TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction

Joseph Gesnouin, Steve Pechberti, Bogdan Stanciulescu, Fabien Moutarde

TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction

Joseph Gesnouin[1,2], Steve Pechberti[1], Bogdan Stanciulescu[2] and Fabien Moutarde[2]
[1] Institut VEDECOM—Versailles, 78000 Versailles, France
[2] Centre de Robotique, MINES ParisTech, Université PSL, 75006 Paris, France

***Abstract*— Understanding the behaviors and intentions of pedestrians is still one of the main challenges for vehicle autonomy, as accurate predictions of their intentions can guarantee their safety and driving comfort of vehicles. In this paper, we address pedestrian crossing prediction in urban traffic environments by linking the dynamics of a pedestrian's skeleton to a binary crossing intention. We introduce TrouSPI-Net: a context-free, lightweight, multi-branch predictor. TrouSPI-Net extracts spatio-temporal features for different time resolutions by encoding pseudo-images sequences of skeletal joints' positions and processes them with parallel attention modules and atrous convolutions. The proposed approach is then enhanced by processing features such as relative distances of skeletal joints, bounding box positions, or ego-vehicle speed with U-GRUs. Using the newly proposed evaluation procedures for two large public naturalistic data sets for studying pedestrian behavior in traffic: JAAD and PIE, we evaluate TrouSPI-Net and analyze its performance. Experimental results show that TrouSPI-Net achieved 76% F1 score on JAAD and 80% F1 score on PIE, therefore outperforming current state-of-the-art while being lightweight and context-free.**

I. INTRODUCTION

The topic of pedestrian crossing prediction has attracted significant interest in computer vision and robotics communities but remains a difficult research topic due to the great variation and complexity of its input data. Although many approaches have been proposed which report interesting results on pedestrian crossing prediction, most of the existing methods may suffer from a large model size and slow inference speed by aggregating multiple forms of perception modalities extracted by additional networks such as background context, optical flow, or pose estimation information [28], [19], [53], [18], [2], [52], [35].

However, in such decisive applications, a desirable action prediction model should run efficiently for real-time usage and should also be robust to a multitude of complexities and conditions. To alleviate this issue, we propose a model using only one additional network to compute poses and disregard the other perception modalities.

Our contributions are summarized in the following:

- We propose TrouSPI-Net: a scene-agnostic, lightweight, multi-branch approach that relies on pose kinematics to predict crossing behaviors. The proposed approach could be applied following the application of any additional network to compute pedestrian body poses and could be easily implemented in any embedded devices with real-time constraints since it only uses standard deep-learning operations in an euclidean grid space.
- We first represent a skeleton sequence as a 2D image-like spatio-temporal continuous representation. As the scale of pedestrians' actions patterns might extend through time and is not limited by a specific temporal resolution, we extract spatio-temporal features by relying on parallel processing of 2D atrous convolutions enhanced with self-attention for multiple dilation rates. This allows TrouSPI-Net to capture features for a given pedestrian action pattern for multiple temporal resolutions.
- We secondly represent a skeleton sequence as its evolution of Euclidean pairwise distances of skeletal joints over time and encode them with U-GRUs [36]: a non-symmetrical bidirectional recurrent architecture designed to exploit the bidirectional temporal context and long-term temporal information for challenging skeletal dynamics having similar patterns but different outputs. This compensates for the inabilities of the first stream in learning temporal patterns invariant to locations and viewpoints.
- Evaluation of TrouSPI-Net has been conducted with the freshly proposed common evaluation criteria [19] on two standard benchmarks for pedestrian behaviors prediction: Joint Attention in Autonomous Driving (JAAD) [34], [33] and Pedestrian Intention and trajectory Estimation (PIE) [32] public data-sets. Architecture variations and branch ablations are also presented to provide insight into our proposed multi-branch approach.

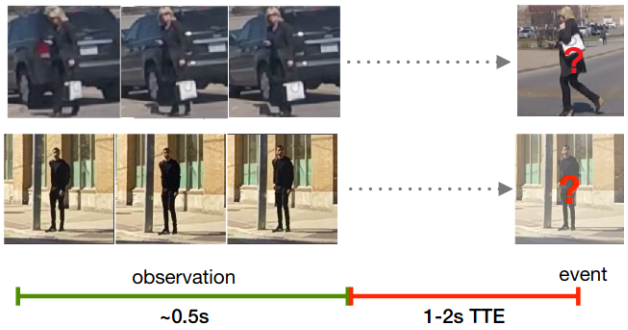

Fig. 1. Pedestrian Action Prediction: the objective is to predict if the pedestrian will start crossing the street at some time t given the observation of length m. Figure adapted from [19].

