

---

# PEDESTRIAN INTENTION PREDICTION: A MULTI-TASK PERSPECTIVE

---

**Smail Ait Bouhsain\***

EPFL, VITA

smail.aitbouhsain@alumni.epfl.ch

**Saeed Saadatnejad\***

EPFL, VITA

saeed.saadatnejad@epfl.ch

**Alexandre Alahi**

EPFL, VITA

alexandre.alahi@epfl.ch

## ABSTRACT

In order to be globally deployed, autonomous cars must guarantee the safety of pedestrians. This is the reason why forecasting pedestrians' intentions sufficiently in advance is one of the most critical and challenging tasks for autonomous vehicles. This work tries to solve this problem by jointly predicting the intention and visual states of pedestrians. In terms of visual states, whereas previous work focused on x-y coordinates, we will also predict the size and indeed the whole bounding box of the pedestrian. The method is a recurrent neural network in a multi-task learning approach. It has one head that predicts the intention of the pedestrian for each one of its future position and another one predicting the visual states of the pedestrian. Experiments on the JAAD dataset show the superiority of the performance of our method compared to previous works for intention prediction. Also, although its simple architecture (more than 2 times faster), the performance of the bounding box prediction is comparable to the ones yielded by much more complex architectures. Our code is available online<sup>2</sup>.

**Keywords** Pedestrian Intention Prediction · Bounding Box Prediction · Multi-task Learning · Autonomous Cars

## 1 Introduction

Future prediction is usually considered as an essential part of intelligence [1]. It becomes more critical in autonomous cars as it can avoid accidents with humans. For example, consider a situation when a pedestrian is next to a street and is going to cross. A non-predictor agent might recognize the pedestrian just when he/she is in front of it and then tries to avoid the collision. However, when a predictor agent looks at the same scene, it goes beyond the pedestrian detection and predicts what happens in the next few seconds. Therefore, it finds out the intention of that person and then, based on that, decides when to stop and when to pass. In the application of autonomous cars, there are different levels of pedestrian prediction, such as its intention and visual states.

**Pedestrian intention prediction.** To guarantee the safety of pedestrians, a self-driving car should not only predict their intention but also predict it sufficiently in advance in order to react accordingly. In most papers of this category [2, 3, 4, 5], either they do intention estimation, or the prediction is performed only for a short horizon as they are considering the current intention. However, we tackle this problem by providing a sequence of predictions of the intention for the next few frames. This leads to a more extended and more accurate prediction.

**Visual states prediction.** There are different levels of state prediction. The problem of trajectory prediction is defined as forecasting a sequence of future positions (x-y coordinates) of a pedestrian given a set of observed past positions. It is not a trivial task for autonomous systems because of the different kinds of paths that humans choose with non-linear speed variation. Numerous methods have been proposed. Some are model-based [6, 7] which are scenario-specific and

---

\*Equal contribution

<sup>2</sup><https://github.com/vita-epfl/bounding-box-prediction>

have low performance in approximating complex functions such as pedestrian trajectory. Other methods are data-driven, using Long Short Term Memory (LSTM) [8, 9] or convolutional neural networks (CNN) [10], or both [11, 12, 13]. These models usually have higher performance in discovering complex patterns, and we follow this approach. Moreover, some works go beyond the trajectory and predict not only future locations (x,y) but also the width and height (w,h). We refer to this task as the bounding box prediction [14, 15]. The accuracy in this task is lower than the trajectory prediction task as this is a harder task. We argue that the ability of visual states prediction is useful in better understanding of pedestrian intention.

In this paper, we propose a data-driven method based on a multi-task learning [16] approach to do efficiently pedestrian intention prediction and pedestrian bounding box prediction simultaneously. We will show how those learned representations share some useful information to enhance the performance of each other. For instance, if we know that a person is going to cross the road, it will be easier to predict its trajectory and vice versa. The network is carefully designed such that it has much fewer parameters while maintaining a higher or comparable accuracy over the state-of-the-art.

