore difference about social interaction and dynamic pattern between multi-type road agents. For example, people will pay more attention to the motor vehicles which have greater speed and inertia than pedestrians. Recently, some models have noticed the heterogeneous road agents in complex traffic scenario [6,7,8]. However, these methods could not determine what messages to transmit between agents appropriately, making it difficult to learn complicated interactions.

Static scene information needs to be considered in trajectory prediction where road agents have different interactions with static environment. For instance, cars drive along the lane, motorcyclist more likely to ride on the side of the road while pedestrians walk on the sidewalk or zebra crossing. In addition, destinations are also greatly oriented by scene. The trajectory of pedestrians will converge to

the entrance of the building, so the entrance has a high probability of being the destination.

In this paper, we propose a novel model called HTFNet, which model the complex traffic scenario by a dynamic heterogeneous graph. Firstly, road agents are modeled as graph with rich relation information. We use heterogeneous graph transformer (HGT) network to learn distinct interactions between nodes by attention and message passing network. Edge and meta-path based parameters are introduced to get adaptive scaling attention. Stacking HGT layers allow our model have a different size of receptive fields. Secondly, we add scene information and dynamic pattern in the process of destinations and trajectories prediction which increase the accuracy and reality of trajectories. Finally, we empirically validate our model on Stanford Drone Datasets and Intersection Drone Dataset. Experimental results show that our model outperforms the baseline.

To summarize, the contributions of this paper are threefold: (i) We model environment with multi-type of agents as dynamic heterogeneous graph, and propose HFTNet to learn heterogeneous attention between nodes. (ii) In the destination and trajectory prediction process, we consider scene information and the dynamic pattern of the agents. (iii) Our model is evaluate in the short term and long term trajectory prediction tasks. The result show that our model can produce more reasonable and accurate trajectories in complex traffic.

2 Related Work

2.1 Trajectory prediction

There is considerable work on trajectory prediction for moving agents. Most of these algorithms developed with a single type of agents (pedestrians or cars) [1,3,9]. Many approaches rely on neural networks (LSTMs or GRUs) to exploit temporal dependencies of time-series. In order to model influence between trajectories, models were designed to account for interactions between one type of road agents (homogeneous interactions). Alahi et al. [5] introduces social pooling to pool nearby pedestrians' hidden feature. Deo et al. [4] use convolutional social pooling to improve pooling process. Another relevant work is STGAT by Huang et al. [10]. STGAT treats each agent as a node of a graph and exploited graph attention network to share information across different pedestrians. Some algorithms consider heterogeneous data: Trajectron++ by Salzmann et al. [7] accounts multiple interacting agents from heterogeneous input data and produce dynamically-feasible trajectory forecasts. TraPHic by Chandra et al. [8] uses a hybrid network LSTM-CNN to predict trajectory, and take into account heterogeneous interactions. The dynamics of a bus-pedestrian interaction differs significantly from a pedestrian-pedestrian or a car-pedestrian. In order to processing dense, heterogeneous traffic scenarios in depth, we model the interactions among different type of agents by using a HGT network.

2.2 Heterogeneous graph neural network

Graph neural networks (GNN) are aimed to process graph structured data and use message passing between the nodes to capture information from its neighborhood with arbitrary depth. Graph convolutional neural network can divide into two categories, namely spectral domain and non-spectral domain. Spectral approaches represent the graph as spectral embedding based on adjacency matrices, while non-spectral approaches use convolutions directly on the graph based on groups of spatially neighbors. Heterogeneous graphs defined as graphs have several kinds of nodes and edges. Heterogeneous graph embedding mainly focuses on meta-path which utilizes the nodes paths to model the context of a node. Meta-path can group the neighbors according to their node types and distances [11]. HAN by Wang et al. [12] proposed the heterogeneous graph attention network, which utilizes meta-paths to model node level and semantic attentions and learn the weights of different neighbors. HGT by Hu et al. [13] uses transformer-like self-attention architecture for learning node representation.

