AJG

Article

Visualization of Pedestrian Interaction through Attention-based Pedestrian Trajectory Prediction

Wen-Xin Qiu*[1] and Takashi Fuse[2]

[1] Graduate Student, Department of Civil Engineering, The University of Tokyo

[2] Professor, Department of Civil Engineering, The University of Tokyo, Email: fuse@civil.t.u-tokyo.ac.jp

*Corresponding author: W-X Qiu <qiu@trip.t.u-tokyo.ac.jp>

Received: November 10, 2023; Accepted: April 8, 2024; Published: May 21, 2024

ABSTRACT

Analysis of pedestrian trajectory from observational data is an important approach to understanding microscopic pedestrian behaviors at the operation level. Based on the understanding, pedestrian simulation and trajectory prediction could facilitate pedestrian space development and pedestrian safety study. The studies can be categorized as conventional approaches and deep learning approaches. The conventional approaches model pedestrian behaviors based on known features, such as avoiding collision, and further improve the knowledge of those features. The deep learning-based approaches learn various features from data and model the pedestrian interactions through designed mechanisms rather than treat them as independent time series data. Although deep learning-based approaches achieved higher accuracies in prediction, the lack of interpretability due to its black-box nature is an obstacle to improving generalizable knowledge of pedestrian behaviors. This study aims to improve the deep learning-based pedestrian trajectory prediction method with the consideration of accuracy, computational cost, and interpretability. A spatial-temporal graph is constructed to model the coordinates and interactions of observed pedestrians. The graph attention network (GAT) is introduced into the proposed approach to obtain attention scores. GAT is effective in the number of learnable parameters, a measure of computational costs, and can handle bidirectional edges. The learned attention scores represent the degree how much a pedestrian is aware of one another, so they can be considered as an explicit quantitative representation of the interactions. With the visualization of the scores, the users, such as space planners or traffic engineers, can perceive how the deep learning model learned the interactions. Our proposed approach is validated on a benchmark dataset, ETH/UCY. Compared to the baseline models, the low computational cost is achieved owing to the efficiency of the GAT; the high accuracy is shown by evaluating average displacement error (ADE) and final displacement error (FDE). Finally, the predictions and the attention scores are visualized to provide an interpretation of pedestrian interaction learned by the deep learning model.

Keywords: Pedestrian trajectory, Deep learning, Spatial-temporal graph, Attention mechanism

1. INTRODUCTION

Studying pedestrian behavior is essential for the design of safe and comfortable walking spaces. For example, simulation of pedestrians in public spaces and analysis of pedestrian conflicts are developed based on the fundamental understanding of pedestrian behaviors. Safe and efficient autonomous vehicle and robot navigation also rely on the understanding of pedestrian behaviors. Recently, the advancement of cameras and image-sensing methods made the observation of pedestrian trajectories easier and more accurate. The centimeter-level observations of the coordinates of pedestrians enable more detailed analyses of microscopic pedestrian behaviors, and it is especially important for operation-level behaviors, which consider every instantaneous decision of a pedestrian.

The studies on operation-level pedestrian behaviors could be categorized into conventional rule-based approaches and deep learning-based approaches. In both types of approaches, it is common to consider that a pedestrian is impacted by the past trajectory itself, the environment, and the other pedestrians or other agents in the surroundings. Conventional rule-based approaches are usually designed to recognize several known behaviors, such as avoiding collision and grouping behaviors. Those behaviors are accumulated by physical or statistical principles, so the effect of each behavior to the outcomes are clear to understand. Deep learning-based approaches mainly learn various high-dimensional features from observed data and could achieve higher prediction accuracy. However, the relationship between the input factors and the outcomes are difficult to understand because of the large amount of the non-linear functions.

From the viewpoint of understanding pedestrian behaviors, deep learning-based approaches provide insufficient interpretation to themselves for improving generalizable knowledge. Those models would not be useful for human users, such as space planners and architectures, because the lack of understanding leads to the low reliability. Thus, designing interpretable deep learning-based pedestrian trajectory modeling methods are desired.

This study aims to design a deep leaning model for pedestrian trajectory prediction that the interactions between pedestrians can be quantitatively interpreted and visualized. Given the observations of pedestrians' trajectories for a constant period of time, the model predicts the future trajectories for the following constant period of time. In this study, only the effects between pedestrians are considered. Those of other agents (vehicles) or the environment are omitted, so the surrounding information is not

required. Both the observation and the prediction are in several seconds, focusing on the operation-level pedestrian behaviors. Three requirements – high prediction accuracy, low computational cost, and high interpretability simultaneously – are expected to be achieved. The requirement of high accuracy indicates sufficient expressive power, low computational cost enables implementation in real applications, and high interpretability enhance the understanding on the model.

