Expert Systems With Applications 225 (2023) 120077

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation

Sarfraz Ahmed [a], Ammar Al Bazi [b], Chitta Saha [a], Sujan Rajbhandari [c], M. Nazmul Huda [d,*]

[a] School of Future Transport Engineering, Coventry University, Priory Street, Coventry, CV1 5FB, West Midlands, United Kingdom
[b] School of Mechanical Engineering, Coventry University, Priory Street, Coventry, CV1 5FB, West Midlands, United Kingdom
[c] DSP Centre of Excellence, School of Computer Science and Electronic Engineering, Bangor University, Bangor, Gwynedd, LL57 2DG, North Wales, United Kingdom
[d] Department of Electronic and Electrical Engineering, Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, London, United Kingdom

ARTICLE INFO

Keywords:
LSTM
Intent prediction
Pose estimation
Tracking
Pedestrian detection

ABSTRACT

There has been a rise of use of Autonomous Vehicles on public roads. With the predicted rise of road traffic accidents over the coming years, these vehicles must be capable of safely operate in the public domain. The field of pedestrian detection has significantly advanced in the last decade, providing high-level accuracy, with some technique reaching near-human level accuracy. However, there remains further work required for pedestrian intent prediction to reach human-level performance. One of the challenges facing current pedestrian intent predictors are the varying scales of pedestrians, particularly smaller pedestrians. This is because smaller pedestrians can blend into the background, making them difficult to detect, track or apply pose estimations techniques. Therefore, in this work, we present a novel intent prediction approach for multi-scale pedestrians using 2D pose estimation and a Long Short-term memory (LSTM) architecture. The pose estimator predicts keypoints for the pedestrian along the video frames. Based on the accumulation of these keypoints along the frames, spatio-temporal data is generated. This spatio-temporal data is fed to the LSTM for classifying the crossing behaviour of the pedestrians. We evaluate the performance of the proposed techniques on the popular Joint Attention in Autonomous Driving (JAAD) dataset and the new larger-scale Pedestrian Intention Estimation (PIE) dataset. Using data generalisation techniques, we show that the proposed technique outperformed the state-of-the-art techniques by up to 7%, reaching up to 94% accuracy while maintaining a comparable run-time of 6.1 ms.

1. Introduction

In recent years, significant improvements have been made in the field of pedestrian detection, with state-of-the-art techniques reaching an accuracy of over 90% (Ahmed et al., 2019a, 2019b; Fang & López, 2018; Galvao et al., 2021). Although, these technologies are yet to achieve human-level accuracy, they have allowed for an increased focus on higher-level tasks such as intent prediction. Intent prediction refers to predicting the future movements of the pedestrians with respect to the vehicle's current path and whether the vehicle will have adequate time to react to the pedestrian's future location. Intent prediction is another crucial aspect for the safety of Vulnerable Road Users (VRUs), especially for autonomous vehicles applications. This is made evident as per a recent report published by Google, which found that most failures tend to occur in busy streets, where 10% of errors are due to incorrect intent prediction of other road users (Google, 2015).

Predicting the future movements and locations of VRUs has been a challenging task, particularly in urban environments (Keller & Gavrila, 2014; Keller et al., 2011; Rasouli et al., 2019; Saleh et al., 2019b). By predicting pedestrian intentions, the vehicle can perform manoeuvres such as slowing down, switching lanes or stopping if need be; thus limiting or even avoiding any vehicle-pedestrian incidents. Techniques, such as dynamical motion modelling and motion planning (Saleh et al., 2017b), have been suggested in the past for pedestrian intent prediction. Although these techniques can be powerful, they both rely on hand-crafting a set of scene-specific features, which in return affects their generalisation in unseen scenes. In more recent studies, techniques based on Deep Learning have been implemented for improved pedestrian detection methods, such as 3D convolution neural network (CNN) and spatio-temporal Long Short-term Memory (LSTM) for behaviour prediction (Liu et al., 2020; Saleh et al., 2019b). According to Razali et al. (2021), methods for predicting future pedestrian intentions can be

* Corresponding author.
E-mail addresses: ahmed157@uni.coventry.ac.uk (S. Ahmed), aa8535@coventry.ac.uk (A.A. Bazi), ab3135@coventry.ac.uk (C. Saha), s.rajbhandari@bangor.ac.uk (S. Rajbhandari), MdNazmul.Huda@brunel.ac.uk (M.N. Huda).

https://doi.org/10.1016/j.eswa.2023.120077
Received 12 March 2022; Received in revised form 15 March 2023; Accepted 6 April 2023
Available online 13 April 2023

either trajectory-based (e.g., Chen et al. (2021), Huang et al. (2021), Li et al. (2020a, 2020b), Quan et al. (2021), Wu et al. (2019)) or treated as a classification task (Gujjar & Vaughan, 2019; Rasouli et al., 2017; Saleh et al., 2019b, 2020). Trajectory-based pedestrian future intent prediction involves using temporal and spatial information to predict future movements of pedestrians. Treating future pedestrian intent prediction as a classification task involves classifying the movements of pedestrians (e.g., crossing, not crossing, standing, walking, etc.). Previously, before intentions can be predicted, the pedestrian needed to be detected. However, in Cheng et al. (2020), an approach that does not require a detector was proposed. This method improves upon the run-time of previous approaches. The time taken by the Autonomous Vehicles to predict pedestrian intention based on raw sensor data and actually performing manoeuvres is a vital consideration. Ideally, these systems should be working in real-time.

In this paper, we propose a novel data-driven approach for multi-scale pedestrian intent prediction. This approach combines trajectory-based and classification tasks discussed previously. Using on a state-of-the-art pose estimator, keypoints are generated. The accumulation of these predicted keypoints over time (along the frames) is fed into an LSTM-based classifier for classifying their crossing behaviour (i.e., whether the pedestrian will cross or not cross). These architectures will be further discussed in Section 3. A challenge facing current pedestrians intent predictors is pedestrian scale variance. Typically, keypoints are more difficult to predict for smaller pedestrians due to their lower resolution. This is caused by the pedestrian being blurred by the background (Kim et al., 2021). For this reason, many previous approaches focused on larger pedestrians, such as in, Fang and López (2018), Fang et al. (2017), Liu et al. (2020). We also consider heavily occluded pedestrians. As far was we are aware, this is the first multi-scale intent prediction approach, that focuses on multiple classes of scales and occlusions. We also implement a bottom-up approach, which does not require a standalone detector. This is to improve upon run-times as compared to similar intent prediction approaches. To the best of our knowledge, this is also the first time the PIE (Rasouli et al., 2019) dataset has been used for evaluation. Previously, most techniques, such as the one previously mentioned and in Section 2, use the JAAD dataset. The PIE dataset is a larger-scale dataset than the JAAD dataset, providing an increased number of unique pedestrian samples. This provides more data for training and evaluation of our proposed model. Methodology in Section 3 will discuss the individual components of the proposed architecture. Sections 4 and 5 will discuss the experimentation and results for the proposed method. We also apply a technique for improved dataset generalisation in Section 5. The original contributions of the paper are as follows:

- An approach for multi-scale pedestrian intent prediction utilising a 2D pose estimation+LSTM architecture.
- Unlike previous methods, the proposed approach also considers small and occluded pedestrians.
- A novel approach using predicted keypoints to generate bounding boxes, which are utilised for the evaluation of datasets that do not contain ground-truth keypoint annotations.
- Exploration of data generalisation techniques for improved general performance over two datasets (i.e., JAAD and PIE datasets).

2. Related works

A pedestrian-centric approach was introduced in Liu et al. (2020), where the pedestrian's pose, location and velocity were used to predict future intentions. Although a simple technique, these features can be useful for inferring pedestrian intent. However, these features do not consider scene context or elemental interactions (e.g., interaction with other pedestrians, zebra-crossings, traffic lights, etc.). Furthermore, head orientation can also be necessary for predicting the pedestrian's intent. Based on this information, the authors proposed an approach for generating a pedestrian-centric dynamic scene graph using an off-the-shelf segmentation model. Using graph convolution techniques, relationships between the pedestrian(s) and the scene context and interactions were recorded with each pedestrian having a corresponding graph. The contextual visual information is aggregated over time to reason temporal relations with the environmental relations. This allows for subtle actions to be captured, which are vital for predicting intent (Liu et al., 2020). This approach outperformed previous state-of-the-art techniques by nearly 9% with an accuracy of 76.98%.

