

Pedestrian Stop and Go Forecasting with Hybrid Feature Fusion

Dongxu Guo, Taylor Mordan and Alexandre Alahi

Abstract—Forecasting pedestrians’ future motions is essential for autonomous driving systems to safely navigate in urban areas. However, existing prediction algorithms often overly rely on past observed trajectories and tend to fail around abrupt dynamic changes, such as when pedestrians suddenly start or stop walking. We suggest that predicting these highly non-linear transitions should form a core component to improve the robustness of motion prediction algorithms. In this paper, we introduce the new task of pedestrian stop and go forecasting. Considering the lack of suitable existing datasets for it, we release TRANS, a benchmark for explicitly studying the stop and go behaviors of pedestrians in urban traffic. We build it from several existing datasets annotated with pedestrians’ walking motions, in order to have various scenarios and behaviors. We also propose a novel hybrid model that leverages pedestrian-specific and scene features from several modalities, both video sequences and high-level attributes, and gradually fuses them to integrate multiple levels of context. We evaluate our model and several baselines on TRANS, and set a new benchmark for the community to work on pedestrian stop and go forecasting.

I. INTRODUCTION

When navigating in populated cities, autonomous vehicles need to anticipate the future movements of pedestrians, who are arguably among the most vulnerable road users, and react accordingly to avoid potential collisions [1], [2]. A large body of work on pedestrian motion prediction uses past observed trajectories to forecast the future locations [3], [4], [5]. These methods are generally effective when the trajectories are rather smooth, and future motions are similar to past ones. However, as the past actions of people may not necessarily imply all of their future movements, trajectory-based methods react poorly to abrupt changes in pedestrian dynamics [3].

We argue that predicting the stops and goes of pedestrians, i.e., the changes between the basic motion states of walking and standing still, can serve as a crucial component for more robust motion forecasting. These transitions are one of the most common and essential aspects of human movement patterns. However, they are highly non-linear in nature, and therefore usually hard to predict [6], [7]. Moreover, stops and goes are often involved in safety-critical traffic scenarios, such as a pedestrian waiting at the curbside and later walking for crossing [8], [9]. Failing to foresee such transitions can lead to catastrophic consequences.

In this paper, we introduce the task of pedestrian stop and go forecasting from the ego-centric view of a moving vehicle, as illustrated in Figure 1. In order to study it, we setup a benchmark with a new dataset and multiple approaches

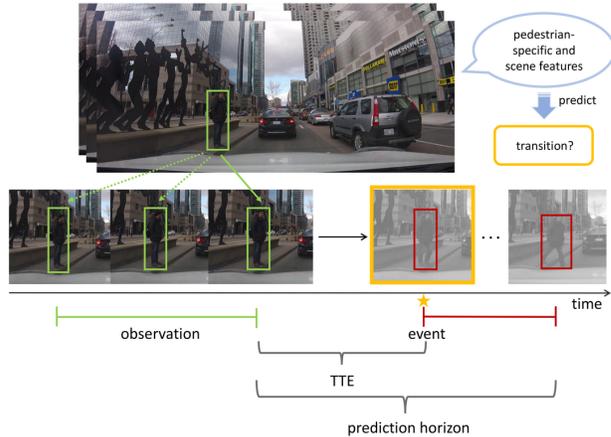


Fig. 1: Illustration of a future transition forecasting. Given past observations, we predict whether a transition in the pedestrian’s motion state will occur within the prediction horizon by reasoning over pedestrian-specific and scene features. Pedestrians are enclosed in green boxes when standing still, and in red boxes when walking.

learned and evaluated on it. 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

Trajectory forecasting is an active area of research for modeling pedestrian behaviors. Many early works [10], [11], [12] primarily focus on developing explicit models about pedestrian movements based on hand-crafted rules. Although they demonstrate some competitive results, these approaches impose strong priors and have limited capacity in capturing complex interactions. In recent years, data-driven

All authors are with École Polytechnique Fédérale de Lausanne (EPFL), Visual Intelligence for Transportation (VITA), CH-1015 Lausanne, Switzerland.

¹Dataset available at <https://github.com/vita-epfl/pedestrian-transition-dataset>.