The proposed deep learning model is based on a spatial-temporal graph structure representing the time series of coordinates of each pedestrian, which provides a clear basis for employing the attention mechanism. The attention mechanism is a type of method that can explicitly and quantitatively model the attention scores, the degree of importance, between elements in a deep learning model. In this study, it is used for representing the attention that each pedestrian pays to the others. The visualization of attention scores is demonstrated for interpreting pedestrian interactions. Despite the visualization of attention mechanisms cannot directly translate to known behaviors such as "avoiding collision with someone" or "following someone," it enables human users to perceive how deep learning models quantitatively model the interactions. It opens the discussions about the characteristics and rationality of the model.

In the following sections, Section 2 briefly introduces pedestrian behavior and trajectory prediction studies, progresses to the interpretation and visualization of those deep learning-based approaches, and focuses on the advantages and difficulties of incorporating attention mechanisms and graph neural networks in prediction models. Section 3 states the framework and elaborates on our proposed method. Section 4 first introduces the benchmark dataset and accuracy metrics used for validation and the quantitative results, shows the visualization of the attention mechanisms, and discusses the interpretation. Section 5 concludes this study.

2. RELATED STUDIES

2.1 Pedestrian Behavior and Trajectory Prediction

Pedestrian behavior and trajectory prediction study on how the trajectories of humans are determined. Based on the observed trajectories and other factors, such as obstacles in the environment or other humans and vehicles, a behavior model is constructed and could predict future trajectories.

2.1.1 Conventional (Rule-Based) Approaches

In conventional pedestrian behavior studies, the objectives are usually to identify various types of behavior and their effect to set some rules to form a behavior model, and the future trajectories can be predicted following the model.

Microscopic and operational-level pedestrian models determine the behaviors of each pedestrian, which generally include avoiding collision, following, and grouping. The two major methods are physical-based methods (Helbing et al., 2002), which consider pedestrians as particles and behaviors as forces, and discrete choice models (Robin et al., 2009), which consider each future step as a choice by the pedestrian itself depending on the factors. Investigating the observed trajectories has improved the knowledge about the behaviors, but the models are limited by the known behaviors.

2.1.2 Deep Learning-Based Approaches

Taking observed trajectories as well as environment images as input and future trajectories as output, deep learning models can generally predict future trajectories with training data by minimizing error functions. Alahi et al. (2016) proposed a pedestrian trajectory prediction problem and its corresponding evaluation data and metrics. Although the problem considering only pedestrian trajectories can be solved as a simple time-series prediction task, they have shown the necessity of modeling the trajectories dependently by introducing the social pooling structure. This study has achieved a higher prediction accuracy than rule-based models or simple time-series prediction models, and it is followed by studies including Social GAN (Gupta et al., 2018) and Reciprocal Net (Sun et al., 2020). The shortage of pooling structure was that the impact from different pedestrians is averaged or only the largest one is considered, which can be conquered by approaches considering pedestrians with separate weights. A major way to learn the weights is to adopt the attention mechanism, which is further introduced in the following. The deep learning-based approaches can learn to produce high-accuracy predictions without prior knowledge of behaviors and are expected to capture high-dimensional features. However, the models do not explain themselves well, thus not producing useful and generalizable knowledge of pedestrian behaviors.

2.2 Interpretation and Visualization of Pedestrian Trajectory Prediction Methods

Deep learning models generally lack interpretability due to the large amount of nonlinear calculation. However, interpretability is quite important for users who access the model in terms of informativeness and transferability (Barredo Arrieta et al., 2020).

To achieve the interpretation, two categories of approaches can be used: (1) making the model itself understandable and (2) applying methods to explain the relationship between the inputs and outputs. The former are also called model-specific methods or interpretable methods, and the latter are also called model-agnostic methods (Molnar, 2023).

2.2.1 Model-Specific Methods

The model-specific methods usually refer

to models such as linear regression, logistics regression, or decision tree, of which the parameters could directly be understood by users. While the methods are too simple so unable to model detailed features, making part of a complex model understandable is also proposed.

Regarding the pedestrian trajectory prediction studies, Vemula et al. (2018) proposed the first pedestrian trajectory prediction model that uses the attention mechanism to show the quantity of each pedestrian focusing on the others and providing the visualization of the quantity in some steps. The attention-based approach concept is adopted and extended in further studies (Huang et al., 2019; Mohamed et al., 2020), but as the amount of attention layer gets larger or the other part of the models gets more complex, the effect of learned attention may become weaker.

