Pedestrian Stop and Go Forecasting with Hybrid Feature Fusion

Sathwik Krishtipati
Binghamton University
Skrisht1@binghamton.edu

Anirudh Nori
Binghamton University
anori1@binghamton.edu

***Abstract*— To enable autonomous driving systems to navigate safely in urban areas, predicting the future movements of pedestrians is crucial. However, current prediction algorithms rely too heavily on past observed trajectories and struggle to account for sudden dynamic changes, such as when pedestrians abruptly start or stop walking. The authors propose that accurately predicting these non-linear transitions should be a key focus to improve the robustness of motion prediction algorithms. They introduce a new task, called pedestrian stop and go forecasting, and release a benchmark dataset called TRANS for studying these behaviors. The TRANS dataset is built from multiple existing datasets with annotations of pedestrian walking motions to capture a variety of scenarios and behaviors. The authors also present a novel hybrid model that combines pedestrian-specific and scene features from multiple modalities to integrate various levels of context. They evaluate their model and several baselines on the TRANS dataset and establish a new benchmark for the research community to strive towards pedestrian stop and go forecasting.**

I. INTRODUCTION

Autonomous vehicles operating in populated cities need to be able to predict the movements of pedestrians, who are one of the most vulnerable groups of road users, and respond appropriately to avoid accidents. Current methods for forecasting pedestrian movements rely on analyzing their previous movements to predict their future location. However, these methods may not be reliable when there are sudden changes in pedestrian dynamics, as past actions do not always accurately reflect future movements. To address this issue, we propose that predicting the transitions between walking and standing still can be an important component of pedestrian movement forecasting. These transitions are a common feature of human movement patterns but are difficult to predict due to their non-linear nature. Furthermore, these transitions are often associated with critical traffic scenarios where the failure to anticipate them can result in severe consequences.

The purpose of this paper is to introduce a new task of predicting when pedestrians will stop and start walking from the perspective of a moving vehicle. This is depicted in Figure 1. To investigate this task, we created a benchmark with a novel dataset and various approaches.

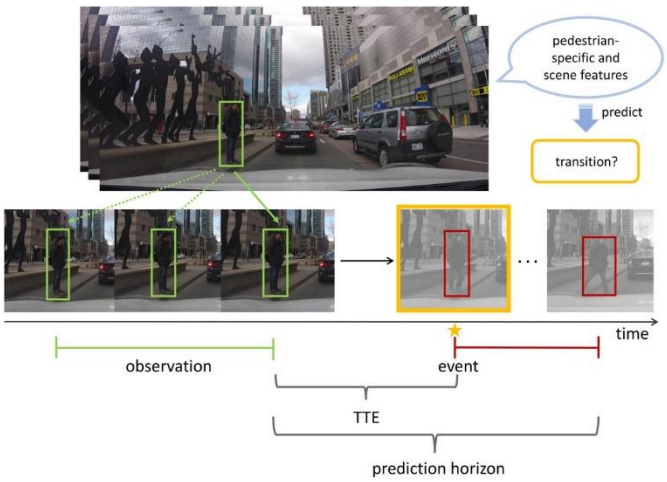


Figure 1 depicts the concept of predicting future transitions in a pedestrian's motion state. The prediction is made by analyzing both pedestrian-specific and scene-related features based on past observations. The prediction horizon is the time frame in which the transition is expected to occur. Pedestrians are represented in green boxes when standing still and red boxes when walking. For this, we first release TRANS, the first large-scale dataset for explicitly studying the stop and go behaviors of pedestrians in urban traffic.[1] It is based on several existing self-driving datasets annotated with pedestrians' walking motions, in order to have diversity in scenarios and environments.

Furthermore, we propose a hybrid model that fuses multi-modal inputs to capture both pedestrian-specific and contextual features in traffic scenes.[2] Our model utilizes feed-forward and recurrent neural networks for spatial-temporal reasoning. We also implement several baselines and analyze the impacts of various design choices. In addition, detailed ablation experiments highlight the importance of contextual cues and temporal dynamics.