In Fang and López (2018), Fang et al. (2017), a fully vision-based (using a monocular camera) approach for pedestrian intent prediction was proposed. The studies proposed a pipeline consisting of an off-the-shelf detector, tracker and poses estimator. The pipeline's output would feed into a classifier to determine whether a pedestrian would cross or not cross. Their approach used a sliding time window for accumulating skeletons generated by the pose estimator. The classifier architecture was based on Random Forest. The aim of the model was to accurately predict whether the pedestrian would cross or not cross. In Fang and López (2018), the authors achieved an accuracy of 88% on the JAAD dataset, which was 25% improvement compared to baseline work in Rasouli et al. (2017). This approach is based purely on monocular-based data and does not require complicated methods of information gathering, such as stereo, optical flow or ego-motion compensation. Also, this method does not require knowledge of head or body orientations. The authors suggest that it is not made clear in previous studies how the pedestrian's head or body orientations would aid in improving the reaction time of an intent prediction system. A previous study in Rasouli et al. (2017) concluded that head-orientated and body-orientated information do not improve effectiveness and the reactionary time for pedestrian intent prediction.

The pedestrian intention estimation (PIE) dataset was introduced in Rasouli et al. (2019). It is a large-scale dataset for pedestrian trajectory prediction. Prior to the PIE dataset, other than the JAAD dataset (Rasouli et al., 2017), there were very few datasets that centred around predicting pedestrian trajectory with the point-of-view from a moving vehicle. Previous datasets for pedestrian trajectory prediction included videos from surveillance camera perspective (Benfold & Reid, 2011; Oh et al., 2011; Zhou et al., 2012) and top-down view (Pellegrini et al., 2009; Robicquet et al., 2016). According to Rasouli et al. (2019), widely used pedestrian detection datasets, such as Dollár et al. (2012), Geiger et al. (2012), Zhang et al. (2017) could possibly be utilised for the purpose of pedestrian intent prediction, however, the datasets do not contain pedestrian behaviour annotations. In 2017, the JAAD dataset was introduced. JAAD is also a large-scale dataset with behavioural information. However, it should be noted that the majority of the pedestrian samples with behaviour annotations are "crossing" samples, meaning the dataset is imbalanced. There is approximately 450 *crossing* samples and 200 *not crossing* samples. This could cause biases when training a model to learn the "crossing" and "not crossing" behaviours. These behaviours are better balanced in the PIE dataset with 512 and 430 of "crossing" and "not crossing" behaviour annotations respectively. The PIE dataset is also a larger dataset than the JAAD dataset, with over 900,000 frames compared to around 82,000. Of these frames, approximately 1800 frames include pedestrians with behaviour annotations while JAAD contains 686 frames with pedestrians with behaviour annotations.

In Minguez et al. (2019), the authors proposed method using balanced Gaussian process dynamical models (B-GPDMs) for pedestrian intent prediction. The B-GPDMs reduce the 3D spatio-temporal information extracted from keypoints over a number of frames to low-dimensional spaces. The proposed approach consists of four distinct models, each for learning a different behaviour. These behaviours include walking, stopping, starting and standing. The authors mention that having a single model to learn multiple pedestrian behaviours provides less accurate intention predictions. The model achieved 80%

accuracy on the CMU Graphics Lab Motion Capture Database (CMU, 2017).

A bottom-up pose estimator was proposed in Cheng et al. (2020), referred to as HigherHRNet. It is a novel approach that utilises high-resolution feature pyramids for representing varying pedestrian scales. The HigherHRNet aims to overcome the challenge of scale variation faced by many bottom-up pose estimation approaches so that keypoints are more accurately localised, particularly for smaller pedestrians. Higher HigherHRNet outperformed the previous state-of-the-art bottom-up approach by 2.5%, reaching the accuracy of 70.5% AP (average precision) on the COCO (Lin et al., 2014) dataset for a medium person. Higher HigherHRNet also outperformed current top-down methods, reaching 67.6% AP on the CrowdPose (Li et al., 2019) dataset. These results demonstrate the quality and robustness of the HigherHRNet for multi-scale pedestrian detection, even in crowded scenes. Refer to Cheng et al. (2020) for further information on bottom-up and top-down approaches.

In this work, we present a novel approach for data-driven pedestrian intent prediction. We take inspiration from the modular architectures introduced in Fang and López (2018), Fang et al. (2017), Saleh et al. (2017b). We implement a tracking-by-detection approach using a 2D pose estimation for generating pedestrian keypoints and LSTM-based classifier to predict pedestrian crossing intentions. Based on the generated keypoints, we can predict the future movements and speed of change of those movements to determine whether a pedestrian is going to cross or not cross. These keypoints represent the evolution of the pedestrian over time (spatio-temporal information), but unlike other similar approaches, this approach does not require any additional environmental or contextual information (traffic lights, traffic signs, relative location of other pedestrians) to predict intentions. In Fang and López (2018), Fang et al. (2017), 14 frames are required for accurate prediction of pedestrian intentions. The papers do not discuss performances of any occluded pedestrian instances that result in unusable frames. In our proposed method, frames where the pedestrian is partially occluded can be skipped without sacrificing the accuracy of making intention predictions. This allows for a more robust and effective method for pedestrian intent prediction. This approach requires only RGB data collected by a monocular camera. We improve upon the limitations of previous works, as far as we know, by (1) introducing a novel multi-scale pedestrian intent predictor (2) improving upon current state-of-the-art techniques in terms of accuracy and (3) implementing a robust design capable of working with limited keypoint information.

3. Methodology

We propose a novel approach using 2D pose estimation with an LSTM-based classifier for multi-scale pedestrian intent prediction. A detailed flowchart of the proposed method is provided in Fig. 1. The pose estimator predicts keypoints for the pedestrians in the image. Based on the predicted keypoints, associated bounding boxes are generated. The bounding boxes are used to track the pedestrian along the images. The tracked pedestrian's keypoints are stored and concatenated. These concatenated keypoints are sent the LSTM-based classifier to predict whether the pedestrian will cross or not cross. As far as we are aware, this implementation has not been used previously. Fig. 2 further illustrates this flowchart. We will now proceed to discuss three main components of the proposed method. These components are the 2D pose estimator, tracking-by-detection and intent classification.

3.1. Pose estimation

The input data is fed into the 2D pose estimator to generate keypoints. Keypoints, also referred to as joints, are points of interest in an image. In this case, these points of interest represent the location of joints, such as shoulders, elbows, ankles etc. The key points themselves provide spatial information in terms of their location in the image, while the changes of these key points over a number of frames represent the temporal information. Over time, predicted keypoints provide spatio-temporal information, which can be used to calculate the trajectory and velocity of the pedestrians. This information is for predicting the crossing intentions of the pedestrian. As the proposed approach in this work is designed for multi-scale pedestrian intent prediction, the HigherHRNet (Cheng et al., 2020) is employed. The HigherHRNet was proposed to overcome the challenge of pedestrian scale variance, particularly, smaller pedestrians. Smaller pedestrians typically are of low-resolution then larger pedestrians, and therefore, can blur into their background. This makes it difficult to detect keypoints for smaller pedestrians. The HigherHRNet architecture will be discussed in Section 4.2.

3.2. Tracking-by-detection

Based on the predicted keypoints, associated bounding boxes can be generated. The process of generating the keypoint-based bounding boxes is discussed in Section 4.3. Bounding boxes provide two useful purposes. The JAAD and PIE datasets are popular datasets for training and evaluation of pedestrian intention prediction models. Although they provide annotations, such as bounding boxes and crossing behaviours, they do not include keypoint information. So, the first purpose of the keypoint-based generated bounding boxes is to compare them to the ground-truth bounding boxes provided by the datasets. By comparing them, the accuracy of the predicted keypoints can be established. This will justify the use of the 2D pose estimator used in Section 4. This allows for fine-tuning the 2D pose estimator further (see Section 4.4). Also, with the use of these generated bounding boxes, a standalone pedestrian detector is not required as in Fang and López (2018), Fang et al. (2017). This leads to a reduced system with less computational cost and complexity, resulting is faster run-times. The comparison of the run-times will be discussed in Section 4. The generated bounding boxes are sent to a tracker. This is the second purpose for using keypoint-based bounding boxes. In this way, the pre-trained 2D pose estimator does not need to be adjusted to learn to predict bounding boxes. Instead, the generated bounding boxes are fed into the tracker.