## 2 Related Work

Most previous works in pedestrian prediction [8, 11, 12] considered pedestrians seen from a bird view. However, in this paper, we predict positions seen from a car camera view. This allows for a more complex representation of pedestrians who can be represented by bounding boxes. Indeed, bounding box center, width, and height can provide information on the position, the orientation of the pedestrian, and its distance from the car. This is valuable for a number of applications and can improve the prediction of pedestrians' intentions as well. This perspective, however, adds a challenge of car movement, making the position of the bounding box vary even if the pedestrian stops.

### 2.1 Pedestrian Intention Prediction

There are model-based approaches using social behavior analysis [2], social force modeling [6] and conditional random fields [3] for pedestrian dynamic prediction. These models are scenario-specific and have low performance in approximating complex functions.

SKLT [4] is taking advantage of extracted skeleton features of pedestrians as input to a Random Forest. [4] proposed a method (CNN(fc6)) using features extracted by a Faster R-CNN, as well as a combination of the two methods (CNN(fc6)+SKLT). ConvNet-Softmax [17] used a simple CNN with softmax activation layer for classification. [18] proposed the same method with an SVM classifier instead of the softmax layer (ConvNet-SVM). [19] replaced the SVM with an LSTM (ConvNet-LSTM) and ST-Dense-Net [20] combined the spatial and temporal features of the observed sequence. In this paper, we show that our method performs better than all these models.

There are also some other works that we could not compare with because of the different setups [21, 5]. [21] used dynamic fuzzy automata and low-level features with boosted random forests. [5] proposed a variant of a variational recurrent neural network (VRNN), incorporating latent variable corresponding to a dynamic state space model. A subtle calculation shows the superiority of our performance over them.

### 2.2 Pedestrian Trajectory Prediction

In trajectory prediction, the input sequence is the set of past observed positions, and the output sequence is a set of predicted positions in the future. Some previous works [8, 11, 9] take advantage of the social neighborhood in order to improve the accuracy of the predictions, while others [11, 12, 13] use the scene features and layout as a predictor as well. While most of these methods use a recurrent neural network, [10] proposed a CNN based human trajectory prediction approach. The go-to method [8] proposed a sequence to sequence LSTM model that combines the observed past positions and the social interactions and conventions involved in humans motion in order to predict their trajectory in crowded scenes.

### 2.3 Pedestrian Bounding Box Prediction

There is a couple of new research performing bounding box prediction. [15] introduces a method (STED) for multiple object bounding box forecasting. It is also a sequence to sequence architecture, combining both observed bounding box coordinates and optical flows in order to predict the future bounding boxes. One disadvantage this network has is the need for optical flows, which are extracted using another neural network called Flownet2 [22]. [15] used two baselines: a dummy Constant-Velocity-Constant-Scale (CV-CS) model always predicting the same observed velocity and a Linear Kalman Filter (LKF) [23] applied to the bounding box prediction problem. It also proposed adaptations of the methods