²Code available at <https://github.com/vita-epfl/hybrid-feature-fusion>.

methods that utilize neural networks to learn interactions have been shown to yield superior performances [3]. Alahi et al. introduce Social-LSTM [13] for predicting pedestrian trajectories in crowded spaces. Social-LSTM deploys a Long Short-Term Memory (LSTM) network for sequential modeling and integrates interactions of nearby pedestrians with a novel social pooling layer. Gupta et al. [14] use a Generative Adversarial Network (GAN) to learn and generate more socially acceptable trajectories. Attention mechanisms [15] are also used to weight the influences of different neighbours on the person of interest [16], [17]. The majority of trajectory prediction methods relies on top-down (bird’s eye) views captured by stationary cameras. Malla et al. [18] explore using action priors from the view of a moving vehicle.

B. Action Recognition and Early Prediction

Before deep learning, methods using hand-crafted features, represented by Improved Dense Trajectories (IDTs) [19], were the state of the art of human activity recognition. Karpathy et al. achieve early action recognition at the frame level using Convolutional Neural Networks (CNNs) [20]. Simonyan et al. [21] introduce a two-stream network where a second CNN is added to learn temporal information in the videos based on optical flow streams. The great success of this last approach [21] inspires following works to jointly model the spatial and temporal information in the videos [22]. Action prediction algorithms are inherently similar to recognition. Some commonly used methods include 3D convolution networks [23], [24], recurrent networks [25], [26], and more recently transformers [27]. Action recognition and early prediction methods have been applied to improve road safety, e.g., with accident estimation [28], anticipating road crossing and pedestrians’ intentions [29], [30], [31], [32], [33], [34], and protecting vulnerable road users [35], [36].

C. Stop/Go Detection and Prediction

Only a few previous works have explored the stop and go behaviors of road users in traffic. Keller et al. [37] detect stopping intentions of pedestrians moving toward the curbside using dense optical flow to predict future paths. Koehler et al. [38] recognize pedestrian intentions to enter a street, to stop, and to bend in using Motion History Images (MHIs), HOG descriptors, and Support Vector Machines (SVMs). Quintero et al. [39], [40] propose an approach to detect and predict pedestrian moving intentions utilizing a Hidden Markov Model (HMM) and body keypoints. Kooij et al. [6], [7] use Switching Linear Dynamics to integrate multiple motion modes into trajectory prediction. Apart from pedestrians, detecting the start intentions of cyclists is also investigated using 3D human pose [41] or MHIs [42]. Despite obtaining solid results, these methods rely primarily on pedestrian-specific features such as position, velocity and body pose. The contextual and environmental cues, which can provide crucial information for long-term prediction, are largely ignored. In contrast, our work anticipates the stop and go behaviors by reasoning over both pedestrian-specific features and context information in the scene.

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

To the best of our knowledge, there are no large-scale, real-world datasets currently available for studying the stops and goes of pedestrians in the context of autonomous driving. Hence, we build TRANS dataset to facilitate the research in this domain. It is built on top of several existing autonomous driving datasets annotated with walking behaviors of pedestrians (see Table I), so that it includes transition samples from diverse traffic scenarios with a unified interface.

A. Benchmark Selection

We augment three existing self-driving datasets, namely JAAD [43], PIE [44] and TITAN [18] for building the benchmark. The three are closely related to our task in the sense that they all provide RGB videos captured from an uncalibrated monocular camera on a moving platform, together with localization and walking annotations for pedestrians. They also provide train, validation, and test splits by video clips.

1) *Joint Attention for Autonomous Driving (JAAD)*: it aims at exploring the road users’ behaviors in pedestrian crossing settings [43]. It consists of 346 short video snippets recorded at 30fps with a dashboard camera under various weather and lighting conditions. JAAD provides 2D bounding boxes for all pedestrians, among which 654 are around potential crossing events and are annotated with walking labels.

2) *Pedestrian Intention Estimation (PIE)*: it is designed for the task of recognizing pedestrians’ intentions of crossing the roads [44]. It is sourced from 6 hours of continuous daytime driving recorded at 30fps by a monocular camera in North America. Labels that indicate motions are available for 1,842 pedestrians close to the road that may potentially interact with the driver.

3) *Trajectory Inference using Targeted Action priors Network (TITAN)*: it is a dataset recently introduced for trajectory forecasting and multi-level action recognition [18]. 10 hours of driving video are recorded at 60fps in densely populated central Tokyo. To construct the final dataset, 700 short clips are extracted and annotated at a sampling frequency of 10Hz. TITAN contains 8,592 unique pedestrians with multiple action labels organized hierarchically by contextual complexity. The labels for primitive actions such as walking, standing still and running are mutually exclusive.

