algorithms

MDPI

Article

Predicting Intentions of Pedestrians from 2D Skeletal Pose Sequences with a Representation-Focused Multi-Branch Deep Learning Network

Joseph Gesnouin [1,2,*], Steve Pechberti [1], Guillaume Bresson [1], Bogdan Stanciulescu [2] and Fabien Moutarde [2]

[1] Institut VEDECOM—Versailles, 78000 Versailles, France; steve.pechberti@vedecom.fr (S.P.); guillaume.bresson@vedecom.fr (G.B.)

[2] Centre de Robotique, MINES ParisTech, Université PSL, 75006 Paris, France; bogdan.stanciulescu@mines-paristech.fr (B.S.); fabien.moutarde@mines-paristech.fr (F.M.)

* Correspondence: joseph.gesnouin@mines-paristech.fr

Received: 30 October 2020; Accepted: 7 December 2020; Published: 10 December 2020

Abstract: Understanding the behaviors and intentions of humans is still one of the main challenges for vehicle autonomy. More specifically, inferring the intentions and actions of vulnerable actors, namely pedestrians, in complex situations such as urban traffic scenes remains a difficult task and a blocking point towards more automated vehicles. Answering the question "Is the pedestrian going to cross?" is a good starting point in order to advance in the quest to the fifth level of autonomous driving. In this paper, we address the problem of real-time discrete intention prediction of pedestrians in urban traffic environments by linking the dynamics of a pedestrian's skeleton to an intention. Hence, we propose SPI-Net (Skeleton-based Pedestrian Intention network): a representation-focused multi-branch network combining features from 2D pedestrian body poses for the prediction of pedestrians' discrete intentions. Experimental results show that SPI-Net achieved 94.4% accuracy in pedestrian crossing prediction on the JAAD data set while being efficient for real-time scenarios since SPI-Net can reach around one inference every 0.25 ms on one GPU (i.e., RTX 2080ti), or every 0.67 ms on one CPU (i.e., Intel Core i7 8700K).

Keywords: skeleton-based action prediction; pedestrian intention prediction; body action; human activity; action and gesture recognition; mobility analysis

1. Introduction

Within the context of autonomous vehicle development and the field of Advanced Driver Assistance Systems (ADAS), determining the pedestrians' discrete intention is mandatory. From this information, their trajectory can be further estimated to understand the pedestrians' next actions or positions, which can greatly reduce the risk of accidents. For instance, knowing the intention of pedestrians to cross the road before they actually set foot on the road would allow the vehicle to warn the driver or automatically perform maneuvers. Therefore, preserving the pedestrians' integrity in a more efficient way than when triggered by an emergency stop once the pedestrians have moved on to the road and become a direct obstacle for the vehicle would be safer for all actors. In such decisive applications, a desirable intention prediction model should run efficiently for real-time usage and should also be robust to a multitude of complexities and conditions (e.g., weather, location).

Human action recognition applied to video is a difficult research topic due to the great variation and complexity of the input data. Currently, the main modalities used for these tasks include RGB videos in their entirety [1–4], optical flow [5–8] and skeleton form modeling [9–12]. The latter requires, prior to the action classification, an approach to estimate the human pose [13]. By focusing our work on the pedestrian skeleton structure shown in Figure 1 and since no background context is included, we are able to represent the invariant information of an action. Moreover, we are also reducing the size of the input data, which makes our approach more efficient once the pedestrian body pose is inferred.

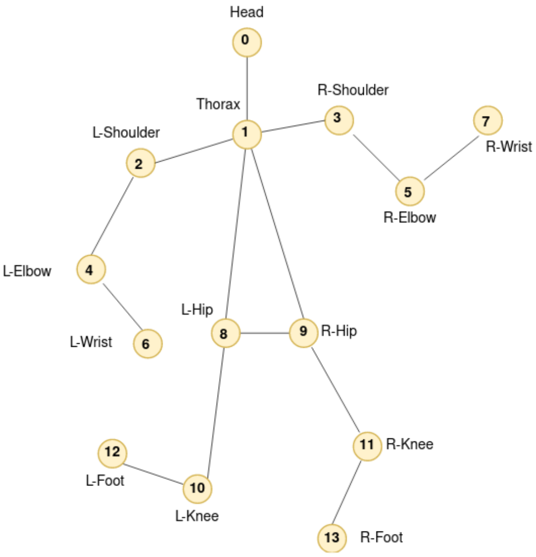

Figure 1. This figure is a non-weighted and undirected graph representation of the human skeleton, with all the selected key-points used for our pedestrian crossing prediction approach. In the present work, only 14 human key-points are chosen from the ones extracted with the Cascaded Pyramid Network (CPN) [14].

In this paper, we propose a real-time and context-invariant approach based on 2D pedestrian body poses to address the Crossing/Not Crossing (C/NC) prediction in realistic driving conditions. We evaluate the proposed approach on the Joint Attention in Autonomous Driving (JAAD) [15,16] public data set, a standard benchmark in the matter of pedestrian behaviors prediction. To proceed, we collect human key-points for all video frames in the JAAD data set thanks to skeleton IDs and associated spatial coordinates defined in JAAD annotations. Thereafter, we develop SPI-Net: a representation-focused multi-branch network that combines Cartesian features and location-invariant geometric skeleton features. Skeleton features are provided by the Cascaded Pyramid Network (CPN) algorithm [14].

The network is divided into two branches: one focuses on the evolution of Euclidean distances relative to certain identified key-points over time, the other focuses on the evolution of the spatial representation of skeletal key-points as a function of time in the Cartesian coordinate system.

The first branch corresponds to the encoder part of an auto-encoder initially trained to reconstruct an action according to the evolution over time of selected key-point distances. We add to the auto-encoder cost function a statistical supervised separability constraint to perform better separation between instances according to their class in the latent space. We obtain, in addition to a reduction of the action representation size, a first draft of class separability in the latent space.

In the second branch, a 2D convolutional network, we represent a skeleton sequence as a pseudo-image. This allows us to extract spatio-temporal features using standard computer vision deep-learning methods.

We then perform a late fusion on those two branches and fine-tune the entire approach in order to evaluate the model. According to experiments, SPI-Net achieved 94.4% accuracy in pedestrian crossing prediction. To the best of our knowledge, SPI-Net is at the moment, the state-of-the-art for the C/NC task on the JAAD data set.

The paper is organized as follows: Section 2 outlines some existing approaches from the literature for skeleton-based action recognition and pedestrians' intentions prediction. Section 3 presents an overview of SPI-Net. Section 4 describes the experiments setups and shows the results of SPI-Net on the JAAD data set. In Section 5, we discuss the results and how they can be interpreted for future works. Finally, Section 6 presents our conclusions.

2. Related Work

In this section, we first review the literature for skeleton-based human action recognition. We then focus on the literature for the prediction of pedestrians' intentions in the context of autonomous driving.

2.1. Skeleton-Based Human Action Recognition

We review here a list of approaches based on "in-depth" learning of action recognition on skeletal data. Those approaches for skeleton-based action recognition can be split into four categories:

- The ones that make use of recurrent cells;
- The ones that make use of convolutional cells;
- The ones that make use of an attention mechanism;
- The ones that do not focus on Euclidean data structure but a graph data structure.

2.1.1. Recurrent Neural Networks (RNN)

In recent years, Recurrent Neural Networks have been the reference approaches for sequence modeling in speech recognition, digital signal processing, video processing and natural language processing. Similarly, most deep-learning approaches for gesture recognition also use recurrent cells such as LSTMs [17] or GRUs [18]. In those approaches, the skeleton is represented in the form of a sequence and state neural networks are applied to it. They allow the exhibition of temporal dynamic behaviors. For instance, Baccouche et al. [19] propose an architecture using LSTMs for action recognition. Avola et al. [20] exploit the geometric characteristics of the angles of the joints learned with an LSTM architecture. From the articulations information, Zhang et al. [21] generate eight geometric indicators and evaluate them with a three-layer LSTM network. Du et al. [10] divide the human skeleton into five parts and then propose a sequential hierarchical approach. Shukla et al. [22] propose a hierarchical recurrent architecture roughly equivalent to Du et al. [10] but reduce the number of joints at the input of the model, some of them being considered superfluous and carrying little information. This reduction in the number of input joints then leads to a reduced set of parameters and reduces the model inference time without degrading the quality of the classifier. Shahroudy et al. [23] use an LSTM approach based on long-term learning of the co-occurrences of joints intrinsically characterizing human actions. Zhang et al. [24] propose an adaptive recurrent network with an LSTM architecture, allowing the network to adapt to the most appropriate end-to-end observational viewpoints in order to manage large variations in the orientation of actions.

2.1.2. Convolutional Neural Network (CNN)

Since recurrent cells are relatively slow and difficult to train and use in real-time compared to convolutional approaches, the latter have become an interesting solution given their advantages in terms of parallel computing, efficiency in learning characteristics and speed. Convolutional approaches can be performed on skeletons represented as pseudo-images, as illustrated in Figure 2, so that standard 2D convolutions can be applied, or any other spatio-temporal version of CNNs such as 3D convolutions. Since skeletal data are small elements, it is possible to organize a sequence of skeletal features chronologically in an image, which retains the original information of the skeletal dynamics.