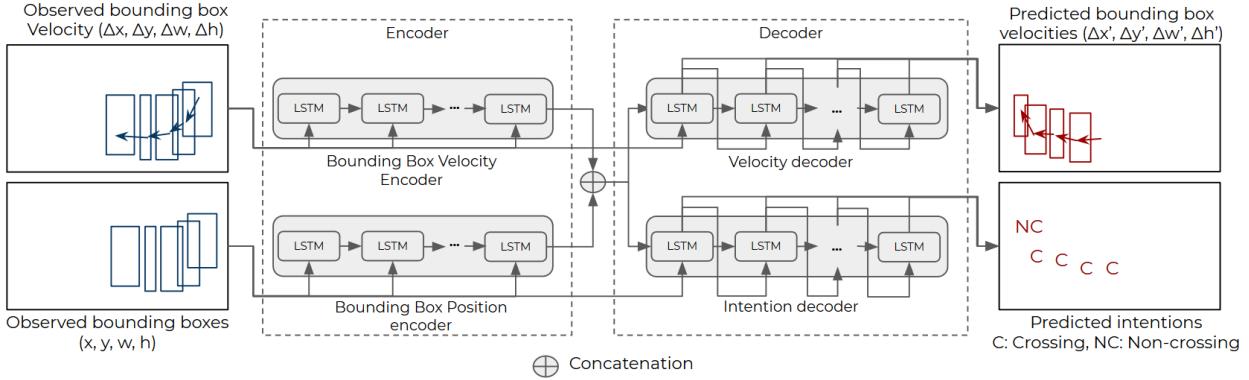


Figure 1: Our proposed multi-task learning network. It is an LSTM based encoder-decoder architecture that leverages the position and dimensions of the observed bounding box as well as their velocity in order to predict the future bounding boxes of pedestrians as well as a sequence of future intentions.

proposed in [24] (DTP-MOF), which used optical flow as input to a CNN for bounding box prediction, and in [25] (FLP-MOF) which combined ego-motion estimation (FlowNet2) and pedestrian pose (OpenPose) as input.

In [14], they took advantage of odometry data for predicting the bounding box of pedestrians under uncertainty. On the other hand, [26] used optical flow and future ego-motion of the viewer car for predicting future car bounding boxes. In both of those works, information about the motion of the car is needed, which is not always available.

### 3 Method

The proposed method (**PV-LSTM**) is a multi-task sequence to sequence LSTM model. It takes as input the velocities and coordinates of observed past bounding boxes and outputs the predicted velocities of the future bounding boxes of the pedestrian, from which the future bounding box coordinates can be computed, as well as the pedestrian's state (crossing/not crossing) in each predicted bounding boxes. A sketch of PV-LSTM is shown in Figure 1, which stands for Position-Velocity-LSTM, as it encodes the position and the velocity of the pedestrian bounding box.

#### 3.1 Problem formulation

Given as sequence  $(B_{t-T_{obs}}, \dots, B_t)$  of bounding boxes of a pedestrian corresponding to the timesteps or frames  $(t-T_{obs}, \dots, t)$  in a video sequence, the task of bounding box prediction is to output the next sequence of bounding boxes  $(B_{t+1}, \dots, B_{t+T_{pred}})$  at the following timesteps  $(t+1, \dots, t+T_{pred})$ , as well as the sequence of the next intentions of the pedestrian  $(I_{t+1}, \dots, I_{t+T_{pred}})$ .

The position of the bounding box of a pedestrian at timestep  $t$  is represented by the coordinates of its center, its width and height  $B_t = (x_t, y_t, W_t, H_t)$ . The intention  $I_t$  is a binary value representing the state of the pedestrian, either crossing or non-crossing, at each predicted timestep.

#### 3.2 Bounding box velocity encoder

Both the position and the velocity of the pedestrian are taken into account to capture the high non-linearity in humans motion. Also, in order to tackle the car movement issue, the relative velocity of the pedestrian from the car perspective is considered. In this paper we use LSTMs [27] given its power in dealing with time sequence data and long term dependencies have been proven.

The bounding box velocity LSTM encodes the velocity of the observed bounding boxes. The latter is represented as the variation between the center positions of successive pairs of bounding boxes and height and width variations  $v_{B_t} = (\Delta x_t, \Delta y_t, \Delta W_t, \Delta H_t) = (x_t - x_{t-1}, y_t - y_{t-1}, W_t - W_{t-1}, H_t - H_{t-1})$ . Given these observed velocities as input, the velocity encoder computes the hidden state of velocity  $v_t$  using:

$$v_t = LSTM_v^{enc}(v_{t-1}, v_{B_t}, \mathbf{W}_v), \quad (1)$$

where  $LSTM_v^{enc}$  is the velocity encoder LSTM, and  $\mathbf{W}_v$  is its weight matrix.