For the detection and tracking components, we utilise the tracking-by-detection approach as in Bewley et al. (2016), Ess et al. (2009), Saleh et al. (2017b), Yu et al. (2016). Which simply means as a pedestrian is detected, they are also tracked. The popularity of this approach is due to its maintaining a high accuracy while not hindering real-time performance. For the tracking, we implement a variation of the SORT (Simple Online and Real-time Tracking) as in Wojke et al. (2018), originally introduced in Bewley et al. (2016). Using Kalman filtering, SORT tracks pedestrians along the frames based on generated bounding boxes. We use the tracker as-is, and therefore, we will not discuss the SORT tracker in further detail in this work. For more information on SORT tracker architecture, please refer to Wojke et al. (2018). Each new tracked pedestrian is given a unique ID. Each time the pedestrian is detected and tracked along the frames, its keypoint information is stored. Once the stored keypoint instances reach a certain number, the information is concatenated and sent to the intent classifier. Each instance refers to a frame. Using the changes in the keypoints over time, we were able to determine the movement of the pedestrian (e.g., the pedestrian moving from left to right or remaining to the left or right with respect to the vehicle) and speed of these movements. Using this information, the future behaviour can be predicted. We focus on "crossing" and "not crossing" behaviours. This can provide an Autonomous Vehicle enough information to decide (e.g., slow down, maintain speed, switch lanes).

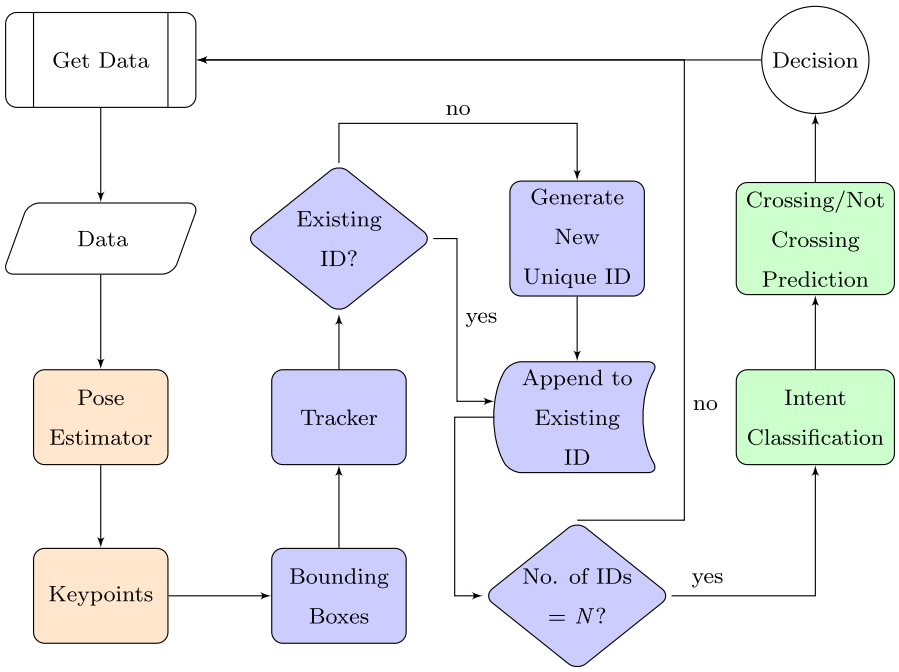
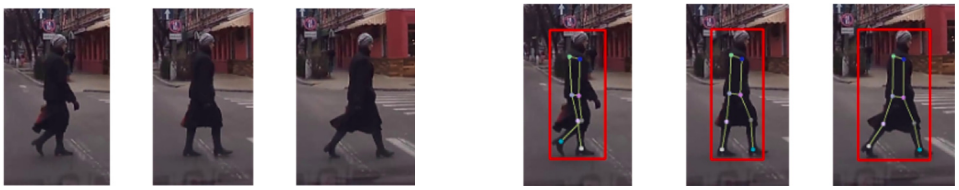


Fig. 1. Flowchart for proposed intent prediction system. The data is fed into the pose estimator, to predict key points and generate bounding boxes. The bounding boxes are used for tracking. When the pedestrian is tracked for N of frames, it is fed into the intent classifier to classify the crossing/not crossing behaviours.



(a) Input frames (b) Pose estimation

(c) Concatenate predicted keypoints

Fig. 2. In this example a pedestrian is crossing the road. Each frame is set 5 frames apart to demonstrate the evolution of the skeleton over time (along frames). In each frame, keypoints are generated from the pose estimator. Using the keypoints, bounding boxes are calculated, allowing to track the pedestrian along the frames. Using the changes of each keypoint, the velocity and direction can be observed to predict the future actions of the pedestrian.

3.3. Intent classification

We use the spatio-temporal information provided by the concatenated keypoints. Both Fig. 4 and Fig. 5 illustrate the skeletons for pedestrians crossing and not crossing, respectively. The skeletons are generated by connecting various keypoints. For illustration purposes, we use only 5 frames with 5 frames between frame in both figures. The figures illustrate the differences between pedestrians crossing and not crossing. The movements of the joints, such as arms and legs vary, and this information is used to predict the pedestrian's future behaviour.

We employ the PV-LSTM architecture introduced in Bouhsain et al. (2020) for intent classification. The original PV-LSTM uses bounding boxes to calculate positional and speed information of a moving pedestrians. This information is passed to a LSTM-based feature extractor to predict crossing intentions based on predicted future bounding boxes. Essentially, it uses bounding boxes along a number of frames, and based on the movements and velocity of those boxes over times, predicts the future bounding boxes. In this way, it predicts the future trajectory and location of a pedestrian. Based on that information, it predicts whether the pedestrian will cross or not cross with respect to the vehicle. This approach is simpler than other similar state-of-the-art approaches as it requires fewer parameters while obtaining a comparable or higher accuracy. This approach achieved an accuracy of 91.45% for multi-tasking, which is 5.32% better than the next best intent prediction model.

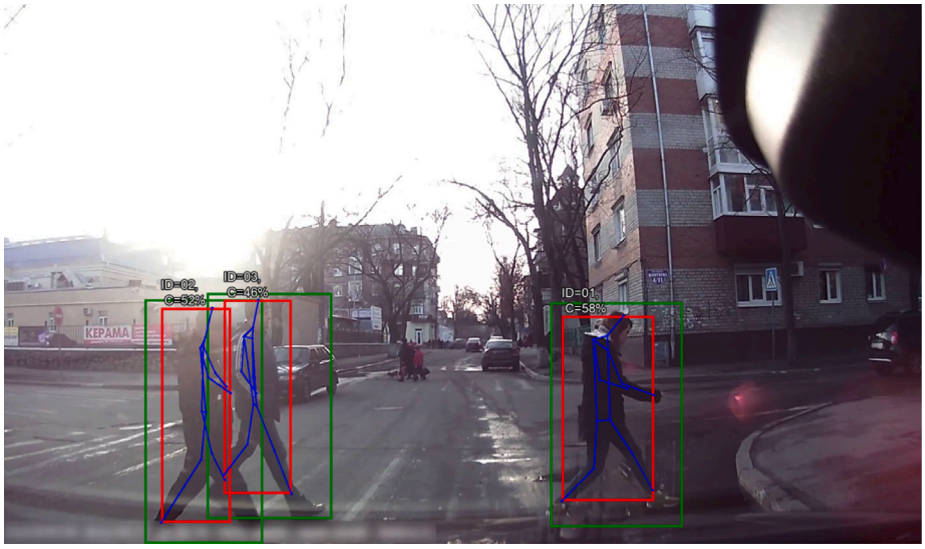


Fig. 3. Multiple pedestrians are being tracked. This is based on the keypoint-generated bounding box (red). The predicted keypoints are connected to create a skeleton for illustration purposes (blue). The ground-truth boxes (green) are for comparison with the generated boxes. The "C" represents the confidence score of the generated bounding box compared to the ground-truth box. The higher the score, the closer the generated bounding box coordinates is to the ground-truth bounding box coordinates (i.e., more accurate). Each tracked pedestrian is given a unique ID to track them along frames.