II. RELATED WORK

A. Trajectory Forecasting

Forecasting the trajectories of pedestrians is a popular research area for understanding their movements. Early studies in this field focused on developing explicit models of pedestrian behaviors based on predefined rules. However, these models had limitations in capturing complex interactions and were constrained by strong priors. In recent years, data-driven methods that use neural networks to learn interactions have shown better results. For instance, Alahi et al. proposed Social-LSTM, which utilizes a Long Short-Term Memory (LSTM) network for sequential modeling and integrates interactions among nearby pedestrians with a social pooling layer. Gupta et al. used a Generative Adversarial Network (GAN) to learn and generate socially acceptable trajectories, while attention mechanisms were employed to weigh the influences of different neighbors on the person of interest. Most trajectory prediction methods rely on top-down (bird's eye) views captured by stationary cameras. However, Malla et al. explored using action priors from the perspective of a moving vehicle.

B. Action Recognition and Early Prediction

Prior to the advent of deep learning, human activity recognition was mainly based on hand-crafted features, with Improved Dense Trajectories (IDTs) [19] being the state-of-the-art approach. Convolutional Neural Networks (CNNs) were later introduced by Karpathy et al. [20] for early action recognition at the frame level. Simonyan et al. [21] developed a two-stream network that included a second CNN to learn temporal information based on optical flow streams. This approach was successful, inspiring subsequent work that aimed to jointly model spatial and temporal information [22]. Action prediction algorithms share similarities with recognition methods, and frequently used approaches include 3D convolution networks recurrent networks and transformers [27]. These methods have been applied to enhance road safety, including accident estimation [28], anticipating road crossing and pedestrians' intentions and protecting vulnerable road users.

C. Stop/Go Detection and Prediction

Previous research on stop and go behaviors of road users in traffic is limited. Some works have focused on detecting stopping intentions of pedestrians moving towards the curbside using dense optical flow [37], recognizing pedestrian intentions to enter a street, to stop, and to bend using Motion History Images (MHIs), HOG descriptors, and Support Vector Machines (SVMs) [38], or detecting and predicting pedestrian moving intentions utilizing a Hidden Markov Model (HMM) and body keypoints [39], [40]. Other works have used Switching Linear Dynamics to integrate multiple motion modes into trajectory prediction [6], [7], or investigated detecting the start intentions of cyclists using 3D human pose [41] or MHIs [42].

However, these methods primarily rely on pedestrian-specific features such as position, velocity, and body pose, while ignoring crucial contextual and environmental cues that could provide crucial information for long-term prediction. In contrast, our work takes into account both pedestrian-specific features and context information in the scene to anticipate stop and go behaviors.

TABLE I: Statistics of our TRANS dataset. *Go*, *Stop*, *Stand*, *Walk* indicate the number of unique pedestrians in corresponding categories. In brackets, we also count the number of events, i.e., stop and go transitions.

Dataset	Go [events]	Stop [events]	Stand	Walk
JAAD [43]	144 [145]	73 [77]	65	416
PIE [44]	397 [482]	528 [622]	697	483
TITAN [18]	339 [381]	398 [439]	1,077	6,233
TRANS	880 [1,008]	999 [1,138]	1,839	7,132