3 Model Design

In this section, we give an overview of our prediction method that uses HGT model interactions between road agents. Our approach HTFNet is visualized in Fig. 1. There are three key components 1) Feature Encoder, 2) Destination Generator, and 3) Interaction and Prediction Module. First, Feature Encoder Module extracts motion, scene and destinations information. Then, Destination Generator Module predicts multi-model destinations of agents by a generative model CAVE [14]. Finally, Interaction and Prediction Module uses HGT network to exchange information between nodes and uses the extracted feature to predict trajectories.

We begin the section by describing problem definition. Then we present details on how the proposed components are adapted to the task.

3.1 Problem Formulation

This paper is focused on a complex traffic scenario involving multi-type road agents, such as cars, pedestrians and bicycles. The scenario can be modeled by dynamic heterogeneous graph denoted as a series of snapshots $\{G\}_{t=1}^{t=N}$. The heterogeneous graph G at each time t represented as $G_t = (\mathcal{V}_t, \mathcal{E}_t, \mathcal{A}, \mathcal{R})$, where the nodes $\mathcal{V}_t$ and edges $\mathcal{E}_t$ change when the dynamic graph evolving. $\mathcal{A}$ and $\mathcal{R}$ are node and edge type sets respectively. Each node $v \in \mathcal{V}_t$ and each edge $e \in \mathcal{E}_t$ are associated with their type by mapping functions $\gamma(v) : \mathcal{V}_t \to \mathcal{A}$ and $\lambda(e) : \mathcal{E}_t \to \mathcal{R}$. In our model, nodes represent road agents while edges represent their interactions.

Meta-relation Meta relation is denoted as $<\gamma(s), \lambda(e), \gamma(k)>$ which describes a relation link between node $s \in \mathcal{V}_t$ to node $k \in \mathcal{V}_t$.

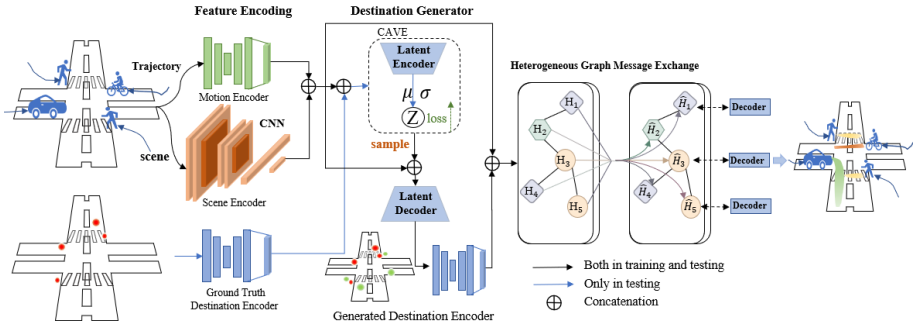

Fig. 1. Overview of model architecture. Our model consists of three components: 1) Feature Encoder, 2) Destination Generator, and 3) Interaction and Prediction Module. The Destination Generator Module combines the dynamic feature of the trajectories and scene information to infer the distribution of destination. The Interaction and Prediction Module use the HGT network to exchange feature between nodes and use the final feature to predict trajectories.

Problem Setup Trajectory forecast is a task of predicting future states of agents. Suppose many types of road agents move in a scene, the model receive as input all the trajectories of agents in a scene for the past t_p time steps. The data represented as a sequence $\{G, \mathcal{S}\}_{t=1}^{t_p}$, where $\mathcal{S}$ is the scene information. In this work, the scene information is obtained by segmenting each scene picture to common categories (lane, sidewalk, lawn, obstacle and entrance). The feature of node $k \in \mathcal{V}_t$ is the location at t time step $\{P_k\}^t = \{(x, y)_k\}^t$. In order to extract dynamic pattern of each agents, the state change of graph is expressed as the displacement of each node $\{\nabla P_k\}_{t=1}^{t_p} = \{(\nabla x, \nabla y)\}_{t=1}^{t_p}$. Our goal is predicting the future state of graph for the next t_f steps $\{G\}_{t=t_p+1}^{t_p+t_f}$.