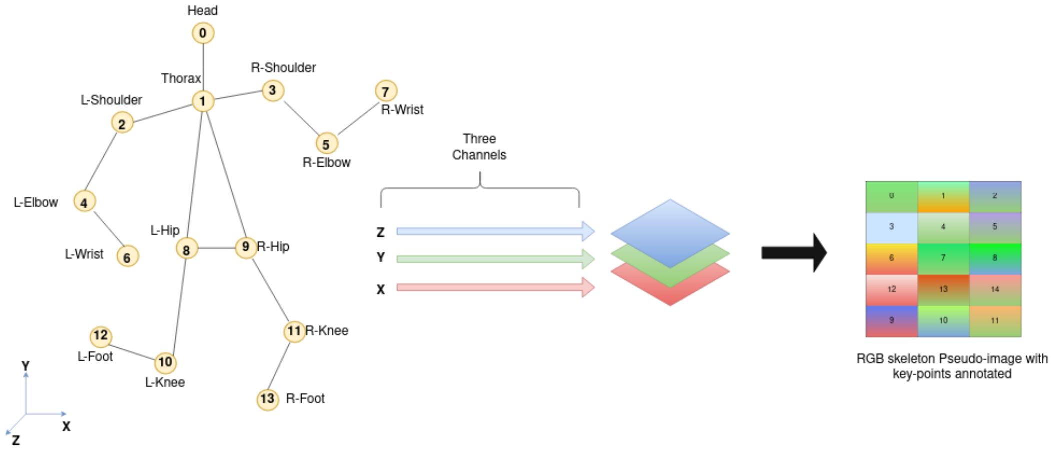

Figure 2. Organization of the 3D skeleton data structure into a three-channel image (RGB).

The general idea of this type of approach is to structure the data in order to give them the expected form (a sequence of images) and thus classify these images using standard computer vision methods. Such motion formalisms to represent skeletal sequences by compact image-like inputs were first proposed by Elias et al. [25] and extended by Sedmidubsky et al. [26] where a special insistence has been given to features representation and data normalization to improve instance indexing.

In the same regard, Ke et al. [27] propose to transform a skeleton sequence into three video clips, the CNN characteristics of the three clips are then merged into a single characteristics vector, which is finally sent to a softmax function for classification. Pham et al. [28] propose to use a residual network [29] with the transformed normalized skeleton in the RGB space as the input. Cao et al. [30] propose to classify the image obtained thanks to gated convolutions. Ludl et al. [31] propose a complete pipeline capable in real-time to detect a human in an image, skeletonize it and determine the action performed using the same encoding format as an RGB image.

By moving away from the image domain while keeping the notion of convolution, other CNN-based approaches use them in 1D format to model sequences: Bai et al. [32] show that convolution networks can match or even surpass the performance of recurrent networks for typical sequential modeling tasks. Therefore, Devineau et al. [33] propose an architecture based on parallel convolutions capable of capturing features at different temporal resolutions. This results in a three-branch convolutional model that takes as input the positions of skeletal joints at different speeds and the distances in pairs between joints. Weng et al. [34] propose a deformable convolutional neural network with one-dimensional convolutions capable of discovering combinations of information-carrying joints to avoid joints in which semantics contribute little to the model.

Recurrent networks and convolutional networks can also be combined. The approach consists of extracting spatial information with convolutive layers, then modeling temporal dynamics with recurrent layers. Thus, Donahue et al. [1] propose to extract visual information from images coming from a video thanks to a 2D CNN then to send them to the input of an LSTM. Li et al. [35] propose a late fusion approach where LSTM and CNN are merged. Ullah et al. [36] propose a bidirectional approach where features obtained by CNN are sent in a bidirectional LSTM [37], connecting two hidden layers from opposite directions to the same output. The output layer can then simultaneously obtain information on past and future states.

2.1.3. Attention Mechanisms

Human perception focuses on the most relevant parts of an image in order to acquire information to understand its semantics. For machine learning, this phenomenon is artificially recreated by a mechanism of attention: conceptually, attention can be interpreted in a broad sense as a vector of weights of importance. In the context of action recognition, attention can be used to weight the importance of certain moments of the action in order to classify it, or to weight the importance of certain skeletal joints. Maghoumi et al. [38] propose to stack GRUs with a global attention mechanism as well as two fully connected layers. Song et al. [39] propose a model based on LSTM and RNN and combine spatial and temporal global attentions: a network focuses on the discriminating articulations of each frame, the other network weights the attention levels of the results for each instant in order to focus on the important frames. Fan et al. [40] use action information from multiple viewpoints to improve recognition performance and provide an attention mechanism for multi-view fusion of skeletons sent to LSTMs. It is also possible to use attention with convolutions. Thus, Hou et al. [41] propose a convolutional network learning different levels of attention for each spatio-temporal feature extracted by the convolution filters for each frame of the sequence.

2.1.4. Geometric Deep-Learning

The evolution over time of the skeleton of the human body can be considered in the form of a dynamic graph. So far, research in deep-learning for action recognition on skeletal data has focused mainly on Euclidean data. The non-Euclidean nature of data in graph format makes the use of basic operations, such as convolution, difficult to perform. However, convolutions have by definition the ability to extract local spatial features and could use the skeleton data structure in graph format for the classification of human actions. Such ability fits perfectly to Graph-type data structures since they are, by definition, locally connected structures: the set of neighbors of a node.

In this way, representing the skeleton in the form of a graph can have the advantage of not exploiting non-existent neighborhood links between joints, but of preserving coherent spatial semantics for the skeleton. Geometric Deep-Learning [42–44] refers to techniques attempting to generalize deep structured neural networks to non-Euclidean domains such as graphs. Wu et al. [45] provide a state-of-the-art on geometric deep-learning and propose a taxonomy to differentiate geometric networks into four categories: recurrent, convolutional, auto-encoder and spatio-temporal. Thus, Zhang et al. [46] propose to apply convolutions on the edges of a graph corresponding to skeletal bones in order to preserve spatial semantics. Yan et al. [47] extend the spatial convolutions of graphs into spatio-temporal convolutions. They propose a convolutional spatio-temporal approach including time-bound joints in the convolutional block in addition to spatially bound joints. Li et al. [48] propose to cumulate spatio-temporal convolutions with an autoregressive–moving-average model. Finally, Si et al. [49] propose to cumulate attention to a CNN-LSTM geometric network, capitalizing all the approaches presented previously in a single network.

2.2. Pedestrian Intention Prediction

The problem of intention prediction of pedestrians from image sequences has gained increasing interest over the past few years. More specifically, in the context of autonomous cars in urban traffic environments, this problem is still an active research area due to its complexity and importance: while action recognition consists of using a complete sequence to label an action, intentions prediction predicts from an incomplete sequence to label an intention (i.e., before the pedestrian crosses).

For intention prediction tasks, it is common to split the intention as a combination of high-level discrete behaviors as well as continuous trajectories describing the expected future movement of the pedestrian. In this paper, we address the Crossing/Not Crossing discrete intention task in realistic conditions. We, therefore, review published state-of-the-art machine-learning approaches for the prediction of pedestrian's discrete intentions in the context of autonomous driving.

Rasouli et al. [15] and Varytimidis et al. [50] formulated the problem as an image classification problem based on Alexnet [51]. Given a single position image of a pedestrian in a traffic scene, they classify whether the pedestrian is crossing or not. Afterward, they extend their model in order to take as input a sequence of consecutive cropped images of the pedestrians before they cross in order to consider the temporal coherence in short-term motions (≈0.5 s). Similarly, Saleh et al. [52] propose to predict the intended actions of pedestrians based on a spatio-temporal DenseNet model. Pop et al. [8]

propose to extract spatial information with convolutive layers, then consider temporal dynamics with recurrent layers and propose a new metric for pedestrians dynamics evaluation: the time to cross (TTC) prediction.

Some works are based on state-of-the-art generative methods in deep-learning, focusing on the future representation of the action, and then classify it in its globality: Gujjar et al. [53] and Chaabane et al. [54] process the crossing actions classification by feeding the predicted frames of their future frame prediction auto-encoder network into a classification network. However, those kinds of approaches have a major drawback: since background context is included, they are noise sensitive. Moreover, predicting future frames of a given scene can be time-consuming considering the type and the structure of the approach proposed, which can be a bit delicate in a real-time scenario.

To overcome these issues, intention prediction based on 2D skeletal pose sequences has also been explored. Most of those approaches are currently based on the assumption that one can link the position of a pedestrian's joints previous to his action to an intention. Consequently, the problem of pedestrian discrete intention prediction is, therefore, dealt with as an action recognition task prior to the action and the action label becomes the intention. Fang et al. [55] combined CNN-based pedestrian detection, tracking and pose estimation to predict the crossing action from monocular images. Marginean et al. [56] and Ghori et al. [57] explore the pedestrian intention prediction task with pose estimation algorithms combined with recurrent networks. However, in [57], sequences in the wild are used, which makes it difficult to evaluate their approach on the JAAD data set. Cadena et al. [58] predict intentions of pedestrians crossing based on 2D skeletal pose estimation and a Graph Convolutional Network that preserves coherent spatial semantics for the pedestrian skeleton.