In another way, Kothari et al. (2021) proposed to predict the probabilities of different categories of behaviors, which is similar to the conventional discrete choice model proposed by Robin et al. (2009), so users may understand the proportion of each type of behavior contributing to the result. Nonetheless, they do not provide any explanations of the mechanism of the model itself. These studies have shown directions to approaching interpretability, though there is still room for improvement in clarifying every part of deep learning models and relating to physical meanings.

2.2.2 Model-Agnostic Methods

Model-agnostic methods are able to explain models regardless of their model structures or the problems being solved. While implementing those methods, users choose which features and data are going to be interpreted. Still, few studies were found to apply model-agnostic methods in pedestrian trajectory prediction. Makansi et al. (2021) and Kalatian & Farooq (2022) adopted Shapley values to explain different features. Kalatian & Farooq experimented with pedestrians walking on roads and showed some relationships between results and surrounding factors. Makansi et al. (2021) studied several pedestrian benchmark datasets and models. Surprisingly, they claimed that not many interactions are actually learned by the deep learning model except for a sports dataset.

2.3 Attention-Based Pedestrian Trajectory Prediction

As mentioned above, the attention-based approaches have the advantage of considering the interactions of different pedestrians separately, and the parameter of attention score could serve as an explanation of how the model learns the effects. In this section, the related attention-based approaches are explored.

Social Attention (Vemula et al., 2018) is an early work using spatial-temporal graph structure to store the data for a clear representation. The model encoded the node features by recurrent neural networks (RNNs) and inserted attention modules on

the edges to calculate both temporal relations and pedestrian interactions. Considering graph neural networks (GNNs) would be more effective, STGAT (Huang et al., 2019) has adopted a graph attention network (GAT) in spatial relationships but still an RNN for temporal relationships, which could be not fast enough because of its recurrent calculation. On the other hand, Social-STGCNN (Mohamed et al., 2020) has adopted a variant of graph convolutional network (GCN) to weight each interaction separately with hand-craft functions and has used temporal convolutional neural networks (CNNs) to construct a fast and lightweight model. However, GCNs are unable to handle bidirectional interactions and, unlike Social Attention and STGAT, do not have learnable weights of attention scores. Thus, the authors propose a model with fast and lightweight prediction parts similar to Social-STGCNN, but also adopt a GAT to learn essential features of interactions of each pedestrian. This study extends the preceding one and further investigates the GAT part and the learned parameters.

3. METHOD

Section 3.1 first briefly describes the framework adopted from the preceding study. Section 3.2 takes a closer look at the GATs adopted and compared with related GNN methods.

3.1 Framework

The proposed model takes the observed trajectories of pedestrians as input and the trajectories to be predicted as output. Formally, they are written as Eq. 1. No other factors such as obstacles or vehicles are considered. The model consists of the following six steps.

$$P = \{p_t^n \mid n \in \{1, \dots, N\},\ t \in \{1, \dots, T_{obs}, T_{obs+1}, \dots, T_{pred}\}\} \quad (1)$$

where

N is the number of pedestrians, and n is the index of pedestrians.

From 1 to T_{obs} are the time steps of observation and from T_{obs+1} to T_{pred} are the time steps of prediction.

t is the index of time.

p_t^n is the (x, y) coordinates of pedestrian n in time step t.

3.1.1 Pre-processing

Deep learning models are prone to gradient explosion or elimination due to the large variation of the raw input, so the input values are processed. In this study, one's own viewpoint of each pedestrian is considered, and the displacement of each step relative to the pedestrian's direction of the first movement is used. The scale of the values is kept because they indicate the actual length.

3.1.2 Building Spatial-Temporal Graph

The graph is constructed as illustrated in Figure 1. Each node stores the displacement of each pedestrian in each time step. Two types of edges – spatial edges and temporal edges – are used to represent the interactions between

pedestrians and the sequential order of a pedestrian. The spatial edges are used for training GAT. They connect all the pairs of pedestrians in the same time step (time section), forming complete graphs.

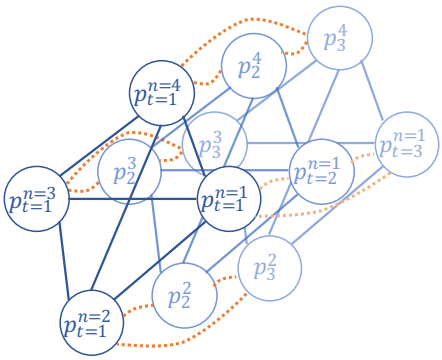

Figure 1. Illustration of the proposed spatial-temporal graph of an example of four pedestrians in three time steps. The spatial edges are shown in blue solid lines and orange temporal edges in orange dotted lines.