Overall, JAAD contains selected short clips centered on road crossings, while PIE focuses on all potential crossings in a more general way. TITAN is even more generic with numerous annotations on pedestrians not interacting with

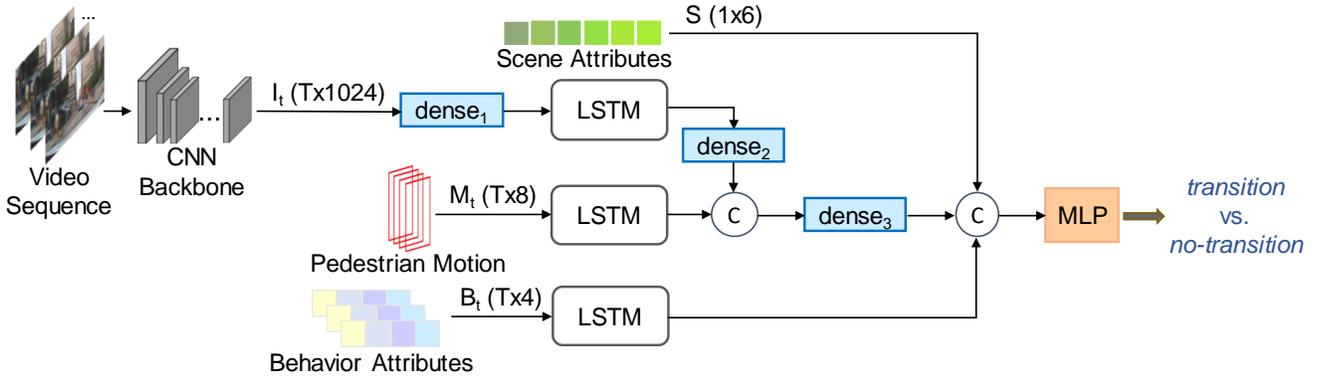


Fig. 2: Overview of our proposed model. With a sequence of T past observations, the model predicts whether a walking/standing pedestrian will stop/go within a time horizon λ . The past observations include video frames and high-level attributes, leading to four input modalities (*Image*, *Motion*, *Behavior* and *Scene* features) that are gradually fused together. A *dense* block consists of a fully connected layer, a ReLU activation function, and dropout. Blocks © represent concatenation. We test several CNN backbones for visual encoding and employ LSTMs for temporal processing.

the traffic. TRANS therefore offers an increasing level of difficulty through the various datasets, by varying the relevance of objective cues related to traffic.

B. Annotation pipeline

To simplify the annotation process, we rely on the original annotations of walking motions in each dataset to detect transitions. We count a state change from *standing* to *walking* as a *Go* candidate, and the opposite as a *Stop* one. In TITAN, the activities like *running* are also viewed as *walking*. We refer to the state before the transition as *pre-state* and the state after as *post-state*. To reduce possible errors by inaccurate labeling and obtain more meaningful samples, we only consider a transition candidate to be valid if the durations of its *pre-state* and *post-state* both last longer than 0.5 seconds.

All unique pedestrians in the original datasets can be categorized into *Walk*, *Stand*, *Stop* and *Go*. *Walk* and *Stand* pedestrians show no transitions throughout the video, whereas *Stop* and *Go* pedestrians exhibit the corresponding transitions. The classes of *Stop* and *Go* are not mutually exclusive since a pedestrian can perform both stops and goes during the same observation. The statistics of datasets are presented in Table I.

After inspecting the detected transitions, we find the majority of stops and goes in JAAD and PIE to be closely related to road crossings. Compared to JAAD, PIE contains more non-crossing transitions and edge cases. However, causes for transitions in TITAN are diverse and often ambiguous.

C. Problem Formulation

We formulate pedestrian stop and go forecasting as a binary classification problem, and illustrate it in Figure 1. Given a sequence of past observations of length T , the objective is to determine whether a given walking/standing pedestrian will stop/go within a time horizon λ . The observations include video frames with additional pedestrian and scene attributes. We assume the pedestrian’s current bounding box and motion state (standing/walking) are known, and the model output is a binary prediction of *transition* vs. *no-transition*. The stops and goes are evaluated as two separate tasks.

