PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction

Amir Rasouli*, Iuliia Kotseruba*, Toni Kunic and John K. Tsotsos
York University, Toronto, Ontario, Canada
{aras,yulia_k,tk,tsotsos}@eecs.yorku.ca

Abstract

Pedestrian behavior anticipation is a key challenge in the design of assistive and autonomous driving systems suitable for urban environments. An intelligent system should be able to understand the intentions or underlying motives of pedestrians and to predict their forthcoming actions. To date, only a few public datasets were proposed for the purpose of studying pedestrian behavior prediction in the context of intelligent driving. To this end, we propose a novel large-scale dataset designed for pedestrian intention estimation (PIE). We conducted a large-scale human experiment to establish human reference data for pedestrian intention in traffic scenes. We propose models for estimating pedestrian crossing intention and predicting their future trajectory. Our intention estimation model achieves 79% accuracy and our trajectory prediction algorithm outperforms state-of-the-art by 26% on the proposed dataset. We further show that combining pedestrian intention with observed motion improves trajectory prediction. The dataset and models are available at `http://data.nvision2.eecs.yorku.ca/PIE_dataset/`.

1. Introduction

In the past decade, we have witnessed a rapid growth in the development of assistive and autonomous driving systems capable of performing various perception, planning and control tasks. Yet these systems still face a major challenge when it comes to driving in highly dynamic urban environments. Aside from perceiving the environment, an intelligent driving system should be capable of comprehending the underlying intentions of other road users and anticipating their forthcoming actions (Figure 1) [33]. This is particularly important when dealing with pedestrians at the point of crossing as they exhibit highly variable behavior patterns [26].

Most current approaches to pedestrian action prediction are trajectory-based [16, 1, 5], meaning that they rely on the past observed motion of the pedestrians and/or vehicle dynamics to predict the future locations of the pedestrians. These approaches, however, are effective when the pedestrians are already crossing or are about to do so, i.e. these algorithms react to an action already in progress instead of anticipating it. For example, scenarios where a pedestrian is standing at the intersection or walking alongside the road prior to crossing can be challenging for trajectory-based approaches. Moreover, the past trajectory of a pedestrian might not necessarily reflect their ultimate objective. For instance, a pedestrian waiting at a bus stop might step on the road to check for the bus. This action might be interpreted as a crossing event by a trajectory-based approach.

A remedy for the common drawbacks of trajectory-based algorithms is to anticipate the action by estimating its underlying cause or intention. Intention estimation allows one to predict a future situation using expected behaviors rather than merely rely on scene dynamics [33]. In the context

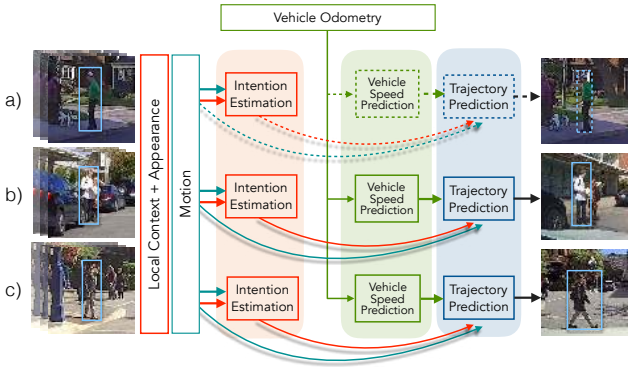

Figure 1. Processing stages for different sources of information required for understanding and predicting pedestrian behavior. Three examples are shown: a) a pedestrian who does not intend to cross, b) a pedestrian who intends to cross but does not cross and c) a pedestrian who intends to cross and crosses the street. Observations of pedestrians' appearance and movement in combination with local context help estimate whether they intend to cross the street. Intention can be used to filter out irrelevant pedestrians (eliminating the need for further processing as shown with dashed lines) and/or to improve trajectory prediction.

*denotes equal contribution

of intelligent driving, a pedestrian's intention reflects their principal goal of crossing the street. The pedestrian might not have any intention to cross (e.g. they could be waiting for a bus, talking to someone, or taking a photo), or they intend to cross and may or may not act on it depending on the traffic conditions. Detecting pedestrians' intentions can potentially reduce the cognitive load of an intelligent driving system allowing it to identify those pedestrians whose actions will be relevant to its own behavior planning. This may also grant such systems a better ability to anticipate pedestrian behavior [33].

In this paper we propose the first large-scale dataset for pedestrian intention estimation and trajectory prediction. The dataset contains several hours of naturalistic video footage of pedestrians in urban environments. In addition to bounding box and behavior annotations, we augment our dataset with human reference data for pedestrian intention estimation established via a large-scale experiment. We propose models for pedestrian intention estimation and trajectory prediction for on-board camera systems.

2. Related works

In the literature various terms such as intention, action and behavior are used to describe what the agent is doing or about to do in the scene. Here, we distinguish intention as the underlying state of mind which cannot be observed but can be inferred from the behavior. This is opposed to actions and, more generally, behaviors, i.e. observable actions such as walking or crossing, for which there is ground truth available.

Action prediction. In the computer vision community there is a large body of works dedicated to video and action prediction [20, 19, 24, 21, 7, 5, 17]. Action (or behavior) prediction algorithms may take different forms such as generating future frames [20, 19, 24, 6], predicting the type of action [15, 21, 7], measuring confidence in the occurrence of an event [27, 37, 10], and forecasting the motion of objects [25, 40, 43, 1, 17, 5, 8].