3.1.3 Spatial Feature Extraction (Encoding)

The spatial feature is first learned by a GAT, which exchanges the information between pedestrians. This step can be considered as each pedestrian observing the others and being affected by them. The interaction refers to the bidirectional effects that each pedestrian feels from others and gives to them. This extraction is done for each time step separately but sharing the learnable parameters.

3.1.4 Temporal Feature Extraction (Encoding)

The temporal feature extraction is done for each pedestrian separately by a sharing-weight CNN. The CNN has a width of three in the temporal dimension, which physically means to consider the effect of only the time steps right before and right after. After the two extraction steps, a feature matrix of each pedestrian is outputted, meaning a summary of the interaction from the others and the past trajectories of itself. Considering a model predicting trajectories as generating sequences, these two feature extraction steps could also be called encoding steps, and the following sequence generation step could also be called as decoding step.

3.1.5 Prediction (Decoding)

Pointwise CNNs are used for generating predictions, which is also considered as decoding the feature to a sequence. Although using simple CNNs rather than a GNN, the permutation equivariance of graph-structure data is preserved. This step does not deterministically produce the predicted coordinates or displacements. Rather, a Gaussian distribution of the possible displacement is assumed, and the parameters of the distribution are predicted.

3.1.6 Sampling from Distribution

The predicted trajectories are sampled from the distribution in the previous step and accumulated. Considering multiple possibilities with some randomness, K trajectories of each pedestrian are sampled.

3.2 Interaction Feature Extraction through GATs

This subsection focuses on the "spatial feature extraction" step using GAT, which

is mentioned above. Firstly, GAT, a category of GNNs, proposed by related studies is introduced. GNNs are effective machine learning methods applied to graphs. The function of GNN in our model extracts and accumulates features of each node, representing a pedestrian in a specific time step, from all the adjacent nodes. The common form of such GNNs can be written as Eq. 2. Figure 2 illustrates an example of a GNN applied on a node.

$$\boldsymbol{h}_i' = \sigma\left(\Sigma_{j\in\{N(i)\cup i\}}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j\right) \quad (2)$$

where

$\boldsymbol{h}_j$ is the feature vector of node j, and $\boldsymbol{h}_i'$ is the vector of node i after update. $N(i)$ is the set of adjacent nodes of node i. $\boldsymbol{W}$ is a learnable weight for adjusting dimensions of node features. α_{ij} can be a hand-crafted or learnable function. $\sigma(\cdot)$ is a non-linear activation function.

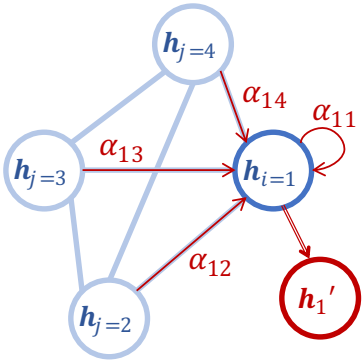

Figure 2. Illustration of applying GNN on a node i.

The GNNs adopted in our model are (a) the original form of GAT (Veličković et al., 2017) and (b) GATv2 improved by Brody et al. (2021). They are designed to learn the attention score α_{ij} shown in Eq. 3 as the weight for accumulation.

$$\alpha_{ij} = softmax(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in N(i)}\exp(e_{ik})} \quad (3)$$

$$\text{(GAT)} \quad e_{ij} = \text{LeakyReLU}(\vec{a}^T\left[\boldsymbol{W}_{GAT}\vec{h_i}||\boldsymbol{W}_{GAT}\vec{h_j}\right]) \quad (4)$$

$$\text{(GATv2)} \quad e_{ij} = \vec{a}^T\text{LeakyReLU}(\mathbf{W}_{GAT}\left[\vec{h_i}||\vec{h_j}\right]) \quad (5)$$

For any node i, the attention score to each adjacent node j is a proportion calculated by the softmax function (the multiple-dimension logistic function) considering all the adjacencies of nodes i to k. Note that every attention score is a scalar. The e_{ij} value is defined as Eq. 4 for GAT and Eq. 5 for GATv2. They both consider the value depends on the features of the adjacent pair, but the original GAT produces identical attention scores regardless of the feature of node i, which is also called query node in attention studies, in some circumstances. GATv2 has proposed to avoid this issue by giving different learnable weights to node i (query node) and node j (key node) and the usage of nonlinear functions. The attention score α_{ij} thus represents each pedestrian of node i paying different degrees of attention to the pedestrian of node j based on their displacements.