IV. HYBRID FEATURE FUSION

Diverse social and environmental factors have been shown to impact pedestrian motions and decision-making in urban traffic [45]. We propose a hybrid model for pedestrian stop and go forecasting that encodes pedestrian-specific features jointly with dynamics and contextual information. Our model relies on both feed-forward and recurrent structures to process multi-modal inputs. The detailed model design is shown in Figure 2 and is described in the following.

Visual Encoding. We process each image with a Convolutional Neural Network (CNN) to extract information about pedestrians and the contexts around them. We implement several CNN backbones that include different levels of context, as shown in Figure 3, to analyze the impact of the context.

First, we compare between no or local contexts around pedestrians. Without context, we crop every image at the pedestrian bounding box and pad it with zeros to make it square. To get local context, we extract a square image patch around the pedestrian by scaling up the corresponding bounding box by 2, then matching the scaled box’s width with its height. For both methods, the patches are fed to a ResNet-18 [46] backbone for feature extraction. We refer to the former as *Crop-Box* and to the later as *Crop-Context*.

We also implement a CNN backbone to extract the visual features from whole images instead of patches. We modify the previous ResNet-18 backbone by inserting a RoI-alignment layer [47] between the fourth and fifth stages to capture broad contexts in the original images. For each frame, the RoI is defined by the pedestrian’s bounding box enlarged following the same procedure as for *Crop-context*. To better preserve the resolution, we also remove the max pooling in the ResNet’s first stage and modify the fifth stage so that the first convolution operates on a 7×7 feature map with a stride of 1. We refer to this design as *RoI-Context*.

After the ResNet backbone, we add a 3×3 convolution to reduce the dimension, and flatten the output to get the visual feature I_t for the frame t .

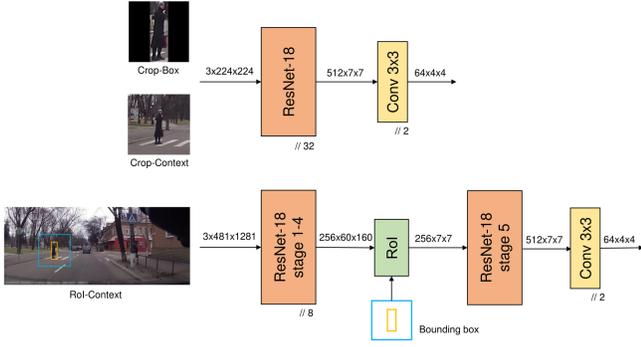


Fig. 3: CNN backbones for visual encoding. For *Crop-Box* and *Crop-context* (top), the inputs are RGB image crops, either by original or enlarged pedestrian bounding boxes, scaled to size 224×224 . The feature extractor is a ResNet-18 backbone followed by a 3×3 convolution. For *RoI-Context* (bottom), the input is the whole image of size 481×1281 , and a RoI-alignment layer, using the enlarged pedestrian bounding box (in blue) as the region proposal, is inserted between the ResNet’s fourth and fifth stages.

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 :

$$P_t = (x_t, y_t, w_t, h_t), \quad (1)$$

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:

$$V_t = \frac{(\Delta x_t, \Delta y_t, \Delta w_t, \Delta h_t)}{\Delta t} = \frac{P_t - P_{t-1}}{\Delta t}. \quad (2)$$

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 :

$$B_t = (b_{walk}^t, b_{look}^t, b_{nod}^t, b_{hand}^t). \quad (3)$$

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:

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

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. Inspired by [50], we adopt 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 the final predictions.

V. EXPERIMENTS

A. Data Preparation

We conduct all experiments on our new TRANS dataset. Examples consist of video sequences of past observations of length T . Each frame possesses a time-to-event (TTE) tag indicating the time gap between this frame and the subsequent stop or go transition. The label of a video sequence (*transition* or *no-transition*) is determined by the TTE tag of its last frame. As crossing-related cases are arguably the most critical for stop and go forecasting, we set the prediction time horizon λ to be 2 seconds, which is the minimum time within which pedestrians make crossing decisions [51]. We choose a relatively large sampling rate of 5fps in the hope of reducing overfitting and speeding up training. Since the pedestrians far away from the ego vehicle are of less interest, we remove the instances whose widths of the pedestrian bounding boxes in the last frames are smaller than 24 pixels. It is important to note that we use ground-truth boxes and attributes as inputs here. However, these would first need to be predicted by another model in practice, and noise in the predictions would likely negatively affect the final results.