III. TRANS DATASET

A. Benchmark Selection:

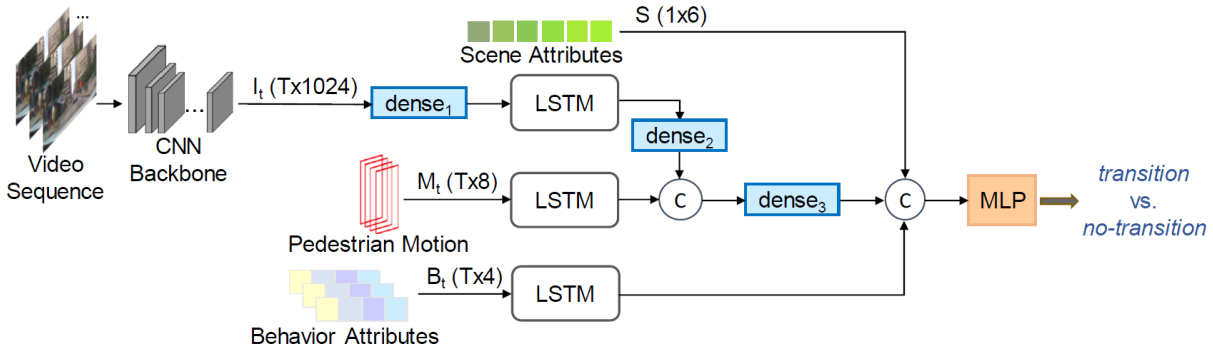
To our knowledge, there are currently no real-world datasets available for studying the stop and go behaviors of pedestrians in the context of autonomous driving. Therefore, we have created the TRANS dataset to facilitate research in this area. The TRANS dataset is built on top of existing autonomous driving datasets, namely JAAD, PIE, and TITAN, which are all related to our task and provide RGB videos captured from an uncalibrated monocular camera, as well as localization and walking annotations for pedestrians.

1) Joint Attention for Autonomous Driving (JAAD):
JAAD contains 346 short video snippets captured under various weather and lighting conditions, with 2D bounding boxes and walking labels provided for pedestrians around potential crossing events.

2) Pedestrian Intention Estimation (PIE):
PIE consists of 6 hours of continuous driving video recorded by a monocular camera, with motion labels available for 1,842 pedestrians close to the road.

*3) Trajectory Inference using Targeted Action priors Network(TITAN):*TITAN is a recently introduced dataset for trajectory forecasting and multi-level action recognition, containing 10 hours of densely populated driving video captured in central Tokyo, with 8,592 unique pedestrians and multiple action labels organized hierarchically by contextual complexity. We have augmented these datasets to build the TRANS dataset, which includes transition samples from diverse traffic scenarios and a unified interface.

Overall, the JAAD dataset comprises selected short clips that are focused on road crossings, while PIE dataset focuses on all potential crossings in a more general manner. On the other hand, TITAN is more generic with numerous annotations on pedestrians who are not interacting with anything.



IV. HYBRID FEATURE FUSION

The proposed model is shown in Figure 2. It takes a sequence of T past observations as input and predicts whether a walking/standing pedestrian will stop/go within a time horizon. The past observations consist of video frames and high-level attributes, resulting in four input modalities (Image, Motion, Behavior, and Scene features), which are gradually fused together. Each dense block contains a fully connected layer, a ReLU activation function, and dropout. The concatenation of the four input modalities is represented by the © symbol. Different CNN backbones are tested for visual encoding, and LSTMs are employed for temporal processing.

A. Annotation pipeline

In order to simplify the annotation process, the original annotations of walking motions in each dataset are used to detect transitions. A transition from standing to walking is considered a "Go" candidate, and the opposite is considered a "Stop" candidate. In TITAN, activities like running are also considered walking. The pre-state is the state before the transition and the post-state is the state after the transition. Only transitions where the duration of both pre-state and post-state is longer than 0.5 seconds are considered valid to reduce labeling errors and obtain more meaningful samples. Pedestrians in the original datasets are categorized into Walk, Stand, Stop, and Go, where Walk and Stand pedestrians do not show transitions, but Stop and Go pedestrians do. Stop and Go are not mutually exclusive since a pedestrian can perform both during the same observation.

B. Problem Formulation

The majority of stops and goes in JAAD and PIE are related to road crossings, while PIE contains more non-crossing transitions and edge cases. The causes of transitions in TITAN are diverse and often ambiguous. The pedestrian stop and go forecasting problem is formulated as a binary classification task, where a sequence of past observations of length T is given, and the objective is to determine whether a given walking/standing pedestrian will stop/go within a time horizon.