Besides, the GNN adopted in Social-STGCNN could be written as Eq. 6 following the definition in Eq. 2. It does not contain learnable parameters in α_{ij}, but uses a hand-craft function and graph Laplacian function, which limits the interaction to distances and is indirectional.

$$\alpha_{ij} = \text{Graph Laplacian}(1/\|p_t^i - p_t^j\|_2 + \epsilon) \quad (6)$$

where ϵ is a small number to avoid division by zero.

4. RESULTS AND DISCUSSION

4.1 Benchmark Datasets

The benchmark datasets used in this study are ETH (Pellegrini et al., 2009) and UCY (Lerner et al., 2007) datasets. Because these datasets mainly consist of pedestrians walking on campus or streets with very few vehicles, they are suitable for studying interactions between pedestrians and ignoring the other agents. The experiments are done through leave-one-out cross-validation by using the pre-processed data provided by Social-STGCNN (Mohamed et al., 2020). There are from 2 to 57 pedestrians in each constructed graph; the scenes of only a single pedestrian are omitted because there is no interaction that can be learned. The number of graphs in each set is shown in Figure 3.

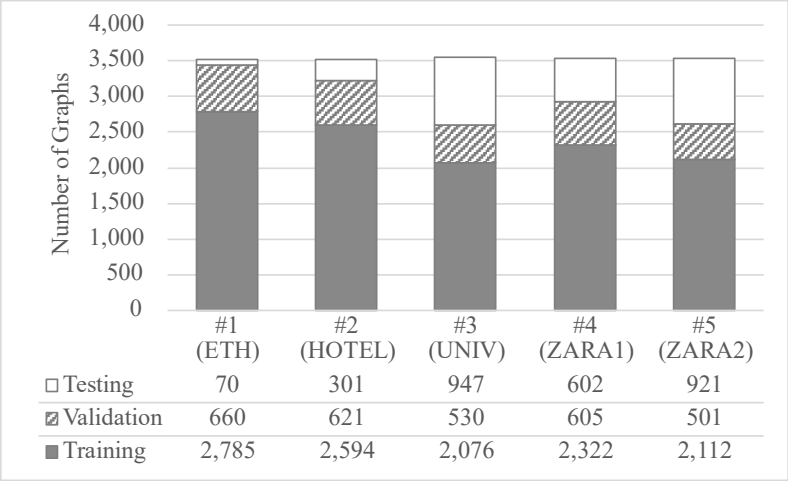

Figure 3. Number of graphs in the pre-processed ETH/UCY dataset. (The names in the parenthesis are the names of testing sets.)

4.2 Accuracy Evaluation

The error metrics of average displacement error (ADE) and final displacement error (FDE) (Alahi et al., 2016), defined in Eq. 7 and 8, are used to evaluate the accuracy of our proposed models. Considering the multiple predicted trajectories, minADE_{20} and minFDE_{20}, which take only the best trajectory among multiple predictions, are used. In this study, we would like to focus on interpreting the mechanism inside the proposed model, so the metrics are only for assessing whether our proposed model is functioning to predict trajectories.

$$ADE = \frac{\sum_{n\in\{1\ldots N\}}\sum_{t\in\{T_{obs+1},\ldots,T_{pred}\}}\left\|\widehat{p_t^n}-p_t^n\right\|_2}{N\times(T_{pred}-T_{obs+1})} \quad \ldots\ldots(7)$$

$$FDE = \frac{\sum_{n\in\{1\ldots N\}}\left\|\widehat{p_t^n}-p_t^n\right\|_2}{N},\ t = T_{pred} \quad \ldots(8)$$

where $\widehat{p_t^n}$ is the prediction value of p_t^n.

The minADE_{20}, minFDE_{20}, and the number of parameters of the models are reported in Figure 4. Figure 4-a lists the minADE_{20} and minFDE_{20} of our

models and the baseline model Social-STGCNN, and also plots those of the related studies which are evaluated on the same dataset. Figure 4-b plots the number of parameters to the minADE_{20} and minFDE_{20}. A logarithm axis is used for the number of parameters (vertical axis) because of the large variance of the numbers. From both figures, our models performed better than the baseline method, which has a similar framework to ours, as well as most of the related studies. Our models also require fewer parameters, indicating the computational efficiency. Compared to the improvement from the related studies, we observed no visually significant differences of our two types of models in minADE_{20} and minFDE_{20} to the number of parameters. We supposed the metrics insensitive to the GAT models because of two reasons. First, the softmax nonlinear function in Eq. 3 make the difference in percentage small, especially when the graphs are complete graphs. Second, the metrics evaluating only the minimum errors are unable to assess all of the multiple predictions, including the variance of the predictions. Nevertheless, their different definition of formulas could reflect different assumptions to the behaviors, and the effect is identified in the following visualizations.