As noted by Ridel et al. [59], a further step to improve the state-of-the-art of pedestrian intention prediction would be the introduction of dynamics and contextual scene information, jointly with pedestrian-specific characteristics. However, the generated context features are not always applicable to current data sets and require different kinds of modalities or hardware constraints [60,61]. In that regard, Liu et al. [62] propose a new data set for pedestrian intention prediction tailored to intent prediction in dense driving scenes. It defines a model based on graph convolutions to represent the spatiotemporal context of the scene where each identified object is presented as a node of a spatiotemporal graph for two different perspectives: pedestrian-centric and location-centric settings graphs. Range et al. [63] propose a multi-task learning model to predict pedestrian actions and crossing intents. They forecast their future path from video sequences based on 2D skeletal pose sequences, context and two geometric features based on head orientation and arms orientation. While we firmly believe that for the task of pedestrian intention prediction, it is better to use pedestrian specific dynamics information and contextual scenes conjointly, we propose in this paper a context-invariant approach based on 2D pedestrian body pose only to address the C/NC task. In that regard, we aim at dividing the task of intention prediction into different sub-tasks in order to optimize how each modality and potential input should be used independently. Therefore, our proposed approach SPI-Net only relies on 2D skeletal pose sequences.

3. Materials and Methods

3.1. Experimental Data Set

Predicting whether or not a pedestrian is going to cross is addressed by the JAAD data set [15,16], which contains 346 videos. In each video, each pedestrian has its individual ID and its actions performed over time, as Figure 3 shows.

To extract the human key-points, we apply the Cascaded Pyramid Network (CPN [14]) algorithm to the ground truth spatial coordinates and individual IDs of each pedestrian provided by the data set. All video frames are normalized to 1280x1024 frame size. We then normalize each key-point $(x, y) \in \mathbb{R}^2$ individually, dividing each coordinate by 1280 and 1024, as shown in Equation (1):

$$x' = \frac{x}{x_{\max}} \quad ; \quad y' = \frac{y}{y_{\max}} \tag{1}$$

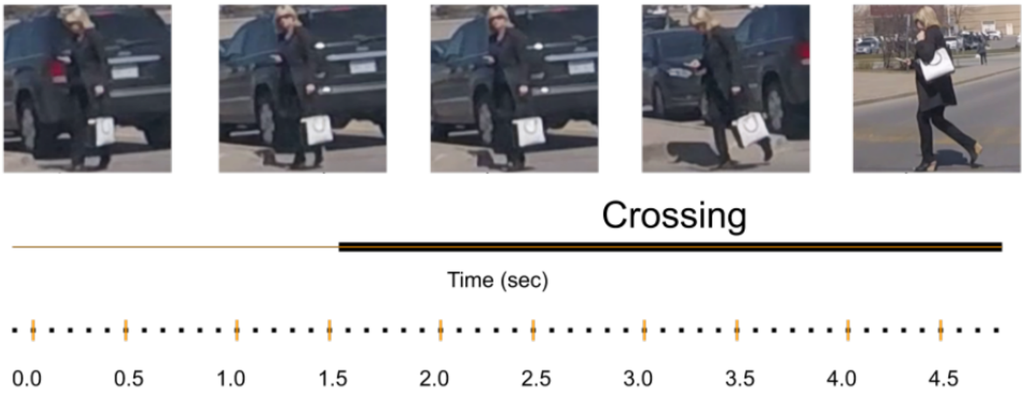

Figure 3. Time line of a crossing pedestrian in the Joint Attention in Autonomous Driving (JAAD) data set.

Such normalization has two benefits: the first one is that data will be ready for neural networks in which weights initialization [64] expects such normalized input (variance ≤ 1), while retaining the spatial information of the pedestrian in the scene.

Subsequently, obtained pedestrian pose sequences are defined as: $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T) \in \mathbf{R}^{T \times K \times d}$, where T is the sequence duration, K is the count of key-points, and d is the dimension of each key-point. All sequences of skeletons are then sampled by a sliding window to keep a fixed size in the form of a 3-dimensional (T, K, d)-shaped tensor where $T = 300$, $K = 14$, and $d = 2$. The majority of the extracted sequences are smaller than the fixed T size of the sliding window, therefore sequences with less than T frames are padded with zeros. Finally, all processed data are introduced as a complete sequence to the SPI-network.

3.2. SPI-Net Architecture

The network architecture of SPI-Net is shown in Figure 4. In the following, we explain our motivation for designing input features and network structures of SPI-Net.

A large majority of current research in skeleton-based action recognition and pedestrian intention prediction focuses mainly on the sequential modeling part of the problem. Moreover, the architectures of the approaches presented for action recognition and intentions prediction have become more and more complex over the years. We think this is because those approaches rely heavily on deep-learning networks to learn informative representations of data itself by adding hidden layers in the architectures.

In this work, we propose to go back to "It is all about embedding and standardization in machine-learning" : once one finds a way to standardize and represent data in a more adequate way, any classifier might be able to obtain good results as long as the input data is informative. By normalizing the input data, creating global-motion features and location-viewpoint invariant features or enforcing certain constraints towards the data representation of designated hidden layers, we send informative-representation ready data to the classification network. It allows us to rely on less hidden layers to learn informative representations of data and therefore reduce the complexity of the network compared to other approaches. Since we choose to rely on a reduced number of hidden layers, we can focus on the inference time of our model, which is mandatory since we take the model speed as one of our priorities.

Finally, we show with SPI-Net's architecture composed of Dense layers and 2D convolutions that combining sequential modeling and taking into consideration the representation and the normalization of an action can be quite effective for the overall accuracy of a network designed for the C/NC task for the JAAD data set.

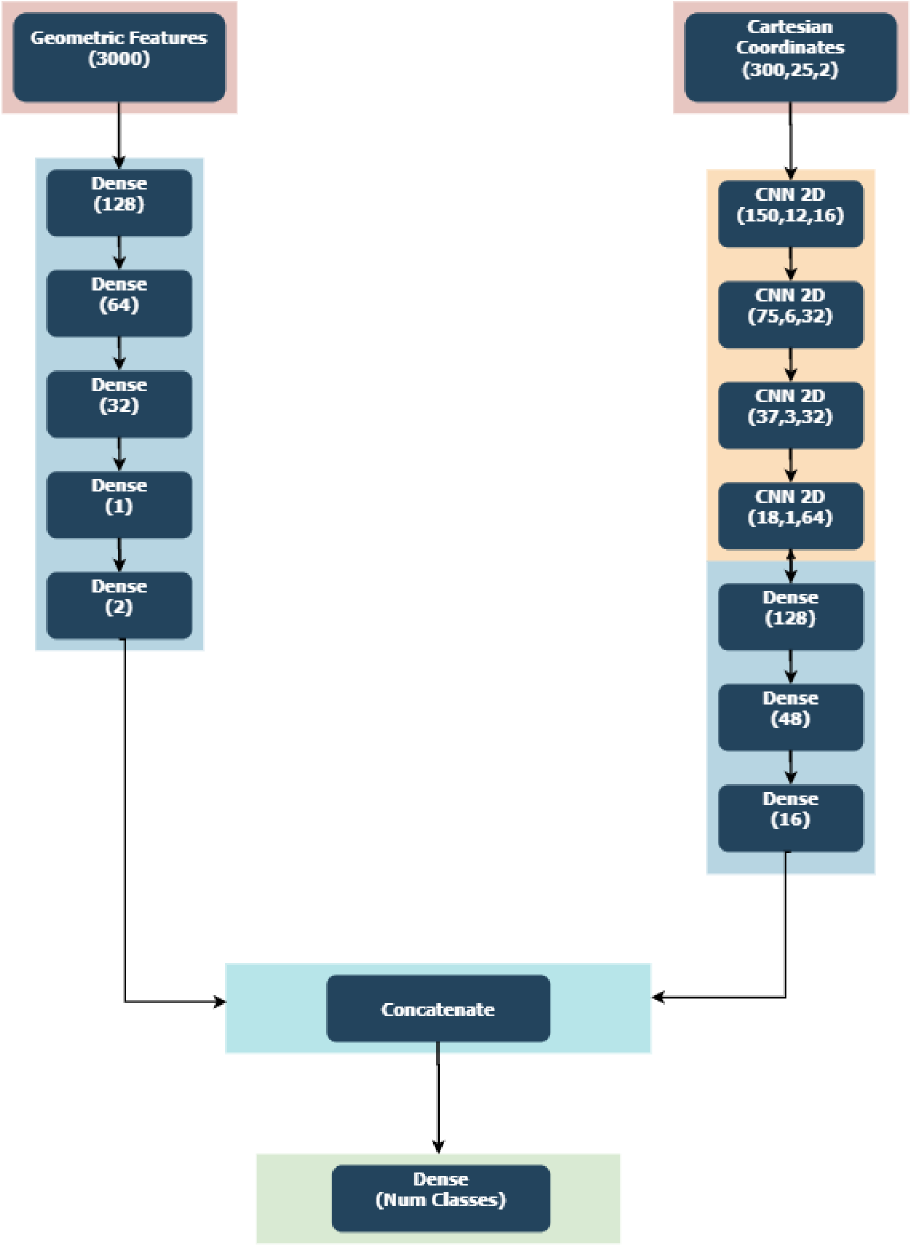

Figure 4. The multi-branch architecture of SPI-Net: the left branch focuses on the evolution of Geometric features relative to certain identified key-points over time. The second one focuses on the evolution of the spatial representation of skeletal key-points as a function of time in the Cartesian coordinate system. CNN 2D Blocks denote one 2D ConvNet layer (kernel size = 3), an AveragePooling layer and a Batchnormalization layer. Other Dense blocks are defined in the same format with a Batchnormalization layer following each Dense layer.