### 3.3 Bounding box position encoder

The position LSTM encodes the positions and dimensions of the bounding boxes throughout the input sequence. It takes as input the observed past bounding boxes coordinates ( $B_{t-T_{obs}}, \dots, B_t$ ), and outputs the updated position hidden state at time  $t$  via:

$$p_t = LSTM_p^{enc}(p_{t-1}, B_t, \mathbf{W}_p), \quad (2)$$

where  $LSTM_p^{enc}$  is the position encoder LSTM, and  $\mathbf{W}_p$  is its weight matrix.

The position encoder is used to extract features and identify movement patterns from the input sequence of positions. These features help with the bounding box prediction but are vital components for intention prediction since the position and orientation of the pedestrian in the scene determine its state and future intentions.

### 3.4 Velocity decoder

First, the position hidden state  $p_t$  and velocity hidden state  $v_t$  are concatenated resulting in one hidden state  $h_t$  grouping all the features:

$$h_t = p_t \oplus v_t$$

To predict the next sequence bounding box velocities of a pedestrian,  $h_t$  is given as an initial hidden state to the velocity decoder  $LSTM_v^{dec}$  which takes as input the last observed velocity  $v_{B_t}$  and outputs the next predicted velocity of the bounding box  $\hat{v}_{B_{t+1}} = (\hat{\Delta}x_{t+1}, \hat{\Delta}y_{t+1}, \hat{\Delta}W_{t+1}, \hat{\Delta}H_{t+1})$ . The equations for the first prediction is computed using:

$$\hat{h}_{t+1}^v = LSTM_v^{dec}(h_t, v_{B_t}, \mathbf{W}_{dv}), \quad (3)$$

Then, this predicted hidden state is fed to a fully connected layer to get the output velocity via:

$$\hat{v}_{t+1} = \mathbf{W}_{ov}\hat{h}_{t+1}^v + b_{ov}, \quad (4)$$

where  $\mathbf{W}_{dv}$  is the weight matrix of the velocity decoder LSTM,  $\mathbf{W}_{ov}$  is the weight matrix of the output layer and  $b_{ov}$  its bias vector.

The next velocities are then computed iteratively while updating the hidden state, giving each time the last predicted one as input to the decoder. The hidden state is updated at time  $t + t'$  using:

$$\hat{h}_{t+t'}^v = LSTM_s^{dec}(\hat{h}_{t+t'-1}^v, \hat{v}_{B_{t+t'-1}}, \mathbf{W}_{ds}), \quad (5)$$

Then the output velocities are computed via:

$$\hat{v}_{B_{t+t'}} = \mathbf{W}_{ov}\hat{h}_{t+t'}^v + b_{ov}, \quad (6)$$

These velocities are used to compute the future position and dimension of the pedestrian's bounding box, by using cumulative addition as described in the equation below:

$$\hat{B}_{t+t'} = \hat{B}_{t+t'-1} + \hat{v}_{B_{t+t'}}, \quad (7)$$

### 3.5 Intention decoder

Similar to velocity decoder, this LSTM takes as initial hidden state the result of the concatenation of the encoders hidden states  $h_t$ . However, the first intention prediction uses as input the last observed position of the bounding box  $B_t = (x_t, y_t, w_t, h_t)$ , and outputs the next predicted state of the pedestrian  $I_{t+1}$  via:

$$\hat{h}_{t+1}^I = LSTM_I^{dec}(h_t, B_t, \mathbf{W}_{di}), \quad (8)$$

$$\hat{I}_{t+1} = \mathbf{W}_{oi}\hat{h}_{t+1}^I + b_{oi}, \quad (9)$$

where  $LSTM_I^{dec}$  is the intention decoder LSTM,  $\mathbf{W}_{di}$  is its weight matrix,  $\mathbf{W}_{oi}$  is the weight matrix of the output layer and  $b_{oi}$  its bias vector.