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

II. RELATED WORK

Currently, the main modalities used for action recognition include RGB videos in their entirety [10], [44], [45], [48], optical flow [41], [54], [38], [29] and pose-based modeling [9], [16], [56], [50], which is more detailed bellow.

A. Pose-based Action Recognition

The detection and pose estimation of humans is the first and necessary step in pose-based action recognition, of which posture analysis is an essential component. Nowadays, pose estimation approaches are not limited to use motion capture systems or depth cameras. RGB data can be used to infer 2D body poses [5], 3D body poses [23] and even track people in real-time [49]. This breakthrough has stimulated the skeletal modality interest since it proved to be sufficient to describe and understand the motion of a given action without any background context. This has made pose-based action recognition preferred over other modalities on a huge amount of real-time scenarios for human action recognition such as human-robot interaction [24], [3], medical rehabilitative applications [25], [8] or pedestrian action prediction [12], [11], [13]. Some commonly used learning architectures for pose-based action recognition include 1D/2D convolutional networks [9], [27], recurrent networks [1], [39], a combination of one of the latter with attention mechanisms [21], [16] or Graph-based models [56], [50].

B. Pedestrian Action Prediction

Pedestrian action prediction formulates the prediction task as a binary classification problem where the objective is to determine if a pedestrian will start crossing in the near future as illustrated in Figure 1. Being a sub-problem within action recognition, most of the existing approaches in the literature rely on the same modalities used for the latter. Preliminary works [34], [46] formulated the problem as a static image classification problem to infer actions in a single image of a pedestrian. Afterward, approaches were designed to consider the temporal coherence in short-term motions of RGB images [37] and combined them with pose-based features [28], [19], [53], increasing the size of their overall approaches drastically since both modalities needed to be extracted. Some works rely on generative models to predict future actions representations which are then sent to a classifier [7], [14]. All those methods present a drawback: they become sensitive to noise, background, and illumination conditions by including scene images in their approaches. To overcome these issues, intention prediction only based on 2D body poses sequences has been explored with various available learning architectures such as convolutions [11], recurrent cells [22], [13], graph-based models [4] and proposed to enhance pose-based approaches by creating features based on body structure to capture different aspects of the data [31], [12]. However, the lack of a common evaluation criterion, of normalized modalities inputs, of a common observation frames selection method, and common prediction horizons made the task of comparing each approach's robustness difficult if not impossible to realize. Lately, common evaluation protocols and modalities inputs [19] were proposed to advance research on pedestrian action prediction further and obtain a fair comparison between all the upcoming methods.

III. METHODOLOGY

Based on the newly proposed evaluation procedures and inputs, we propose a new model for pedestrian action prediction based on 2D body poses: TrouSPI-Net, which is a largely modified and significantly improved version of the SPI-net architecture [12]. The diagram of the model is shown in Figure 2 and the implementation details follow below.

A. Extracting spatio-temporal features via parallel atrous convolutions on pseudo-images

Pedestrian body poses sequences are defined as a vector:

$$\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m) \in \mathbf{R}^{m \times N \times d} \tag{1}$$

where m is the sequence duration, N is the count of key-points, and d is the dimension of each key-point. All sequences of skeletons are then sampled in the form of a 3-dimensional (m,N,d)-shaped tensor representing a 2D image-like spatio-temporal continuous representation of the sequence of poses. The horizontal axis of each pseudo-image represents the key-points axis while the vertical axis represents the time axis. (x,y) dimensions of each key-point are then mapped to $RG(B)$ channels.

By using a 2D-convolution-ready representation format, we extract multi-scale spatio-temporal features using standard computer-vision methods such as atrous convolutions and enhance the feature extraction modules by using Convolutional Block Attention Module (CBAM) [47] for self-attention mechanisms in each branch. Since sequences are represented as pseudo-images, CBAM blocks act as self-attention mechanisms for time and space conjointly.

Each of the pseudo-images is directly fed to three parallel branches. All three branches present a similar architecture designed for single-scale spatio-temporal feature extraction. In each branch, the pseudo-image is passed to an atrous CBAM block, illustrated in Figure 2, followed by a pooling layer. This process is repeated two more times. The difference between the three atrous CBAM blocks resides in the value of the dilation rate fixed in each branch. Having three different dilation rates for the spatio-temporal convolution layers allows the network to directly work at different time resolutions while staying at the same spatial resolutions. Moreover, compared to using different kernel sizes for each convolution, working with atrous convolution does not harm the model size. The outputs of the three branches extracting multi-scale spatio-temporal features are then summed into a single vector for later stages.