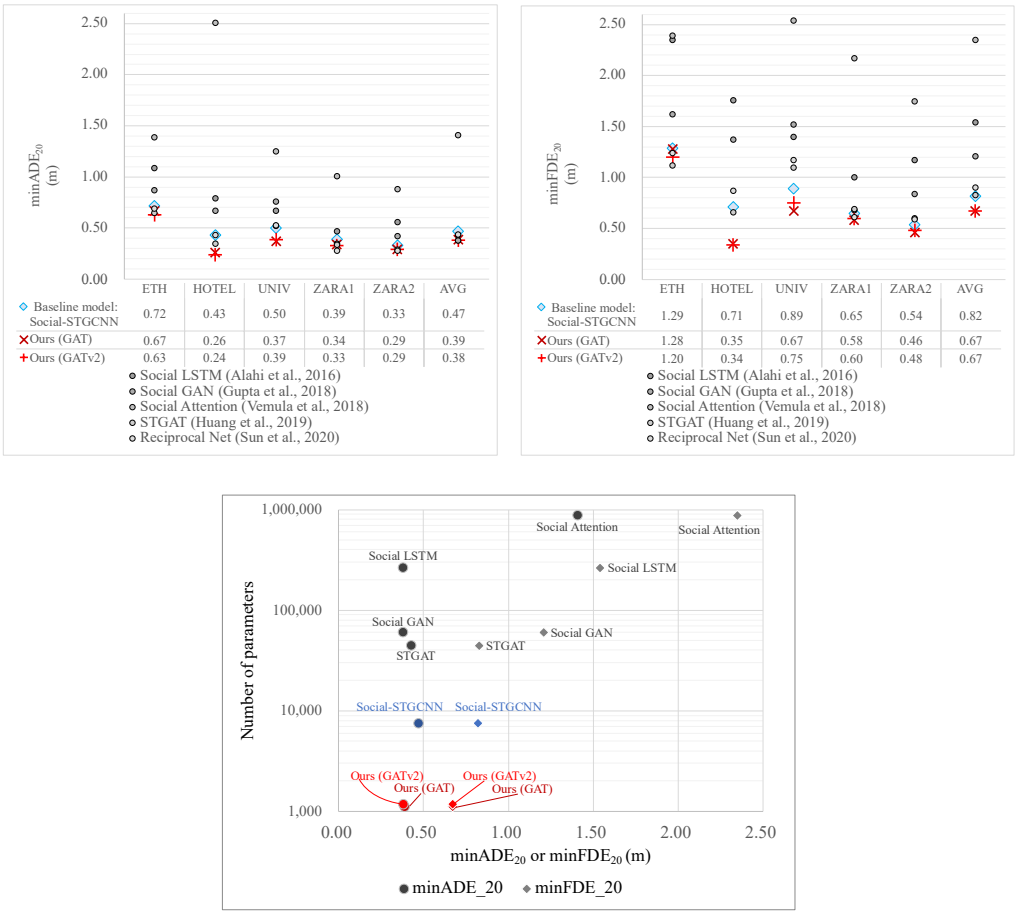

Figure 4. Comparison of (a: top) accuracy and (b: bottom) number of parameters of our models and related studies

