PePScenes: A Novel Dataset and Baseline for Pedestrian Action Prediction in 3D

Amir Rasouli **Tiffany Yau** **Peter Lakner** **Saber Malekmohammadi**

Mohsen Rohani **Jun Luo**

Noah's Ark Laboratory
Huawei, Canada
{amir.rasouli,tiffany.yau1, peter.lakner,saber.malekmohammadi,
mohsen.rohani,jun.luo1}@huawei.com

Abstract

Predicting the behavior of road users, particularly pedestrians, is vital for safe motion planning in the context of autonomous driving systems. Traditionally, pedestrian behavior prediction has been realized in terms of forecasting future trajectories. However, recent evidence suggests that predicting higher-level actions, such as crossing the road, can help improve trajectory forecasting and planning tasks accordingly. There are a number of existing datasets that cater to the development of pedestrian action prediction algorithms, however, they lack certain characteristics, such as bird's eye view semantic map information, 3D locations of objects in the scene, etc., which are crucial in the autonomous driving context. To this end, we propose a new pedestrian action prediction dataset created by adding per-frame 2D/3D bounding box and behavioral annotations to the popular autonomous driving dataset, nuScenes. In addition, we propose a hybrid neural network architecture that incorporates various data modalities for predicting pedestrian crossing action. By evaluating our model on the newly proposed dataset, the contribution of different data modalities to the prediction task is revealed. The dataset is available at https://github.com/huawei-noah/PePScenes.

1 Introduction

One of the major challenges faced by autonomous driving systems is predicting road users' behavior, in particular, pedestrians as they exhibit a diverse set of actions [1] influenced by various environmental and social factors [2]. In the context of driving, behavior prediction is commonly actualized in terms of forecasting the future trajectories of road users. However, as the recent developments in this field suggest, prediction of higher-level actions of road users, e.g. pedestrian crossing actions, can be beneficial for trajectory forecasting and motion planning [3, 4, 5, 6, 7].

In recent years, a number of pedestrian action prediction algorithms have been introduced [8] many of which were trained and evaluated on existing pedestrian behavior datasets [9, 10, 5, 11]. These datasets, however, are limited since they do not contain information such as 3D maps of environments, 3D locations of objects, etc. necessary for prediction in the context of autonomous driving systems.

In this paper, we introduce a novel dataset for pedestrian crossing action and dense trajectory prediction for autonomous driving applications. Our dataset contains new per-frame bounding box

Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

and behavioral annotations for the nuScenes dataset [12]. The annotations are added to 3D as well as 2D data making it suitable for various applications in the autonomous driving domain.

Furthermore, we propose a hybrid baseline model that uses multi-modal data inputs to predict pedestrian crossing action. We train and evaluate the proposed model on our new dataset and show how different modalities of data contribute to prediction accuracy.

2 Related Works

2.1 Datasets

Pedestrian behavior prediction can take two forms: implicit where pedestrian trajectories are forecasted and explicit where pedestrian actions are predicted. There are many existing datasets that cater to trajectory prediction for different domains such as surveillance [13, 14, 15], anomaly detection [16, 17, 18], and intelligent driving [19, 20, 21]. However, the choices for pedestrian action prediction are more limited. There are a few datasets that provide rich behavioral tags along with temporally coherent spatial annotations that can be used for pedestrian action prediction in the driving context. One of the early datasets is Joint Attention in Autonomous Driving (JAAD) [11] which consists of 346 video clips annotated with 2D bounding boxes for pedestrians and behavioral tags for a subset of them along with the ego-vehicle driver's actions. A major drawback of this dataset is the lack of ego-motion information which is vital for prediction from a moving camera perspective. A more recent dataset, Pedestrian Intention Estimation (PIE) [5], rectifies this issue by providing the ego-vehicle motion parameters in addition to more samples, annotations for all relevant objects (besides pedestrians), and pedestrian intention information obtained by conducting a human experiment. There are two other datasets similar to PIE, namely Trajectory Inference using Targeted Action priors Network (TITAN) [9] and Stanford-TRI Intent Prediction (STIP) [10] both of which provide 2D bounding box and pedestrian behavior annotations. These datasets, however, are available under very restrictive terms of use. VIENA^2 [22] is another action anticipation dataset which contains only simulated video sequences collected from a computer game.