Formally, let $h^{(l,\beta)}(m,n)$ represent the input of the l-th atrous CBAM block of the β branch, $K^{(l,\beta)}$ be the number of feature maps, $W_k^{(l,\beta)}(i,j)$ the k-th convolution filter of the l-th convolution in the β branch with the length and the width of m and n, $b_k^{(l,\beta)}$ the bias shared for the k-th filter map,

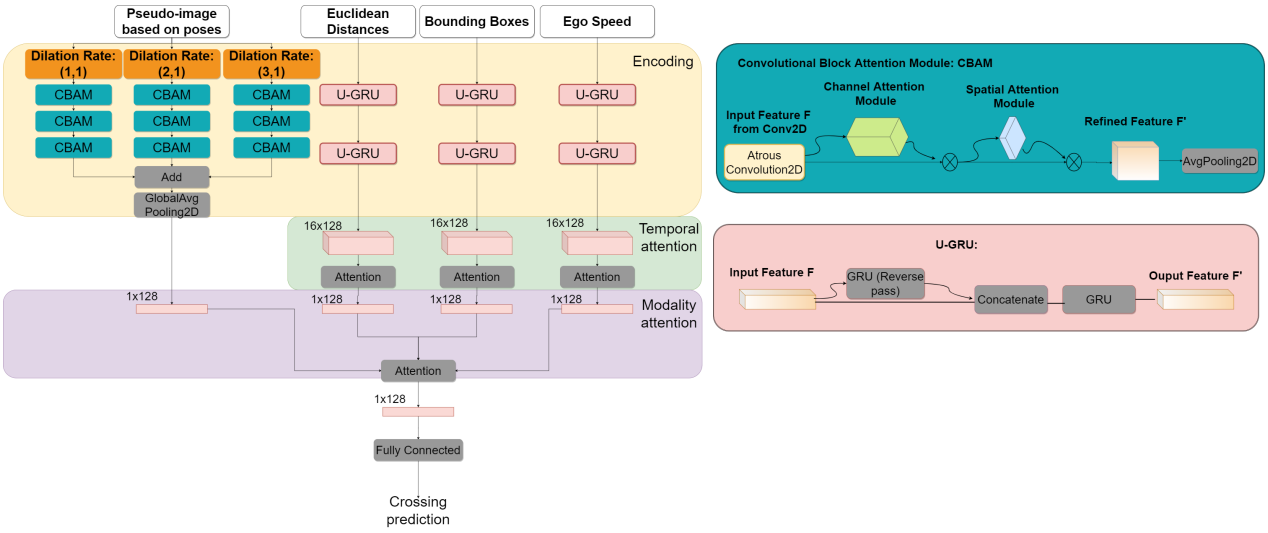

Fig. 2. **The network architecture of TrouSPI-Net:** Its inputs consist in a sequence of 2D body poses transformed into a pseudo-image, relative pairwise distances of skeletal joints, bounding boxes, and ego-vehicle speed. U-GRUs encode every feature except pseudo-images, and each is fed into a temporal attention block. Pseudo-images are processed by parallel atrous CBAM [47] blocks with different dilation rates and then added into a single vector in order to make the size of the pseudo-images block equal to the size of the U-GRUs outputs. Modality attention is then applied to the outputs of each branch, and the weighted outputs are fed into the fully connected layer. **U-GRU blocks:** the first GRU layer does the reverse pass, we then concatenate its output with the input data and finally compute the second GRU layer's output with a forward pass. **CBAM blocks:** given an intermediate feature map extracted by atrous 2D convolutions, the module sequentially infers attention maps along two separate dimensions: channel and spatial.

$(r_1^{(l,\beta)}, r_2^{(l,\beta)})$ the dilation rates and σ an activation function. The intermediate feature map $F(m,n)$ obtained by atrous 2D convolutions of the $l+1$-th CBAM block is calculated as:

$$F(m,n) = \sigma\left(\sum_{k=1}^{K}\sum_{i=1}^{M}\sum_{j=1}^{N} h^{(l,\beta)}(m + r_1 \times i, n + r_2 \times j) \times W(i,j) + b\right) \tag{2}$$

Where $K = K^{(l,\beta)}$, $(r_1, r_2) = (r_1^{(l,\beta)}, r_2^{(l,\beta)})$, $W = W_k^{(l,\beta)}$ and $b = b_k^{(l,\beta)}$. The ouput of the CBAM block $h^{(l+1,\beta)}(m,n)$ is then computed by sequentially inferring a 1D channel attention map $\mathrm{M_c}$ and a 2D spatial attention map $\mathrm{M_s}$ following the original recommendations of the CBAM paper [47] and as illustrated in Figure 2:

$$\begin{aligned} \mathbf{F}'(m,n) &= \mathbf{M_c}(\mathbf{F}(m,n)) \otimes \mathbf{F}(m,n) \\ h^{(l+1,\beta)}(m,n) &= \mathbf{M_s}(\mathbf{F}'(m,n)) \otimes \mathbf{F}'(m,n) \end{aligned} \tag{3}$$

where $\otimes$ denotes element-wise multiplication. Finally, the output $h^{(l+1,\beta)}(m,n)$ serves as the input of the batch normalization and pooling layer that directly follow the atrous CBAM block.

