Visual Reasoning using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention

Tina Chen[†], Renran Tian[‡], and Zhengming Ding[♯]
[†]Department of Electrical & Computer Engineering, IUPUI, USA
[‡]Department of Computer Information Technology, IUPUI, USA
[♯]Department of Computer Science, Tulane University, USA
chen289@iupui.edu, rtian@iupui.edu, zding1@tulane.edu

Abstract

Autonomous vehicles being able to anticipate rather than just react to pedestrian behavior is vital for the harmonious existence of the two on the road. Previous methods for predicting pedestrian crossing intention from the ego-view relied on bounding box location, and if any, limited visual features for their prediction. However, decisions made on the road by drivers and pedestrians are heavily dependent on context, which should be taken into account when trying to predict what pedestrians on the road intend to do. In this paper, we propose using rich visual features in graph convolutional autoencoders to encode the relationship between the pedestrian and its surrounding objects to reason their crossing intention. To further improve prediction results, we also incorporate pedestrian bounding boxes and human pose estimation in the prediction module. Our model differs in that we consider the effects other road objects/agents have on the pedestrian through visual reasoning of those objects/agents. We evaluate our model's performance using balanced accuracy and F1-score to show that we are able to outperform the state-of-the-art. Our model is able to predict crossing intention with 0.79 balanced accuracy, and is able to predict particularly better for cases where the pedestrian has no crossing intention. The code for our model is released at `https://github.com/chen289/Visual-GCN`.

1. Introduction

Human drivers are able to navigate complicated driving scenarios in urban environments where there are constant streams of vehicles, pedestrians, and other road users. For vehicles and pedestrians to harmoniously co-exist on the road, and share the road efficiently, there are sets of rules that both follow. However, this is not a perfect world, and sometimes the rules are not sufficient or are disregarded.

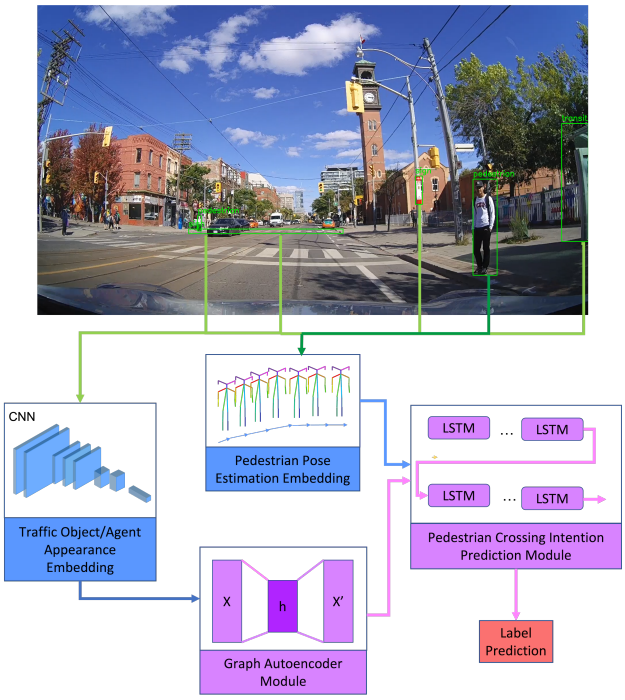

Figure 1. A block diagram of the multiple steps in our pedestrian crossing intention model, which reasons rich visual features from the ego-view to reason the pedestrian's crossing intention. Visual features of relevant traffic objects/agents are embedded using a graph autoencoder to model the relationship the pedestrian has with their surroundings. Human pose estimation of the pedestrian along with the graph autoencoder embedding are fed into an LSTM encoder-decoder to predict the pedestrian's crossing intention.

When this happens and ambiguous situations arise, drivers make decisions based on their intuition to avoid dangerous scenarios.