3.2 Destinations Predict

Our model use CVAE to generate multiple possible destinations $\widehat{D}$ by sampling destinations from the estimated distribution. The input data of this model include the destination $D = (x, y)^{t_p+t_f}$, location information $\{P_k\}_{i=1}^{t_p}$, scene information $\mathcal{S}$ and displacement $\{\nabla P_k\}_{t=1}^{t_p} = \{(\nabla x, \nabla y)\}_{t=1}^{t_p}$. First, we extract temporal information from history state:

$$F_t = MLP(\{P_k\}_{t=1}^{t_p}), \tag{1}$$

Because different types of road agents have different dynamic patterns which can influence the prediction of destinations, we use the dynamic pattern encoder to encode displacement of trajectory which is aimed at extracting information about speed, acceleration and direction:

$$F_p = MLP(\{\nabla P_k\}_{t=1}^{t_p}), \tag{2}$$

F_t and F_p together form the motion feature of agents. Trajectories are significantly oriented by scene context. We use the CNN encoder to extract scene information:

$$F_s = CNN(\mathcal{S}_i). \tag{3}$$

And the destination encoder are used to extract feature of destination D:

$$F_d = MLP(D_k). \tag{4}$$

Then CAVE are used to learn the distribution of destination. We combine temporal feature F_t, dynamic feature F_p and scene feature F_s as observation feature which denoted as X while destination feature F_d denoted as Y. With the conditional distribution $p(Y, z|X) = p(Y|z, X)p(z|X)$, deep neural network (parametrized by θ) are used to approximate $p(z|X)$ and $p(Y|z, X)$. We refer to $p_\theta(z|X)$ as prior network and $p_\theta(Y|z, X)$ as the decoder network. The generative process of Y is:

1. Sample a latent variable z from the prior network $p_\theta(z|X)$.
2. Generate destinations $\widehat{D}_i$ through the response decoder $p_\theta(Y|z, X)$.

The goal of CVAE is maximizing the conditional log likelihood $\log p_\theta(Y|X)$ and can be trained by maximizing the variational lower bound of the conditional log likelihood. Finally the loss function can be represented as:

$$L_{CVAE}(\theta, \phi; Y, X) = -KL(q_\phi(z|Y, X)||p_\theta(z|X)) + E_{q_\phi(z|X,Y)}[\log p_\theta(Y|z, X)]. \tag{5}$$

3.3 Heterogeneous Graph Message Exchange

In a dense and complex traffic scene, different types of road agents have various types of interactions. In order to share information across them, we consider each road agent as a node of heterogeneous graph and use HGT network [13] to exchange messages. HGT network allows for aggregating information from neighbors by assigning different attention to different types of nodes. It also put different attention within same type. For example, vehicles and bicycles extract different information from agents in front of them and on their sides to avoid collisions.

There are two stages in the HGT. Firstly, heterogeneous multi-head attention mechanism are used to calculate attention and messages. Then we use heterogeneous message passing framework to exchange information between nodes. HGT is constructed by stacking HGT layers, which allow HGT have a different size of receptive fields. The structure of a single HGT layer is show in Fig. 2. We concat trajectory feature F_t, dynamic pattern feature F_p, scene feature F_s and generated destination feature F_{gd} as H and input them to the Heterogeneous Graph Message Exchange Module:

$$H = concat(F_t, F_p, F_s, F_{gd}), \tag{6}$$

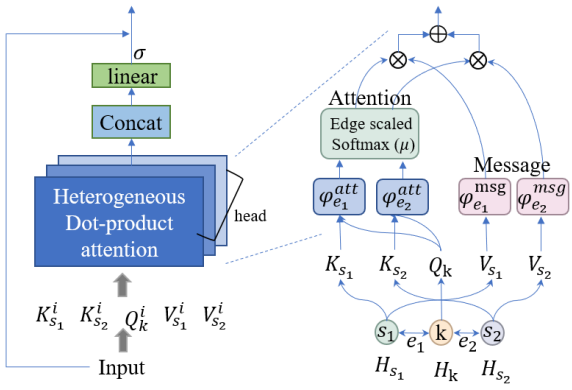

Fig. 2. Overview of architecture of Heterogeneous Graph Transformer

For a meta relation $<\gamma(s), \lambda(e), \gamma(k)>$, the influence from source node s to the target node k are calculated by a meta relation based attention network. We map trajectory hidden feature of target node k and source node s into Query vector and Key vector respectively and then calculate their dot product as attention:

$$K^i(s) = \boldsymbol{K}^i_{\gamma(s)}(H^{l-1}[s]), \tag{7}$$

$$Q^i(k) = \boldsymbol{Q}^i_{\gamma(k)}(H^{l-1}[k]), \tag{8}$$

$$Att^{h_i}(s,e,k) = (K^i(s)\varphi^{att}_{\lambda(e)}Q^i(k)^T).\frac{\mu_{<\gamma(s),\lambda(e),\gamma(k)>}}{\sqrt{d}}, \tag{9}$$

where $Att^{h_i}(s,e,k)$ means one of multi-head attention. H^l means the output of l-th HGT layer which is also the input of the $(l+1)$-th layer. In order to distinguish edge type and meta relation, a matrix $W_{\lambda(e)}$ for each edge type and a parameter $\mu_{<\gamma(s),\lambda(e),\gamma(k)>}$ for meta relation are used to donate their difference. Then we concatenate different representation subspaces of attention and make them through softmax procedure:

$$ATT_{HGT}(s,e,k) = \underset{\forall s \in N(k)}{Softmax}(\underset{i\in[1,h]}{\Vert} Att^{h_i}(s,e,k)). \tag{10}$$

After we got the multi-head attention, the message passing process can be computed in a similar way. We use a matrix $\varphi^{msg}_{\lambda(e)}$ to distinguish different edge type. Then all heads of message are aggregated:

$$Msg^{hi}(s,e,k) = \boldsymbol{M}^i_{\gamma(s)}(H^{l-1}[s])\varphi^{msg}_{\lambda(e)}, \tag{11}$$

$$MSG_{HGT}(s,e,k) = \underset{i\in[1,h]}{\Vert} Msg^{h_i}(s,e,k). \tag{12}$$

After getting heterogeneous multi-head attention and message, we need to aggregate them to get the final target node feature representation:

$$\widetilde{H}^{l}[k] = \sum_{\forall s \in N(k)} (ATT_{HGT}(s, e, k) \cdot MSG_{HGT}(s, e, k)), \tag{13}$$

following the residual connection, the output of the l-th layer is

$$H_{[k]}^{l+1} = \sigma(\boldsymbol{A}_{\gamma(k)} \widetilde{H}_{[k]}^{l}) + H_{[k]}^{l}, \tag{14}$$

$\boldsymbol{A}_{\gamma(k)}$ is a linear function mapping target note's vector back to its node type-specific distribution.

3.4 Trajectory Prediction

The final trajectory feature $\widehat{H}$ of each node are passed through the prediction decoder and get the future trajectory:

$$\{P_k\}_{t=t_p+1}^{t_p+t_f} = MLP(\widehat{H_k}). \tag{15}$$

To train HTFNet, we use the following losses:

$$L = \mathcal{L}_{CVAE}(\theta, \phi; Y, X) + \|\widehat{\{P\}}_{t=t_p+1}^{t_p+t_f} - \{P\}_{t=t_p+1}^{t_p+t_f}\|^2, \tag{16}$$

where $\mathcal{L}_{CVAE}(\theta, \phi; Y, X)$ measures destination error and $\|\widehat{\{P\}}_{t=t_p+1}^{t_p+t_f} - \{P\}_{t=t_p+1}^{t_p+t_f}\|^2$ measures how far the generated trajectories from the ground truth.

4 Experimental and Evaluation

4.1 Dataset

Our model is evaluated on the Stanford Drone Datasets (SDD) [15] and the Intersection Drone Dataset (InD) [16]. We perform long term and short term prediction on them respectively.