In our experiments, we have three branches: low resolution, medium resolution, high resolution branches $r_1^{(l,\beta)} \in [1;3]$, $r_2^{(l,\beta)} = 1$, $\beta \in [1;3]$, three atrous CBAM blocks and pooling layers in each branch: $l \in [1;3]$. $K^{(l,\beta)} = 64$ feature maps for each layer. Each convolution uses 3x3 kernels and is followed by a batch normalization layer. All the neurons use the LeakyRelu activation function: $\sigma(x) = max(0.2\mathrm{x}, \mathrm{x})$, with the exception of the $\mathrm{M_c}$ and $\mathrm{M_s}$ neurons which use the same hyper-parameters settings than the original CBAM paper.

B. Modeling Location-viewpoint Invariant Features via U-GRUs

To extract skeletal pose kinematic features invariant to locations and viewpoint, we represent a pose sequence as its evolution of skeletal joints relative Euclidean distances over time with the Joint Collection Distances (JCD) feature [51]. The JCD feature is then flattened to a vector of dimension $m * \binom{N}{2}$. Euclidean distances features are then processed with two U-GRUs blocks as illustrated in Figure 2.

Compared to regular Bidirectional GRUs, where the output layer can get information from past and future states simultaneously but are most sensitive to the input values around time t, in U-GRUs, past and future interact but in a limited way. U-GRUs allow the model to accumulate information while knowing which part of the information will be useful in the future and therefore exploit long-term temporal patterns on invariant locations and viewpoint skeletal dynamics. This compensates for the inabilities of the first pseudo-images stream to learn long-range temporal patterns and therefore acts as a regularizer for time and space.

Similarly, context features such as bounding box positions and ego-vehicle speed are processed in parallel through the same U-GRUs architecture.

C. Combining all the features branches

Following the successful application of temporal attention and modality attention in multi-modal approaches for pedestrian action prediction, we finally apply the same temporal attention and modality attention mechanisms used in PCPA [19] to all our features branches to fuse them effectively. Nonetheless, the nature of the inputs merged in TrouSPI-net is entirely different compared to the initial multi-modal

TABLE I

EVALUATION RESULTS FOR BASELINE AND STATE-OF-THE-ART MODELS AND THEIR VARIANTS ON PIE AND JAAD DATA-SETS. DASHED LINES SEPARATE DIFFERENT TYPES OF ARCHITECTURES. MODALITIES CORRESPOND TO THE TYPE OF NETWORKS USED IN THE GIVEN APPROACH, MODEL PARAMS CORRESPONDS TO THE SIZE OF THE NETWORK COMPILED ON THE BENCHMARK [19] WITH ADDITIONAL COSTS (OPTICAL FLOW, BODY POSE, RGB FEATURES) ALREADY EXTRACTED.