The next intentions at the next timesteps are then predicted iteratively while updating the hidden state. The last predicted state is fed to an embedding fully connected layer before being given as input to the decoder. The hidden state is updated at time  $t + t'$  using:

$$\hat{E}_{t+t'-1} = \mathbf{W}_{embedding}\hat{I}_{t+t'-1} + b_{embedding}, \quad (10)$$

$$\hat{h}_{t+t'}^I = LSTM_s^{dec}(\hat{h}_{t+t'-1}^I, \hat{E}_{t+t'-1}, \mathbf{W}_{di}), \quad (11)$$

Then the output intentions are computed via:

$$\hat{I}_{t+t'} = \mathbf{W}_{oi} \hat{h}_{t+t'}^T + b_{oi}. \quad (12)$$

where  $\mathbf{W}_{embedding}$  is the weight matrix of the embedding layer and  $b_{embedding}$  is the bias vector.

The output intentions are then fed to a softmax activation layer in order to compute the probability of each possible outcome.

### 3.6 Implementation details

The dimension of the hidden states of LSTMs is 256. Our network is implemented in Pytorch and trained using the mean square error loss for bounding box prediction and the binary cross-entropy loss for intention prediction. We leverage the Adam optimizer starting at a learning rate of 0.0001, updated by an adaptive scheduler. All the models are trained for 100 epochs on one NVIDIA GTX-1080-Ti GPU.

## 4 Experiments

### 4.1 Datasets

The Joint Attention in Autonomous Driving (JAAD) dataset [28] consists of 346 high-resolution videos ( $1920 \times 1080$ ) taken from a car camera perspective. These video clips show various typical driving scenarios in urban areas. The dataset has bounding box and action annotations for all pedestrians in the video sequences. These videos are split into shorter ones of fixed length for each pedestrian and then divided into observed data and future data. Sequences generated from the first 300 videos (20000 sequences) are our training set, and the ones from the last 46 (8000 sequences) are our testing set. This ensures that testing and training scenes do not have any intersection.

We also use the Citywalks dataset [15] in order to compare our performance in pedestrian bounding box prediction with theirs. It is not made for autonomous driving purposes. We use it mainly for comparison purposes. It contains first-person perspective videos showing pedestrians in various public areas. The dataset has been annotated using Mask-RCNN network for bounding box tracking.

### 4.2 Baselines and Evaluation Metrics

We compare the performance of our model to the following baselines as well as other published works used for pedestrian bounding box and intention prediction:

- **Pedestrian intention prediction** We use the results reported in [20] in order to evaluate the performance of our model for pedestrian intention prediction. We also compare our performance with [4].
- We add a baseline (**Trajectory-LSTM**) in which the bounding box prediction  $(x, y, w, h)$  is replaced by trajectory prediction  $(x, y)$  in order to verify the effectiveness of our approach.
- **Pedestrian bounding box prediction:**
  - **Evaluation on JAAD:** For comparison with the following baselines, the models take as input 18 observations (0.6 seconds) and outputs 18 predictions (0.6 seconds) :
    - \* **Position-LSTM:** This model predicts the position of future bounding boxes directly given the observed past ones.
    - \* **Velocity-LSTM:** This network takes as input the velocities of bounding boxes only and outputs the predicted future velocities. This is similar to the ablation of the position encoder and the intention decoder of our model.
    - \* **Scene-PV-LSTM:** It adds a scene features encoder to our proposed model, by combining Resnet-50, another CNN for dimensionality reduction and an LSTM. The scenes are rendered by drawing a red bounding box around the pedestrian in question.
  - **Evaluation on Citywalks:** As there is no previous work doing the same experiments in JAAD, the model is evaluated on Citywalks. We test how our model deals with long term predictions and compare the performance with those reported in the recent work [15] on their proposed Citywalks dataset.

We use the average displacement error (ADE) and the final displacement error (FDE) as evaluation metrics for the performance of the center prediction of the future bounding boxes. We evaluate the performance of the dimensions prediction using the average and final intersection over union (AOIU, FIOU). For the pedestrian intention prediction task, we use accuracy as the evaluation metric.

Table 1: The accuracy performance of the intention prediction on JAAD dataset. Each method takes as input 14 observations and makes one future intention prediction.