The observations include video frames with additional pedestrian and scene attributes, and the model output is a binary prediction of transition vs. no-transition, with stops and goes evaluated as separate tasks.

The impact of social and environmental factors on pedestrian decision-making in urban traffic has been studied before. This paper proposes a hybrid model for predicting when pedestrians will stop or go, which takes into account pedestrian-specific features as well as contextual information. The model uses both feed-forward and recurrent structures to process multi-modal inputs. The visual encoding part of the model processes each image using a Convolutional Neural Network (CNN) to extract information about pedestrians and their surroundings. The CNN is implemented with several backbones that capture different levels of context. The model uses three different CNN backbones: Crop-Box, Crop-Context, and RoI-Context, to extract visual features. The Crop-Box method crops every image at the pedestrian bounding box and pads it with zeros to make it square. The Crop-Context method extracts a square image patch around the pedestrian by scaling up the corresponding bounding box by 2, and then matching the scaled box's width with its height. The RoI-Context method extracts visual features from the whole image using a modified ResNet-18 backbone with a RoI-alignment layer. After the ResNet backbone, a 3x3 convolution is added to reduce the dimension, and the output is flattened to get the visual feature for each frame.

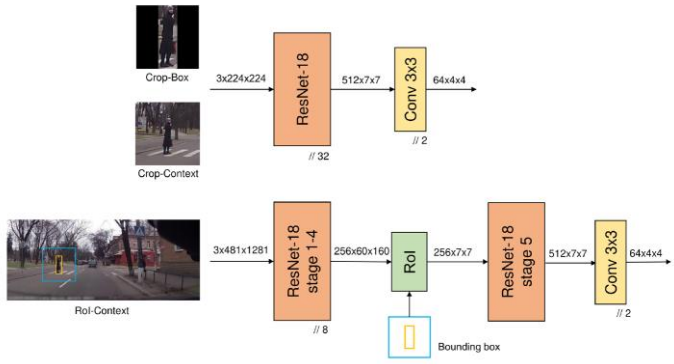


Figure 3 illustrates the different CNN backbones used for visual encoding in the proposed model. The top part shows the Crop-Box and Crop-Context methods where the inputs are RGB image crops of the original or enlarged pedestrian bounding boxes. The ResNet-18 backbone is used for feature extraction followed by a 3x3 convolution. On the other hand, the bottom part shows the RoI-Context method where the input is the whole image of size 481x1281. A RoI-alignment layer is inserted between the fourth and fifth stages of the ResNet, using the enlarged pedestrian bounding box (in blue) as the region proposal.

Motion Encoding. We encode the motions of pedestrians by collecting their positions and velocities. A 4D vector P_t represents a pedestrian's position at each time step t:

$$P_t = (x_t, y_t, w_t, h_t), \tag{1}$$

where (x_t, y_t) are the x-y coordinates of the center of the pedestrian's bounding box, and w_t, h_t are the box's width and height. The velocity V_t at time step t is then defined as the change in position from the previous frame $t-1$, with a time difference of Δt between both frames:

$$\begin{aligned} V_t &= (\Delta x_t, \Delta y_t, \Delta w_t, \Delta h_t) \\ &= \frac{P_t - P_{t-1}}{\Delta t}. \end{aligned} \tag{2}$$

The position and velocity vectors together form our motion feature vector $M_t = (P_t, V_t)$ at each time step t.

Behavior Encoding. Specific body language may reflect pedestrians' will to communicate with the drivers or imply future motions. We use binary attributes to indicate three non-verbal behaviors: looking, nodding, and hand gestures. Additionally, we include the walking state, making the behavioral feature B_t a 4D binary vector at each frame t:

$$B_t = (b^t_{walk}, b^t_{look}, b^t_{nod}, b^t_{hand}). \tag{3}$$

These behavioral attributes are only available in JAAD and PIE. Although hand gestures can cover a wide range of meanings, we only use a binary attribute as done in these datasets. Having further distinctions would likely lead to better and more confident predictions when gestures are recognized.