For autonomous vehicles (AVs) to successfully navigate on the road, they must be able to make critical decisions

based on the continuously changing driving environment. Pedestrians are one of the most vulnerable road users so how AVs interact with pedestrians is important. Most current methods approach this with pedestrian trajectory prediction to determine whether the pedestrian will cross in front of the ego-vehicle. They do so by observing previous trajectory or behavior and predict future trajectory using recurrent neural networks (RNNs). Predicting pedestrian trajectory is not a naturalistic way for AVs to interact with pedestrians. Pedestrians do not operate with specific steps in mind, and human drivers don't interpret pedestrian behavior down to precise movements. Rather, pedestrians have end goals, and perform activities to reach those end goals, such as crossing the street.

When there is potential for the ego-vehicle and pedestrian's path to intersect, instead of predicting trajectory, we can predict on the pedestrian's intention to cross. This is more similar to how human drivers navigate around pedestrian activity. Drivers first determine whether the pedestrian wants to cross the street based on the pedestrian's behavior, and if the answer is yes, then the driver uses a combination of the surrounding environment and the pedestrian's behavior to determine whether the pedestrian will cross the street in front of the ego-vehicle.

To create a system more closely resembling how human drivers interact with pedestrians, we propose an intention prediction model that takes into consideration not only the pedestrian's behavior, but also elements of the ego-scene. Unlike existing works that only rely on bounding boxes or pose estimation, which doesn't take into consideration the reason for the movements, we utilize pedestrian bounding box and pose estimation in conjunction with the pedestrian and the surrounding objects' visual appearance to provide further context for motivation. Furthermore, we model the pedestrian's and surrounding objects' visual appearance using graph convolution to model the pedestrian's relationship with their surroundings. To reduce computation cost and time, we train a graph autoencoder to encode the relationships into a lower dimension. Fig. 1 shows the multiple steps and modules in our proposed prediction model.

We test our model on the Pedestrian Intention Estimation (PIE) [17] dataset, and show that our model outperforms the state-of-the-art in pedestrian crossing intention prediction. In particular, our use of visual reasoning from the pedestrians and objects from the whole ego-view greatly improves prediction on no crossing intention cases. In summary, the contributions in our paper is twofold: 1) We propose a neural network that predicts pedestrian crossing intention through visual reasoning of the ego-scene. 2) We improve upon the state-of-the-art pedestrian crossing intention prediction methods.

2. Related Work

2.1. Pedestrian Trajectory Prediction

Many works have been proposed to predict pedestrian trajectory in dense areas using human-human interaction models. Social-LSTM [1] uses a "Social" pooling layer to share hidden states between nearby pedestrians to learn their interaction behaviors. Social-GAN [5] improves performance through adversarial training against a recurrent discriminator to predict multiple trajectories, and uses a "global" pooling layer to model social interactions for all pedestrians in the scene. These methods are better suited for static, bird's-eye view cameras where the focal point of the camera does not change.

For on-board cameras, not only is the scene constantly changing, but also the relative size and distance of objects to the ego-vehicle as the ego-vehicle is also moving. Bhattacharyya et al. [3] uses a two-stream Bayesian RNN to jointly predict the ego-vehicle's future movement and speed with the pedestrian's future trajectory. Rasouli et al. [17] builds upon Bhattacharyya et al. by using LSTMs, and adding a pedestrian intention prediction branch to improve trajectory prediction. Trajectory prediction can be used to infer intention prediction, but only up to a certain point into the future. Since trajectory prediction is such a precise prediction of the pedestrian's future movements, it is susceptible to change with the passing of every time step. This is not how human driver's interpret other road users' actions. Human drivers do not plan out the potential trajectory pedestrians may take, but rather interpret pedestrian end goals, and the types of actions the pedestrian will take to reach that end goal. For long-term prediction, and mutual understanding between road users, intent should be the preferable prediction.

2.2. Pedestrian Intention Prediction

Pedestrian intention prediction research has gained more attention in recent years with the release of the Joint Attention in Autonomous Driving (JAAD) [18] and Pedestrian Intention Estimation (PIE) [17] datasets. Rasouli et al. proposed JAAD to study pedestrian crossing behavior in traffic scenes. To this end, JAAD is annotated with pedestrian behavior (e.g., crossing/not crossing, walking, standing), pedestrian bounding boxes, pedestrian attributes, pedestrian appearance, environment tags, and ego-vehicle action. Later, Rasouli et al. proposed PIE which not only contains the same types of annotations as JAAD, but additionally annotated with more traffic-related bounding boxes, and conducted a subject research survey to quantitatively measure pedestrian crossing intention. PIE is the first dataset of its kind to define intention as something other than action labels.