3.2.1. Geometric Features Branch

According to the universal approximation theorem [65], any bounded function can be approximated as well as one wants with a shallow Neural Network containing only one hidden layer. As such, one may even use a trivial feed-forward neural network such as a Multi-Layer Perceptron (MLP) to model sequences, like any other type of data.

For the Geometric Features branch, we use the simplest form of an auto-encoder: a trivial feed-forward non-recurrent neural network to reconstruct an action according to the evolution of the Euclidean distances of five given key-points over time: Torso, Left and Right Shoulders, Left and Right

Knees. The given key-points were selected in order to extract specific information for the model such as pedestrian's orientation or pedestrian's dynamics over time.

A considerable amount of literature has been published on modeling pedestrian's attention towards the environment as an input to infer their crossing intention [66–72], mainly by focusing on specific key-points such as the head and more specifically its orientation. Rasouli et al. [70] show that across all the possible forms of attention and communication a pedestrian could use, the most notable one is to look in the direction of the approaching vehicle: for a collision incoming within the next few seconds, pedestrians always tend to look at the vehicle before crossing [16]. Therefore, such head orientation input is not necessarily useful for the particular task of intention prediction since it is almost always recurrent information. In that regard, Schulz et al. [69] report that head detection is not particularly useful for the particular task of intention prediction. Similar results were reported in [15]: specifically focusing on the head for modeling pedestrian's attention does not seem to bring better performance for the task of intention prediction.

Key-points such as elbows or wrists were considered as well in order to capture specific attention behaviors of pedestrians relying on hand gestures to communicate their intention of crossing to the driver. However, it has been shown that pedestrians mainly use explicit communication such as hand gestures to signal gratitude or dissatisfaction following the driver's action [71]. Such a specific gesture happens too late for our current intention prediction task as the pedestrian would be already either crossing or not at that time.

In fact, Schneemann et al. [72] discovered that evident attention indicators used by humans for inferring crossing intentions such as the head orientation of pedestrians are not always sufficient. Even more, they concluded that *"a lack of information about the pedestrian's posture and body movement results in a delayed detection of the pedestrians changing their crossing intention"*.

In conformity with this conclusion, we chose to capture different information for the Geometric features branch. Instead of extracting pedestrian's awareness features towards its environment, we try to capture pedestrian's orientation features and pedestrian's dynamics features over time based on relative distances of their key-points.

Therefore the torso and shoulders key-points are preferred over the head, elbows or wrists to model the pedestrian orientation towards his environment. Furthermore, knee key-points are taken into consideration in order to determine if the given pedestrian is walking or standing in the scene and therefore capture the dynamics.

By selecting a lower amount of key-points than the ones available in the complete body structure, we reduce the inference time of the Geometric features branch without degrading its quality for classification.

To avoid redundancy in our distances matrix and to minimize the geometric branch input size, we use the Joint Collection Distances (JCD) [35,73] feature to represent our vector of distances over time. This gives us a one-dimensional distance vector as our branch input of size equal to 3000 for each sequence: $T * \binom{nb_{keypoints}}{2} = 300 * \binom{5}{2}$.

We add to the reconstruction cost function of the auto-encoder a statistical supervised constraint specific to the separability of classes with a Linear Discriminant Analysis. This allows to condition the projection of the instances in the latent space upon their class. We then obtain, in addition to a reduced representation of the action, a first draft of the separability of the classes in the latent space. We extract the encoder part of the trained auto-encoder and evaluate its classification ability, as shown in Figure 5.

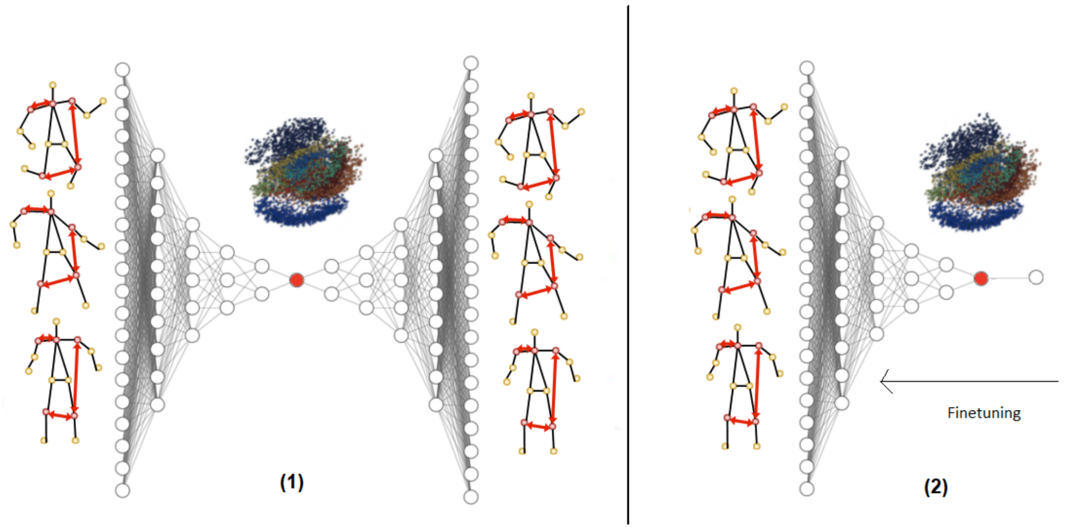

Figure 5. Pipeline of the approach for the Geometric branch: (**1**) we train an auto-encoder to reconstruct a sequence representing an action according to the evolution over time of the distances (represented by the red arrows) of selected keypoints (Torso, Left and Right Shoulders, Left and Right Knees). We also add a constraint specific to the separability of classes in the latent space. (**2**) We then extract the weights of the encoder part up to the bottleneck represented in red and add a classifier, which transforms the encoder part into a pre-trained network on the data for action classification.

More specifically, we obtain an auto-encoder with a separability constraint term that focuses on two completely different pieces of information in the data. One in an unsupervised manner, the other in a supervised manner:

- The inherent structure of the data captured in an unsupervised manner thanks to the reconstruction of the auto-encoder and its abstraction ability. Some of the important and discriminating information in the data set would then be retained.
- The separability of classes thanks to Linear Discriminant Analysis projection of the instances in the latent space.

Formally, we define the problem as follows:

$$\min_{\theta_1,\theta_2}, \left\|\mathbf{X} - g_{\theta_2}\left(f_{\theta 1}(\mathbf{X})\right)\right\|^2 \tag{2}$$

Equation (2) is the usual reconstruction function of an auto-encoder with **X** a data matrix, θ_1, θ_2 the parameters of the encoder and decoder blocks and $f(), g()$ are, respectively, the transition functions such that:

$$\begin{aligned} f_{\theta 1} &: \mathbf{X} \rightarrow \mathcal{F} \\ g_{\theta_2} &: \mathcal{F} \rightarrow \mathbf{X} \end{aligned} \tag{3}$$

where $\mathcal{F}$ is the feature space, which can be regarded as a compressed representation of the input matrix **X**. Throughout that study, we refer to $\mathcal{F}$ as the bottleneck or the latent space of the auto-encoder.

We then add a statistical supervised constraint specific to the separability of classes in the cost function: with **S** being the projection matrix of the instances in the latent space obtained with a linear discriminant analysis (LDA) and λ a weighting parameter as presented in Equation (4):

$$\min_{\theta_1,\theta_2,\mathbf{S}} \left\|\mathbf{X} - g_{\theta_2}\left(f_{\theta_1}(\mathbf{X})\right)\right\|^2 + \lambda \left\|f_{\theta_1}(\mathbf{X}) - \mathbf{S}_{f_{\theta 1}}(\mathbf{X})\right\|^2 \tag{4}$$

The training method is a simple iterative algorithm, optimizing an appropriate objective function. This algorithm is based on two updating steps according to the scheme written in Algorithm 1:

```
Algorithm 1: Auto-encoder with statistical separability constraint training algorithm
  Input: data matrix $\mathbf{X}$, ground truth labels y, weighting parameter $\lambda$, loss threshold $\varepsilon$
  Initialization of the encoder and decoder parameters $\theta_1$ and $\theta_2$;
  while $\left\|\mathbf{X} - g_{\theta_2}\left(f_{\theta_1}(\mathbf{X})\right)\right\|^2 + \lambda \left\|f_{\theta_1}(\mathbf{X}) - \mathbf{S}_{f_{\theta_1}}(\mathbf{X})\right\|^2 > \varepsilon$ do
      Update $\theta_1$ and $\theta_2$ using the auto-encoder.;
      Update S using Linear Discriminant Analysis on $f_{\theta 1}(\mathbf{X})$ data matrix and y.;
  end
  Result: $\theta_1$, parameters of the encoder block
```

We choose the value of λ for the supervised separability constraint part empirically, by modifying its value for different trainings and evaluate its gain for later stages.