The major drawback of the existing pedestrian action prediction datasets is the lack of information, such as the semantic map of the environment, 3D coordinates, etc. all of which are necessary for developing algorithms for autonomous driving systems. Given that our proposed dataset is built on an existing autonomous driving dataset, all required types of data are available.

2.2 Behavior Prediction in Driving

The dominant approach to predicting road users' behaviors is to forecast their future trajectories [23, 24, 25, 26, 27]. However, recent evidence suggests that predicting high-level actions of road users can benefit various planning tasks both directly [3, 4, 28] and indirectly, e.g. via improving trajectory forecasting [9, 5, 6, 7].

Many algorithms have been proposed for pedestrian action prediction. A subset of these algorithms relies on feedforward architectures [3, 29, 28, 11]. Some of these models predict actions directly by classifying various components in the scene [29], some predict from intermediate features such as pedestrian head orientation [11], and others generate future scene representations which are used to classify future actions [28]. Opposed to such unimodal approaches are recurrent architectures that benefit from a combination of different data modalities, such as images, ego-motion information, poses, trajectories, etc. to make predictions [30, 10, 4, 22].

While feedforward networks are very powerful for capturing the spatiotemporal representations of the scenes, recurrent networks provide flexibility for combining multi-modal data with different dimensionalities. In our proposed approach, we take advantage of both of these architectures in a hybrid framework that uses both convolutional layers for processing image data and recurrent networks for encoding trajectories and ego-motion information.

3 PePScenes Dataset

The proposed dataset is a set of additional 2D/3D bounding box and behavioral annotations to the existing nuScenes dataset [12]. Although the main goal of creating this dataset was for pedestrian action prediction, the newly added annotations can be used in various tasks such as tracking, trajectory prediction, object detection, etc. We refer to the new data as **Pe**destrian **P**rediction on nu**Scenes** (PePScenes).

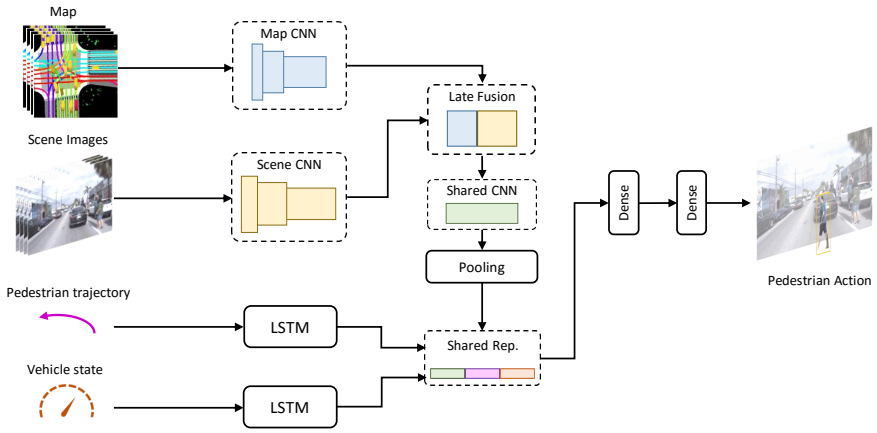

Figure 1: An overview of the proposed architecture. The model relies on four different input modalities: semantic maps, scene images, trajectories and ego-vehicle states. Visual features are processed with two sets of Conv2D layers followed by a late-fusion Conv2D layer for joint processing. Trajectories and ego-vehicle states are processed using two LSTMs the output of which are concatenated to visual features to form a shared representation which is fed into consecutive dense layers to make predictions.

Annotations. nuScenes has 1000 segments (i.e. data sequences) out of which annotations for 850 are available online. We added bounding box annotations for all the existing objects in the annotated portion of the dataset. However, for behavioral annotations, we only chose a subset of samples that, 1) appear in front of the ego-vehicle, 2) have or appear to have an intention of crossing (e.g. they are not far away on a sidewalk), and 3) are observable for at least a few frames prior to making crossing decision. Given these criteria, we added behavioral labels to 719 unique pedestrian tracks. The overall statistics of the proposed dataset can be found in Table 1.