When JAAD was the only dataset with naturalistic driv-

ing and pedestrian action behaviors annotated, the intention prediction methods were predicting future action. Gujjar and Vaughan [4] use a three-dimensional convolutional encoder with a convolutional LSTM (convLSTM) decoder to generate future images based on observed images to classify future crossing action. Varytimidis et al. in [22] uses convolutional neural networks to extract features of the pedestrian's head for head orientation estimation, and features of the legs for motion estimation to classify crossing behavior. Fussi-Net [16] uses early, late, and combined fusion of bounding boxes and pose estimation to reduce false positives and predict crossing intention.

2.3. Visual Features for Behavior Reasoning

At the start of using RNNs for pedestrian trajectory or intention prediction, the pedestrian's location via a single point representation or bounding box coordinates were the only inputs into the neural networks. By excluding the context of the rest of the image, a lot of information regarding the motivation behind the pedestrian's behavior is lost, this is especially true for traffic scenes where the environment is constantly changing. Liang et al. [10] improved prediction results on ETH & UCY [15, 9] and ActEV/VIRAT [14, 2] by considering visual features of both the pedestrian and the pedestrian's surrounding to encode both person-scene and person-object relationships. Rasouli et al. in [17] expands the size of the pedestrian bounding box to capture the space around the pedestrian, and uses convLSTMs to extract features from the expanded bounding box to provide context for intention prediction. SF-GRU [19] uses pedestrian visual features and context around the pedestrian to predict crossing action using stacked GRUs.

2.4. Neural Networks on Graphs and Scene Modeling

Graph neural networks are a useful method for modeling irregular data in the non-Euclidean domain. In more recent works, graph convolutional networks (GCNs) have generalized the convolutional operation on unstructured graph data [8]. Rather than shifting a filter across an image, graph convolution aggregates neighboring node information using the adjacency matrix. GCNs have already been applied to many other domains. Joshi et al. in [6] use GCNs to optimize the Travelling Salesman Problem. Works [7, 23] apply GCNs to molecular structures to better understand them and generate novel structures for drug design. In our domain, Social-STGCNN [13] uses GCNs to model the social interactions between pedestrians, and the temporal connections in a spatio-temporal graph to predict future trajectories. In [12], Liu et al. uses segmentation and graph convolution to create scene graphs across frames to model the spatiotemporal relationship the pedestrian has with other object instances to predict crossing intention. Different from these works, we use extracted visual features as features of the graph node, whereas [13, 12] do not consider the visual features of the pedestrian or scene objects. Furthermore, we use a graph autoencoder to learn the graph representation in a lower dimension, which allows us to cut down on computation time and cost.

3. The Proposed Method

The imagery and perspective from the ego-view is constantly changing. Human drivers continuously process these changes, and make decisions based on what they see. Other road users (e.g., pedestrians, cyclists) and objects (e.g., cars, traffic signs) all affect how the ego-vehicle and other road users behave. This motivates us to use relevant visual information from the entire scene to predict pedestrian intention.

Problem Formulation. We define intention and the task at hand similarly to PIE [17]. Intention, $I \in \{0, 1\}$, is measured through aggregated subject responses that are rescaled to be in the range of [0, 1]. Thus, we train intention prediction as a binary classification task. For each pedestrian, p, our system observes bounding boxes, B_p, human pose estimation, E_p, and visual features of other objects, F, from time 1 to T_{obs} to predict the pedestrian's crossing intention.

3.1. Overall Network Architecture

Fig. 2 shows the overall network architecture of our proposed model. Rather than relying only on previous trajectory information to predict future intention, we extract pedestrian and object visual features, and model their relationships through graph convolutional networks. In summary, our model has the following key components:

- **Pedestrian pose module** extracts human pose estimation using the bounding box sequence.
- **Graph autoencoder module** uses convolutional graph autoencoder network and visual feature representations of pedestrians and objects to model the relationships between them.
- **Intention prediction module** leverages the visual information and embeddings from the two previous modules to predict pedestrian crossing intention with an LSTM encoder-decoder.