Model Name	**Model Variants**	**Model Params (Additional Costs)**	*PIE*					*JAAD$_{behavior}$*					*JAAD$_{all}$*				
			ACC	**AUC**	**F1**	**P**	**R**	**ACC**	**AUC**	**F1**	**P**	**R**	**ACC**	**AUC**	**F1**	**P**	**R**
Static	VGG16 [43]	14.7M	0.71	0.60	0.41	0.49	0.36	0.59	0.52	0.71	0.63	0.82	0.82	0.75	0.55	0.49	0.63
	Resnet50 [15]	23.6M	0.70	0.59	0.38	0.47	0.32	0.46	0.45	0.54	0.58	0.51	0.81	0.72	0.52	0.47	0.56
ATGC [34]	AlexNet	58.3M	0.59	0.55	0.39	0.33	0.47	0.48	0.41	0.62	0.58	0.66	0.67	0.62	**0.76**	**0.72**	0.80
ConvLSTM [40]	VGG16	0.001M (VGG)	0.58	0.55	0.39	0.32	0.49	0.53	0.49	0.64	0.64	0.64	0.63	0.57	0.32	0.24	0.48
	ResNet50	0.001M (Resnet)	0.54	0.46	0.26	0.23	0.29	0.59	0.55	0.69	0.68	0.70	0.63	0.58	0.33	0.25	0.49
SPI-Net [12]	CNN MLP	0.1M (OpenPose)	0.66	0.54	0.30	0.35	0.27	0.58	0.55	0.66	0.67	0.65	0.81	0.72	0.52	0.48	0.58
SingleRNN [18]	LSTM	1.4M (2*VGG,OpenPose)	0.83	0.77	0.67	0.70	0.64	0.58	0.54	0.67	0.67	0.68	0.65	0.59	0.34	0.26	0.49
	GRU	1.0M (2*VGG,OpenPose)	0.81	0.75	0.64	0.67	0.61	0.51	0.48	0.61	0.63	0.59	0.78	0.75	0.54	0.44	0.70
MultiRNN [2]	GRU	1.8M (2*VGG,OpenPose)	0.83	0.80	0.71	0.69	0.73	0.61	0.50	0.74	0.64	0.86	0.79	0.79	0.58	0.45	0.79
StackedRNN [53]	GRU	2.6M (2*VGG,OpenPose)	0.82	0.78	0.67	0.67	0.68	0.6	**0.6**	0.66	**0.73**	0.61	0.79	0.79	0.58	0.46	0.79
HierarchicalRNN [52]	GRU	3M (2*VGG,OpenPose)	0.82	0.77	0.67	0.68	0.66	0.53	0.5	0.63	0.64	0.61	0.80	0.79	0.59	0.47	0.79
SFRNN [35]	GRU	2.6M (2*VGG,OpenPose)	0.82	0.79	0.69	0.67	0.70	0.51	0.45	0.63	0.61	0.64	0.84	0.84	0.65	0.54	**0.84**
C3D [44]	RGB	78M	0.77	0.67	0.52	0.63	0.44	0.61	0.51	0.75	0.63	**0.91**	0.84	0.81	0.65	0.57	0.75
I3D [6]	RGB	12.3M	0.80	0.73	0.62	0.67	0.58	0.62	0.56	0.73	0.68	0.79	0.81	0.74	0.63	0.66	0.61
	Optical flow	12.3M (FlowNet2)	0.81	0.83	0.72	0.60	**0.9**	0.62	0.51	0.75	0.65	0.88	0.84	0.80	0.63	0.55	0.73
TwoStream [42]	VGG16	134.3M (FlowNet2)	0.64	0.54	0.32	0.33	0.31	0.56	0.52	0.66	0.66	0.66	0.60	0.69	0.43	0.29	0.83
PCPA [19]	Temp. +mod. attention	31.2M (C3D,OpenPose)	0.87	0.86	0.77	-	-	0.58	0.5	0.71	-	-	**0.85**	**0.86**	0.68	-	-
TrouSPI-Net (ours)	CBAM attention block	1.5M (OpenPose)	**0.88**	**0.88**	**0.80**	0.73	0.89	**0.64**	0.56	**0.76**	0.66	**0.91**	**0.85**	0.73	0.56	0.57	0.55
	SE attention block	1.5M (OpenPose)	**0.88**	0.87	**0.80**	**0.77**	0.84	**0.64**	0.55	**0.76**	0.65	**0.91**	0.82	0.77	0.58	0.49	0.70

PCPA [19] architecture. While PCPA [19] merges inputs such as sequences of RGB camera images processed by 3D convolution and poses processed via simple recurrent networks without spatio-temporal coherence of body actions, TrouSPI-Net was designed to operate without needing additional RGB scene-context and uses different body poses representations that were encoded to treat the spatial and the temporal information of body action for different time resolutions.

For each feature extracted by U-GRUs: we apply temporal attention [19] to weight the relative importance of frames in the observation relative to the last seen frame. We then apply modality attention [19] to the weighted outputs of the U-GRUs features and the output of the pseudo-images stream. This fuses inputs from multiple modalities into a single representation by weighted summation of the information from individual modalities. The output of the modality attention block is finally passed to a dense layer for prediction.

IV. EXPERIMENTS

To evaluate the presented multi-branch approach and several variations of its architecture, we conducted experiments on two large public data-sets for studying pedestrian behaviors in traffic: JAAD [34], [33] and PIE [32]. JAAD contains 346 clips and focuses on pedestrians intending to cross, PIE contains 6 hours of continuous footage and provides annotations for all pedestrians sufficiently close to the road regardless of their intent to cross in front of the ego-vehicle and provides more diverse behaviors of pedestrians.

A. Evaluation Setup

We base our experiments on the newly proposed evaluation criteria [19] with common evaluation protocols, splits and normalized modalities inputs. As provided in the new benchmark, observation data for each pedestrian is sampled so that the last frame of observation is between 1s and 2s before the crossing event. We report the results using regular classification metrics: accuracy, AUC, precision, recall and F_1-score given by $F_1 = \frac{2\times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$.

TABLE II

ARCHITECTURE VARIATIONS AND ABLATION STUDIES FOR TROUSPI-NET ON PIE DATA-SET.