Bounding boxes. nuScenes contains LIDAR scans and camera images recorded at 20 and 12Hz respectively. The existing bounding box annotations of nuScenes, however, are at 2Hz which is fairly sparse, especially for the task of pedestrian behavior prediction. As a result, we augmented spatial annotations of pedestrians and all objects at 10Hz. We interpolated the bounding boxes between two consecutive original annotations using the global coordinates of pedestrians in the environment. To better align the new bounding boxes with the actual samples, we used a 2D detection algorithm, RetinaNet [31] pre-trained on COCO [32], to first localize pedestrians in the images and then use the detected boxes to adjust the locations of added bounding boxes according to the projection of 3D bounding boxes on the image plane. In the end, we randomly sub-sampled a portion of the data and manually evaluated them to assure the quality of newly added boxes.

Behavioral labels. Behavioral labels for crossing actions were added to a subset of pedestrians. Each of the unique pedestrian samples has an object-level annotation indicating whether at a given point in the sequence they will cross the road in front of the ego-vehicle. In addition, for each frame, we also include the current crossing state of the pedestrian, i.e. whether they are currently crossing or not by specifying the start and end time of crossing events. For samples that eventually cross the road, a label is added to specify the critical point in time when crossing starts.

4 Proposed Model

Problem statement. We formulate pedestrian action prediction as an optimization process in which the goal is to learn distribution $p(A_i^{t+m}|SC_o, M_o, L_o, V_o)$ for some pedestrian $1 < i < n$ where $A_i^{t+m} \in \{0, 1\}$ is pedestrian crossing action at some time $t + m$ in the future. Predictions are based on observed scenes $SC_o = \{sc^1, sc^2, ..., sc^t\}$, changes in the semantic map of the environment $M_o = \{m^1, m^2, ..., m^t\}$, the pedestrian's observed trajectory $L_o = \{l^1, l^2, ..., l^t\}$ and the ego-vehicle states $V = \{v^1, v^2, ..., v^t\}$.

Architecture. As mentioned earlier, we employ a hybrid approach to encode different input modalities (see Figure 1). We use rasterized maps encoded as 3-channel images similar to [33]. The map is of size 30×30 meters centered around the ego-vehicle. As for scene images, we use the entire scene images from the forward center camera resized to 300×300 pixels. Both map and scene

Table 1: The overall statistics of the annotations. The numbers under *New* column refer to the newly added annotations and under *Original*, the existing nuScenes annotations.

Annt.	*New*	*Original*	Total
# Ped. with beh.	719	-	719
# Cross. peds	149	-	149
# Non-cross peds.	570	-	570
# Per-frame beh. annt.	63.4K	-	63.4K
# Ped. box annt.	845K	222K	1.06M
# Other box annt	3.58M	944K	4.52M
Annt. frame rate	10Hz	2Hz	10Hz

Table 2: The performance of the proposed model trained and tested on the new PePScenes dataset. Our model is evaluated with different input modalities.

Method		Acc	AUC	F1	Prec
LSTM		0.78	0.54	0.20	0.39
SF-GRU [4]		0.86	0.60	0.31	0.39
Ours	Scene	0.80	0.58	0.26	0.24
	Map	0.82	0.55	0.26	0.23
	Map+Scene	0.85	0.62	0.35	0.38
	Map+Scene+Traj	0.86	0.62	0.34	0.43
	All	**0.87**	**0.71**	**0.48**	**0.47**

image sequences are stacked channel-wise and fed into two separate sets of Conv2D layers with sizes $\{[32, 3, 3], [64, 3, 2], [128, 3, 2]\}, \{[64, 3, 3], [128, 3, 2], [256, 3, 2]\}$ respectively where values in order stand for *[number of filters, kernel size, stride]*. The final outputs of map and scene conv layers are concatenated and fed into a single Conv2D layer, $[512, 3, 1]$, followed by a global average pooling to generate visual representations.

For trajectories, we use $[x, z]$ coordinates of pedestrians in the environment and ego-vehicle state represented by velocity $[v_x, v_y, v_z]$. Both trajectories and ego-vehicle states are processed using two LSTMs with 128 cells. The final shared representation is formed by concatenating the output of the LSTMs and visual representations. The shared representation is then fed into two dense layers, with dropout of 0.5 in between, to predict actions. For learning, we use binary cross-entropy loss function.

5 Evaluation

Data. We split the data into train/test sets with a ratio of $70/30$ while maintaining the ratios of positive and negative samples consistent. Following [4], we clip sequences up to the first frame of crossing events. In cases where no crossing occurs, we select the last frame in the center-view camera where the pedestrian is visible. We choose an observation length of 0.5 seconds (or 5 frames at 10Hz) and sample sequences from each pedestrian track between 1 to $2s$ to the event of crossing with an overlap of 50% between each sample.