Model	Accuracy (%)
CNN(fc6) [4]	70.0
SKLT [4]	88.0
CNN(fc6)+SKLT [4]	87.0
Trajectory-LSTM 4.2	86.16
Single-task PV-LSTM (ours)	89.67
Multi-task PV-LSTM (ours)	<b>91.48</b>

Table 2: The run-time performance of the network on JAAD dataset.

Model	Runtime (ms)
ConvNet-Softmax [17]	28ms
ConvNet-SVM [18]	27ms
ConvNet-LSTM [19]	40ms
C3D [19]	27ms
ST-Dense-Net [20]	10ms
Trajectory-LSTM 4.2	4.9ms
Single-Task PV-LSTM (ours)	<b>4.1ms</b>
Multi-Task PV-LSTM (ours)	<b>4.7ms</b>

## 4.3 Results

### 4.3.1 Quantitative Results

The quantitative results of our experiments in pedestrian intention prediction can be found in Table 1 and 2. The ones in pedestrian bounding box prediction are in Tables 3 and 4. We evaluate the performance of each task without considering the other head (*Single-task*) and with the whole network (*Multi-task*).

For comparison with all different models presented in [4], we modified the experiment setups as in 1. As Table 1 shows, the accuracy of our model in intention prediction outperforms all the baselines [4]. This is because our model is not limited to one prediction. It takes advantage of multiple predictions in the future to determine with more accuracy the immediate future intention of the pedestrian. This suggests that gaining more accuracy is possible without needing more features such as skeleton or scene features. As they did not provide timings, we cannot compare our run-time.

From Tables 1 and 2, we conclude that bounding box prediction (**Multi-task PV-LSTM**) instead of trajectory prediction (Trajectory-LSTM) could successfully improve the intention prediction without penalizing the run-time.

Table 3 shows the short-term prediction results of our model in bounding box prediction compared to the baselines. Our method has a high performance. It has a very low ADE of 9 pixels in the resolution of  $1920 \times 1080$ , which is less than 1% error. The high AIOU and FIOU performances show that the model precisely predicts not only the centers but also the whole bounding boxes.

Tables 1 and 3 show that multi-task learning improves the overall performance by introducing features from the predicted bounding boxes. Training both head networks simultaneously by back-propagating the weighted sum of the two losses helps both intention and bounding box prediction.

Table 4 reports the results of bounding box prediction on the Citywalks dataset. We followed all implementation details mentioned in [15] to compare fairly. The duration of the observation and prediction period are mentioned in the second column of the table. **PV-LSTM** has a comparable performance on ADE and FDE, capturing the dependencies of the motion of the bounding box center. However, the AIOU and FIOU results are worse than the baselines reported in the previously mentioned work. This shows that our method struggles with the dimensions of the bounding box for longer predictions. The high performance of our model on shorter sequences of the same dataset confirms this.

Table 3: Comparison of the performances of our models and the baselines on the JAAD dataset for the pedestrian bounding box prediction task. All the results reported are the evaluation metrics of the models on the testing set. ADE and FDE are in pixels in which the lower is better.

Model	ADE	FDE	AIoU (%)	FIoU (%)
Position-LSTM <b>4.2</b>	31.5	46.5	39.0	29.9
Velocity-LSTM <b>4.2</b>	9.28	15.4	74.6	63.0
Scene-PV-LSTM <b>4.2</b>	9.78	15.9	74.1	61.89
Single-task PV-LSTM (ours)	<b>9.18</b>	15.35	74.9	63.2
Multi-task PV-LSTM (ours)	9.19	<b>15.22</b>	<b>75.2</b>	<b>63.3</b>

Table 4: Comparison of the performance of our models with the STED network as well as the proposed baselines in [15] on the Citywalks dataset. We also show our results for a shorter input and output sequence length (mentioned in the second column in seconds) on the same dataset. ADE and FDE are in pixels in which the lower is better.