Stanford Drone Datasets The SDD dataset is a heterogeneous dataset which consists of following road agents categories: pedestrians, skateboarders, bikers, cars, carts and buses. To make the category more general, we combine car, cart, bus into one type and denote as vehicle. In total, there are eight unique top view scenes recorded by drone, 60 videos and 10300 unique trajectories. For the eight unique scenes, we segment each of them to 5 categories: lane, sidewalk, lawn, obstacle and entrance. For short term prediction, we follow the dataset split defined in TrajNet benchmark [17] which used in prior work [18,19,20]. We sample the data at FPS=2.5 and the length of input sequence is $n_p = 8$ (3.2 seconds) while the length of output sequence is $n_f = 12$ (4.8 seconds).

Intersection Drone Dataset InD dataset is a dataset contains 11500 real trajectories of vehicles and vulnerable road users recorded at German Intersections. The dataset includes vehicles, pedestrians and bicyclists. We use location ID 4 for testing and we segment the scene based on lanes and sidewalks. The raw video is recorded in FPS=25, and we sample them to FPS=1. For long term prediction, the length of observation time and prediction time respectively are $n_p = 5$ seconds and $n_f = 30$ seconds.

4.2 Implementation Details

In order model interactions between road agents and scene, we annotate scene segmentation. The scene images are downsampled and resized to the same size. All types of road agent are considered which will increase the difficulty of prediction. The entire network are trained end to end by ADAM optimizer and we use 2 layers HGT network to handle interaction between road agents.

Evaluation Metrics As in prior works, our model for trajectory forecasting is evaluated with following two error metrics:

Average Displacement Error(ADE) : The root mean square error (RMSE) of all the predicted positions and real positions during the prediction time.

$$ADE = \frac{\sum_{i=1}^{N} \sum_{t=t_p+1}^{t_p+t_f} \|\widehat{P_k^t} - P_k^t\|_2}{N * t_f}. \tag{17}$$

Final Displacement Error(FDE) : The RMSE distance between the final predicted positions at the end of the predicted trajectory and the corresponding true location.

$$FDE = \frac{\sum_{i=1}^{N} \|\widehat{P}_k^{t_p+t_f} - P_k^{t_p+t_f}\|_2}{N}. \tag{18}$$

Both on the SDD and inD, the error are measured in the pixel space.

4.3 Baselines

We compare HTFNet with prior approaches on SDD dataset and consider multimodal trajectory prediction. The following are the baselines we considered:
Social GAN [21]: In this approach SeqtoSeq model are used to encode motion histories and predict future trajectories. The outputs of LSTM are the generator of GAN. The diverse predictions are evaluated against with ground truth by the discriminator.
DESIRE [22]: This model use CAVE to generate diverse set of hypothetical future trajectories. Then use a scoring-regression module rank every prediction. A feedback mechanism further increases the prediction accuracy.

SoPhie [18]: This model predicts future trajectories based on GAN and consider two sources of information, which are history trajectories and scene context information. This model proposes physical attention and social attention to model interaction between agents.
CF-VAE [20]: This model use Conditional Flow Variational Autoencoder(CF-VAE) to learning multi-modal trajectories distributions. And it also proposed posterior regularization and condition regularization to stabilize training.
P2TIRL [23]: This model use MaxEnt IRL to infer goals and paths by learning rewards and it also defined coarse 2-D grid over the scene to predict trajectories.
PECNET[24]: This model first use CVAE to predict multi-modality destination of road agents. Then predict trajectories based on these destinations.

4.4 Quantitative Evaluation

Short Term Forecasting Results Table 1 shows the ADE and FDE values of HTFNet and prior approaches on SDD for short term setting. We predict 20 possible trajectories, and then select minimum value of ADE and FDE for comparison. Our proposed model achieves 11.84 and 20.46 in terms of ADE_{20} and FDE_{20}, which outperforms the baseline. This can be reasonably expected, since we add heterogeneous interaction between different types of road agents and we incorporate static scene context, which help to find the reasonable destination.

Table 1. Short term comparison of recent methods on SDD. We report two metrics: minADE and minFDE of the trajectory with least error among K=20 predicted trajectories. The units of ADE/FDE are pixels.