3.2. Pedestrian Pose Module

In this module, we use an off-the-shelf human pose estimation framework [21] trained on the MS COCO [11] keypoint detection dataset to extract 17 human keypoints of every pedestrian in the PIE dataset. In contrast to other methods that only use the bounding box coordinates, by also using pose, we gain information on the pedestrian's

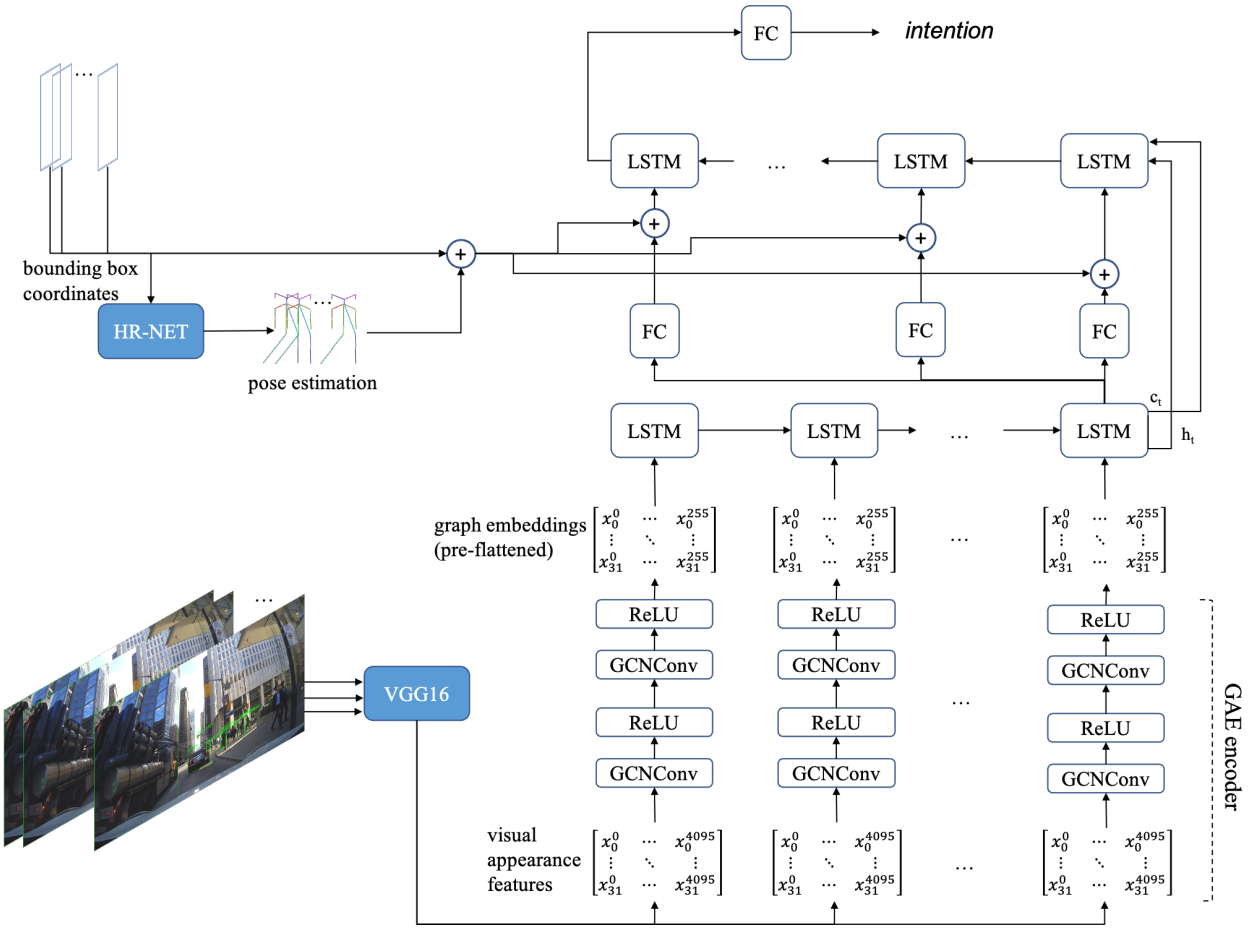

Figure 2. Overall network architecture for our proposed pedestrian crossing intention prediction model. This architecture combines HRNet for human pose estimation, VGG16 for visual appearance feature extraction, graph convolutional autoencoder for encoding graphs, and an LSTM encoder-decoder for predicting intention. This model observes 15 frames to predict the crossing intention for the future 45 frames. $\oplus$ denotes concatenation of features, c_t is the LSTM cell state, and h_t is the LSTM hidden states.

posture and gaze. Pose can tell us the pedestrian's movement status (e.g., walking, standing, crouching), and gaze the pedestrian's focal point. Both of which are indicators of the pedestrian's intent.

For 1,800 pedestrians in PIE, we extracted 740,901 unique poses for the dataset. For the pedestrian our model is predicting intention on, we have the human keypoint representation of shape $T_{obs} \times 32$. There are 17 keypoints in the MS COCO format, and each point has an x and y coordinate, so a complete pose estimation will have 32 values. Pose and bounding box coordinates are concatenated together to form a person pose embedding of shape $T_{obs} \times 36$. Each bounding box coordinate is in the format $(x_{tl}, y_{tl}, x_{br}, y_{br})$. As shown in Fig. 2, the person pose embeddings are the input features for the decoder LSTMs in the intention prediction module.

3.3. Graph Autoencoder Module