Behavior and trajectory estimation. Algorithms which predict the occurrence of certain events, e.g. crossing the street, use information such as the road structure, pedestrian head orientation [27] and pose [10], or scene dynamics [37]. Although these algorithms are ideal for providing situational awareness, they do not give any information regarding the future location of objects that can potentially be helpful for trajectory planning. Some algorithms construct the future scenes from which they either calculate optical flow and scene motion [19] or directly localize the objects of interest [6]. These algorithms, however, are very susceptible to occlusion and degrade rapidly with increasing temporal prediction duration.

Trajectory-based algorithms rely on observation of pedestrians' past motion history and predict the location of the pedestrians in the future using contextual information such as 3D depth [14, 34, 39], social interactions [25, 42, 1, 31, 41], the ego-vehicle dynamics and the scene structure [16, 17, 5]. In an on-board camera setting, however, accurate depth information may not be available in many driver assistance systems. Social interactions may also be difficult to infer without a top-down view of the scene as pointed out in [5].

The state-of-the-art on-board pedestrian trajectory estimation in [5] uses a two-stream encoder-decoder scheme which combines encoding of observed bounding box locations and the ego-vehicle's odometry to predict the future bounding box locations of pedestrians over one second into the future. This approach uses the last observed visual information to estimate the future odometry of the vehicle, however, does not take into consideration any visual features of the pedestrians to predict their trajectories.

Intention estimation. In the computer vision and robotics literature the term intention is often used in the context of action classification or path refinement. In [11, 27], the authors assume that pedestrians want to cross and decide whether the crossing takes place in front of the vehicle and when. Intention, defined as the potential goal (destination) of pedestrians, is used to refine predicted trajectories [3, 29, 2, 30]. These approaches rely heavily on motion history of the pedestrians and predict the trajectory of every individual. To the best of our knowledge, there is only one previous work that defines pedestrian crossing intention as their principal goal to cross [33]. The authors propose to infer pedestrian crossing intention from their movement patterns and their proximity to various road elements, e.g. curbside, bus stop, ego-vehicle lane. Their algorithm, however, does not contain a perception mechanism and relies on ground truth information for reasoning.

Datasets. A number of datasets for trajectory prediction contain videos collected from a top-down view [18, 25, 22, 31] or surveillance camera perspective [23, 4, 45]. There are relatively fewer datasets that are specifically catered for pedestrian behavior prediction from a moving vehicle perspective. Publicly available pedestrian detection datasets [9, 12, 44] can potentially be used for such a purpose, however, they lack necessary characteristics such as ego-vehicle information [9], temporal correspondence [44], or enough pedestrians samples with long tracks [12]. These datasets also do not include any form of pedestrian behavior annotations that can be used for action prediction. A recently proposed dataset, JAAD [27], contains a large number of pedestrian samples with temporal correspondence, a subset of which are annotated with behavior information. However, for the purposes of intention estimation and trajectory prediction, this dataset has a number of drawbacks. The dataset does not have ego-vehicle information, the videos are divided into short discontinuous chunks, and the majo-

rity of pedestrian samples with behavioral annotations have the intention of crossing.

Contributions. This paper offers the following four contributions: **1)** A large-scale pedestrian intention estimation (PIE) dataset which includes hours of video footage of pedestrians at various types of crosswalks collected using a calibrated on-board camera. The dataset contains annotations necessary for perception and visual reasoning, including bounding boxes for traffic objects, pedestrian intentions and actions, pedestrian attributes (e.g. gender, age), road boundaries, and ego-vehicle information (e.g. GPS, speed, heading angle). **2)** A human baseline for pedestrian intention estimation established by conducting in-lab and large-scale online experiments involving human subjects of different ages and driving backgrounds. This information provides us with an estimate of pedestrians' crossing intention that can serve for both training and evaluation of intention estimation algorithms. **3)** A novel algorithm that combines past trajectory information and local visual context for predicting pedestrians' intention of crossing. **4)** A trajectory prediction algorithm that achieves state-of-the-art performance on PIE and JAAD datasets, and shows how various contextual information can impact the accuracy of predicted trajectories.

3. PIE Dataset

3.1. Data

The PIE dataset consists of over 6 hours of driving footage captured with calibrated monocular dashboard camera Waylens Horizon equipped with 157° wide angle lens. All videos are recorded in HD format (1920×1080 px) at 30 fps. The camera was placed inside the vehicle below the rear-view mirror. For convenience, videos are split into approx. 10 minute long chunks and grouped into 6 sets. The entire dataset was recorded in downtown Toronto, Canada during daytime under sunny/overcast weather conditions.

Our dataset represents a wide diversity of pedestrian behaviors at the point of crossing and includes locations with high foot-traffic and narrow streets as well as wide boulevards with fewer pedestrians. PIE provides long continuous sequences and annotations for a wide range of applications.

Annotations. For each pedestrian close to the road that can potentially interact with the driver we provide the following annotations: bounding boxes with occlusion flags, as well as crossing intention confidence and text labels for pedestrians' actions ("walking", "standing", "looking", "not looking", "crossing", "not crossing"). Each pedestrian has a unique id and can be tracked from the moment of appearance in the scene until going out of the frame. An occlusion flag is set to partial occlusion if between 25 and 75% of the pedestrian is not visible and to full if $> 75\%$ of the pedestrian is not visible. Crossing intention confidence is a numeric score estimated from human reference data (see Section 3.2).