Once the training of the auto-encoder has been performed, we recover the weights of the encoder part: θ_1 and add perceptrons with a softmax activation function right after the bottleneck. In order to use the encoder as a classifier, we train the given modified network with the Categorical Cross-Entropy Loss Function:

$$\text{Loss} = -\sum_{i=1}^{\textit{output size}} y_i \cdot \log \hat{y}_i \tag{5}$$

We evaluate the capability of classification of that branch alone for the C/NC task and we then concatenate the two branches to evaluate the approach as its whole.

3.2.2. Cartesian Coordinates Features Branch

As the Geometric branch only takes as input relative Euclidean distances between key-points, the Geometric branch is location-viewpoint invariant. Hence, it does not contain any global spatial motion information of the pedestrian. Solely using the Geometric feature branch is therefore unsubstantial as it does not take any information about the spatial information of the pedestrian in the scene. To overcome this issue, we develop a Cartesian Coordinates features branch that is made to retain such spatial information. Moreover, the Geometric features branch treats no explicit sequential modeling at all, but only treats the question of representation of an action in the embedding. Our Cartesian Coordinates features branch is therefore designed to extract both spatial and temporal features: features that are not explicitly learned in the Geometric branch.

Since we take the model speed as one of our priorities, we use a 2D-convolution-ready representation format of the sequence to represent human pose sequences allowing us to extract spatio-temporal features using standard computer-vision deep-learning methods. Human pose sequences are converted to a 2D image-like spatio-temporal continuous representation based on a spatial joint reordering trick [74,75] called Tree Structure Skeleton Image (TSSI) [76]. Such representation preserves both spatial and temporal relationships by repeating the joints and re-indexing them. TSSI is described in more detail in Figure 6.

Since a sequence is represented with a 3-dimensional $(300, 14, 2)$-shaped tensor, we can easily apply the TSSI normalization [76] on the input and transform the original sequences into a multi-channel redundant image of shape $(300, 25, 2)$. A few sequences of pedestrian actions in the TSSI-format are plotted with their ground truth intentions in Figure 7 for illustration.

We then classify these images using standard computer vision deep-learning methods, such as in [27–31], while preserving spatial and temporal relationships.

Therefore, after the normalization of its input, the second branch corresponds to any other image classifier based on convolutions and pooling blocks for features extractions and fully-connected layers at later stages of the network. Similarly to the Geometric features branch, we evaluate the capability of discrimination of that branch alone for the C/NC task and we then concatenate the two branches and evaluate the approach as its whole.

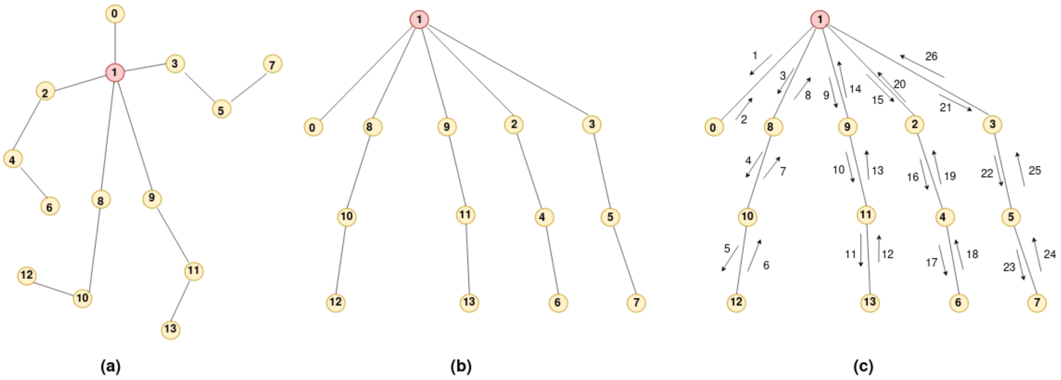

Figure 6. (**a**) Joints of the skeleton of a human body with the initial data structure. The visiting order of the nodes is incremental: 0-1-2-3-...-13. (**b**) The skeleton is transformed into a tree structure. (**c**) The tree can be unfolded into a chain in which the order of visit of the nodes maintains the physical relationship of the joints: 1-0-1-8-10-12-10-8-1-9-11-13-11-9-1-2-4-6-4-2-1-3-5-7-5-3.

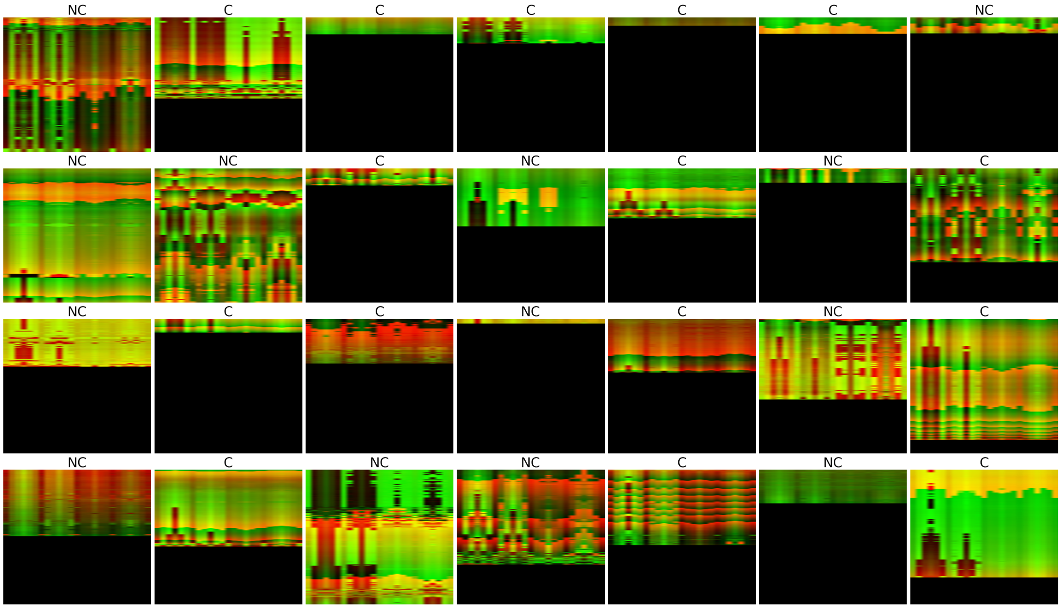

Figure 7. Twenty-eight different ground-truth sequences represented in a 3-dimensional (300,25,2)-shaped tensor after the TSSI normalization. The horizontal axis of each Tree Structure Skeleton Image (TSSI) sequence is the keypoints axis. The vertical axis of each TSSI sequence is the time axis. The x, y dimensions are mapped to RG(B) channels for visualization. The axes are kept fixed and the aspect is adjusted so that the data fit in the axes. Ground truth labels C/NC represent the Crossing or not Crossing future action of the pedestrian.

4. Results

4.1. Evaluation Setup

We use the same methodology, splits and evaluation protocol as Cadena et al. [58] for the C/NC prediction task on the JAAD data set: to perform pedestrian crossing prediction, only crossing labels are used, other labels such as drivers information or context are currently omitted and might be used for future research.

Every pedestrian with a crossing marker along their timeline is taken as a positive sample, if not, it is taken as a negative sample. Afterward, all positives samples are divided into two categories, the ones preceding the crossing stage and the ones taking action during the crossing stage. Only the ones preceding the crossing stage are considered. All frames with annotation are then taken from the starting time of the action to time n. They are then sampled with a sliding window of frame size $T = 300$. This procedure results in 927 crossing samples, 1855 non-crossing samples and 697 preceding the crossing samples. Only the remaining 697 prior to crossing positive samples and the 1855 negative samples are used. To avoid redundancy and bias in the data, only the last three steps of a single pedestrian sample are taken from the sliding window if the event is longer than the fixed T frames. It results in 322 positive and 182 negative samples being retained. All samples are then divided into training and test sets. According to Fang et al. [55] splits, we use the first 250 videos for training and the last 96 videos for testing. Since the number of positive examples is greater than the number of negative examples, some positive examples are discarded to maintain a balanced data set. The final data set consists of 240 examples equally distributed between C/NC labels in the training data set and 124 examples equally distributed between C/NC labels in the test data set.