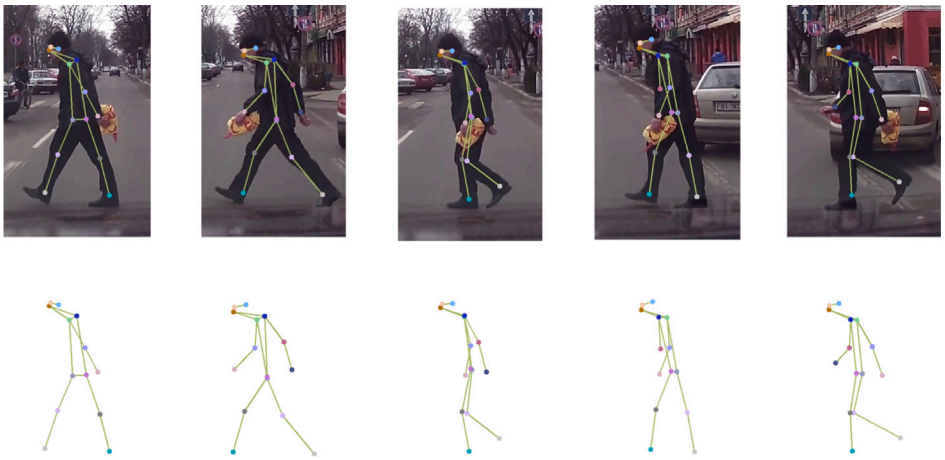


Fig. 4. Example of a pedestrian crossing in front of the vehicle.

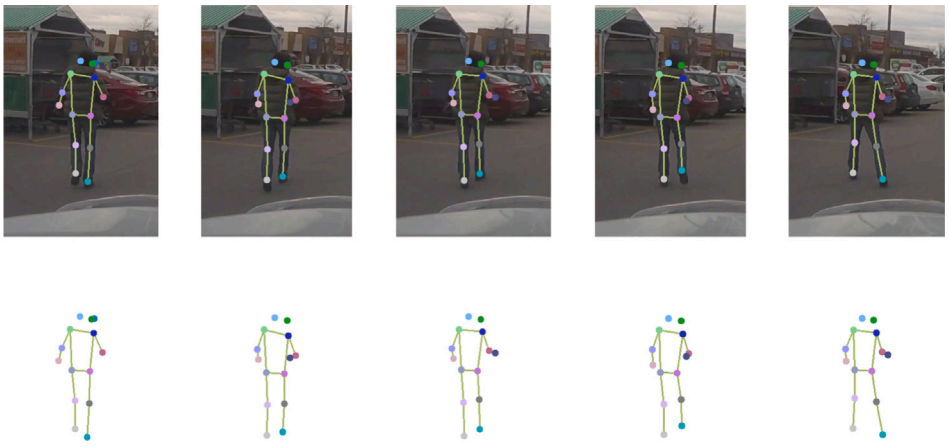


Fig. 5. Example of pedestrian walking perpendicular to vehicle.

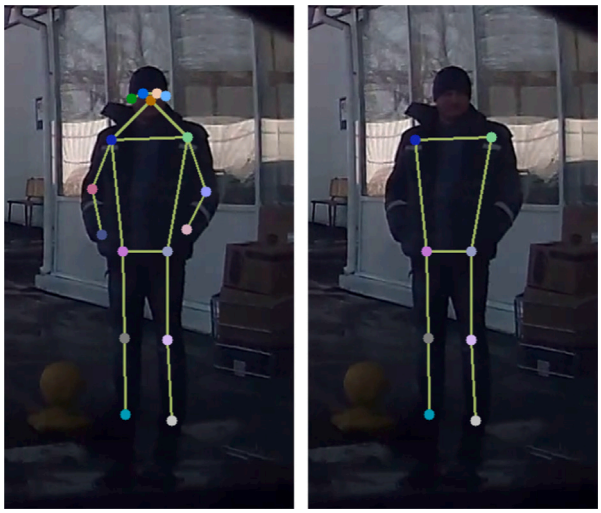


Fig. 6. (*left*) Skeleton generated using all keypoints. (*right*) Skeleton using only shoulders, hips, knees and ankles.

In this work, we customise the PV-LSTM to use keypoints instead of the bounding boxes. We found that keypoints offer more useful and accurate information in terms of the speed and position of the pedestrians. The evolution of pedestrian joints provides information of the joints with respect to other joints (e.g., left and right shoulders, left and right knees), and the speed at which those joints move help to predict the trajectory. (Refer to Fig. 6, where the aforementioned key points are used to generate the pedestrian skeleton.)

4. Experimentation

4.1. Datasets

For the task for training and evaluation, the JAAD (Rasouli et al., 2017) and PIE (Rasouli et al., 2019) datasets were utilised. The JAAD dataset has been widely used for pedestrian intent prediction task, particularly crossing predictions, as in Bouhsain et al. (2020), Fang and López (2018), Gesnouin et al. (2020), Liu et al. (2020), Yang et al. (2021). Unlike other publicly available large datasets, the JAAD dataset provides behavioural information, such as crossing behaviour, as well as bounding boxes and occlusion information. The JAAD dataset consists of 346 videos with a resolution of 1920 × 1080 running at 30 fps (frames per second). It consists of 2200 unique pedestrian samples with 337,000 corresponding bounding boxes in a total of 8200 frames. There are 81 ”not crossing” and 234 ”crossing” samples (when omitting heavily occluded samples). PIE dataset was introduced by the same authors as the JAAD dataset. It is a larger-scale dataset then the JAAD dataset. Like the JAAD dataset, the PIE dataset also consists of videos of pedestrians portrayed naturally. The dataset also includes bounding box information for the pedestrians as well as behavioural information. There is a total of 911,000 frames, of which 293,000 are annotated with 1800 unique pedestrian samples and 740,000 bounding boxes. Both datasets also include contextual information, such as road width, traffic signs and traffic lights. Although this information is not used in our proposed method, this can be used for further improvement. In this work, we focus on the JAAD dataset as it has been widely used for pedestrian intent prediction. This is to provide direct comparability with previous works. We also use the PIE dataset as it is provided by the same authors as the JAAD dataset. It is of the same format as JAAD in terms of image resolution and behavioural annotations, however it is a larger dataset.

4.2. HigherHRNet architecture

There are two main approaches for pose estimation: top-down and bottom-up and. The top-down approach uses a detector to locate the pedestrians in an image. For each detected pedestrian, pose estimation is applied within boxing bound coordinates to generate keypoints. Examples of the top-down approaches can be found in Fang and López (2018), Fang et al. (2017). In contrast, bottom-up approach predicts the keypoints in an image and then groups them together for each pedestrian. This means that the bottom-up approach does not require a standalone detector. This makes the bottom-up approach faster, allowing for the capability of real-time application (Cheng et al., 2020). The ability for the bottom-up approach to reach real-time run-times is critical, especially for time-dependent applications, such as Autonomous Vehicles. Therefore, in this work we focus on the HigherHRNet, which is both a bottom-up approach and a multi-scale pose estimator.

The HigherHRNet employs the HRNet (Sun et al., 2019) as the backbone. The HRNet architecture consists of a high-resolution branch, which subsequently followed by additional branches in parallel. The additional branches are $1/2$ the resolution of the lowest resolution of the current branches. In this way, as the network branches increase, the number of parallel branches also increases, providing different resolutions while preserving all the resolutions. According to Cheng et al. (2020), for smaller pedestrians, the resolution of the heat-map is vital. However, most current pedestrian pose estimators predict keypoints using Gaussian-smoothing techniques and a Gaussian kernel for each keypoint. Although this approach is useful during training, it causes an uncertainty as to the precision localisation of the predicted keypoints. This can be determinate for smaller pedestrians. As Cheng et al. (2020) mentions, a simple solution could be to reduce the Gaussian kernel's standard deviation. However, they found this approach negatively impacted the performance of the keypoint prediction. Therefore, for the HigherHRNet, Cheng et al. (2020) proposed using multiple high-resolutions feature maps without changing the standard deviation for each feature map for predicting heat-maps. This was achieved using a deconvolution module, which generated high-resolution feature maps. Deconvolution module uses both the features maps and heat-maps generated by the HRNet and outputs feature maps with double the resolution. This provides a feature pyramid, with two resolutions, one from the HRNet and one from the deconvolution module. The benefit of this architecture is that if higher resolution is required, more deconvolution modules can simply be added. For smaller pedestrians, typically larger resolution features maps were required.