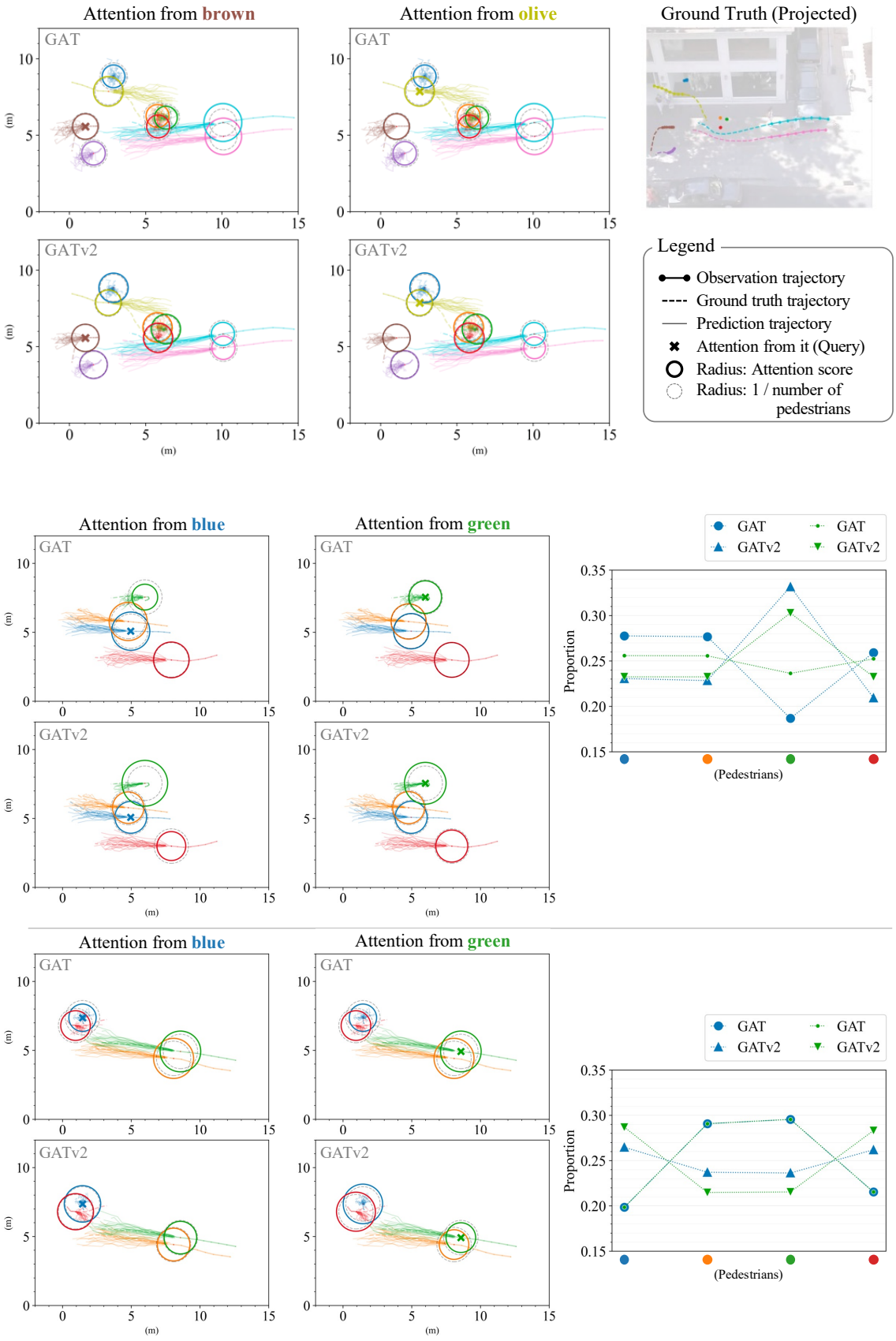

Figure 5. Visualization of prediction results and attention scores

4.3 Visualization of Interactions

Some of the results are visualized in Figure 5. The following introduces the meaning of the figures, taking the first set of results in Figure 5 as an example. The upper right figure shows the observed and ground truth with the background image of the observed scene. Each color represents a pedestrian; the colors have no specific meaning. The four figures on the left shows the prediction results of two models and the attention scores. The colors are corresponded to those in the ground truth figure. The thick solid lines are the observed trajectories, the thin solid lines are the predictions, and the dashed lines are the ground truth of predictions. Each column shows the same prediction result. The upper row is the model adopting the original GAT, and the bottom row is the model adopting GATv2. The circles show the attention scores. The "x" symbol denotes the pedestrian concerned with, which is the query node in GATs. The solid-line circles show the attention scores, which is proportional to the radius of that circle. The colors of solid-line circles are also corresponded to the same color of trajectories. For each color solid-line circle, a gray dashed-line circle represents the radius if the attention scores are equal for all the pedestrians. In this nine-pedestrian scene, the radius represents 11.11%. The first and second column shows the same prediction results but the attention scores focusing on different pedestrians.

Comparing two columns in each set of Figure 5, the GAT model tends to give the very similar set of the attention scores regardless of the query nodes, where the sizes of circles are almost the same pattern on the left and right figure. The values of attention scores may not be exactly the same but have the same order of quantities. To emphasize this effect, the attention scores are plotted on the right in the second and third sets of Figure 5. Brody et al. (2021) figure out this "static attention" problem and proposed the GATv2 method to alleviate it. The static attention problem is due to using the same weight, the $\boldsymbol{W}_{GAT}$ in Eq. 4, for the query and key, $\overrightarrow{h_i}$ and $\overrightarrow{h_j}$. It results in the same order of quantities for attention values. GATv2 uses different weights for the query and key, so the attention scores can be different except the model learns the same value. In this study, the original GAT means each pedestrian perceives the others in the same pattern, regardless of itself. In other words, each pedestrian attracts the same degree of others' attention. While in past studies, interactions are believed to be dependent on different pairs of pedestrians, such as relative directions and distances, the GAT model could not reflect this basic assumption. Thus, from this visualization, the GATv2 model was found more suitable to model pedestrian interactions.