Scene Encoding. JAAD and PIE provide six high-level semantic attributes that form a coarse, static representation S of the traffic scene:

$$S = (s_{tl}, s_{in}, s_{de}, s_{si}, s_{td}, s_{md}), \tag{4}$$

where s_{tl} denotes the number of traffic lanes, binary variables s_{in}, s_{de}, s_{si} indicate whether the scene is at an intersection, and whether this intersection is designated (with a zebra crossing or a traffic signal) or signalized, and s_{td}, s_{md} show the traffic direction (one-way or two-way) and pedestrian motion direction (lateral or longitudinal). These semantic attributes are not available in TITAN.

Temporal Processing. Recurrent Neural Networks (RNNs) have proven to be highly successful in sequential learning [48]. Long Short-Term Memory (LSTM) networks, as variations of RNNs, address the problem of vanishing gradients and long-term dependency in modeling long sequences [49]. We propagate visual, motion and behavior features through separate LSTMs for temporal processing, and obtain the hidden states at the final time step.

Hybrid Fusion. Adopted a hybrid fusion strategy where inputs and intermediate features are individually processed and then concatenated in a gradual fashion, as observed in Figure 2. We use dense layers to process features and reduce dimensions. The joint representation of all input modalities is then fed to a Multi-Layer Perceptron (MLP) to yield.

V. EXPERIMENTS

A. Data preparation

We performed experiments using our TRANS dataset which consists of video sequences of past observations with a time-to-event (TTE) tag for each frame indicating the time difference between that frame and the next stop or go transition. The label of a video sequence is determined by the TTE tag of its last frame. We set the prediction time horizon to 2 seconds, as this is the minimum time within which pedestrians make crossing decisions. To reduce overfitting and speed up training, we use a sampling rate of 5fps. We removed instances where the widths of the pedestrian bounding boxes in the last frames were smaller than 24 pixels as they are of less interest. It is important to note that we use ground-truth boxes and attributes as inputs, but in practice, they would need to be predicted by another model, and any noise in the predictions would likely affect the final results negatively.

B. Models

In this study, the proposed model is compared to several baseline models that are divided into three groups: Static, Video, and Hybrid. The Static baselines use a single image frame and a visual encoder (CB, CC, or RC) with a fully connected layer to generate the predictions without using LSTM. The Video baselines, on the other hand, use LSTM for temporal encoding in addition to the visual encodings used in the Static baselines. The Hybrid baselines incorporate high-level attributes as input in addition to videos. The first Hybrid model uses two input modalities (Images and pedestrian Motion) available in all three datasets and its architecture is consistent with the design in Figure 2. The full Hybrid model also includes Behavioral and Scene attributes. It is important to note that these baselines are compared with the proposed model to evaluate its performance.

C. Implementation details

The image processing backbones used are ResNet-18. Vanilla LSTMs with tanh activation are used for recurrent networks. The hidden state sizes for encoding Images, Motion, and Behavior features are 256, 64, and 16 respectively. Three embedding dense layers are used with sizes 256, 128, and 64. The MLP for final prediction has three fully connected layers of sizes 86, 86, and 1.

A dropout rate of 0.2 is applied to dense layers for regularization. T = 5 is the input observation sequence length if not specified. Training is done in two stages. In the first stage, the visual encoder CNN's weights are obtained by training the corresponding Static baseline RoI-Context for the same classification task. The images are augmented with random horizontal flipping, cropping out of the top third, resizing to 481 x 1281, random color jittering, and random grayscale conversion. The ResNet backbone is then frozen, and the other parts of the model are trained. Adam optimizer is used with a batch size of 8, a learning rate of 1 x 10^-4, and weight decay of 1 x 10^-5. A binary cross-entropy loss function is used. During training, early stopping is employed, and the number of epochs for convergence varies for each dataset. To compensate for data imbalance, the over-represented class is randomly sampled during training.