4.3. Keypoint-based bounding box coordinates

As discussed, the JAAD and PIE datasets do not include ground-truth keypoint However, the datasets do include ground-truth bounding box information. Therefore, the keypoints predicted using the HigherHRNet cannot be evaluated with ground-truth annotations. To overcome this challenge, we propose a technique whereby the predicted keypoints are used to generate bounding boxes. By comparing the generated bounding boxes with the ground-truth bounding boxes, the quality of the bounding boxes can be evaluated. In this way, the predicted keypoints are evaluated, as the generated boxes are calculated using the predicted keypoints. We will now proceed to discuss this proposed approach.

The predicted keypoints by the HigherHRNet are in the format of either 14×3 or 17×3 for CrowdPose and COCO formats, respectively. The rows represent the x, y coordinates and confidence score (see Fig. 7). We omit any columns with a confidence score less than 0.8. Using the x and y coordinates remaining keypoints, the associated bounding box is generated for each predicted pedestrian. The bounding box (1) is calculated by (2)–(5) based on the predicted keypoints. These generated bounding boxes can be compared with the ground-truth bounding boxes provided by the datasets. Thus, the predicted

$$\begin{bmatrix} x_1 & y_1 & score \\ x_2 & y_2 & score \\ \vdots & \vdots & \vdots \\ x_n & y_n & score \end{bmatrix}$$

Fig. 7. HigherHRNet keypoint predictions format.

Table 1
Multi-scale settings.

Setting	Height(h)[a]	Occlusion(o)[b]
Reasonable	$h \geq 50$	$o \leq 75\%$
Reasonable (small)	$50 \leq h \leq 75$	$o \leq 75\%$
Heavy occlusion	$h \geq 50$	$o \geq 75\%$
All	$h \geq 20$	$o \leq 75\%$

[a]Pixel height of pedestrian.
[b]Visibility of pedestrian.

Table 2
Multi-scale performance.

Setting	JAAD	PIE
Reasonable	89%	91%
Reasonable (small)	78%	81%
Heavy occlusion	40%	43%
All	87%	89%

keypoints were evaluated without ground-truth keypoint information. As far as we are aware, this approach is a novel approach for keypoint evaluation without the need for ground-truth keypoints.

$$box = [left, top, right, bottom] \tag{1}$$

$$left = \min(\left[x_1, x_2 \dots x_n\right]) \tag{2}$$

$$right = \max(\left[x_1, x_2 \dots x_n\right]) \tag{3}$$

$$top = \min(\left[y_1, y_2 \dots y_n\right]) \tag{4}$$

$$bottom = \max(\left[y_1, y_2 \dots y_n\right]) \tag{5}$$

Both the JAAD and PIE datasets have a very minimal number of smaller pedestrians. Therefore, we implemented a strategy to scale down the images to 30% of the original size. We found that sufficient samples were not generated if we used a large-scale value. This method generated more smaller pedestrian samples without significantly reducing the quality of the images. Using this strategy, we were able to increase the number of smaller pedestrian samples by approximately a magnitude of 7 for the JAAD dataset. For the PIE dataset, we nearly doubled the amount of smaller pedestrian samples.

4.4. Training & evaluation protocols

We fine-tune the pre-trained Higher HigherHRNet using images from the JAAD and PIE datasets as the pedestrians in these datasets are more challenging to detect than the COCO dataset. As discussed, the JAAD and PIE datasets do not have ground-truth keypoint annotations. Therefore, we implemented the following strategy. We evaluated the keypoints using the generated boxes (Section 4.3) and comparing them with the ground-truth boxes. Those keypoints with an accuracy of over 80% (threshold) were used to fine-tune the HigherHRNet model. As we did this, accuracy of those boxes with previously lower accuracy improved. We this several times, each time adding further images. We stopped this process when the accuracy of the model achieved 70% as this was close to the multi-scale results in the original work (see Cheng et al. (2020)). We used the default training settings for the HigherHRNet provided in Cheng et al. (2020). The model was trained for a total of 300 and Adam optimiser with a base learning rate set to 1e−3. After 200 epochs, the learning rate is dropped to 1e−4 and after 260, it was further dropped to 1e−5.

Once the HigherHRNet was fine-tuned, we proceeded to train the PV-LSTM model for intent prediction. We followed the settings in Bouhsain et al. (2020) for the JAAD dataset for data splitting training and evaluation. There are a total of 346 videos, where 300 were used for training and the remaining 46 were used for testing. We also split the PIE dataset in a similar manner. The PIE dataset is a larger dataset and is split into sets, with each set containing several videos. There are a total of 6 sets, we used the first 4 sets of training and the 2 remaining sets for testing. For training, we used the Adam optimiser and set the initial learning rate at 1e−4 and use an adaptive scheduler to automatically reduce the learning rate when the loss began to plateau. The model was trained for 100 epochs on an NVIDIA RTX 2080 Ti GPU. We set the hidden states of the PV-LSTM model to 256.

During the training of the custom PV-LSTM model, we follow the settings from Fang et al. (2017) and Fang and López (2018), where they use specific keypoints instead of using all the predicted keypoints. This is because not all the predicted keypoints are useful for the task of intent prediction. The keypoints for the shoulder, hips, knees and ankles are the most important keypoints as the legs perform the act of walking/stopping while the shoulders provide a global orientation of these pedestrians. However, unlike in Fang and López (2018), Fang et al. (2017), we do not connect the key points to create a skeleton. The skeleton was used to provide information, such as distances and angles between pairs of key points. Instead, we feed the raw key points into the custom PV-LSTM model. The changes of the position and the rate of those changes along the frames provided useful information for predicting pedestrian crossing intentions. For example, if the keypoints are moving from left to right with a high velocity, the pedestrian is most likely crossing (refer to Fig. 4). If the key points remain to the left or right of the vehicle with a slower velocity, the pedestrian is most likely moving perpendicular to the vehicle and does not intend to cross in front of the vehicle (refer to Fig. 5). To further illustrate this point, the figures in this section and following sections will include pedestrian skeletons based on the predicted keypoints. These skeletons are strictly for illustrative purposes for visualising the movement of the keypoints with respect to one another from frame to frame. During training and evaluation, the proposed technique requires only the raw keypoints.

5. Results & discussions

5.1. Comparison with state-of-the-art

We use the settings from Dollár et al. (2012) and Zhang et al. (2017) for multi-scale intent prediction evaluation (see Table 1). Table 2 compares the accuracy for different multi-scale settings as set in Table 1. The accuracy is calculated by (6), as in Bouhsain et al. (2020). *True Positives* refers to those correctly predicted samples. Accuracy calculated in throughout this section uses this metric, unless stated otherwise. Our proposed method performs with a high level of accuracy for both the reasonable and reasonable (small) categories. As such, we cannot compare our results with current state-of-the-art approaches. Even combining all the settings yields high accuracy results, with JAAD and PIE achieving 87% and 89%, respectively. We compare our results and settings with current state-of-the-art techniques in Table 3. The results in Table 3 compares the results for the *reasonable* setting. As far as we are aware, the approach proposed in this paper is the first for multi-scale intent prediction. Previous papers exclude smaller and heavily occluded pedestrians, focusing on medium to larger pedestrians with minimal occlusion (i.e., *reasonable* setting.) We use 12 frames,

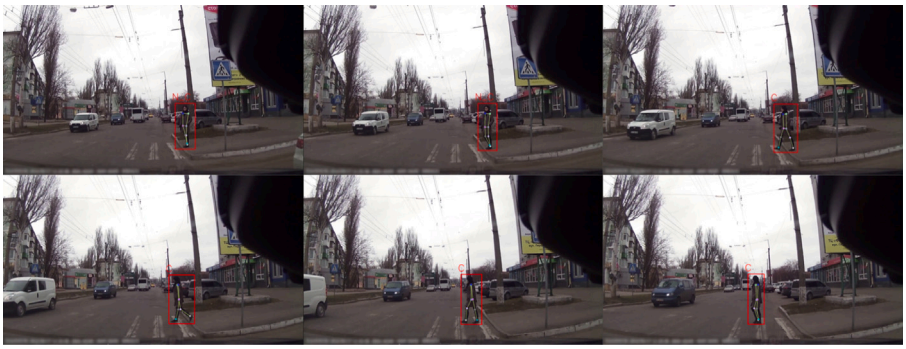


Fig. 8. This figure shows a pedestrian going from a state of not crossing to crossing. The frames go from left to right with each frame being a time-step of 10 frames from the previous frame. The bounding box (red rectangle) is generated using the predicted keypoints (circles of assorted colours).