B. Models

We compare the performance of the proposed model to a series of baselines, grouped into *Static*, *Video* and *Hybrid*.

Static. Static baselines take single image frames as input, and directly output the classification results. We combine the visual encoders mentioned previously (*Crop-Box* (CB), *Crop-Context* (CC), *RoI-Context* (RC)) with a fully connected layer to yield the predictions (i.e., without using LSTM).

Video. Video baselines extract the visual features from video sequences and use LSTMs for temporal encoding. They use the same visual encodings as *Static* baselines.

Hybrid. Hybrid baselines use high-level attributes as input in addition to videos. The first model utilizes two input modalities available in all three datasets: *Images* (I) using a *RoI-Context* (RC) encoder, and pedestrian *Motion* (M). Its architecture is consistent to the design in Figure 2. Our full model also uses *Behavioral* (B) and *Scene* (S) attributes.

C. Implementation Details

All backbones for image processing are ResNet-18 [46]. For recurrent networks, we use vanilla LSTMs with *tanh* activation, and the sizes of hidden states for encoding *Images*, *Motion*, and *Behavior* features are set to 256, 64, and 16 respectively. We set the sizes of the three embedding dense layers to 256, 128 and 64 in order. The MLP for final prediction consists of three fully connected layers of size 86, 86 and 1. We apply dropout with a rate of 0.2 in dense layers for regularization. If not specified, we set the input observation sequence length to $T = 5$.

Training consists of two stages. First, the weights of the visual encoder CNN are obtained by training the corresponding *Static* baseline *RoI-Context* for the same classification task. At this training stage, we augment the images with random horizontal flipping, cropping out of the top third, resizing to 481×1281 , random color jittering, and random grayscale conversion. We then freeze the ResNet backbone and train other parts of the model. The models are trained with Adam [52] optimizer with a batch size of 8, learning rate of $1 \cdot 10^{-4}$, and weight decay of $1 \cdot 10^{-5}$. We use a binary cross entropy (BCE) loss function. We employ early stopping during training and the number of epochs for convergence varies for each dataset. To compensate for the data imbalance, we randomly sample the over-represented class during training.

D. Evaluation Metrics

We use Average Precision (AP) to evaluate the models’ performances. As each dataset is unbalanced toward *no-transition*, we calculate the AP on a balanced test set where negative instances are randomly sampled. To reduce the variance introduced by sampling, we conduct 10 randomized trials and report the averaged results for each model.

E. Quantitative Results

Table II summarizes the stop and go forecasting results with AP metric. Our full model achieves the best performance on JAAD and PIE. In particular, for go forecasting, compared to the best *Video* baseline, the full model improves AP by 9.5 points on JAAD and 5.5 points on PIE. The improvements are expected as the high-level attributes of the pedestrians and traffic scenes have strong correlations with crossing, which is the primary cause for go transitions in JAAD and PIE. Interestingly, the improvements in stop forecasting are less noticeable, which may suggest the stops of pedestrians are less correlated with the behavioral and semantic attributes.

On the three datasets, the *Hybrid* model that fuses pedestrian motion features with visual cues outperforms all *Video* baselines. Comparing the results of *Static* baselines, we see that adding visual representation of the context improves

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).

Model	Modalities	Go			Stop		
		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	–

AP by a large margin. Using global context yields better results than local context on JAAD and PIE but is inferior on TITAN. The notable improvements in AP from *Static* baselines to *Video* ones demonstrate the benefits of using sequential models for temporal reasoning.

On all the datasets, the models’ performances for go prediction are generally better than for stop. This could be caused by the fact that, in a typical crossing scenario, go predictions may benefit more from specific high-level attributes of the pedestrian behaviors and the scene, such as body languages and the existence of designated crossing, while the social cues for stops can sometimes be ambiguous or not present in our attributes. As a result, we can see a wide gap between the stop and go predictions on JAAD, which is mainly about crossing scenarios. On the other hand, the performance gap shrinks on PIE, which contains more non-crossing cases, and the results of these two tasks are close on TITAN, which focuses less on crossing.

F. Qualitative Analysis

Figure 4 displays qualitative example predictions of our proposed full *Hybrid* model on JAAD and PIE datasets. We can see that the stops and goes of pedestrians at crossroads remain challenging to predict, partially due to the lack of ego vehicle speed and states of the traffic lights. In addition, sudden changes of moving direction, weather conditions (e.g., rainy, snowy), and irregular cases like construction workers can also negatively impact the predictions.