4.2. Implementation Details

As our SPI-Net implementation relies on multiple networks being trained independently and then concatenated for fine-tuning, we firstly here present our entire training setup to obtain SPI-Net:

- Training the Geometric features branch:
    - Training the auto-encoder with a separability constraint term: We use a standard feed-forward non-recurrent MLP, the dimensions of which are $(3000) \rightarrow (128) \rightarrow (64) \rightarrow (32) \rightarrow (1) \rightarrow (32) \rightarrow (64) \rightarrow (128) \rightarrow (3000)$. We use a value of fixed $\lambda = 5$ for the LDA constraint term ponderation in the modified reconstruction cost function. To address the vanishing gradient problem, each perceptron in the given auto-encoder network uses the LeakyRelu [77] activation function. For regularization purposes, we use Dropout [78] ($p = 0.5$), L_2 regularization with $\lambda = 1^{-1}$ and batch normalization [79] after each layer. We choose Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) [80] as the optimizer, with an annealing learning rate that drops from 1^{-3} to 1^{-8}. In order to obtain a good separability in the latent space with the LDA separability constraint, we choose to send all the training examples at once for the auto-encoder training and select a batch size of 240.
    - Training the Encoder part for classification: we recover the encoder part of the auto-encoder, then train a classifier with weights initialized via the auto-encoder. We use the same values of the Adam optimizer for training. We, however, divide the training set into 30 batches of size 8. We use *ReduceLROnPlateau* with a factor of 0.2 and patience of 10.
- Training the Cartesian features branch: The Cartesian features branch is composed of four 2D-convolution blocks composed of 2D-convolution layers (kernel size = 3×3). Similarly to the auto-encoder, we use the LeakyRelu activation function, L_2 regularization with $\lambda = 1^{-4}$ and a Dropout value of 0.5. Each convolution layer is then followed by a Batch Normalization layer and an Average Pooling layer. The fully connected layers following the spatio-temporal features extraction done by convolutions is then completely similar to any other Dense layer of the Geometric feature branch for hyper-parameters tuning. We choose Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate that drops from 1^{-2} to 1^{-8} and *ReduceLROnPlateau* with a factor of 0.5, patience of 5, cooldown of 5 and a batch size of 8.
- Concatenating the branches: We then remove the classification layer of each branch and concatenate those two networks deprived of their last layer into a single one. It allows us to keep the previously learned weights of each network independently. We then add a classification layer in which the weights are initialized randomly after the concatenated layer of the obtained network. Finally, we fine-tune the entire network, from pre-trained weights to the randomly initialized classification layer. We get our presented SPI-Net: a late fusion and fine-tuned version of the Geometric and Cartesian features branches. As proposed in [81], we increase the batch size

over time during the training and therefore fine-tune the approach with two different trainings on the same SPI-network with two different batch sizes. For the first training, we use Adam with a learning rate that drops from 9^{-3} to 5^{-8}, *ReduceLROnPlateau* with a factor of 0.5, patience of 25 and a batch size of 8. For the second one, we use Adam with a learning rate that drops from 9^{-8} to 5^{-18} and *ReduceLROnPlateau* with a factor of 0.5, patience of 25 and a batch size of 240.

4.3. Results on JAAD Data Set

In ablation studies, we first explore how each SPI-Net branch contributes to the intention prediction performance. We, therefore, explore how the LDA constraint for the Geometric branch or the spatial joint reordering trick impact the intention prediction performance on JAAD. Therefore, both Geometric and Cartesian branches results are presented in Tables 1 and 2, Figure 8. The crossing prediction results of the overall SPI-Net approach on the JAAD data set are then presented in Table 3. Finally, more details about each branch and SPI-Net are listed in their respective confusion matrices for the C/NC task in JAAD data set in Table 4.

Table 1. Intention prediction accuracies of the Geometric branch alone, for different encodings of the sequences of inter-keypoints distances.

Method	Accuracy
LDA on Geometric features branch input	51.6%
LDA on the classic Encoder ($\lambda = 0$)	53.2%
LDA on the regularized Encoder ($\lambda = 5$)	54.0%
Encoder (He initialization [64])	66.9%
Encoder with a classic auto-encoder ($\lambda = 0$)	68.5%
Encoder with a regularized auto-encoder ($\lambda = 5$)	69.4%

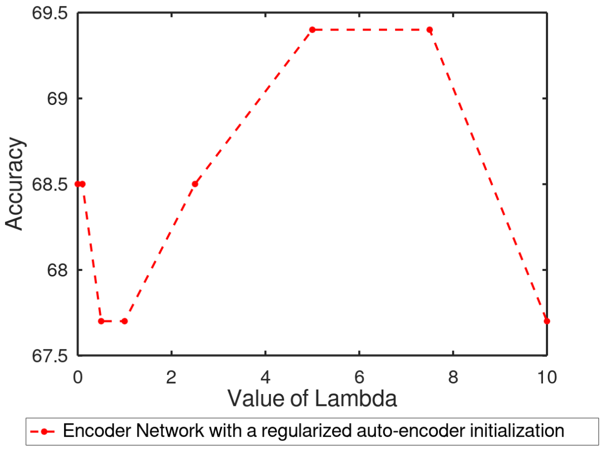

Figure 8. Intention prediction accuracy of the Geometric branch alone, as a function of its λ parameter.

Table 2. Ablation studies: classification accuracy of the Cartesian branch for pedestrian intention prediction for the C/NC task in JAAD.

Method	Accuracy
Cartesian feature branch without spatial joint reordering trick	83.1%
Cartesian feature branch with spatial joint reordering trick	88.7%

Table 3. Classification accuracies for pedestrian intention prediction for the C/NC task in JAAD. CPN [14], Alphapose [82] and Openpose [83] stand for the use of human pose estimation algorithms used by the skeleton-based features method. We have also included the results reported in [15,50], where CNN features are based on a non-fine-tuned AlexNet [51] and Context refers to features of the environment, not of the pedestrian itself.

Method	Accuracy
Alexnet + Context [15]	63.0%
Alexnet + SVM [50]	74.4%
Alphapose + LSTM [56]	78.0%
Res-EnDec [53]	81.0%
ST-DenseNet [52]	84.76%
auto-encoder + Prediction[54]	86.7%
Openpose + Keypoints [55]	88.0%
Alexnet + SVM + Context [50]	89.4%
CPN + GCN [58]	91.9%
CPN + Geometric branch ($\lambda = 5$)	69.4%
CPN + Cartesian branch	88.7%
CPN + SPI-Net ($\lambda = 5$)	**94.4%**

Table 4. Confusion matrix of the JAAD data set obtained by each branch of SPI-Net and SPI-Net on JAAD for the C/NC task.

	Geometric Branch		**Cartesian Branch**		**SPI-Net**	
Ground Truth	**Crossing**	**Not Crossing**	**Crossing**	**Not Crossing**	**Crossing**	**Not Crossing**
Crossing	37	25	57	5	60	2
Not Crossing	16	46	9	53	5	57

Overall, although SPI-Net is not that complex in its architecture, Table 3 shows that it outperforms by more than 2.5% the current state-of-the-art approach [58] based on CPN [14] for pedestrian discrete intention prediction task on the JAAD data. The confusion matrices in Table 4 also shows that SPI-Net accuracy is similar on both action classes, which demonstrates its ability to adapt to intra-class variation for skeleton-based dynamics.

5. Discussion and Future Works

- **Ablation studies:** From Table 1, we figure that solely using the Geometric features branch alone cannot produce a satisfactory performance for the C/NC task: since most of the prior to crossing actions are strongly correlated to global spatial motion of the pedestrian in the scene, the usage of only relative Euclidean distances between key-points is missing necessary information such as spatial dynamics or sequential modeling. However, the Geometric features branch still seems to capture some information only relative to the orientation and dynamics of the skeleton in the data without explicit temporal modeling or global spatial information. For this study, it was of interest to investigate if using the data projected into the latent space provided more information compared to the initial Geometric features input without fine-tuning the entire approach and updating the weights of the network. Table 1 shows that, by using the same binary classifier on the projected data in the bottleneck obtained from a classical auto-encoder, a simple LDA finds slightly more meaning in the data than the initial Geometric features input. Moreover, the latent space representation obtained by our regularized auto-encoder seems to be a little bit more informative than a regular auto-encoder latent space representation. In Figure 8, we evaluate the correspondence between the value of λ for the supervised separability constraint part and prediction accuracy. Afterward, we evaluate the necessity of using a pre-trained encoder network for classification initialized with an auto-encoder training. By comparing the results from the same network with He's weights initialization [64] prior to any auto-encoder training to the entire geometric branch approach, we show that using an auto-encoder to initialize the network's weights helps to a certain extent the network's accuracy. From Table 2, we can deduce that by taking into consideration both spatial and temporal features in the Cartesian coordinate system, we obtain better results than by only considering relative distances of given key-points of the pedestrian skeleton. We can also conclude that the usage of the Tree Structure Skeleton Image (TSSI) [76] normalization improves the results of the Cartesian branch for the C/NC task considerably. Such normalization is therefore relevant as it only changes the size of the image input and therefore does not change the network's architecture much while becoming better for the task it was designed for. Finally, Table 3, shows that by merging and fine-tuning both Geometric and Cartesian features branches into a single network, we can achieve better results for the C/NC task than by considering each branch independently.