Table 3
Performance comparison.

Method	Description	Dataset	No. Frames	Accuracy
AlexNet[a]	Ground-truth bounding boxes + Environmental	JAAD	15	63%
3D CNN[b]	Predicted bounding boxes	JAAD	20	85%
PV-LSTM[c]	Predicted bounding boxes	JAAD	14	82%
2D pose estimation[d]	Predicted keypoints	JAAD	14	88%
This work	Predicted keypoints	JAAD	12	89%
This work	Predicted keypoints	PIE	12	**91**%

[a]Rasouli et al. (2018).
[b]Saleh et al. (2019b).
[c]Bouhsain et al. (2020).
[d]Fang and López (2018).

Table 4
Run-time performance comparison.

Method	Performance
ConvNet-Softmax (Saleh et al., 2017a)	28 ms
ConvNet-SVM (Rasouli et al., 2017)	27 ms
ConvNet-LSTM (Carreira & Zisserman, 2017)	40 ms
C3D (Carreira & Zisserman, 2017)	27 ms
ST-Dense-Net (Saleh et al., 2019b)	10 ms
This work	6.1 ms
Trajectory-LSTM (Bouhsain et al., 2020)	4.9 ms
Multi-Task PV-LSTM (Bouhsain et al., 2020)	4.9 ms

which is 2 less than the next best method. This reduces the amount of input data required to predict the pedestrian's crossing intentions. Although the model did not perform as well for the heavy occlusion categories, the performance for the "all" setting (see Table 1) is highly accurate as there was a limited number of heavily occluded pedestrians. We will look to improve this category in Section 5.2. Refer to Fig. 8 for illustration of the predictions made by our proposed method.

$$accuracy = \frac{True\ Positives}{Total\ Samples} \tag{6}$$

In Table 4, we compare the run-time of our architecture with similar state-of-the-art techniques. We calculate a single run-time based on the amount of time it takes for the model to predict keypoints, track and predict the crossing intention for a pedestrian over 12 frames. We take the average run-time for all intent predictions made by the model. We found that we could achieve comparable run-time, even considering our proposed method uses more data points. However, as illustrated in Table 2, these data points provide improved multi-scale intent prediction results. Our proposed method is only 1.1 ms slower than the approach in Bouhsain et al. (2020), while outperforming. It should be noted that the technique in Bouhsain et al. (2020) uses ground-truth bounding boxes to make intent predictions. Whereas, our proposed method consists of keypoint prediction, tracking and intent classification.

5.2. Data generalisation

In Hasan et al. (2020), a novel approach for data generalisation was proposed. The authors referred to this approach as progressive training. Progressive training involves using multiple datasets to improve model generalisation. Generalisation refers to the ability to maintain a high level of performance when seeing previously unseen data. In Hasan et al. (2020), progressive training was used for pedestrian detection. By combining numerous widely used and publicly available datasets. In some cases, they were able to achieve improvements of approximately 10%. We applied the same progressive training to improve upon the previous results (see Table 2). Overall (all categories), we improved the accuracy by 3% and 2% for JAAD and PIE evaluation sets, respectively, compared to results from Table 2. (See Table 5 for full results.) Although the pipeline improved accuracy results, we predict that more datasets would further improve results, particularly for smaller and heavy occluded pedestrians.

5.3. Performance, efficiency and robustness

As illustrated in Table 3, the proposed method outperformed previous state-of-the-art intent prediction techniques while also reducing the amount of spatio-temporal information (i.e., requiring fewer frames) to achieve those levels of performance. When operating an autonomous vehicle in public, speed and efficiency (run-time) is vital, so having an approach that is capable of making accurate predictions in a short amount of time can significantly improve the safety for both the passengers as well as other road users.

In terms of robustness, although the model requires a specific number of frames to make predictions, those frames are not required to be directly subsequent to one another. This is achieved by utilising a variation of the SORT tracker which implements a convolutional neural network (CNN) as introduced by Wojke et al. (2018). This version of the SORT tracker has been pre-trained on a large-scale person re-identification dataset to overcome issues of tracking a pedestrian tracking through occlusions along frames. This means that even if the

Table 5
Dataset generalisation benchmarking.

Method	JAAD			
	Reasonable	Reasonable (small)	Heavy occlusion	All
JAAD	89%	78%	40%	87%
PIE	91%	79%	39%	89%
JAAD + PIE	90%	81%	61%	**90%**
JAAD → PIE	89%	78%	59%	89%
PIE → JAAD	89%	79%	61%	89%
Method	PIE			
	Reasonable	Reasonable (small)	Heavy occlusion	All
JAAD	92%	78%	65%	92%
PIE	91%	81%	43%	89%
JAAD + PIE	95%	82%	69%	**94%**
JAAD → PIE	94%	79%	69%	93%
PIE → JAAD	93%	80%	71%	92%

+ refers to merging the datasets and → refers to pre-training on the first dataset and fine-tuning on the second dataset.

pedestrian is not visible along some frames, those frames are skipped. Once the pedestrian is again visible in the frames, the tracking continues. As long as the number of keypoint instances is 12, the model is able to make predictions. Even in extreme circumstances, such as the heavy occlusion setting, our approach achieves an accuracy of 40% and 43% for JAAD and PIE datasets, respectively. This accuracy is further improved by progressive training by between 20%–30%. In these cases, several frames may be skipped and yet the MS-PIP model was able to achieve accuracy of 71%. This demonstrates the robustness of our approach. We are unaware if other techniques, such as those in Table 5, which also provide such robustness. However, we should mention that this type of robustness can affect the accuracy of intention prediction as if too many frames are missing, it could lead to a lack of usable information to accurately predict whether a pedestrian is going to cross or not cross.

It is worth noting that our approach is a multi-task approach, unlike some of the previous works that we previously discussed. In works, such as Fang and López (2018), Fang et al. (2017), they predicted the intentions of a single pedestrian as a time. This is referred to a single-task approach. However, our multi-task approach considers all the pedestrians in the image (see Fig. 3). This approach is more complex as pedestrians could be varying sizes moving at different speeds and in different directions. There are two distinct reasons why we employed the multi-task approach in this work. The first being that the HigherHRNet is not designed for single-task pose estimation. This is due to the architecture of the bottom-up technique we have previously discussed. The bottom-up approach predicts keypoints for all the pedestrians in the image and then groups the keypoints per pedestrian. This meant we could not focus on one pedestrian at time. The second reason is speed (i.e., run-time). We aim for real-time performance. Single-task intent prediction is slower than multi-task as each pedestrian in an image is considered individually, adding to the run-time. Whereas, considering multiple pedestrians are a time improves on run-time as all the pedestrians in the image are considered at the same time. We discuss and compare the run-time of our proposed method with other techniques in Section 5.1.

5.4. Limitations

As discussed in Section 4.1, there are two relevant datasets which contain pedestrian behaviour annotations: the JAAD and PIE datasets. The JAAD has been widely utilised, whereas the PIE dataset is newer and have not be employed as widely. Due to this, we are limited by the quality of these datasets. Although we have presented state-of-the-art results for multi-scale pedestrian intent prediction, we believe our proposed method could potentially perform better. This is because we are limited by smaller and heavily occluded samples in both datasets. As most previous methods focus on *reasonable* pedestrians, it is understandable that the datasets focus on these settings. Another aspect to be considered is the run-time. Although the performance is comparable to similar technique, it can still be further improved. This could be resolved by removing the dependence on generating keypoint-based bounding boxes for tracking. An approach that uses the keypoints to track the pedestrians is an aspect that will be investigated in future works.