D. Evaluation results

The performance of our models is evaluated using Average Precision (AP). Since each dataset is biased towards negative instances, we evaluate the AP on a balanced test set where negative instances are randomly sampled. To reduce the impact of sampling variability, we perform 10 randomized trials and report the average results for each model.

E. Quantitative results

Table II provides a summary of the stop and go forecasting results measured in terms of Average Precision (AP) metric. According to the results, our full model performs the best on JAAD and PIE datasets. When compared to the best Video baseline, the full model shows a 9.5-point and 5.5-point improvement in AP for go forecasting on JAAD and PIE, respectively. This improvement is expected since the high-level attributes of pedestrians and traffic scenes have a strong correlation with crossing, which is the main cause for go transitions in these datasets. However, the improvements in stop forecasting are not as significant, which could suggest that the stops of pedestrians are less correlated with the behavioral and semantic attributes. Additionally, the Hybrid model that combines pedestrian motion features with visual cues outperforms all Video baselines on the three datasets. Furthermore, compared to the Static baselines, adding visual representation of the context improves the performance of the models.

		Go			Stop		
Model	Modalities	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Static	I (CB)	54.3	52.0	56.2	52.5	53.1	56.4
	I (CC)	70.4	59.1	61.4	57.3	61.1	60.3
	I (RC)	73.3	61.2	60.9	58.7	62.5	59.1
Video	I (CB)	60.6	56.4	58.6	57.2	59.4	58.7
	I (CC)	73.6	61.8	63.2	61.4	63.3	61.5
	I (RC)	76.4	64.7	62.9	62.9	64.2	61.7
Hybrid	IM (RC)	80.6	66.5	**65.1**	64.7	64.9	**63.6**
	IMBS (RC)	**85.9**	**70.2**	–	**67.8**	**65.4**	–

TABLE II: Results in Average Precision (AP, %) on TRANS dataset. Modalities are Images (I), Motion (M), Behavior (B) and Scene (S). Visual contexts are Crop-Box (CB), Crop-Context (CC) and RoI-Context (RC).

The Hybrid model, which combines pedestrian motion features and visual cues, performs better than all Video baselines on all three datasets. Adding visual representation of the context significantly improves Average Precision (AP) compared to Static baselines. Global context yields better results on JAAD and PIE but not on TITAN.

The improvements from Static baselines to Video models demonstrate the advantages of using sequential models for temporal reasoning. Across all three datasets, go prediction performs better than stop prediction, possibly due to the fact that high-level attributes such as body language and designated crossing are more relevant for go transitions. There is a significant gap in performance between stop and go predictions on JAAD, which is primarily focused on crossing scenarios. In contrast, the gap is smaller on PIE, which has more non-crossing cases, and the results of both tasks are similar on TITAN, which focuses less on crossing.

F. Qualitative results

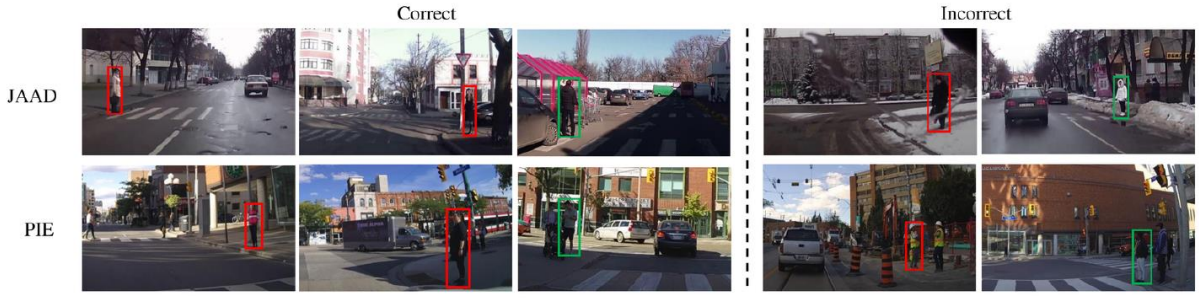
The proposed full Hybrid model's qualitative example predictions on the JAAD and PIE datasets are displayed in Figure 4. The stops and goes of pedestrians at crossroads continue to be difficult to predict, partly due to the absence of ego vehicle speed and traffic light states. Moreover, abrupt changes in movement direction, weather conditions (such as rainy or snowy conditions), and uncommon occurrences such as construction workers can all have a negative impact on the predictions.