	PIE	JAAD
# of frames	911K	82K
# of annotated frames	293K	75K
# of pedestrians	1.8K	2.8K
# of pedestrians with behavior annot.	1.8K	686
# of pedestrian bboxes	740K	391K
Avg. pedestrian track length	401	140
Pedestrian intention	yes	no
Ego-vehicle sensor information	yes	no
Scene object annotations	bboxes+text	text

Table 1: Properties of the PIE dataset compared to JAAD.

Spatial annotations are provided for other relevant objects in the scene, including infrastructure (e.g. signs, traffic lights, zebra crossings, road boundaries) and vehicles that interact with pedestrians of interest[1].

Using an on-board diagnostics (OBD) sensor synchronized with the camera we provide GPS coordinates and vehicle information, such as accurate speed and heading angle, for each frame of the video.

Table 1 summarizes the properties of PIE and JAAD datasets. JAAD has bounding box annotations for all pedestrians, which makes it suitable for detection and tracking applications. However, it lacks accurate vehicle information, spatial annotations for traffic objects and pedestrian intentions which are vital for pedestrian action prediction.

3.2. Human experiment

As mentioned in Section 2, research in the field of pedestrian behavior understanding largely focuses on the problem of action and behavior prediction, while the topic of intention estimation remains relatively unaddressed. Partly this is due to the fact that establishing ground truth for crossing intention is infeasible since it would require interviewing people on the street and observing their actions after the vehicle passed by them [33]. However, this data is necessary for identifying and focusing on the most relevant pedestrians on the street, pedestrian behavior understanding and prediction, including trajectory estimation. In order to determine human reference data for samples in the PIE dataset we conducted a human experiment described below.

Experiment description. The experiment involved watching short videos from the PIE dataset. We asked the participants to observe a single pedestrian highlighted in the first few seconds and, after viewing each video *once*, answer the following question: "Does this pedestrian **want** to cross the street?". The options were set on a 5-interval scale (the outer intervals for definite 'yes' or 'no' and 3 intervals expressing varying degrees of uncertainty in between).

[1]We used the CVAT tool (`https://github.com/opencv/cvat`) for all spatial annotations and behavior labels.

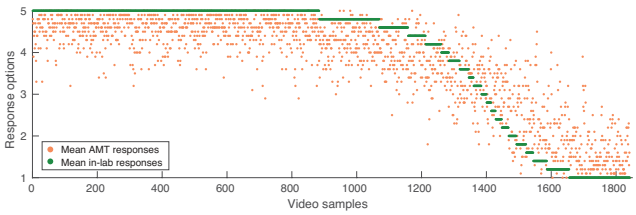

Figure 2. A plot of average responses to the question "Does this pedestrian want to cross?" for each of the 1842 video samples containing a single pedestrian of interest. Answer option 5 is selected for the presence and option 1 for the absence of crossing intention respectively. Answer options in between represent various levels of uncertainty. In-lab and AMT responses are shown as green and red dots respectively. Average responses are sorted in descending order for clarity.

Videos used in the experiment were generated for each of the 1842 labeled pedestrians in the PIE dataset. Using GPS information and vehicle speed we created short clips showing $\approx$ 3s before the vehicle reaches $1.5 - 3$s time-to-event. In cases when ego-vehicle was stationary the video was cropped 3s before the pedestrian began crossing. The first and the last frames of each video clip were frozen for 4s to allow the subjects to get familiar with the scene. The pedestrian of interest was highlighted with a red arrow pointing down for the duration of freeze-frames in the beginning and at the end of the video.

Procedure. We first ran the experiment in a lab setting with 5 subjects (ages $27 - 62$) each of whom viewed the entire set of 1842 videos. We then repeated the same experiment on Amazon Mechanical Turk (AMT) to gather additional 10 answers per video. For the AMT experiment we grouped videos into sets of 10 for each HIT (Human Intelligence Task). We limited our study to participants residing in Canada and the USA to ensure that they are familiar with the rules of the road, signs, road delineation, etc. and reduce any cultural bias. In total, we collected $27,630$ responses from over 700 subjects (ages $19 - 88$).

Results. A plot of aggregated responses from lab and AMT participants is shown in Figure 2. Since ground truth data was not available, we focused on analyzing the agreement among subjects to validate our results. First, we computed intraclass correlation coefficient (ICC), a measure of inter-rater consistency, commonly used to analyze subjective responses from a large population of raters in the absence of ground truth data [35]. Despite an inherent degree of subjectivity of estimating pedestrian intention, the measured ICC[2] is 0.97 and 0.93 for the lab and AMT subjects respectively, which suggests a very high degree of agreement within both groups of raters (ICC $=$ 1 for absolute agreement). The slightly lower agreement among the AMT workers is likely due to the much larger and diverse group of subjects and the presence of factors that we could not control for (e.g. viewing conditions, distractions, etc.).

Despite some noise present in the AMT data, the Pearson correlation coefficient between the average responses of the lab and AMT subjects is 0.90 suggesting that both groups answer similarly. For instance, 14 out of 15 raters agreed on the same answer in nearly 17% of cases. On the other hand, there were only 10 cases in the entire dataset where raters did not reach an agreement with respect to the pedestrian's intention, resulting in an average score of exactly 3 ('Not sure'). The samples in question included pedestrians who were close to the curb or already stepped onto the road but were distracted, e.g. by their phone or by interacting with another person. Bus stops in close proximity to the pedestrian crossings were another source of confusion, making it difficult to distinguish between pedestrians waiting for the bus and those waiting to cross. However, the number of these borderline cases was very low ($\approx 3\%$).