5.5. Ablation study

To justify the settings used in Section 4, we perform an ablation study. This ablation study involved the replacing and/or removing certain aspects from the proposed method and adjusting model parameters. In this way, we demonstrate the effectiveness of the contributions made by this work. During discussions of the results, we focus on reasonable settings as other similar models focus on this setting. This provided a direct comparability with those techniques. For this section, only the PIE dataset it utilised. It provides a larger number of samples as well as more balanced crossing and not crossing instances. The following aspects will be ablated:

1. Compare the various pre-trained HighHRNet architectures and hyperparameters
2. Number of keypoints used by the LSTM model to predict pedestrian speed and velocity
3. Number of frames used for prediction
4. Hidden States for LSTM model

Defaults Settings For our results in Section 5, we initialised the HigherHRNet model with the COCO-w48 backbone, where w48 refers to the model capacity. Based on the keypoints predicted by the HigherHRNet model, we only used specific keypoints (i.e., shoulders, hips, knees and ankles). These 8 keypoints are sent to the custom PV-LSTM model for intention classification.

1: Pose Estimator Weights The HigherHRNet includes pre-trained model weights trained the COCO dataset. The accuracy is calculated by (6) (see 5.1). The model has two parameter capacities, w32 and w48 (i.e., number of parameters in the model). For more information with regards to these model capacities, refer to Cheng et al. (2020). In 6, we compare these capacities. We were unable to achieve comparable results using the HigherHRNet trained on the CrowdPose dataset when compared to the COCO dataset. Therefore, we do not include the HigherHRNet pre-trained on CrowdPose in 6. We found that the larger capacity COCO backbone generally provides better results, when compared to the smaller capacity backbone with reduced accuracy by at least 1%.

2: Keypoints In Table 6, we compare the performance of our proposed method using all the keypoints and using only the keypoints that represent shoulder, hips, knees and ankles (as in Section 5). We found that using all the keypoints (i.e. 17 for COCO) negatively affects the model accuracy. We suspect that this is caused by keypoints, such as elbows, eyes, etc., do not provide useful information with regards to the crossing behaviour of the pedestrians and actually over-complicate the task.

3: Number of Frames The number of frames refers to the number of videos frames used as input and output. We evaluate 16 frames and 10 frames to compare with 12 frames used in 5. 16 frames provide improved results over 10 frames, which can be attributed to the amount of information provided by the increased number of frames. However, these are negligible and do not justify the increased number of frames when compared to the 12 frames used in 3. Increased frames results in increased run-time, which affects the model to function in real-time.

4: PV-LSTM Settings We adjust the hyper-parameters to verify that the settings used in the final results are the optimal for the proposed method. For this aspect, we focus mainly on the hidden state. The hidden state is what allows the PV-LSTM to store information, functioning

Table 6
Ablation results.

Model weights	No. Keypoints	No. Frames	Hidden states	Reasonable	Reasonable (small)	Heavy occlusion	All
COCO (w32)	17	16	128	87%	71%	39%	86%
COCO (w32)	8	16	128	88%	74%	40%	87%
COCO (w32)	17	10	128	83%	69%	37%	84%
COCO (w32)	8	10	128	85%	71%	39%	86%
COCO (w32)	17	16	512	85%	72%	40%	86%
COCO (w32)	8	16	512	87%	73%	41%	88%
COCO (w32)	17	10	512	84%	73%	38%	86%
COCO (w32)	8	10	512	88%	76%	39%	87%
COCO (w48)	17	16	128	88%	77%	40%	87%
COCO (w48)	8	16	128	90%	79%	41%	89%
COCO (w48)	17	10	128	84%	74%	38%	85%
COCO (w48)	8	10	128	86%	77%	40%	86%
COCO (w48)	17	16	512	86%	76%	39%	86%
COCO (w48)	8	16	512	89%	80%	40%	91%
COCO (w48)	17	10	512	87%	75%	39%	84%
COCO (w48)	8	10	512	90%	80%	41%	90%

as the memory unit. It stores the information from the concatenated keypoints. As this allows for the PV-LSTM to predict future pedestrian crossing behaviours, the optimal size of the hidden state is vital. It needs to be able to store enough information to make predictions, while also not being too large or complex, thereby slowing the overall intent prediction system with unnecessary computational complexity.

6. Conclusions and future works

Advancements in pedestrian detection are only the first step in ensuring safety for pedestrians. In this paper, we propose multi-scale pedestrian intent prediction approach. The proposed approach comprises of a combination of 2D pose estimation and LSTMs for predicting pedestrian crossing behaviours. Based on concatenated keypoints over time generated by the pose estimator, the LSTM is able to predict whether a pedestrian will cross or not cross with respect to the Autonomous Vehicle. The concatenated keypoints represent the changes to the pedestrian's joints over time, thus providing the LSTM-based classifier relevant spatio-temporal information to make accurate intention predictions. Based on our proposed method, we have outperformed previous state-of-the-art techniques, achieving 88% and 92% for JAAD and PIE datasets, respectively, while maintaining a comparable runtime of 6.1 ms. For the multi-scale implementation, we achieved an overall accuracy of 87% for the JAAD and 93% for the PIE dataset when including smaller pedestrians (20-pixel height) and heavily occluded pedestrians (visibility <75%). We also apply a technique for data generalisation referred to as progressive training, which provides improved generalisation over multiple datasets. We improved our initial multi-scale results by 3% and 1% for the JAAD and PIE datasets, respectively, while outperforming previous methods by up to 7%.

6.1. Future works

For future works, we intend to utilise colour images with 3D CNNs for improved performances. Colour images provide significant scenes and contextual information, such as traffic lights, traffic signs and other objects, adding to further spatio-temporal information. However, this information increases the overall complexity to our proposed method. However, it may also provide significant gains in accuracy and robustness. Annotations, such as looking, nodding or waving may also be considered for improving the accuracy. Cyclist intent prediction is another aspect that would be useful to explore. Concurrent pedestrian and cyclist detection would provide further safety for both the vehicle's passengers and other road users. However, both the JAAD and PIE datasets focus on pedestrians and do not consider cyclists. There are other datasets, such as in Saleh et al. (2021, 2019a), but those datasets do not provide crossing or not crossing annotations.

CRediT authorship contribution statement

Sarfraz Ahmed: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft. **Ammar Al Bazi:** Writing – review & editing, Supervision. **Chitta Saha:** Writing – review & editing, Supervision. **Sujan Rajbhandari:** Writing – review & editing. **M. Nazmul Huda:** Project administration, Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Only publicly available data were used.

References

Ahmed, S., Huda, M. N., Rajbhandari, S., Saha, C., Elshaw, M., & Kanarachos, S. (2019). Pedestrian and cyclist detection and intent estimation for autonomous vehicles: A survey. *Applied Sciences (Switzerland)*, *9*(11), 1–38.

Ahmed, S., Huda, M. N., Rajbhandari, S., Saha, C., Elshaw, M., & Kanarachos, S. (2019). Visual and thermal data for pedestrian and cyclist detection. In *LNAI: vol. 11650, TAROS 2019: Towards autonomous robotic systems* (pp. 223–234). Springer International Publishing.

Benfold, B., & Reid, I. (2011). Stable multi-target tracking in real-time surveillance video. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 3457–3464). IEEE.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In *Proceedings - International conference on image processing, ICIP. 2016-Augus* (pp. 3464–3468). IEEE, http://dx.doi.org/10.1109/ICIP.2016.7533003.

Bouhsain, S. A., Saadatnejad, S., & Alahi, A. (2020). Pedestrian intention prediction: A multi-task perspective. In *European association for research in transportation conference*.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299–6308).

Chen, K., Song, X., & Ren, X. (2021). Pedestrian trajectory prediction in heterogeneous traffic using pose keypoints-based convolutional encoder-decoder network. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(5), 1764–1775.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). HigherhrNet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 5385–5394).