G. Ablation studies

We have already established the significance of utilizing multiple modes of input and temporal processing. In order to delve further into the contributions of different input modalities and the effect of observation length T, we have conducted ablation experiments. Our experiments are divided into two parts:

Modalities: We have assessed the impact of individual features by experimenting with various combinations of input modalities for the Hybrid models. We have observed that incorporating Behavior (B) and Scene (S) attributes into Motion (M) information enhances the performance.

The improvements are particularly noteworthy for go forecasting, with an increase of up to 23.2 points on JAAD and 7.5 points on PIE on the AP metric. Table III summarizes the results.



(a) Go transitions

(b) Stop transitions

Fig. 4: Qualitative results of our full proposed Hybrid model on JAAD [43] and PIE [44] datasets. The predictions for future transitions and non-transitions are indicated by red and green boxes respectively. The results are grouped by Go (top) and Stop (bottom) forecasting with correct predictions on the left and incorrect ones on the right.

TABLE III: Ablation study in Average Precision (AP, %) on the choice of modalities for Hybrid models. Modalities are Images (I), Motion (M), Behavior (B) and Scene (S). Visual contexts are Crop-Context (CC) and RoI-Context (RC).

	Go		Stop	
Modalities	JAAD	PIE	JAAD	PIE
S	74.2	55.1	53.3	54.2
M	61.5	59.8	59.4	60.6
I (CC)	73.6	61.8	61.4	63.3
I (RC)	76.4	64.7	62.9	64.2
IM (CC)	78.4	65.1	63.4	63.5
IM (RC)	80.6	66.5	64.7	64.9
MBS	84.7	67.3	62.5	64.7
IMBS (CC)	85.2	69.5	67.2	**65.7**
IMBS (RC)	**85.9**	**70.2**	**67.8**	65.4

It is important to note that when combined with high-level attributes, the performance differences between local (Crop-Context) and global (RoI-Context) visual contexts are not significant. In Table IV, the impact of the length of observation sequences on prediction performance is studied.

The results show that the predictions improve with an increase in the number of observations. However, at some point, the performance improvements reach saturation as evidenced by the stagnation or decrease in AP when the input length is extended from 10 to 15 frames. This behavior is expected as earlier frames should be less correlated with the later transitions.

TABLE IV: Ablation study in Average Precision (AP, %) on the length T of observation sequences. Modalities are Images (I), Motion (M), Behavior (B) and Scene (S). Visual contexts are RoI-Context (RC)

		Go		Stop	
Model	T	JAAD	PIE	JAAD	PIE
Video – I (RC)	1	72.5	61.8	55.8	61.3
	5	76.4	64.7	62.9	64.2
	10	**76.9**	65.8	**63.4**	**66.1**
	15	74.8	**66.2**	60.7	65.7
Hybrid – IMBS (RC)	1	73.6	62.6	59.7	62.0
	5	85.9	70.2	67.8	65.4
	10	**86.7**	71.5	**68.4**	**67.9**
	15	85.1	**71.9**	64.3	67.2

VI. CONCLUSIONS

This paper presented the problem of predicting pedestrian stop and go movements, which is crucial for understanding their trajectories and ensuring their safety. To promote research on this topic, the authors created TRANS, a large dataset for pedestrian stop and go forecasting from a vehicle perspective. The dataset was constructed by combining several existing datasets to cover a range of scenarios and environments.