The PIE dataset contains 898 examples of people who intend to but do not cross, 512 pedestrians with the intention to cross who eventually cross in front of the vehicle and 430 pedestrians with no crossing intention. Interestingly, there are only 2 samples where the pedestrian crossed the street but responses from human subjects did not indicate crossing intention. Since this type of false negative is a potential safety concern, it is reassuring that human participants are particularly good at interpreting others' intentions.

4. Method Descriptions

In this work we address the problem of pedestrian behavior prediction on two levels: *Early anticipation* in the form of estimating pedestrians' intention of crossing and *trajectory prediction* as late forecasting of the future trajectory of pedestrians based on observed scene dynamics. The former primarily serves as a refinement procedure to change the focus of an intelligent system to those pedestrians that matter, or potentially will interact with the vehicle. Intention estimation may also benefit trajectory prediction by implying the types of motion patterns that are more probable in the scene. For instance, someone with no intention of crossing will not perform a lateral movement across the street in front of the vehicle.

4.1. Pedestrian Intention Estimation

We represent pedestrian intention for each sample as an average response of human experiment participants, rescaled to range $[0, 1]$ and rounded. Then we define the task as a binary classification problem of predicting whether the pedestrian i has an intention of crossing the street $int_i \in \{0, 1\}$ given a partial observation of local visual context around pedestrian $C_{obs} = \{c_i^{t-m}, c_i^{t-m+1}, ..., c_i^t\}$ and

[2]We use ICC(3,k) and ICC(1,k) for lab and AMT data respectively. The first measure assumes that a fixed number of raters k (in this case $k = 5$ for in-lab participants) rate all targets and the second measure assumes that a subset of k raters ($k = 10$) from a large population rates all targets. Ratings are aggregated across raters in both cases.

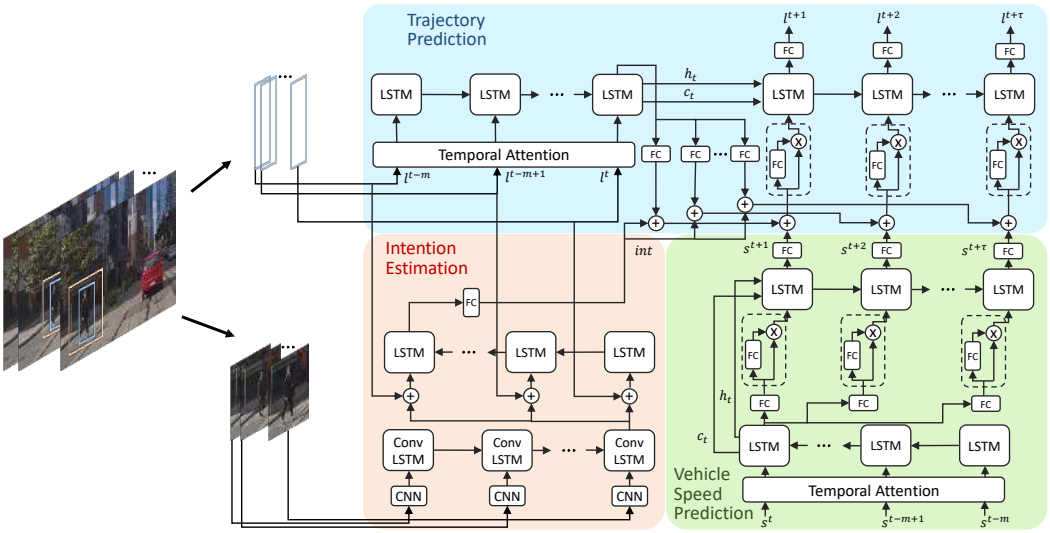

Figure 3. The proposed intention estimation and trajectory prediction framework. The system receives as input a sequence of images and the current speed of the ego-vehicle. The intention estimation model's encoder receives as input a square cropped image around the pedestrians, produces some representation which is concatenated with their observed locations (bounding box coordinates) before feeding them to the decoder. The speed model predicts future speed using an encoder-decoder scheme followed by a series of self-attention units. The location prediction unit receives location information as encoder input and the combination of encoder representations, pedestrian intention and future speed as decoder input, and predicts future trajectory. In the diagram, FC refers to fully-connected layers, $s^{1:m}$. $\bigoplus$ to concatenation operation and $s^{1:m}$. $\bigotimes$ to element-wise multiplication. Location, intention and speed are denoted by l, int and s respectively.

trajectory $L_{obs} = \{l_i^{t-m}, l_i^{t-m+1}, ..., l_i^t\}$, where l is a 2D bounding box around the pedestrian defined by top-left and bottom-right points ($[(x_1, y_1), (x_2, y_2)]$).

It has been shown that pose, implicitly encoded in the appearance (e.g. whether the person is leaning forward or turned towards the road), immediate local surroundings (e.g. location relative to the curb) and motion, convey vital information about the intention to cross. Other context elements, such as street signs, traffic signals as well as the behavior of the ego-vehicle, may influence pedestrian's actions, e.g. whether they will attempt to cross, but will not have an effect on their initial intention to cross the street.

For the task of the intention estimation we employ an RNN encoder-decoder architecture (see Figure 3), where encoder receives a sequence of feature representations corresponding to the image areas around the detected pedestrian. The output of the encoder is then concatenated with the sequence of bounding box coordinates which capture pedestrian dynamics. We use a binary cross-entropy loss function for training.