Training. We trained the model end-to-end using RMSProp [34] optimizer with batch size of 8 and learning rate of 5×10^{-5} for 50 epochs. To compensate for data imbalance, we used class weights based on the ratio of positive and negative samples.

Metrics. For evaluation purposes common binary classification metrics as in [4] are used including $accuracy$, Area Under the Curve (AUC), $F1$, and $precision$.

5.1 Crossing Prediction

We compare the performance of the proposed model to a baseline LSTM model trained only on trajectories and state-of-the-art crossing prediction algorithm, SF-GRU [4]. For a fair comparison, we use global coordinates and velocity instead of 2D bounding box coordinates and the ego-vehicle speed originally used in SF-GRU. In addition, to highlight the contributions of different data modalities to the prediction task, different subsets of the proposed model are evaluated on PePScenes. We refer to these subsets based on the types of input modalities that are used.

As shown in Table 2, when relying merely on dynamics features such as trajectory the model performs poorly. By combining visual features with dynamics information, the results on all metrics show improvements. The best results are achieved on all metrics when all sources of information are included as shown by the performance of the proposed model.

6 Conclusion

We proposed a novel dataset for pedestrian behavior prediction by augmenting the nuScenes dataset with more than 60K behavioral and 4 million bounding box annotations. This is the first dataset that provides high-level action annotations on 3D data for research in pedestrian behavior prediction. In addition, new dense annotations in the dataset are suitable for tasks such as tracking, detection, trajectory prediction, etc. We also proposed a hybrid model for pedestrian crossing prediction and showed how a combination of different data modalities can improve the accuracy of prediction.

References

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *Intelligent Vehicles Symposium (IV)*, 2017.

[2] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[3] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *WACV*, 2020.

[4] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked RNNs," in *BMVC*, 2019.

[5] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019.

[6] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *CVPR*, 2019.

[7] S. Casas, W. Luo, and R. Urtasun, "IntentNet: Learning to predict intention from raw sensor data," in *CORL*, 2018.

[8] A. Rasouli, "Deep learning for vision-based prediction: A survey," *arXiv:2007.00095*, 2020.

[9] S. Malla, B. Dariush, and C. Choi, "TITAN: Future forecast using action priors," in *CVPR*, 2020.

[10] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[11] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCVW*, 2017.

[12] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[13] J. Liang, L. Jiang, K. Murphy, T. Yu, and A. Hauptmann, "The garden of forking paths: Towards multi-future trajectory prediction," in *CVPR*, 2020.

[14] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *ECCV*, 2016.

[15] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer graphics forum*, vol. 26, no. 3, pp. 655–664, 2007.

[16] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *ICCV*, 2013.

[17] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *CVPR*, 2018.

[18] C. C. Loy, T. Xiang, and S. Gong, "Modelling multi-object activity by gaussian processes," in *BMVC*, 2009.

[19] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3D tracking and forecasting with rich maps," in *CVPR*, 2019.

[20] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.

[21] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012.

[22] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "VIENA: A driving anticipation dataset," in *ACCV*, 2019.

[23] L. Fang, Q. Jiang, J. Shi, and B. Zhou, "TPNet: Trajectory proposal network for motion prediction," in *CVPR*, 2020.

[24] M. Liang, B. Yang, W. Zeng, Y. Chen, R. Hu, S. Casas, and R. Urtasun, "PnPNet: End-to-end perception and prediction with tracking in the loop," in *CVPR*, 2020.

[25] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "CoverNet: Multimodal behavior prediction using trajectory sets," in *CVPR*, 2020.

[26] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *CVPR*, 2019.

[27] N. Rhinehart, K. M. Kitani, and P. Vernaza, "R2P2: A reparameterized pushforward policy for diverse, precise generative path forecasting," in *ECCV*, 2018.

[28] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *ICRA*, 2019.

[29] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet," in *ICRA*, 2019.

[30] K. Mangalam, E. Adeli, K.-H. Lee, A. Gaidon, and J. C. Niebles, "Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision," in *WACV*, 2020.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *CVPR*, 2017.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[33] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *ICRA*, 2019.

[34] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp, coursera: Neural networks for machine learning," 2012.