The authors then proposed a deep learning model that combines video sequences with high-level attributes of pedestrians and their surroundings through a hybrid feature fusion. They evaluated the proposed model and several baseline models on the TRANS dataset and conducted experiments to analyze the impact of different components and design choices.

VII. ACKNOWLEDGMENTS

We would like to thank Professor Adnan Siraj Rakin for his guidance and Researchers at Honda Research Institute for providing us the access to the TITAN dataset.

REFERENCES

[1] Pedestrian Stop and Go Forecasting with Hybrid Feature Fusion by Dongxu Guo, Taylor Mordan and Alexandre Alahi arXiv:2203.02489v1 [cs.CV] 4 Mar 2022

[2] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf, "A literature review on the prediction of pedestrian behavior in urban scenarios," in *Proceedings of the International Conference on Intelligent Transporta- tion Systems (ITSC)*, 2018, pp. 3105–3112.

[3] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 2021.

[4] P. Kothari, B. Sifringer, and A. Alahi, "Interpretable social anchors for human trajectory forecasting in crowds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 556–15 566.

[5] Y. Liu, Q. Yan, and A. Alahi, "Social nce: Contrastive learning of socially-aware motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 118–15 129.

[6] J. F. P. Kooij, N. Schneider, and D. M. Gavrila, "Analysis of pedestrian dynamics from a vehicle perspective," in *IEEE Intelligent Vehicles Symposium Proceedings (IV)*, 2014, pp. 1445–1450.

[7] J. F. Kooij, F. Flohr, E. A. Pool, and D. M. Gavrila, "Context-based path prediction for targets with switching dynamics," *International Journal of Computer Vision (IJCV)*, vol. 127, no. 3, pp. 239–262, 2019.

[8] K. Jayaraman, D. M. Tilbury, X. J. Yang, A. K. Pradhan, and L. P. Robert, "Analysis and prediction of pedestrian crosswalk behavior dur- ing automated vehicle interactions," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6426–6432.

[9] H. Razali, T. Mordan, and A. Alahi, "Pedestrian intention prediction: A convolutional bottom-up multi-task approach," *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103259, 2021. [Online].

[10] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 0, pp. 4282–4286, 1995.

[11] J. van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 05 2008, pp. 1928–1935.

[12] C. Burstedde, K. Klauck, A. Schadschneider, and J. Zittartz, "Simula- tion of pedestrian dynamics using a two dimensional cellular automaton," *Physica A-statistical Mechanics and Its Applications*, vol. 295, pp. 507– 525, 2001.

[13] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, 2016.

[14] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pp. 2255–2264, 2018.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, Red Hook, NY, USA, 2017, p. 6000–6010.

[16] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "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 (CVPR)*, 2019, pp. 1349–1358.

[17] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph trans- former networks for pedestrian trajectory prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[18] S. Malla, B. Dariush, and C. Choi, "TITAN: Future forecast using action priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 186–11 196.

Team member -1

Name: Sathwik Krishtipati
B Number: B00979769
Git Profile: https://github.com/skrisht1

Team member -2

Name: Anirudh Nori
B Number: B00978464
Git profile: https://github.com/anirudhnori7

Github link to project:

Sathwik-Krishtipati/hybrid-feature-fusion (github.com)

Steps to run the code:

Dataset available at

https://github.com/vita-epfl/pedestrian-transition-dataset

1. JAAD, PIE datasets could be forked to our github or to be downloaded locally.

2. We have obtained the TITAN dataset required for this project by contacting the researchers at Honda Research Institute.

Link to download: https://usa.honda-ri.com/titan

3. Steps to install dependencies and datasets, to train the model, to evaluate are given in the below link.

https://github.com/Sathwik-Krishtipati/hybrid-feature-fusion#readme