G. Ablation Study

We have discussed the importance of multi-modal fusion and temporal processing. We now conduct ablation experiments to further investigate the contributions of different input modalities and the impact of the length T of the observations.

1) *Choice of Modalities:* We assess the contributions of individual features in Table III by experimenting with different combinations of input modalities for *Hybrid* models. Adding *Behavior* (B) and *Scene* (S) attributes to *Motion* (M) information boosts the performance. The improvements on AP metric are particularly important for go forecasting, by up to 23.2 points on JAAD and 7.5 points on PIE. Adding *Image* (I)

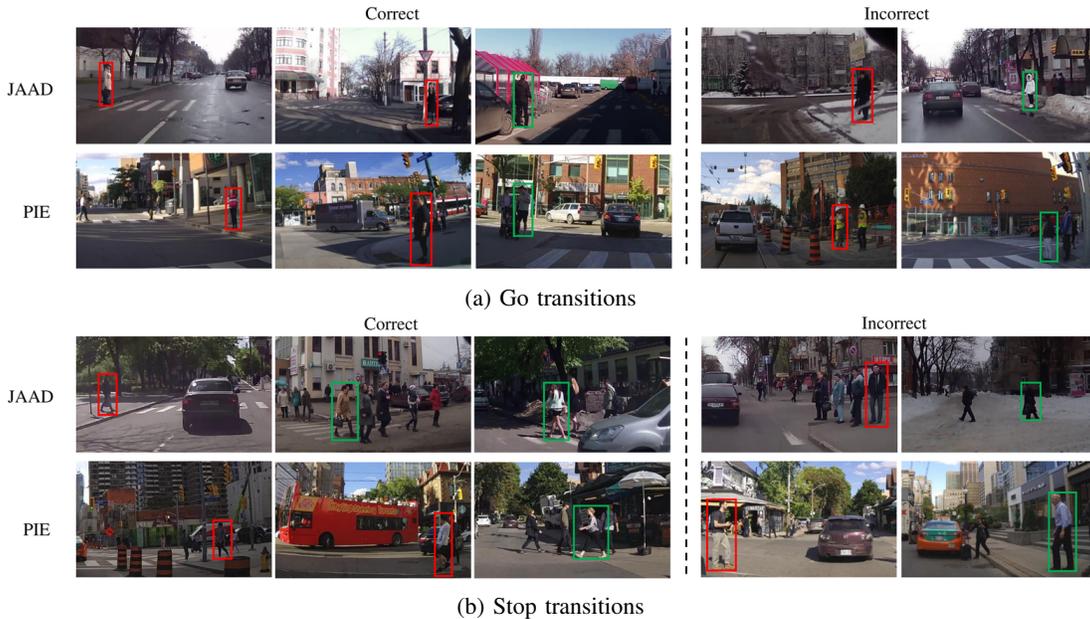


Fig. 4: Qualitative results of our full proposed *Hybrid* model on JAAD [43] and PIE [44] datasets. The predictions for future transitions and non-transitions are indicated by red and green boxes respectively. The results are grouped by *Go* (top) and *Stop* (bottom) forecasting with correct predictions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

Modalities	Go		Stop	
	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

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).

Model	T	Go		Stop	
		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

features further enhances the models’ performances. It should be noted that when combined with high-level attributes, the performance gaps between local (*Crop-Context*) and global (*RoI-Context*) visual contexts are narrow.

2) *Impact of the Length of Observation Sequences*: In Table IV, we study how the prediction performance evolves as we change the length T of observation sequences. Overall, the predictions improve when more observations come in. However, the performance improvements reach saturation at some point, as evidenced by the stagnation or decrease in AP when we extend the input length from 10 to 15 frames. This behavior is expected as earlier frames should be less correlated with the later transitions.

VI. CONCLUSIONS

In this paper, we have introduced the problem of pedestrian stop and go forecasting. Accurately predicting these highly

non-linear transitions is crucial to understand pedestrians’ trajectories and guarantee their safety. To foster research on this problem, we have also set up a new benchmark for the community. For this, we have released TRANS, the first large-scale dataset for pedestrian stop and go forecasting from a vehicle perspective. It is based on several existing datasets in order to contain various scenarios and environments. We have then introduced a new deep learning model leveraging video sequences and high-level attributes about pedestrians and contextual scenes through a hybrid feature fusion, and have evaluated it, along with multiple baselines, on TRANS. We have finally performed extensive experiments to analyse the impact of all the components and design choices.