Model	(input, output)	ADE	FDE	AIoU (%)	FIoU (%)
CV-CS [15]	(1,2)	31.6	57.6	46.0	21.3
LKF [23]	(1,2)	32.9	59.0	43.9	20.1
DTP-MOF [15]	(1,2)	27.3	49.2	49.6	25.1
FPL-MOF [15]	(1,2)	29.3	51.0	44.9	22.6
STED [15]	(1,2)	26.0	<b>46.9</b>	<b>51.8</b>	<b>27.5</b>
Single-task PV-LSTM (ours)	(1,2)	<b>25.2</b>	49.9	40.2	20.3
Single-task PV-LSTM (ours)	(0.3,0.6)	10.22	17.17	73.2	62.8

### 4.3.2 Qualitative Results

Figure 2 shows some qualitative results in the dataset. The details are described in its caption. For predictions involving the pedestrians on the road, the predicted intentions are correct for all scenarios. The same thing can be observed for pedestrians on the sidewalk.

We observe that the bounding box center prediction is very accurate. However, the dimensions prediction starts with a very high IOU, which decreases at each timestep. These illustrations are compatible with the quantitative results that we discussed.

One observed challenge in this problem is varying pedestrian bounding boxes across frames, even if the pedestrians are not moving. Our model performs well in these cases, as shown in the last scenario of Figure 2. It shows the ability of the model to capture car speed.

## 5 Conclusion

We have presented a method for performing pedestrian intention and visual states prediction in a multi-task learning approach. We have shown that our model outperforms previously proposed methods as well as state of the art in pedestrian intention prediction, and has a competitive performance for pedestrian state (bounding box) forecasting. We demonstrated that predicting the velocity of the bounding boxes instead of the position, taking advantage of the observed positions and speed of bounding boxes, as well as using a multi-task learning architecture, could improve the accuracy of the pedestrian intention prediction. It is also worth mentioning that our model has high accuracy while being simple with two times faster runtime.

Although our method yields better results than previous works, it still needs to be improved in longer period prediction. In future work, we will explore how much the visual scene information can improve the predictions without adding much complexity. Moreover, we will evaluate the proposed method on more datasets in order to validate its potential for deploying in the real-world.

## Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 754354.