Model Variants (Additional Costs)	**Params**	**ACC**	**AUC**	**F1**
TrouSPI-Net without euclidean distances	1.4M	0.87	0.85	0.78
TrouSPI-Net without parallel atrous branches	0.8M	0.86	0.80	0.72
TrouSPI-Net GRUs	1.3M	0.85	0.80	0.72
TrouSPI-Net BiGRUs	1.6M	0.86	0.82	0.75
TrouSPI-Net without attention Block	1.4M	0.87	0.85	0.78
TrouSPI-Net with SE attention Block	1.5M	**0.88**	0.87	**0.80**
TrouSPI-Net	1.5M	**0.88**	**0.88**	**0.80**
TrouSPI-Net with two modalities (C3D)	30.2M	**0.88**	0.87	**0.80**

In architecture variations and branch ablations studies, we explore how each TrouSPI-Net component contributes to the pedestrian action prediction performance by removing one component while keeping others unchanged. We also explore the performance of CBAM blocks in the pseudo-image stream by comparing them to similar self-attention blocks designed for 2D convolutions: Squeeze and Excitation method (SE blocks) [17]. Finally, we explore the impact of adding a second modality to TrouSPI-Net by using 3D convolutions [44] on the local box feature available in the data-set.

B. Implementation Details

We use U-GRUs with 64 hidden units for encoding all features, except the pseudo-image. L2 regularization of 0.001 is added to the final dense layer and a dropout of 0.5 is added after the attention block. The number of observation frames m is set to 16. Body poses extracted by OpenPose [5] and proposed in the benchmark [19] are sampled in the form of a 3-dimensional (16,18,2)-shaped tensor for the pseudo-images stream and 2-dimensional (16,153)-shaped tensor for the U-GRUs stream. The ego-vehicle speed feature is used only in the PIE data-set and omitted in JAAD. To compensate for the significant class imbalance, we apply class weights inversely proportional to the percentage of samples of each class in each data-sets. We train the model with Ranger Optimizer: a combination of Lookahead ($k = 6, \alpha = 0.5$) [55] and Radam

[20], binary cross-entropy loss and batch size set to 8. We train for 80 epochs with learning rate set to 5.0e-05 for PIE and 5.0e-06 for JAAD.

C. Discussion

The results of the final TrouSPI-Net model are presented in Table I. Results are most improved compared to State-of-the-Art on the PIE data-set, where accuracy is increased by 1%, AUC by 2% and F_1-score by 3% compared to PCPA [19], a model with two perception modalities: RGB images and poses. On JAAD, our model performs comparably if not better with state-of-the-art across some metrics. This leads us to believe that approaches using only one additional network to compute perception modalities can be competitive with approaches that combine multiple.

A comparison of F_1-scores between our approach and the best-performing methods that exist at this day shows that our approach offers better F_1-scores for two out of three benchmarks. It shows that TrouSPI-Net is more balanced than other approaches for the task of pedestrian crossing prediction. Finally, results obtained by TrouSPI-Net on $JAAD_{all}$ should be taken with a pinch of salt since the data-set considers all the visible pedestrians who are far away from the road and are not crossing. Since pose estimation algorithms are still struggling with scale to extract informative poses for people at the back of a scene, TrouSPI-Net does not manage to extract discriminating features because of the low quality of the poses extracted and relies mainly on other features to realize its inference. This explains its lower performance compared to the two other benchmarks. However, it should not be considered as an issue since those pedestrians are not directly interacting with the vehicle in any way. If they were to become a danger in the future, they would have to step closer to it, and therefore pose estimation algorithms should be able to extract informative poses.

*1) **Architecture variations and branch ablations**:* Removing the parallel atrous branches from the pseudo-image stream leads to a degradation of the performance indicators (Acc, AUC, F_1) on PIE data-set by respectively, 2%, 8% and 8%. Similarly, removing the stream acting as a regularizer with relative distances degrades the performance indicators by respectively 1%, 3% and 2%. Therefore, we can highlight the importance of the three parallel branches to extract spatio-temporal features for different time scales and the importance of the euclidean distances stream to act as a regularizer for the overall approach performance.

Secondly, we evaluate the importance of using a spatio-temporal attention module over the parallel pseudo-images extraction module. We first disregard spatio-temporal attention completely in the given pseudo-images stream and then replace CBAM blocks [47] with SE blocks [17]. Experimental results show that removing the attention-enhanced 2D atrous convolutions degrades the performance indicators by respectively 1%, 3%, 2%, whereas replacing CBAM blocks [47] by SE blocks [17] do not drastically impact TrouSPI-Net's performance and even increases it across some metrics according to Table I. In conclusion, introducing a spatio-temporal attention module over the parallel features extraction module seems to improve our model performance. Future studies could fruitfully explore this further by introducing a custom spatio-temporal attention module specifically designed for the parallel pseudo-images extraction module.