In this module, we extract visual features from all objects and pedestrians in an image, and model their relationship to the main pedestrian through graph convolutional networks. We further reduce the dimensions of the graph by training a graph autoencoder to decrease feature size and computation costs.

Object Features. To be able to understand the scene and the pedestrian's reasoning, we use a pre-trained classification algorithm [20] for visual feature extraction. Using the bounding box annotations in PIE, we extract visual features for every object and pedestrian according to the bounding box coordinates. We use the results from the first fully connected layer of the classification algorithm instead of the last layer to get a feature vector of size (1×4096) for each pedestrian and object in the frame. We will use these feature vectors to represent each object in a graph representation of

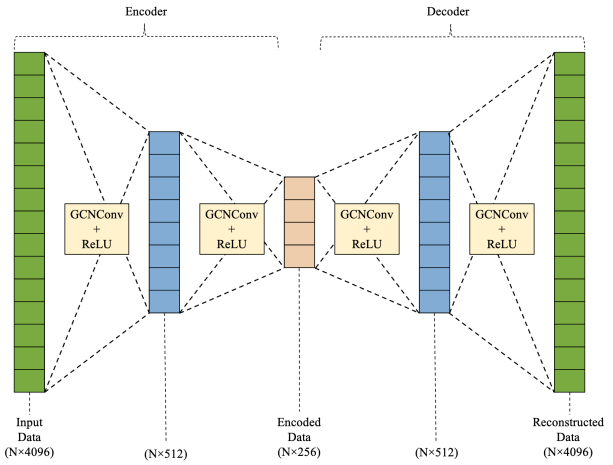

Figure 3. Configuration of the graph autoencoder with convolutional layers. The encoder embeds the input, X, into a lower dimension. The decoder attempts to reconstruct the embedding back into the original input. For our intention prediction module, we use the embedded data as input.

the scene.

Graph convolution. Using the same graph shape as [12], we represent pedestrians and objects in a frame as nodes in the graph. The center node is the pedestrian we are predicting intention on, with edges connecting to all other objects and pedestrians in the scene. The node features are the extracted visual features of the object the node is representing. Following [8], we define layer propagation as:

$$f(X^{(l)}, A) = \sigma(AX^{(l)}W^{(l)}), \tag{1}$$

where the rows of X are the feature vectors of the nodes, A is the adjacency matrix, W is the learnt weight matrix, l is the layer, and σ is the activation function.