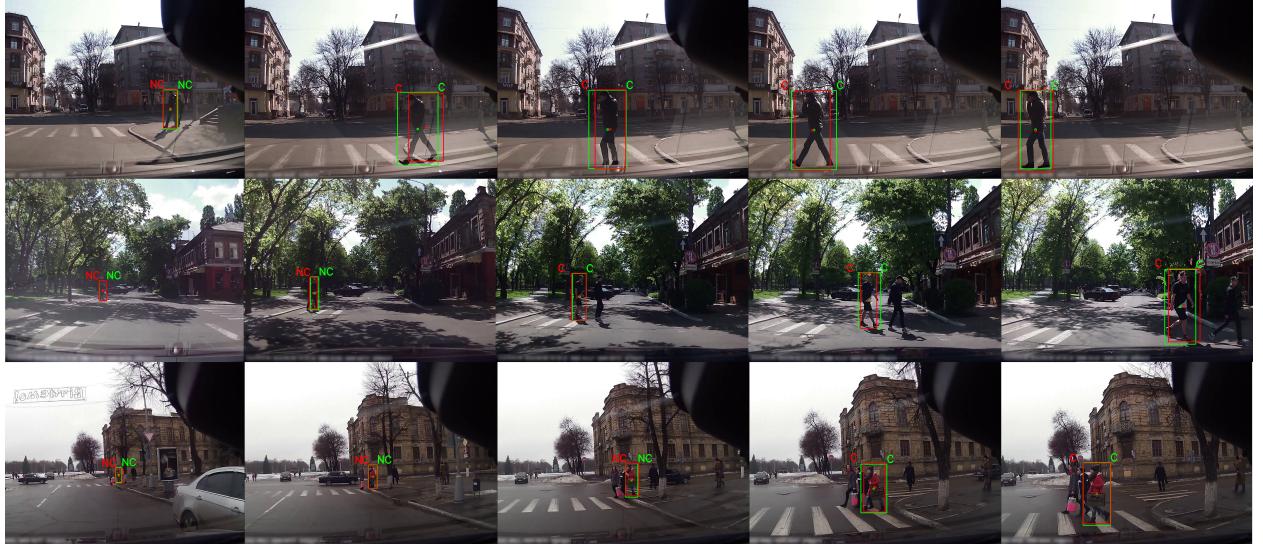


Figure 2: Visualisation of the pedestrian bounding box and intention prediction in different scenarios. Red rectangle/text: Predicted bounding box/intention | Green rectangle/text: Ground truth bounding box/intention | C : Crossing | NC: Non crossing. Each frame represents a time step, going from left to right. The first row shows how our model acts at points where the state of the pedestrian changes from crossing to non-crossing. The second shows a whole crossing sequence. The third shows how our network handles non-moving pedestrians when the car is moving.

## References

- [1] Andreja Bubic, D. Yves Von Cramon, and Ricarda Schubotz. Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4:25, 2010. [1](#)
- [2] Daphne Evans and Paul Norman. Predicting adolescent pedestrians’ road-crossing intentions: an application and extension of the theory of planned behaviour. *Health education research*, 18(3):267–277, 2003. [1, 2](#)
- [3] Satyajit Neogi, Michael Hoy, Weng Chaoqun, and Justin Dauwels. Context based pedestrian intention prediction using factored latent dynamic conditional random fields. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, 2017. [1, 2](#)
- [4] Zhijie Fang and Antonio M López. Is the pedestrian going to cross? answering by 2d pose estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1271–1276, 2018. [1, 2, 5, 6](#)
- [5] Michael Hoy, Zhigang Tu, Kang Dang, and Justin Dauwels. Learning to predict pedestrian intention via variational tracking networks. In *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3132–3137, 2018. [1, 2](#)
- [6] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. [1, 2](#)
- [7] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You’ll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE 12th International Conference on Computer Vision*, pages 261–268, 2009. [1](#)
- [8] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. [2](#)
- [9] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5275–5284, 2018. [2](#)
- [10] Nishant Nikhil and Brendan Tran Morris. Convolutional neural network for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. [2](#)
- [11] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194, 2018. [2](#)

- [12] Huynh Manh and Gita Alaghband. Scene-lstm: A model for human trajectory prediction. *arXiv preprint arXiv:1808.04018*, 2018. [2](#)
- [13] Lituan Wang, Lei Zhang, and Zhang Yi. Trajectory predictor by using recurrent neural networks in visual tracking. *IEEE transactions on cybernetics*, 47(10):3172–3183, 2017. [2](#)
- [14] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4194–4202, 2018. [2, 3](#)
- [15] Oliver Styles, Victor Sanchez, and Tanaya Guha. Multiple object forecasting: Predicting future object locations in diverse environments. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 690–699, 2020. [2, 5, 6, 7](#)
- [16] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, Jul 1997. [2](#)
- [17] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Early intent prediction of vulnerable road users from visual attributes using multi-task learning network. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3367–3372, 2017. [2, 6](#)
- [18] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017. [2, 6](#)
- [19] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. [2, 6](#)
- [20] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9704–9710, 2019. [2, 5, 6](#)
- [21] Joon-Young Kwak, Byoung Chul Ko, and Jae-Yeal Nam. Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime. *Infrared Physics & Technology*, 81:41–51, 2017. [2](#)
- [22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2462–2470, 2017. [2](#)
- [23] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960. [2, 7](#)
- [24] Oily Styles, Arun Ross, and Victor Sanchez. Forecasting pedestrian trajectory with machine-annotated training data. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 716–721, 2019. [3](#)
- [25] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7593–7602, 2018. [3](#)
- [26] Yu Yao, Mingze Xu, Chiho Choi, David J Crandall, Ella M Atkins, and Behzad Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9711–9717, 2019. [3](#)
- [27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. [3](#)
- [28] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016. [5](#)