4.2. Pedestrian Trajectory Prediction

We address the problem of future trajectory prediction as an optimization process in which the objective is to learn the distribution $p(L_{pred}|L_{obs}, S_{pred}, Int_i)$ for multiple pedestrians $1 \leq i \leq n$, where $L_{pred} = \{l_i^{t+1}, l_i^{t+2}, ..., l_i^{t+\tau}\}$ are the predicted trajectories of pedestrians, $L_{obs} = \{l_i^{t-m}, l_i^{t-m+1}, ..., l_i^t\}$ are the observed locations of pedestrians, $S_{pred} = \{s^{t+1}, s^{t+2}, ..., s^{t+\tau}\}$ refers to predicted future speed of the ego-vehicle, and Int_i is the crossing intention of pedestrian i estimated by the intention estimation stream. The locations, l are 2D bounding boxes around pedestrians defined by top-left and bottom-right corner points $[(x_1, y_1), (x_2, y_2)]$

As depicted in Figure 3, the proposed model is based on an RNN encoder-decoder architecture where the inputs to the encoder are the observed locations of pedestrians for some time t and the output of the decoder is the future trajectory prediction up to time $t + \tau$. We use two types of attention: a *temporal attention* module applied to the encoder inputs and a *self-attention* unit applied to the decoder inputs. The former focuses on finding the most relevant information (key frames) in the observed sequence, whereas the latter is applied at feature-level and focuses on the parts of the encoding representation that are relevant to current prediction. The self-attention units are preceded by embedding units for dimensionality reduction of encodings. The final predictions are generated by a linear transformation of the decoder's output.

The vehicle speed estimation stream follows a similar scheme, except it learns $p(S_{pred}|S_{obs})$, where S_{obs} refers to observed speed of the vehicle up to time t. At training time, both sequence prediction models use a mean squared error loss function defined as $MSE = \frac{1}{N}\sum_{j=1}^{\tau} ||loc_i^{t+j} - \hat{loc}_i^{t+j}||$.

5. Empirical evaluation

5.1. Implementation

Intention Estimation. We use Convolutional LSTM with 64 filters and kernel size of 2×2 with stride 1 as encoder and for decoder an LSTM with 128 hidden units, $tanh$ activation, dropout of 0.4 and recurrent dropout of 0.2. VGG16 [36] (without fc layers) pretrained on ImageNet [32] is used to encode image features. We experiment with two different types of visual information. The first is img_{bbox} which is input image cropped to the size of the bounding box, resized so that the larger dimension matches the VGG input size of 224×224 and padded with zeros to preserve the aspect ratio. The second type of input is local context around the pedestrian ($img_{context}$) which is input image cropped to $2\times$ the size of the bounding box, squarified and resized to 224×224.

Trajectory Prediction. We use LSTMs with 256 hidden units and $softsign$ activation in our trajectory and speed prediction streams. Compared to $tanh$ activation, we observed faster training and performance improvement of up to 5% when using $softsign$ activation. The embedding layer in the trajectory prediction stream is a fully connected network with 64 output nodes and no dropout.

5.2. Datasets

Pedestrian Intention Estimation (PIE). There are 1842 pedestrian samples divided into train, test and validation sets with the ratios of 50%, 40% and 10% respectively. We sample the tracks with an overlap ratio of 0.5. For trajectory prediction training, the tracks below the minimum length of 2 seconds (observation + prediction) are discarded. We use the OBD sensor readings for speed information.

JAAD [27]. For trajectory prediction evaluation using only bounding boxes we use pedestrian tracks from the JAAD dataset. Given the smaller number of samples and shorter tracks in this dataset, we use all pedestrian samples with overlap ratio of 0.8. We use the same train/test split as in [28], excluding the low-resolution and low-visibility videos (13 out of 346) from the evaluation.

Training. Models are trained separately and combined at the test time. Intention and trajectory models are trained using RMSProp [38] optimizer with learning rate of 10^{-5} and 10^{-2} respectively. The intention model was trained for 300 epochs using a batch size of 128 with $L2$ regularization of 0.001. We trained the trajectory model for 60 epochs using a batch size of 64 with $L2$ regularization of 0.0001.

Metrics. For intention estimation we report $accuracy$ and $F1$-score defined as $2 \cdot precision \cdot recall / (precision + recall)$. The following metrics are used for evaluation of the proposed trajectory prediction algorithm: MSE over bounding box coordinates [5], C_{MSE} and CF_{MSE} which are the MSEs of the center of the bounding boxes averaged over the entire predicted sequence and only the last time step ($t + \tau$) respectively. All results of the bounding box predictions are in pixels.

Method	Input data	acc	$F1$
LSTM	loc	0.63	0.73
LSTM$_{ed}$	loc	0.67	0.76
	img_{bbox}	0.60	0.78
PIE$_{int}$	img_{bbox}	0.69	0.79
	$img_{context}$	0.71	0.82
	$img_{bbox} + loc$	0.73	0.82
	$img_{context} + loc$	**0.79**	**0.87**

Table 2: Pedestrian intention estimation results for various combinations of input data: loc - bounding box coordinates, img_{bbox} - image cropped to the size of bounding box, and $img_{context}$ - image cropped to $2\times$ size of the bounding box to show local context.

Pedestrian intention estimation. Table 2 summarizes the results of various models trained on different combinations of input data over 0.5s of observation. The following models are used in the evaluation: a vanilla LSTM trained on normalized bounding box coordinates (loc) as a baseline, an LSTM encoder-decoder (LSTM$_{ed}$) trained on normalized bounding box coordinates or img_{bbox} and the proposed model PIE$_{int}$ trained on 4 different types of input data, img_{bbox}, $img_{context}$, $img_{bbox} + loc$ and $img_{context} + loc$.