- **Inference time:** SPI-Net manages to preserve the information in a Euclidean grid space while keeping a coherent skeleton spatial structure and only uses 2D-convolutions and Dense layers in its architecture. It allows us to keep the advantages of convolutions in terms of parallel computing while capturing spatial and temporal features. Moreover, since SPI-Net only uses classic deep-learning operations, it could be easily implemented in any Deep Learning frameworks and also in any neural hardware solution like Intel Movidius©, or FPGA without redefining any operations. For this study, we implement it by Keras [84] backend in Tensorflow [85]. Therefore, the knowledge of the optimization of euclidean data structure networks proposed in both libraries is conserved compared to approaches based on Graph Networks where basic operations need to be redefined and one might lose speed efficiency in the process. Since our SPI-Net approach has only ∼0.57 M parameters, its speed can reach around one inference every 0.25 ms on one GPU (i.e., RTX 2080ti), or every 0.67 ms on one CPU (i.e., Intel Core i7 8700K), which is roughly 100 times faster than the current state-of-the-art Graph Convolutional Network [58] approach, the average speed of which takes 23 ms on two GPUs (i.e., two GTX 1080).
  Referring to human reaction times, visual skeletal representations are known to be sufficient for humans to describe and understand biological motion, specifically in the case of human motions [86] (i.e., walking, running). It comforts us in the idea of only working with skeleton-based models rather than image-based models. Thompson et al. [87] documented that the average reaction time to detect visual stimuli is approximately 180–200 ms: Kemp et al. [88] show that a visual stimulus takes around 20–40 ms to reach the brain, which leads to an average of 140–180 ms "inference time" for a human once the data reached the brain. SPI-Net relies on the Cascaded Pyramid Network (CPN) algorithm [14] to extract pose sequences before determining pedestrians' intentions. Therefore, one could argue that the pose extraction feature should be compared to the time to reach the brain information and SPI-Net should be compared to the average "inference time" for a human once the data reached the brain. Since the CPN took approximately 60 ms per frame to extract pedestrian poses, the overall approach is roughly two to three times faster than the average human reaction time to a stimulus.
- **Image sampling and pose estimation method:** Necessary step of an intention prediction model of which the analysis of the posture is an essential component. One major drawback of our work is to rely on pose estimation algorithms. However, similarly to the OSI model, our approach relies on independent implementations of methods for specific tasks. It leads to a practical methodology: interchanging the pose estimation algorithms does not compromise the SPI-Net approach. Currently, one of the main limitations of a 2D pose estimation is the ability to deal with pedestrian occlusions in a two-dimensional space. Therefore, in order to improve the pose detection, the question of adding a third dimension may arise. Currently, the methods for estimating 3D poses are much less mature than those for 2D pose estimation. One of the main reasons, to this day, has been the lack of reliable data sets available [89]. However, our pipeline makes it easy to keep up with the state-of-the-art in this field without completely disrupting the SPI-Net approach for intention prediction. Compared to image-based approaches, if major advances are made in the computer vision field and more specifically for pose estimation, SPI-Net could still be relevant.

- **Temporal tracking of pedestrians:** In the real world, there are usually more pedestrians on the streets passing and occluding each other, which requires sophisticated mechanisms not only for their detection but for their temporal tracking without mixing their identity over time. In order to compare SPI-Net to the literature on JAAD, our current approach completely omits such issue and relies on the ground truth spatial coordinates and individual IDs of each pedestrian provided by the data set. To address a better follow-up of the protagonists in the scene and to avoid mixing the dynamics of two protagonists due to a change of camera angle, future research will focus on building an end-to-end framework based on unlabeled coordinates of pedestrians, temporal tracking of pedestrians and SPI-Net for intention prediction. Current research direction tends to evaluate the benefits of using a pose estimation model sequentially based on pose matching for tracking [90–93] compared to a frame by frame pose estimation model [14,94–96] combined with more naive identifications approaches [97,98] that are supposedly faster.
- **Data set size:** A recurrent barrier to using deep-learning is small data sets. Even though JAAD is one of the most complete data sets for pedestrians intents, the number of instances present in the data set is still undersized to use the generalization ability of neural networks to its finest. In the present work, we had to focus a lot on regularization techniques present in the literature to avoid over-fitting. It is, therefore, necessary to extend the total number of instances for such task. Our model is directly extensible to other input formats with different 2D or 3D skeletal data structures: the proposed approach can therefore be applied to a broader family of applications that discover the intentions of moving subjects. However, to use the generalization ability of neural networks to its finest on such small data sets, future research will focus on proposing a tool to enrich the existing databases on human skeletal dynamics by combining both Geometric features and Cartesian features in order to generate skeleton dynamics that are coherent both spatially and sequentially.
- **Continuous intention prediction of pedestrians:** SPI-Net showed that one could link the dynamics of a pedestrian to its discrete intention faster and better. Consequently, future research will focus on using SPI-Net to build a multi-modal architecture taking as input information such as skeleton, image semantic segmentation and qualitative information where discrete intention prediction could be used to infer the continuous trajectories describing the future movement of the pedestrian and therefore propose an intention prediction of pedestrians framework for both discrete and continuous intention prediction.

6. Conclusions

In this work, we have introduced a new real-time representation-focused multi-branch deep-learning skeleton-based approach for the task of discrete intention prediction of pedestrians in urban traffic environments. We propose to go back to "It is all about embedding and standardization in machine-learning" and put great emphasis on finding a way to standardize and represent data in a more adequate way for 2D skeletal pose sequences based models.

By normalizing the input data based on physical world constraints of the body structure, creating features in different coordinate systems allowing to capture different aspects of the data or enforcing certain constraints towards the data representation of designated hidden layers, we send informative-representation ready data to the classification network, which allows us to rely on fewer hidden layers to learn informative representations of data. Our SPI-Net approach has achieved remarkable results: 94.4% accuracy, i.e., 2.5% more than the current state-of-the-art for the Crossing or Not Crossing prediction task on the JAAD data set.

Furthermore, since we choose to rely on a reduced number of hidden layers, we can focus on the inference time of our model, which is mandatory since we take the model speed as one of our priorities: SPI-Net speed can reach around one inference every 0.25 ms on one GPU (i.e., RTX 2080ti), or every 0.67 ms on one CPU (i.e., Intel Core i7 8700K), which makes it highly effective for the task of predicting discrete intentions of pedestrians and directly applicable to embedded devices with real-time constraints.

Author Contributions: Conceptualization, J.G., S.P., G.B., B.S. and F.M.; formal analysis, J.G. and S.P.; project administration, G.B., B.S. and F.M.; software, J.G.; supervision, S.P., G.B., B.S. and F.M.; writing—original draft, J.G., S.P. and G.B.; writing—review and editing, J.G., S.P., G.B., B.S. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors acknowledge the infrastructure and support of the "PELOPS" unit and the interdisciplinary R&D department of the Vedecom institute. The authors would like to thank Anna Venia, Marie Morel and Philéas Samir for proofreading the article. Finally, the authors sincerely thank the anonymous reviewers whose suggestions helped improve and clarify this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
2. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. *arXiv* **2014**, arXiv:1412.0767.
3. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef] [PubMed]
4. Wu, C.Y.; Zaheer, M.; Hu, H.; Manmatha, R.; Smola, A.J.; Krähenbühl, P. Compressed Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
5. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 2014; pp. 568–576. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.749.5720&rep=rep1&type=pdf (accessed on 9 December 2020).
6. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726.
7. Sevilla-Lara, L.; Liao, Y.; Güney, F.; Jampani, V.; Geiger, A.; Black, M.J. On the integration of optical flow and action recognition. In Proceedings of the German Conference on Pattern Recognition, Stuttgart, Germany, 9–12 October 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 281–297.
8. Pop, D.; Rogozan, A.; Chatelain, C.; Nashashibi, F.; Bensrhair, A. Multi-Task Deep Learning for Pedestrian Detection, Action Recognition and Time to Cross Prediction. *IEEE Access* **2019**.10.1109/ACCESS.2019.2944792. [CrossRef]
9. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 588–595.
10. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
11. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. *arXiv* **2016**, arXiv:1607.07043.
12. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *arXiv* **2018**, arXiv:1801.07455.
13. Chen, Y.; Tian, Y.; He, M. Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods. *arXiv* **2020**, arXiv:2006.01423.

14. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-Person Pose Estimation. *arXiv* **2017**, arXiv:1711.07319.
15. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. 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, Venice, Italy, 22–29 October 2017; pp. 206–213.
16. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Agreeing to cross: How drivers and pedestrians communicate. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Redondo Beach, CA, USA, 11–14 June 2017; pp. 264–269.
17. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
18. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
19. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
20. Avola, D.; Bernardi, M.; Cinque, L.; Foresti, G.L.; Massaroni, C. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Trans. Multimed.* **2018**, *21*, 234–245. [CrossRef]
21. Zhang, S.; Liu, X.; Xiao, J. On geometric features for skeleton-based action recognition using multilayer lstm networks. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; IEEE: New York City, NY, USA, 2017; pp. 148–157.
22. Shukla, P.; Biswas, K.K.; Kalra, P.K. Recurrent neural network based action recognition from 3D skeleton data. In Proceedings of the 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Jaipur, India, 4–7 December 2017; IEEE: New York City, NY, USA, 2017; pp. 339–345.
23. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
24. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2117–2126.
25. Elias, P.; Sedmidubsky, J.; Zezula, P. Motion Images: An Effective Representation of Motion Capture Data for Similarity Search. In *Similarity Search and Applications*; Amato, G., Connor, R., Falchi, F., Gennaro, C., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 250–255.
26. Sedmidubsky, J.; Elias, P.; Zezula, P. Effective and Efficient Similarity Searching in Motion Capture Data. *Multimed. Tools Appl.* **2018**, *77*, 12073–12094. [CrossRef]
27. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3288–3297.
28. Pham, H.H.; Khoudour, L.; Crouzil, A.; Zegers, P.; Velastin, S.A. Learning to recognise 3D human action from a new skeleton-based representation using deep convolutional neural networks. *IET Comput. Vis.* **2018**, *13*, 319–328. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Cao, C.; Lan, C.; Zhang, Y.; Zeng, W.; Lu, H.; Zhang, Y. Skeleton-Based Action Recognition with Gated Convolutional Neural Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3247–3257. [CrossRef]
31. Ludl, D.; Gulde, T.; Curio, C. Simple yet efficient real-time pose-based action recognition. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; IEEE: New York City, NY, USA, 2019; pp. 581–588.
32. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
33. Devineau, G.; Moutarde, F.; Xi, W.; Yang, J. Deep learning for hand gesture recognition on skeletal data. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: New York City, NY, USA, 2018; pp. 106–113.