Methods	minADE	minFDE
Social-GAN(k=20)	27.25	41.44
DESIRE(k=5)	19.25	34.05
SoPhie(k=20)	16.27	29.38
CF-VAE(k=20)	12.60	22.30
P2TIRL(k=20)	12.58	22.07
HTFNet(ours)(k=20)	**11.84**	**20.46**

Long Term Forecasting Results We evaluate our model on InD for longer term trajectory prediction which observe 5 seconds and predict 30 seconds. Table 2 shows the results on the InD. Our approach performs better than S-GAN and PECNET and achieve 12.42 for ADE and 25.53 for FDE. Our model has better performance for pedestrian and bicycle categories. This is because in long term prediction, pedestrian and bicycle categories have more complex interaction with scene context and other road agents. Our model adds scene context and heterogeneous transformer network can predict more reasonable destinations and trajectories. In addition, our model encodes dynamic pattern for different types

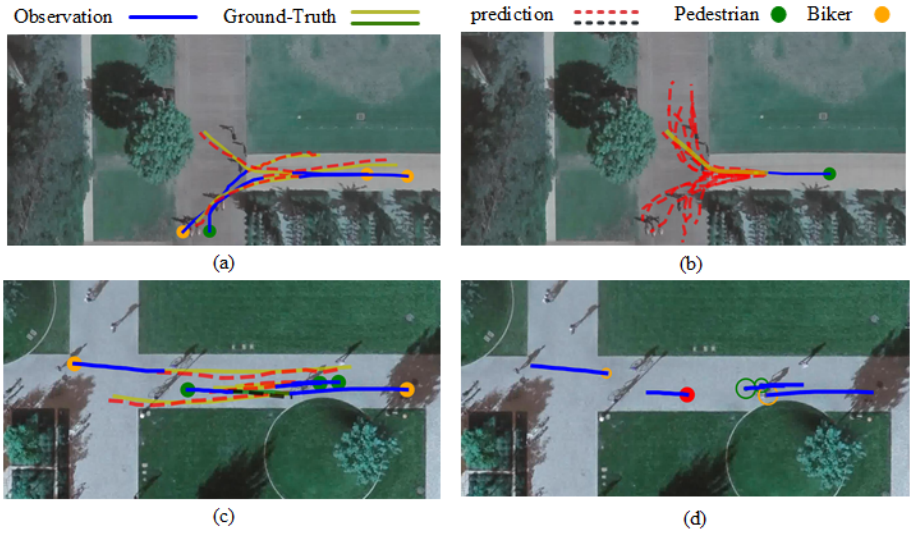

Fig. 3. Trajectory visualization. Fig (a) show that the distribution of trajectories is related to the static scene. Fig (b) show that multi-modal destinations are also greatly oriented by scene semantics. Fig (c) and (b) show that the agent will react differently to different types of road agents. In Fig (b), red dotted agent pay more attention biker with higher speed.

of road agents which can solve the different characteristics of speed, acceleration and inertia.

Table 2. Long term comparison of recent methods on InD. minADE and minFDE are the least error among K=20 predicted trajectories. The units of ADE/FDE are pixels.

Metric	type	S-GAN	PECNet	HTFNet(ours)
minADE	ped	43.59	29.88	**26.47**
	bicycle	112.33	125.95	**96.58**
	car	8.27	**6.65**	7.10
	bus	**42.97**	44.67	43.20
	average	17.33	12.98	**12.42**
minFDE	ped	82.32	55.98	**48.17**
	bicycle	216.26	261.37	**210.83**
	car	17.94	**16.60**	16.67
	bus	92.55	**70.04**	78.78
	average	34.31	27.24	**25.53**

4.5 Ablation Study

In order to develop an understanding of which component influence performance, an ablation study is performed in Table 3.

Dynamic Pattern: To further evaluate the usefulness of dynamic pattern, experiments on removing dynamic pattern feature are conducted. After removing dynamic pattern feature in destination and trajectories prediction, the model achieves 26.84 on FDE metric and 13.04 on ADE metric. This shows that extracting dynamic patterns has a certain effect on the accuracy of the model.

Heterogeneous Interaction: In order to evaluate the usefulness of HGT network, we remove heterogeneous transformer network before prediction final trajectories. The result id shown in Table 3. After remove this module, the performance of the model has dropped a bit. Because the interaction between road agents is not considered, their trajectories have no tendency to avoid each other.