The baseline LSTM achieves 63% accuracy. In comparison, LSTM encoder-decoder (LSTM$_{ed}$), performs better using the same information, however, it does worse using only img_{bbox} even though it has a higher $F1$-score. This can be due to the fact that pedestrian appearance in the absence of dynamics is not informative enough.

PIE$_{int}$ overall performs better than the other two models on all input types. Its performance on appearance features (img_{bbox}) and motion data (loc) is approx. 4% above the baseline performance. Adding local context ($img_{context}$) offers a small performance improvement. This suggests that, despite using different representations, motion or appearance features on their own may not be effective in estimating intention. As expected, combining different sources of information results in improved performance. We see that motion improves intention estimation on samples that are relatively far away or occluded, where visual information is unreliable. However, in situations where the pedestrian was more visible, their pose and context elements were also very important. Overall, the combination of appearance, local context and motion offer the most advantage boosting the final accuracy to 79%. Figure 4 shows some examples of the proposed algorithm's performance.

Trajectory prediction. We begin by evaluating the proposed model using only location (bounding box) information. For this purpose we report the results on the following models: two baseline models, a linear Kalman fil-

Method	*PIE*					*JAAD*				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	1s	1.5s	1.5s	1.5s	0.5s	1s	1.5s	1.5s	1.5s
Linear	123	477	1365	950	3983	223	857	2303	1565	6111
LSTM	172	330	911	837	3352	289	569	1558	1473	5766
B-LSTM[5]	101	296	855	811	3259	159	539	1535	1447	5615
PIE$_{traj}$	**58**	**200**	**636**	**596**	**2477**	**110**	**399**	**1248**	**1183**	**4780**

Table 3: Location (bounding box) prediction errors over varying future time steps. MSE in pixels is calculated over all predicted time steps, C_{MSE} and CF_{MSE} are the MSEs calculated over the center of the bounding boxes for the entire predicted sequence and only the last time step respectively.

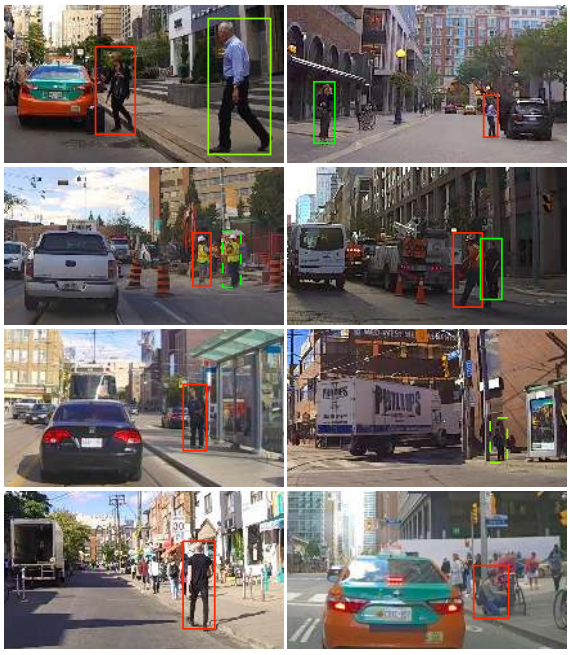

Figure 4. Results of pedestrian intention estimation overlaid on top of frames from the PIE dataset (cropped for better visibility). Bounding boxes are colored depending on the presence (green) or absence (red) of crossing intention as detected by our model. Dashed bounding boxes represent incorrectly estimated intention.

ter [13] and a vanilla LSTM model, state-of-the-art algorithm, Bayesian LSTM [5] (B-LSTM), and the proposed model PIE$_{traj}$. Each model is trained and tested on 0.5s (15 frames) observation, and predicts trajectories over 0.5, 1 and 1.5 seconds in future.

Table 3 summarizes the results of the predictions using only bounding box information. As shown in the table, the proposed method achieves state-of-the-art performance on all metrics, by up to 26% on the PIE dataset and 18% on JAAD compared to B-LSTM. The performance of all models is generally poorer on the JAAD dataset which can be partially attributed to the smaller number of samples, scales and shorter tracks all of which reduce the diversity of the dataset. The deterioration of linear model performance for long-term predictions indicates the complexity of human motion patterns that cannot be explained with simple linear interpolation. As expected, the performance of all models is generally better on bounding box centers due to the fewer degrees of freedom.

Method	MSE			
	0.5s	1s	1.5s	last
Linear	0.87	2.28	4.27	10.76
LSTM	1.50	1.91	3.00	6.89
PIE$_{speed}$	**0.63**	**1.44**	**2.65**	**6.77**

Table 4: Speed prediction errors over varying time steps on the PIE dataset. *Last* stands for the last time step. The results are reported in km/h.

Context in trajectory prediction. We first evaluate the proposed speed prediction stream, PIE$_{speed}$, by comparing this model with two baseline models, a linear Kalman filter and a vanilla LSTM model. We use MSE metric and report the results in km/h. Table 4 shows the results of our experiments. The linear model achieves reasonable performance in short-term which is better than the vanilla LSTM over 0.5s. This indicates that the speed variation often is insignificant in short-term, especially in urban environments which is the case in the proposed PIE dataset. In long-term, however, LSTM-based models perform significantly better. The proposed PIE$_{speed}$ achieves the best performance by up to 10% over vanilla LSTM model.