Graph Autoencoder. For each frame, the feature vector for each node in the graph is shape (1×4096), and the maximum number of nodes, N, that can be in the graph is 32. To lower computational cost and time, we use the graph convolution layers in an autoencoder configuration. The proposed autoencoder is trained to learn to reconstruct the scene graphs so that we can reduce the size of the scene graphs from $(N \times 4096)$ to $(N \times 256)$. The autoencoder configuration is shown in Fig. 3.

3.4. Intention Prediction Module

As previously discussed, there are two types of inputs for our intention prediction module. The input for the LSTM encoder is the embedding from the graph autoencoder. Each sequence is represented by an embedding of shape $(T_{obs} \times 32 \times 256)$ that is flattened before being fed into the LSTM encoder. The input for the LSTM decoder is the concatenation of the pedestrian bounding box, pose estimation, which has a shape of $(T_{obs} \times 36)$, and the representations from the LSTM encoder.

The LSTM decoder is initialized with the last hidden state of the LSTM encoder. The encoder representations are calculated from the encoder state and a fully connected layer. Pedestrian intention is calculated from the decoder state and a fully connected layer. The predicted intention is a probability of crossing that is rounded to 0 for no crossing intention, and 1 for having crossing intention.

4. Evaluation

4.1. Implementation

Graph Autoencoder. For the feature vectors, x, we use VGG16 to extract visual features from the pedestrians and objects. We use a two-layer GCN that performs two propagations in the forward pass to embed our X from $(N \times 4096) \rightarrow (N \times 512) \rightarrow (N \times 256)$. We use ReLu activations for each convolutional layer, learning rate of 0.001, and train for 50 epochs. To train our graph autoencoder, we use the same training set as designated by PIE's split. We implement our graph convolutional autoencoder using deep learning library PyTorch Geometric (PyG) whose source code is at `https://github.com/rusty1s/pytorch_geometric`.

Intention Prediction. Both the encoder and decoder of our intention prediction module uses LSTMs with 128 hidden units, softsign activation, 0.4 dropout, and 0.2 recurrent dropout. HRNet is used to extract the human pose estimation used for the decoder input.

4.2. Datasets

PIE. Our model is trained and tested on the Pedestrian Intention Estimation (PIE) dataset. It is a public dataset collected by researchers at York University [17] with 6 hours of naturalistic driving data on urban streets recorded at 30 fps. PIE is the first of its kind to measure pedestrian intention quantitatively through aggregating subject responses in their research study. Subjects were asked to watch a pedestrian in a video clip collected from the ego-vehicle, and answer on a five-point scale whether the pedestrian wants to cross the street. 15 subjects viewed the 1,842 video clips, and their answers were aggregated and re-scaled to [0,1] to use as the probability of crossing. Additional annotations in PIE include pedestrian and object bounding boxes, pedestrian behavior, pedestrian demographics, object attributes, and ego-vehicle sensor data.

Following [17], our model observes 15 frames (0.5 seconds), and predicts the pedestrian's crossing intention for the future 45 frames (1.5 seconds), meaning each sample is 60 frames (2 seconds). We also follow the same train, validation, and test set splits as [17] to ensure fair comparison of

Method	BA	F1 ($Y = 1$)	F1 ($Y = 0$)
PIE	0.61	**0.87**	0.36
Ours	**0.79**	0.83	**0.55**

Table 1. Comparison between PIE and our proposed model using balanced accuracy, and F1-score for each class as evaluation metrics.

Method	Avg. Accuracy	Avg. F1
PIE	0.62	0.70
Ours	**0.79**	**0.78**

Table 2. Results from randomly sampling testing set to create a balanced dataset. Comparing the average accuracy and F1-score from randomly sampling 10 times.

our results. PIE has 432 video clips with pedestrians with no crossing intention, and 1,410 video clips with pedestrians with crossing intention. This is converted to 12,274 samples with no crossing intention, and 49,341 samples with crossing intention. Due to the dataset being imbalanced $(0 \ll 1)$, we use evaluation metrics and sampling methods that will correct for the imbalance.