Table 3. Ablation study on effectiveness of dynamic pattern module and heterogeneous interaction. DP represents dynamic pattern while HT represents heterogeneous interaction.

DP	×	√	×	√
HT	×	×	√	√
ADE	16.22	14.72	13.04	12.42
FDE	30.90	25.22	26.84	25.53

4.6 Qualitative Evaluation

Scene context and multi-model trajectories We present out qualitative results in Fig. 3. Our model can extract the contour information of the scene, and then follow the interaction with the scene and other road agents. In Fig. 3 (a), we can see that the distribution of trajectories and destination points will avoid obstacles and lawns. In Fig. 3 (b), we can see that when predicting the possible future destination points of the trajectory, our model considers the scene information.

Road agents interaction When different types of pedestrians meet, there will be different types of interaction. In Fig. 2 (c)(d), we visualize the trajectories and attention values of the heterogeneous transformer network. When oncoming pedestrians and bicycles, pedestrians will have different levels of attention to moving objects at different speeds. We can see that the red dotted pedestrian pay more attention on road agents in front itself. And it also has different attention on different type of agents.

5 Conclusion

In this work we propose a model for motion prediction. This model first generates possible future destinations, then predict trajectories based on multi-model destinations. We add scene information and dynamic pattern in the forecasting process. In order to model different interaction between different types of road agents, we use HGT to model their interaction between each other. We evaluate our model on two datasets and prove that our model can predict a reasonable and accurate future trajectory.

References

1. Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S.: Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: European Conference on Computer Vision. pp. 507–523. Springer (2020)
2. Amirian, J., Hayet, J.B., Pettré, J.: Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
3. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. arXiv preprint arXiv:2104.01528 (2021)
4. Deo, N., Trivedi, M.M.: Convolutional social pooling for vehicle trajectory prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1468–1476 (2018)
5. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961–971 (2016)
6. Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W., Manocha, D.: Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6120–6127 (2019)
7. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. arXiv preprint arXiv:2001.03093 (2020)
8. Chandra, R., Bhattacharya, U., Bera, A., Manocha, D.: Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8483–8492 (2019)
9. Pan, J., Sun, H., Xu, K., Jiang, Y., Xiao, X., Hu, J., Miao, J.: Lane attention: Predicting vehicles' moving trajectories by learning their attention over lanes. arXiv preprint arXiv:1909.13377 (2019)
10. Huang, Y., Bi, H., Li, Z., Mao, T., Wang, Z.: Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6272–6281 (2019)
11. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: Scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 135–144 (2017)
12. Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Yu, P.S.: Heterogeneous graph attention network. In: The World Wide Web Conference. pp. 2022–2032 (2019)

13. Hu, Z., Dong, Y., Wang, K., Sun, Y.: Heterogeneous graph transformer. In: Proceedings of The Web Conference 2020. pp. 2704–2710 (2020)
14. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28**, 3483–3491 (2015)
15. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: European conference on computer vision. pp. 549–565. Springer (2016)
16. Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L., Eckstein, L.: The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections (2019)
17. Becker, S., Hug, R., Hübner, W., Arens, M.: An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. CoRR **abs/1805.07663** (2018), `http://arxiv.org/abs/1805.07663`
18. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1349–1358 (2019)
19. Zhao, T., Xu, Y., Monfort, M., Choi, W., Baker, C., Zhao, Y., Wang, Y., Wu, Y.N.: Multi-agent tensor fusion for contextual trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12126–12134 (2019)
20. Bhattacharyya, A., Hanselmann, M., Fritz, M., Schiele, B., Straehle, C.N.: Conditional flow variational autoencoders for structured sequence prediction. arXiv preprint arXiv:1908.09008 (2019)
21. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264 (2018)
22. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 336–345 (2017)
23. Deo, N., Trivedi, M.M.: Trajectory forecasts in unknown environments conditioned on grid-based plans. arXiv preprint arXiv:2001.00735 (2020)
24. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: European Conference on Computer Vision. pp. 759–776. Springer (2020)