Earlier we argued that pedestrian intention can serve as an early prediction stage in addition to trajectory prediction. Here, we examine whether estimating pedestrians' intention of crossing can improve trajectory prediction. We report the results on our trajectory prediction model PIE$_{traj}$ which receives as input the context information provided by PIE$_{speed}$ and PIE$_{int}$. We report the results on 0.5s observation and 1.5s prediction.

As shown in Table 5, conditioning trajectory prediction on pedestrian intentions can improve the results by up to 4%. This is due to the fact that intention may imply certain patterns of motion. For instance, someone with the intention of crossing might have a lateral movement across the street whereas someone without intention might stand still. As one would expect, the ego-vehicle's speed improves the trajectory prediction, and when combined with pedestrian intention, the best results are achieved with more than 11% improvement over baseline using only bounding boxes.

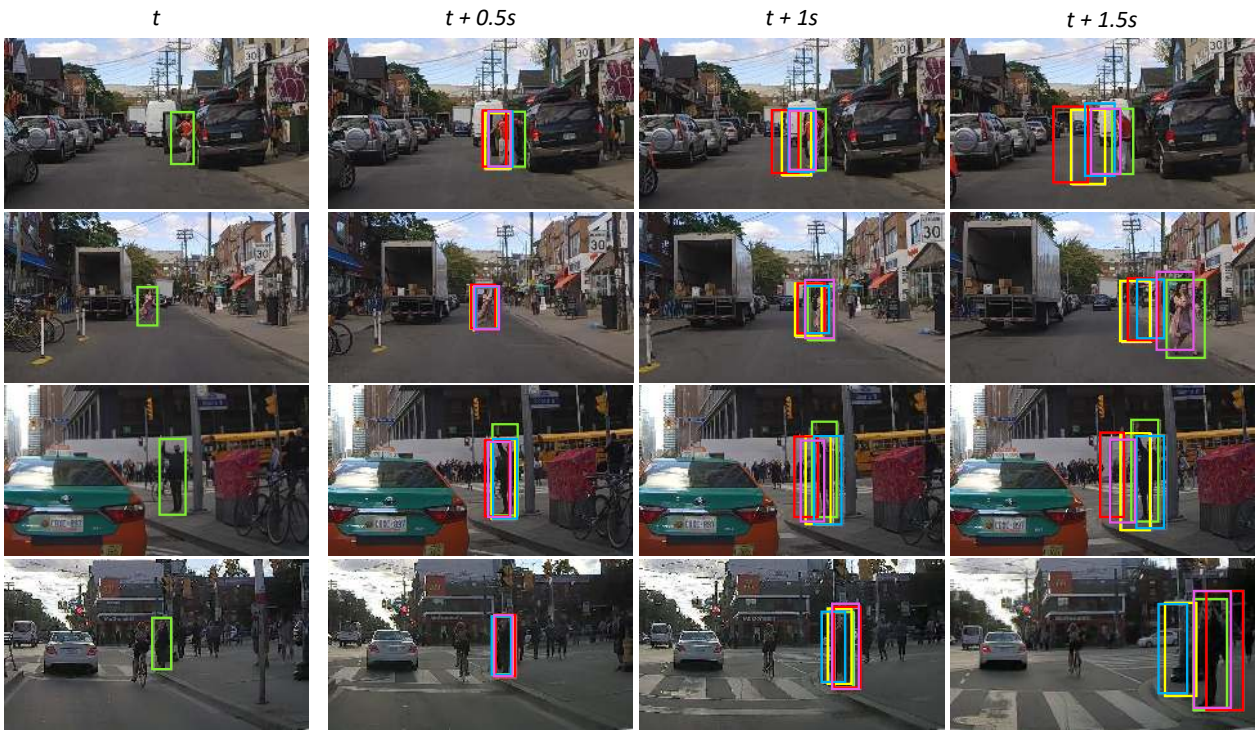

Figure 5. Examples of trajectory prediction algorithm using the proposed model PIE$_{traj}$ with different input combinations. The color and model combinations are: loc (yellow), loc+PIE$_{int}$ (blue), loc+PIE$_{speed}$ (red), and loc+PIE$_{int}$+PIE$_{speed}$ (purple). Ground truth annotations are shown in green. The sequences depict different traffic scenarios. From top to bottom: A man leaving his vehicle, a woman crossing the street, a man hailing a taxi, and a woman waiting to cross.

Method	Input	MSE	C_{MSE}	CF_{MSE}
PIE$_{traj}$	loc	636	596	2477
	loc+PIE$_{int}$	611	570	2414
	loc+PIE$_{speed}$	572	535	2204
	loc+PIE$_{int}$+PIE$_{speed}$	**559**	**520**	**2162**
	$loc + int + speed$	473	435	1741

Table 5: Location (bounding box) prediction errors of the proposed model PIE$_{traj}$ on 0.5s observation and 1.5s prediction using different inputs. loc, int and $speed$ stand for location, intention and vehicle speed. PIE$_{int}$ and PIE$_{speed}$ are the outputs of the intention and vehicle speed estimation models. MSE is reported in pixels and calculated over all predicted time steps. C_{MSE} and CF_{MSE} are the MSEs over the center of the bounding boxes for the entire predicted sequence and only the last time step respectively.

Figure 5 illustrates the performance of our proposed algorithm using different contextual information on the PIE dataset. Even though speed has a dominant effect in improving trajectory prediction it may also fail in certain cases, when the vehicle is stationary or when the pedestrian has no intention of crossing.

6. Conclusion

We presented a novel large-scale dataset for studying pedestrian crossing intention and behavior with extensive multimodal annotations for visual reasoning tasks. Since there is no ground truth data for crossing intention, we conducted a large-scale experiment to determine human reference data for this task. Our data shows that a large number of human experiment subjects have a high degree of agreement in their answers.