Some characteristics are also noticed from the visualization. For example, the GAT model seemed to produce smaller attention scores for pedestrians of smaller displacements, and larger attention scores

for larger displacements. Conversely, the same trend was not noticeable in the GATv2 model. Besides, the query pedestrians of small displacements (the green one in the second set and the blue one in the third set of Figure 5) usually have more equal attention to themselves and others. It is guessed that none of the pedestrians could create a large impact on pedestrians who are not heading in a specific direction.

5. CONCLUSION

This paper presents a deep learning and attention-based approach for predicting pedestrian trajectories. The proposed models are efficient with lower parameters and more accurate than the baseline model. The GATs employed in this approach enable the quantification and visualization of pedestrian interactions. Different from conventional pedestrian modeling methods, the interactions between all pairs of pedestrians modeled by the GATs consider all the pedestrian pairs in the scene simultaneously. It is expected to improve the understanding of scene and to facilitate the navigation of autonomous vehicles and robots. For instance, a robot can not only consider avoiding collision with pedestrians in current time step, but also keep tracking of any pedestrian that is affected to moving toward it, and even warn someone that was not sufficiently caring about the robot.

By visualizing and investigating the interactions, different trends in attention scores due to different types of GATs were found. The visualized attention scores also clearly illustrated the issue of similar scores of the GAT method. Through the experiments and visualizations, it is discovered that the model using the original GAT could not represent the interactions suitably. It could be seen as an improvement in interpretability that allows users to determine whether this issue is contrary to the desired assumptions.

Nevertheless, although the attention scores have been visualized, the reasons for the trends in attention scores remain unclear. Therefore, interpreting the causal effects of pedestrian interactions is a future task.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S., 2016. Social LSTM: Human trajectory prediction in crowded spaces. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961-971. doi: 10.1109/CVPR.2016.110

Barredo Arrieta, A., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., et al., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp. 82-115. doi: 10.1016/j.inffus.2019.12.012

Brody, S., Alon, U., & Yahav, E., 2021. How attentive are graph attention

networks? Retrieved from https://arxiv.org/abs/2105.14491

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A., 2018. Social GAN: socially acceptable trajectories with generative adversarial networks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2255-2264. doi: 10.1109/CVPR.2018.00240

Helbing, D., Farkas, I. J., Molnar, P., & Vicsek, T., 2002. Simulation of pedestrian crowds in normal and evacuation situations. In: *Pedestrian and Evacuation Dynamics (PED)*.

Huang, Y., Bi, H., Li, Z., Mao, T., & Wang, Z., 2019. STGAT: modeling spatial-temporal interactions for human trajectory prediction. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6271-6280. doi: 10.1109/ICCV.2019.00637

Kalatian, A., & Farooq, B., 2022. A context-aware pedestrian trajectory prediction framework for automated vehicles. *Transportation Research Part C: Emerging Technologies*, 134, 103453. doi: 10.1016/j.trc.2021.103453

Kothari, P., Sifringer, B., & Alahi, A., 2021. Interpretable social anchors for human trajectory forecasting in crowds. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15556-15566. doi: 10.1109/cvpr46437.2021.01530

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

Makansi, O., von Kügelgen, J., Locatello, F., Gehler, P. V., Janzing, D., Brox, T., & Schölkopf, B., 2021. You mostly walk alone: analyzing feature attribution in trajectory prediction. Retrieved from https://arxiv.org/abs/2110.05304

Mohamed, A., Qian, K., Elhoseiny, M., & Claudel, C., 2020. Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14412-14420. doi: 10.1109/CVPR42600.2020.01443

Molnar, C., 2023. Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/

Pellegrini, S., Ess, A., Schindler, K., & 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 (ICCV)*, pp. 261-268.

Robin, Th., Antonini, G., Bierlaire, M., & Cruz, J., 2009. Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1), pp. 36-56. doi: 10.1016/j.trb.2008.06.010

Sun, H., Zhao, Z., & He, Z., 2020. Reciprocal learning networks for human trajectory prediction. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7414-7423. doi: 10.1109/CVPR42600.2020.00744

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y., 2017. Graph attention networks. Retrieved from https://arxiv.org/abs/1710.10903

Vemula, A., Muelling, K., & Oh, J., 2018. Social Attention: modeling attention in human crowds. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4601-4607. doi: 10.1109/ICRA.2018.8460504