CMU (2017). *CMU Graphics Lab Motion Capture Database*. Carnegie Mellon University.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Ess, A., Leibe, B., Schindler, K., & van Gool, L. (2009). Robust multiperson tracking from a mobile platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1831–1846. http://dx.doi.org/10.1109/TPAMI.2009.109, URL: https://www.vision.rwth-aachen.de/media/papers/ess08cvpr.pdf.

Fang, Z., & López, A. M. (2018). Is the pedestrian going to cross? Answering by 2D pose estimation. In *IEEE intelligent vehicles symposium, proceedings. 2018-June* (pp. 1271–1276). http://dx.doi.org/10.1109/IVS.2018.8500413.

Fang, Z., Vázquez, D., López, A., Fang, Z., Vázquez, D., & López, A. M. (2017). On-board detection of pedestrian intentions. *Sensors*, *17*(10), 2193.

Galvao, L. G., Abbod, M., Kalganova, T., Palade, V., & Huda, M. N. (2021). Pedestrian and vehicle detection in autonomous vehicle perception systems-a review. *Sensors*, *21*(21), 1–47.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 3354–3361). IEEE.

Gesnouin, J., Pechberti, S., Bresson, G., Stanciulescu, B., & Moutarde, F. (2020). Predicting intentions of pedestrians from 2d skeletal pose sequences with a representation-focused multi-branch deep learning network. *Algorithms*, *13*(12), 1–23. http://dx.doi.org/10.3390/a13120331.

Google (2015). Google self-driving car testing report on disengagements of autonomous mode. In *Google auto LLC. Vol. 92. No. December* (pp. 249–255).

Gujjar, P., & Vaughan, R. (2019). Classifying pedestrian actions in advance using predicted video of urban driving scenes. *Icra*, *2019-May*, 2097–2103.

Hasan, I., Liao, S., Li, J., Akram, S. U., & Shao, L. (2020). Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the conference on computer vision and pattern recognition* (pp. 1–10).

Huang, Z., Hasan, A., Shin, K., Li, R., & Driggs-Campbell, K. (2021). Long-term pedestrian trajectory prediction using mutable intention filter and warp LSTM. *IEEE Robotics and Automation Letters*, *6*(2), 542–549.

Keller, C. G., & Gavrila, D. (2014). Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, *15*, 494–506.

Keller, C. G., Hermes, C., & Gavrila, D. M. (2011). Will the pedestrian cross? Probabilistic path prediction based on learned motion features. In *LNCS: vol. 6835, Lecture notes in computer science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 386–395). http://dx.doi.org/10.1007/978-3-642-23123-0{_}39.

Kim, J. U., Park, S., & Ro, Y. M. (2021). Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 3030–3039). http://dx.doi.org/10.1109/ICCV48922.2021.00304.

Li, X., Liu, Y., Wang, K., & Wang, F. Y. (2020). A recurrent attention and interaction model for pedestrian trajectory prediction. *IEEE/CAA Journal of Automatica Sinica*, *7*(5), 1361–1370.

Li, Y., Lu, X. Y., Wang, J., & Li, K. (2020). Pedestrian trajectory prediction combining probabilistic reasoning and sequence learning. *IEEE Transactions on Intelligent Vehicles*, *5*(3), 461–474.

Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. S., & Lu, C. (2019). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition. 2019-June* (pp. 10855–10864).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *ECCV 2014. Vol. 8693. No. June* (pp. 740–755).

Liu, B., Adeli, E., Cao, Z., Lee, K. H., Shenoi, A., Gaidon, A., & Niebles, J. C. (2020). Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, *5*(2), 3485–3492.

Minguez, R. Q., Alonso, I. P., Fernandez-Llorca, D., & Sotelo, M. A. (2019). Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition. *IEEE Transactions on Intelligent Transportation Systems*, *20*(5), 1803–1814.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C. C., Lee, J. T., Mukherjee, S., Aggarwal, J. K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., ... Desai, M. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition. Vol. 2* (pp. 3153–3160). IEEE.

Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 261–268). IEEE.

Quan, R., Zhu, L., Wu, Y., & Yang, Y. (2021). Holistic LSTM for pedestrian trajectory prediction. *IEEE Transactions on Image Processing*, *30*, 3229–3239.

Rasouli, A., Kotseruba, I., Kunic, T., & Tsotsos, J. (2019). PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE international conference on computer vision. 2019-Octob* (pp. 6261–6270).

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE intelligent vehicles symposium* (pp. 264–269). IEEE, http://dx.doi.org/10.1109/IVS.2017.7995730.

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2018). Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, *3*(1).

Razali, H., Mordan, T., & Alahi, A. (2021). Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation Research Part C (Emerging Technologies)*, *130*(June), Article 103259.

Robicquet, A., Sadeghian, A., Alahi, A., & Savarese, S. (2016). Learning social etiquette: Human trajectory understanding in crowded scenes. In *Lecture notes in computer science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): vol. 9912 LNCS*, (pp. 549–565).

Saleh, K., Abobakr, A., Hossny, M., Nahavandi, D., Iskander, J., Attia, M., & Nahavandi, S. (2021). Fast intent prediction of multi-cyclists in 3D point cloud data using deep neural networks. *Neurocomputing*, *465*, 205–214. http://dx.doi.org/10.1016/j.neucom.2021.09.008.

Saleh, K., Abobakr, A., Nahavandi, D., Iskander, J., Attia, M., Hossny, M., & Nahavandi, S. (2019). Cyclist intent prediction using 3D LIDAR sensors for fully automated vehicles. In *2019 IEEE intelligent transportation systems conference* (pp. 2020–2026). Institute of Electrical and Electronics Engineers Inc. http://dx.doi.org/10.1109/ITSC.2019.8917291.

Saleh, K., Hossny, M., & Nahavandi, S. (2017). Early intent prediction of vulnerable road users from visual attributes using multi-task learning network. In *2017 IEEE international conference on systems, man, and cybernetics, SMC 2017. 2017-Janua* (pp. 3367–3372).

Saleh, K., Hossny, M., & Nahavandi, S. (2017). Intent prediction of vulnerable road users from motion trajectories using stacked LSTM network. In *2017 IEEE 20th international conference on intelligent transportation systems* (pp. 327–332). IEEE, http://dx.doi.org/10.1109/ITSC.2017.8317941.

Saleh, K., Hossny, M., & Nahavandi, S. (2019). Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal dense net. In *Proceedings - IEEE international conference on robotics and automation. 2019-May* (pp. 9704–9710).

Saleh, K., Hossny, M., & Nahavandi, S. (2020). Contextual recurrent predictive model for long-term intent prediction of vulnerable road users. *IEEE Transactions on Intelligent Transportation Systems*, *21*(8), 3398–3408.

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition. 2019-June* (pp. 5686–5696).

Wojke, N., Bewley, A., & Paulus, D. (2018). Simple online and realtime tracking with a deep association metric. In *Proceedings - International conference on image processing. 2017-Septe* (pp. 3645–3649).

Wu, J., Woo, H., Tamura, Y., Moro, A., Massaroli, S., Yamashita, A., & Asama, H. (2019). Pedestrian trajectory prediction using BiRNN encoder–decoder framework. *Advanced Robotics*, *33*(18), 956–969.

Yang, J., Gui, A., Wang, J., & Ma, J. (2021). Pedestrian behavior interpretation from pose estimation. In *IEEE conference on intelligent transportation systems, proceedings. 2021-Septe* (pp. 3110–3115). Institute of Electrical and Electronics Engineers Inc. http://dx.doi.org/10.1109/ITSC48978.2021.9565098.

Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., & Yan, J. (2016). POI: Multiple object tracking with high performance detection and appearance feature. In *LNCS: vol. 9914, ECCV: European conference on computer vision* (pp. 36–42). Springer Verlag, http://dx.doi.org/10.1007/978-3-319-48881-3{_}3.

Zhang, S., Benenson, R., & Schiele, B. (2017). CityPersons: A diverse dataset for pedestrian detection. In *Proceedings - 30th IEEE conference on computer vision and pattern recognition. 2017-Janua* (pp. 4457–4465).

Zhou, B., Wang, X., & Tang, X. (2012). Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 2871–2878). IEEE.