Finally, we evaluate the importance of U-GRUs by replacing them with GRUs and Bidirectional GRUs. Table II results show that both modified approaches lead to a degradation of the performance indicators by respectively 3%, 8%, 8% and 2%, 6%, 5%. Therefore, we can highlight the importance of U-GRUs to exploit the bidirectional temporal and long-term contexts compared to other state-of-the-art approaches designed to capture sequential features. It also leads us to believe that an effective pedestrian action prediction model should focus on both long-term dependencies and multi-scale short temporal features to be effective.

TABLE III

ARCHITECTURE COMPARISON OF FLOATING POINT OPERATIONS PER SECOND (FLOPS) IN MILLIONS, CUDA MEMORY USAGE (CMU) IN MEGABYTES AND WEIGHTS MEMORY REQUIREMENTS (WMR) IN MEGABYTES. RGB FEATURES EXTRACTED BY CNNS ARE TAKEN IN CONSIDERATION DURING COMPUTATIONS.

Model(Additional Costs)	FLOPS (Mio.)	CMU (MB)	WMR (MB)
VGG16 [43]	29.4	72.1	56.1
Resnet50 [15]	47.0	47.0	90.0
ConvLSTM [40] (VGG)	29.5	93.5	56.2
SingleRNN [18] (2 VGG)	65.3	145.3	60.0
MultiRNN [2] (2 VGG)	71.6	146.0	63.0
StackedRNN [53] (2 VGG)	76.3	146.8	66.0
SFRNN [35] (2 VGG)	73.6	146.5	64.5
C3D [44]	156.0	182.6	297.5
I3D [6]	24.6	334.1	46.9
PCPA [19] (C3D)	220	320.2	414.9
SPI-net [12]	**0.3**	**2.5**	**0.3**
TrouSPI-Net (ours)	3.0	6.8	5.4
TrouSPI-Net with two modalities (C3D)	216.7	322.6	412.9

*2) **Using a second perception modality with TrouSPI-Net**:* One of the main advantages of using a scene-agnostic model using such sparse perception modality instead of aggregating multiple perception modalities is the smaller model size leading to an easier deployment into embedded devices, as table III shows. Moreover, when combined with 3D convolutions of cropped images, including the pedestrians, TrouSPI-Net's computational costs grown dramatically without gaining any performance on PIE data-set as tables II and III show. This may be considered a further validation of pose-based only networks for Pedestrian Action Prediction as lightweight models designed for embedded devices with real-time constraints, which do not need additional context input to work effectively. While this affirmation is established for pedestrians where pose estimation inferences are possible and with limited occlusions, the question remains open for scenes with very high occlusions between pedestrians, occlusions between pedestrians and scene objects, or abnormal behaviors such as crowd movement. For those cases, implementing a way to treat Static RGB images effectively as a context feature might still prove important.

*3) **The drawbacks of relying on additional networks to extract perception modalities**:* Although other models

also apply additional networks to extract multiple perception modalities such as pose, flow or background context and the proposed approach beats the state-of-the-art while being smaller in comparison according to Tables I and III, its application also relies on one additional algorithm to operate. If TrouSPI-Net was to be implemented outside of the JAAD and PIE benchmark, one would have to add to TrouSPI-Net's size the pose extraction model used to compute the pose information. In our case, OpenPose [5] was used to compute the inputs available in [19]. Therefore, the overall approach is $\sim$ 53.5M parameters. However, it leads to a practical methodology as interchanging the additional approaches to extract poses does not jeopardize the TrouSPI-Net approach. Contrary to image-based approaches, if improvements such as inference time or average precision by key-points were made in the field of pose estimation, TrouSPI-Net could still be applicable without any modification.

Moreover, the proposed benchmark [19] currently omits a major issue for pedestrian intention prediction: temporal tracking of pedestrians to avoid mixing identities over time. Such questions are rarely raised and approaches mainly rely on the ground-truth IDs of each pedestrian. However, such concerns are mandatory to easily transpose the pedestrian action prediction approaches into real-life scenarios without pedestrians' ground-truth IDs.

In TrouSPI-Net's case, to ensure a better follow-up of the protagonists in the scene and avoid mixing the identities of two protagonists, one could for example replace OpenPose [5] by pose estimation networks sequentially based on pose matching for tracking [49], [26], [30]. Such a substitution would provide the TrouSPI-Net model every modality it needs to work in a non-controlled environment with only one additional network: body poses, handcrafted body poses features, bounding boxes positions of the pedestrians and their respective individual ID's.