We proposed a baseline model for pedestrian intention estimation and by evaluating various input data combinations we showed that local context in conjunction with pedestrian motion are good predictors for crossing intention. In addition, we proposed a trajectory prediction for an onboard camera. Our model outperforms the state-of-the-art by a significant margin. We show that conditioning the trajectory prediction on pedestrian intention and ego-vehicle speed further improves the results.

In future work pedestrian intention estimation can be further improved by including explicit pose and social interactions. Likewise, trajectory estimation can benefit from other sources of information, such as traffic dynamics, signals and road structure, all of which affect future pedestrian actions.

Acknowledgements. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the NSERC Canadian Robotics Network (NCRN), the Air Force Office for Scientific Research (USA), and the Canada Research Chairs Program through grants to JKT.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016.

[2] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online POMDP planning for autonomous driving in a crowd. In *ICRA*, pages 454–460, 2015.

[3] Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. Intention-aware motion planning. In *Algorithmic foundations of robotics X*, pages 475–491. Springer, 2013.

[4] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464, 2011.

[5] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *CVPR*, pages 4194–4202, 2018.

[6] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. ContextVP: Fully context-aware video prediction. In *ECCV*, pages 781–797, 2018.

[7] Lei Chen, Jiwen Lu, Zhanjie Song, and Jie Zhou. Part-activated deep reinforcement learning for action prediction. In *ECCV*, pages 421–436, 2018.

[8] Nachiket Deo and Mohan M. Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *CVPRW*, pages 1581–1589, 2018.

[9] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *CVPR*, pages 304–311, 2009.

[10] Zhijie Fang and Antonio M López. Is the pedestrian going to cross? answering by 2D pose estimation. In *Intelligent Vehicles Symposium (IV)*, pages 1271–1276, 2018.

[11] Zhijie Fang, David Vázquez, and Antonio López. On-board detection of pedestrian intentions. *Sensors*, 17(10):2193, 2017.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[13] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[14] Christoph G Keller, Christoph Hermes, and Dariu M Gavrila. Will the pedestrian cross? Probabilistic path prediction based on learned motion features. In *Joint Pattern Recognition Symposium*, pages 386–395, 2011.

[15] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *CVPR*, pages 1473–1481, 2017.

[16] Julian Francisco Pieter Kooij, Nicolas Schneider, Fabian Flohr, and Dariu M Gavrila. Context-based pedestrian path prediction. In *ECCV*, pages 618–633, 2014.

[17] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. DESIRE: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017.

[18] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Computer graphics forum*, 26(3):655–664, 2007.

[19] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion gan for future-flow embedded video prediction. In *ICCV*, pages 1744–1752, 2017.

[20] Chaochao Lu, Michael Hirsch, and Bernhard Scholkopf. Flexible spatio-temporal networks for video prediction. In *CVPR*, pages 6523–6531, 2017.

[21] Tahmida Mahmud, Mahmudul Hasan, and Amit K. Roy-Chowdhury. Joint prediction of activity labels and starting times in untrimmed videos. In *ICCV*, pages 5773–5782, 2017.

[22] Barbara Majecka. Statistical models of pedestrian behaviour in the forum. Master's thesis, School of Informatics, University of Edinburgh, 2009.

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

[24] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *ECCV*, pages 5773–5782, 2018.

[25] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009.

[26] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *Intelligent Vehicles Symposium (IV)*, pages 264–269, 2017.

[27] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In *ICCVW*, pages 206–213, 2017.

[28] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. It's not all about size: On the role of data properties in pedestrian detection. In *ECCVW*, pages 210–225, 2018.

[29] Eike Rehder and Horst Kloeden. Goal-directed pedestrian prediction. In *ICCVW*, pages 50–58, 2015.

[30] Eike Rehder, Florian Wirth, Martin Lauer, and Christoph Stiller. Pedestrian prediction by planning using deep neural networks. In *ICRA*, pages 1–5, 2018.

[31] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, pages 549–565, 2016.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[33] Friederike Schneemann and Patrick Heinemann. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *IROS*, pages 2243–2248, 2016.

[34] Andreas Th Schulz and Rainer Stiefelhagen. Pedestrian intention recognition using latent-dynamic conditional random fields. In *Intelligent Vehicles Symposium (IV)*, pages 622–627, 2015.

[35] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

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

[37] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *CVPR*, pages 3521–3529, 2018.

[38] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp, COURSERA: Neural networks for machine learning. *University of Toronto, Technical Report*, 2012.

[39] Benjamin Völz, Karsten Behrendt, Holger Mielenz, Igor Gilitschenski, Roland Siegwart, and Juan Nieto. A data-driven approach for pedestrian intention estimation. In *Intelligent Transportation Systems Conference (ITSC)*, pages 2607–2612, 2016.

[40] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851, 2016.

[41] Hao Xue, Du Q Huynh, and Mark Reynolds. SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In *WACV*, pages 1186–1194, 2018.

[42] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR*, pages 1345–1352, 2011.

[43] YoungJoon Yoo, Kimin Yun, Sangdoo Yun, JongHee Hong, Hawook Jeong, and Jin Young Choi. Visual path prediction in complex scenes with crowded moving objects. In *CVPR*, pages 2668–2677, 2016.

[44] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A diverse dataset for pedestrian detection. In *CVPR*, pages 3213–3221, 2017.

[45] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *CVPR*, pages 2871–2878, 2012.