34. Weng, J.; Liu, M.; Jiang, X.; Yuan, J. Deformable pose traversal convolution for 3d action and gesture recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 136–152.
35. Li, C.; Wang, P.; Wang, S.; Hou, Y.; Li, W. Skeleton-based action recognition using LSTM and CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; IEEE: New York City, NY, USA, 2017; pp. 585–590.
36. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access* **2017**, *6*, 1155–1166. [CrossRef]
37. Schuster, M.; Paliwal, K.K.; General, A. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
38. Maghoumi, M.; LaViola, J.J., Jr. DeepGRU: Deep gesture recognition utility. In Proceedings of the International Symposium on Visual Computing, Lake Tahoe, NV, USA, 7–9 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 16–31.
39. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
40. Fan, Z.; Zhao, X.; Lin, T.; Su, H. Attention-Based Multiview Re-Observation Fusion Network for Skeletal Action Recognition. *IEEE Trans. Multimed.* **2019**, *21*, 363–374. [CrossRef]
41. Hou, J.; Wang, G.; Chen, X.; Xue, J.H.; Zhu, R.; Yang, H. Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
42. Gori, M.; Monfardini, G.; Scarselli, F. A new model for learning in graph domains. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, 2005, Montreal, QC, Canada, 31 July–4 August 2005; IEEE: New York City, NY, USA, 2005; Volume 2; pp. 729–734.
43. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [CrossRef]
44. Bronstein, M.M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18–42. [CrossRef]
45. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A comprehensive survey on graph neural networks. *arXiv* **2019**, arXiv:1901.00596.
46. Zhang, X.; Xu, C.; Tian, X.; Tao, D. Graph Edge Convolutional Neural Networks for Skeleton Based Action Recognition. *arXiv* **2018**, arXiv:1805.06184.
47. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018.
48. Li, C.; Cui, Z.; Zheng, W.; Xu, C.; Yang, J. Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition. *arXiv* **2018**, arXiv:1802.09834.
49. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
50. Varytimidis, D.; Alonso-Fernandez, F.; Duran, B.; Englund, C. Action and intention recognition of pedestrians in urban traffic. *arXiv* **2018**, arXiv:1810.09805.
51. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F.; Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
52. Saleh, K.; Hossny, M.; Nahavandi, S. Real-time Intent Prediction of Pedestrians for Autonomous Ground Vehicles via Spatio-Temporal DenseNet. *arXiv* **2019**, arXiv:1904.09862.
53. Gujjar, P.; Vaughan, R. Classifying Pedestrian Actions In Advance Using Predicted Video Of Urban Driving Scenes. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 2097–2103. [CrossRef]
54. Chaabane, M.; Trabelsi, A.; Blanchard, N.; Beveridge, R. Looking Ahead: Anticipating Pedestrians Crossing with Future Frames Prediction. *arXiv* **2019**, arXiv:1910.09077.

55. Fang, Z.; López, A.M. Is the Pedestrian going to Cross? Answering by 2D Pose Estimation. *arXiv* **2018**, arXiv:1807.10580.
56. Marginean, A.; Brehar, R.; Negru, M. Understanding pedestrian behaviour with pose estimation and recurrent networks. In Proceedings of the 2019 6th International Symposium on Electrical and Electronics Engineering (ISEEE), Galati, Romania, 18–20 October 2019; pp. 1–6. [CrossRef]
57. Ghori, O.; Mackowiak, R.; Bautista, M.; Beuter, N.; Drumond, L.; Diego, F.; Ommer, B. Learning to Forecast Pedestrian Intention from Pose Dynamics. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1277–1284. [CrossRef]
58. Gantier, R.; YANG, M.; Qian, Y.; Wang, C. Pedestrian Graph: Pedestrian Crossing Prediction Based on 2D Pose Estimation and Graph Convolutional Networks. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 2000–2005. [CrossRef]
59. Ridel, D.; Rehder, E.; Lauer, M.; Stiller, C.; Wolf, D. A Literature Review on the Prediction of Pedestrian Behavior in Urban Scenarios. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3105–3112. [CrossRef]
60. Xie, D.; Shu, T.; Todorovic, S.; Zhu, S.C. Learning and inferring "dark matter" and predicting human intents and trajectories in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1639–1652. [CrossRef]
61. Wei, P.; Liu, Y.; Shu, T.; Zheng, N.; Zhu, S. Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 6801–6809. [CrossRef]
62. Liu, B.; Adeli, E.; Cao, Z.; Lee, K.H.; Shenoi, A.; Gaidon, A.; Niebles, J.C. Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction. *arXiv* **2020**, arXiv:2002.08945.
63. Ranga, A.; Giruzzi, F.; Bhanushali, J.; Wirbel, E.; Pérez, P.; Vu, T.H.; Perrotton, X. VRUNet: Multi-Task Learning Model for Intent Prediction of Vulnerable Road Users. *arXiv* **2020**, arXiv:2007.05397.
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv* **2015**, arXiv:1502.01852.
65. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **1991**, *4*, 251–257. [CrossRef]
66. Rehder, E.; Kloeden, H.; Stiller, C. Head detection and orientation estimation for pedestrian safety. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; IEEE: New York City, NY, USA, 2014; pp. 2292–2297.
67. Köhler, S.; Goldhammer, M.; Zindler, K.; Doll, K.; Dietmeyer, K. Stereo-vision-based pedestrian's intention detection in a moving vehicle. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, Spain, 15–18 September 2015; IEEE: New York City, NY, USA, 2015; pp. 2317–2322.
68. Flohr, F.; Dumitru-Guzu, M.; Kooij, J.F.; Gavrila, D.M. A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 1872–1882. [CrossRef]
69. Schulz, A.T.; Stiefelhagen, R. Pedestrian intention recognition using latent-dynamic conditional random fields. In Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV), Seoul, Korea, 28 June–1 July 2015; IEEE: New York City, NY, USA, 2015; pp. 622–627.
70. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Towards Social Autonomous Vehicles: Understanding Pedestrian-Driver Interactions. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 729–734. [CrossRef]
71. Dey, D.; Terken, J. Pedestrian interaction with vehicles: Roles of explicit and implicit communication. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg, Germany, 24–27 September 2017; pp. 109–113.
72. Schneemann, F.; Heinemann, P. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 2243–2248.
73. Yang, F.; Sakti, S.; Wu, Y.; Nakamura, S. Make Skeleton-based Action Recognition Model Smaller, Faster and Better. *arXiv* **2019**, arXiv:1907.09658.

74. Baradel, F.; Wolf, C.; Mille, J. Human Activity Recognition with Pose-driven Attention to RGB. In Proceedings of the BMVC 2018—29th British Machine Vision Conference, Newcastle, UK, 2–6 September 2018; pp. 1–14.
75. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
76. Yang, Z.; Li, Y.; Yang, J.; Luo, J. Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences. *arXiv* **2018**, arXiv:1801.10304.
77. Maas, A.L. Rectifier Nonlinearities Improve Neural Network Acoustic Models. 2013. Available online: https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf (accessed on 9 December 2020).
78. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
79. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
80. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
81. Smith, S.L.; Kindermans, P.J.; Ying, C.; Le, Q.V. Don't Decay the Learning Rate, Increase the Batch Size. *arXiv* **2017**, arXiv:1711.00489.
82. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. *arXiv* **2016**, arXiv:1612.00137.
83. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv* **2016**, arXiv:1611.08050.
84. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 9 December 2020).
85. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Available online: tensorflow.org (accessed on 9 December 2020).
86. Johansson, G. Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **1973**, *14*, 201–211. [CrossRef]
87. Thompson, P.; Colebatch, J.; Brown, P.; Rothwell, J.; Day, B.; Obeso, J.; Marsden, C. Voluntary stimulus-sensitive jerks and jumps mimicking myoclonus or pathological startle syndromes. *Mov. Disord. Off. J. Mov. Disord. Soc.* **1992**, *7*, 257–262. [CrossRef] [PubMed]
88. Kemp, B.J. Reaction time of young and elderly subjects in relation to perceptual deprivation and signal-on versus signal-off conditions. *Dev. Psychol.* **1973**, *8*, 268. [CrossRef]
89. Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; Wang, X. 3D Human Pose Estimation in the Wild by Adversarial Learning. *arXiv* **2018**, arXiv:1803.09722.
90. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient Online Pose Tracking. *arXiv* **2018**, arXiv:1802.00977.
91. Ning, G.; Huang, H. LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking. *arXiv* **2019**, arXiv:1905.02822.
92. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. *arXiv* **2018**, arXiv:1804.06208.
93. Raaj, Y.; Idrees, H.; Hidalgo, G.; Sheikh, Y. Efficient Online Multi-Person 2D Pose Tracking With Recurrent Spatio-Temporal Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
94. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
95. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards Accurate Multi-person Pose Estimation in the Wild. *arXiv* **2017**, arXiv:1701.01779.
96. Iqbal, U.; Gall, J. Multi-Person Pose Estimation with Local Joint-to-Person Associations. *arXiv* **2016**, arXiv:1608.08526.
97. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York City, NY, USA, 2017; pp. 3645–3649.

98. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.