Evaluation Metrics. As previously mentioned, PIE is heavily imbalanced with 4 times more crossing than not crossing samples. To correct the imbalance, we use three error metrics for pedestrian crossing intention prediction:

1. *Balanced Accuracy (BA):* Accuracy is used to measure binary classifiers, and balanced accuracy is used when the dataset is imbalanced as it accounts for both positive and negative classes. Balanced accuracy is the average of sensitivity and specificity.

$$BA = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{2}$$

2. *F1-Score:* Measures the balance between precision and recall. We will report the F1 score for each class to evaluate the performance for both crossing and not crossing cases.

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

3. *Accuracy with Random Sampling:* To create a balanced dataset, we randomly sample the test set to achieve an even split between the two classes. This equates to 8,904 samples for each class to create a balanced 17,808 sample test set. We evaluate the balanced test set with accuracy. We randomly sample 10 times and take the average accuracy.

4.3. Results and Analysis

Baseline Model. We compare our model with the most recent method proposed in PIE. We train PIE's intention prediction module, which uses context visual features around the pedestrian and the pedestrian bounding boxes in a convLSTM encoder-LSTM decoder model.

Main Results. Table 1 compares the crossing intention prediction results between PIE and our model. The "BA" column computes the balanced accuracy of the predictions, indicating how well each model performed. The "F1" columns calculate the balance between recall and precision for each class.

Comparing BA, our model outperforms PIE by 18 points. The F1-scores give insight into how our model outperforms PIE by such a large margin. PIE is able to predict slightly better on crossing cases, but our model performs far better on not crossing cases compared to PIE. However, for both models, it is still more difficult to predict not crossing cases correctly. This is most likely due to there not being enough no crossing cases in the training set.

Additionally, our random sampling results in Table 2 show our model consistently outperforms PIE when we have a balanced dataset. Our model is able to predict crossing intention with 17 more points in accuracy, and also have an 8 point higher F1-score. Our model is able to use contextual visual features to better predict when pedestrians have no intention to cross the street.

5. Conclusion

In this paper we present a pedestrian crossing intention prediction model that utilizes pedestrian and object visual appearances to reason the pedestrian's motivation. We first preprocess the data from PIE to extract human pose estimation and visual feature representations. Then we encode the pedestrian's relationship with their surrounding pedestrians and objects using a graph convolutional autoencoder network. To predict crossing intention, we utilize an LSTM encoder-decoder on the pedestrian bounding boxes, human pose estimation, and visual features graph embedding. We demonstrate that our model is able to outperform the state-of-the-art through visual reasoning of the surrounding environment.

In future work, datasets can benefit from collecting more no crossing intention cases. Both crossing and no crossing intention cases are critical for pedestrian safety, but it is far harder to predict no crossing intention cases, and that may be due to a shortage in samples. To further expand visual feature extraction of the scene, we can utilize scene segmentation to gather a holistic view from the ego-vehicle.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzad Godil, David Joy, Andrew Delgado, Alan Smeaton, Yvette Graham, et al. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*, 2018.

[3] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4194–4202, 2018.

[4] Pratik Gujjar and Richard Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2097–2103. IEEE, 2019.

[5] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[6] Chaitanya K Joshi, Thomas Laurent, and Xavier Bresson. An efficient graph convolutional network technique for the travelling salesman problem. *arXiv preprint arXiv:1906.01227*, 2019.

[7] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[9] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[10] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. 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*, pages 5725–5734, 2019.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[12] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[13] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. 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*, pages 14424–14432, 2020.

[14] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[15] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, pages 452–465. Springer, 2010.

[16] Francesco Piccoli, Rajarathnam Balakrishnan, Maria Jesus Perez, Moraldeepsingh Sachdeo, Carlos Nunez, Matthew Tang, Kajsa Andreasson, Kalle Bjurek, Ria Dass Raj, Ebba Davidsson, et al. Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network. *arXiv preprint arXiv:2005.07796*, 2020.

[17] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[18] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–213, 2017.

[19] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582*, 2020.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[22] Dimitrios Varytimidis, Fernando Alonso-Fernandez, Boris Duran, and Cristofer Englund. Action and intention recognition of pedestrians in urban traffic. In *International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 676–682. IEEE, 2018.

[23] Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *arXiv preprint arXiv:1806.02473*, 2018.