VII. ACKNOWLEDGMENTS

We would like to thank Valeo and EPFL for funding our work, and the reviewers for their helpful comments.

REFERENCES

- [1] A. Alahi, M. Bierlaire, and M. Kunt, "Object detection and matching with mobile cameras collaborating with fixed cameras," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008.
- [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 Transportation 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. Gavrilu, "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. Gavrilu, "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 during 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]. Available: <https://www.sciencedirect.com/science/article/pii/S0968090X21002710>
- [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, "Simulation 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 transformer 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.
- [19] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3551–3558.
- [20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 2014, pp. 1725–1732.
- [21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, 2014.
- [22] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.
- [23] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 1, pp. 221–231, 2013.
- [24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.
- [25] Y. Shi, B. Fernando, and R. Hartley, "Action anticipation with RBF kernelized feature mapping RNN," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [26] Y. Abu Farha, A. Richard, and J. Gall, "When will you do what?-anticipating temporal occurrences of activities," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5343–5352.
- [27] R. Girdhar and K. Grauman, "Anticipative video transformer," *arXiv preprint arXiv:2106.02036*, 2021.
- [28] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 273–280.
- [29] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 2243–2248.
- [30] B. Völz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "A data-driven approach for pedestrian intention estimation," in *2016 IEEE 19th international conference on intelligent transportation systems (itsc)*. IEEE, 2016, pp. 2607–2612.
- [31] D. Varytimidis, F. Alonso-Fernandez, B. Duran, and C. Englund, "Action and intention recognition of pedestrians in urban traffic," in *2018 14th International conference on signal-image technology & internet-based systems (SITIS)*. IEEE, 2018, pp. 676–682.
- [32] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9704–9710.
- [33] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," in *British Machine Vision Conference (BMVC)*, 2019.
- [34] Y. Yao, E. Atkins, M. J. Roberson, R. Vasudevan, and X. Du, "Coupling intent and action for pedestrian crossing behavior prediction," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.
- [35] V. Kress, S. Schreck, S. Zernetsch, K. Doll, and B. Sick, "Pose based action recognition of vulnerable road users using recurrent neural networks," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 2723–2730.
- [36] T. Mordan, M. Cord, P. Pérez, and A. Alahi, "Detecting 32 pedestrian attributes for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 2021.
- [37] C. Keller and D. Gavrilu, "Will the pedestrian cross? a study on pedestrian path prediction," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, vol. 15, no. 2, pp. 494–506, 2013.
- [38] S. Koehler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer, "Stationary detection of the pedestrian's intention at intersections," *IEEE Intelligent Transportation Systems Magazine*, vol. 5, no. 4, pp. 87–99, 2013.
- [39] R. Quintero, I. Parra, J. Lorenzo, D. Fernández-Llorca, and M. A. Sotelo, "Pedestrian intention recognition by means of a hidden markov model and body language," in *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–7.
- [40] R. Quintero, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, pp. 1–12, 2018.
- [41] V. Kress, J. Jung, S. Zernetsch, K. Doll, and B. Sick, "Pose based

- start intention detection of cyclists,” in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 2381–2386.
- [42] S. Zernetsch, V. Kress, B. Sick, and K. Doll, “Early start intention detection of cyclists using motion history images and a deep residual network,” *arXiv preprint arXiv:1803.02242*, 2018.
- [43] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.
- [44] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6262–6271.
- [45] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, “Joint attention in autonomous driving (JAAD),” *arXiv preprint arXiv:1609.04741*, 2020.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [47] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
- [48] Z. C. Lipton, J. Berkowitz, and C. Elkan, “A critical review of recurrent neural networks for sequence learning,” *arXiv preprint arXiv:1506.00019*, 2015.
- [49] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, 2014.
- [50] A. Rasouli, T. Yau, M. Rohani, and J. Luo, “Multi-modal hybrid architecture for pedestrian action prediction,” *arXiv preprint arXiv:2012.00514*, 2020.
- [51] S. Schmidt and B. Färber, “Pedestrians at the kerb – recognising the action intentions of humans,” *Transportation Research Part F-traffic Psychology and Behaviour*, vol. 12, pp. 300–310, 2009.
- [52] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the International Conference for Learning Representations (ICLR)*, 2015.