V. CONCLUSIONS

We introduced a new lightweight multi-branch neural network to predict pedestrians' actions using only one additional network to extract perception modalities: 2D pedestrian body poses. The proposed TrouSPI-Net model largely extends and improves the SPI-Net [12] approach in several ways. First, we introduce parallel processing branches to allow the architecture to access different time resolutions with atrous convolutions enhanced with self-attention mechanisms. Secondly, we apply U-GRUs on the evolution of relative Euclidean body distances over time, which acts as a regularizer of the first stream for both time and space. We then extend the U-GRUs approach as one baseline method to consider long-term temporal coherence and process each sequence of context features such as bounding box positions or ego-vehicle speed with U-GRUs. Finally, following the newly proposed evaluation procedures and benchmarks for JAAD and PIE (two challenging pedestrian action prediction data-sets), our experimental results show that TrouSPI-Net achieved 76% F1 score on JAAD and 80% F1 score on PIE, therefore outperforming current state-of-the-art. This shows that using only body poses can outperform approaches that combine multiple networks to extract different perception modalities. Subsequently, our model inherits interesting properties such as being completely invariant to any scene-background context, leading to a lightweight approach focusing only on the pedestrian's movement. Therefore, we believe that TrouSPI-Net could be an interesting baseline to easily compare to for future works aiming at developing a pose-only based model for pedestrian intention prediction and has the potential to improve many other human action recognition or prediction tasks.

VI. ACKNOWLEDGEMENT

The authors acknowledge the infrastructure and support of the PELOPS unit and the interdisciplinary R&D department of the Vedecom institute. The authors would also like to thank Ahmet Erdem, Thomas Gilles and Raphaël Rozenberg for the helpful discussions regarding U-GRUS, Marie Morel and Emmanuel Doucet for their fruitful comments and corrections on the manuscript.

REFERENCES

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer, 2011.

[2] A. Bhattacharyya, M. Fritz, and B. Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4194–4202, 2018.

[3] J. Bujalance Martin and F. Moutarde. Real-time gestural control of robot manipulator through deep learning human-pose inference. In *International Conference on Computer Vision Systems*, pages 565–572. Springer, 2019.

[4] P. R. G. Cadena, M. Yang, Y. Qian, and C. Wang. Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2000–2005, 2019.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[7] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2297–2306, 2020.

[8] Y.-J. Chang, S.-F. Chen, and J.-D. Huang. A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in developmental disabilities*, 32(6):2566–2570, 2011.

[9] G. Devineau, F. Moutarde, W. Xi, and J. Yang. Deep learning for hand gesture recognition on skeletal data. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 106–113. IEEE, 2018.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[11] Z. Fang and A. M. López. Is the pedestrian going to cross? answering by 2d pose estimation. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1271–1276. IEEE, 2018.

[12] J. Gesnouin, S. Pechberti, G. Bresson, B. Stanciulescu, and F. Moutarde. Predicting intentions of pedestrians from 2d skeletal pose sequences with a representation-focused multi-branch deep learning network. *Algorithms*, 13(12):331, 2020.

[13] O. Ghori, R. Mackowiak, M. Bautista, N. Beuter, L. Drumond, F. Diego, and B. Ommer. Learning to forecast pedestrian intention from pose dynamics. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1277–1284, 2018.

[14] P. Gujjar and R. Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2097–2103, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang. Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[17] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[18] I. Kotseruba, A. Rasouli, and J. K. Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1688–1693, 2020.

[19] I. Kotseruba, A. Rasouli, and J. K. Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1258–1268, 2021.

[20] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv e-prints*, page arXiv:1908.03265, Aug. 2019.

[21] M. Maghoumi and J. J. LaViola Jr. Deepgru: Deep gesture recognition utility. In *International Symposium on Visual Computing*, pages 16–31. Springer, 2019.

[22] A. Marginean, R. Brehar, and M. Negru. Understanding pedestrian behaviour with pose estimation and recurrent networks. In *2019 6th International Symposium on Electrical and Electronics Engineering (ISEEE)*, pages 1–6, 2019.

[23] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[24] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, and A. Cherubini. Towards Real-Time Physical Human-Robot Interaction Using Skeleton Information and Hand Gestures. In *IROS: Intelligent Robots and Systems*, pages 1–6, Madrid, Spain, Oct. 2018.

[25] H. Mousavi Hondori and M. Khademi. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of medical engineering*, 2014, 2014.

[26] G. Ning, J. Pei, and H. Huang. Lighttrack: A generic framework for online top-down human pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1034–1035, 2020.

[27] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. Learning to recognise 3d human action from a new skeleton-based representation using deep convolutional neural networks. *IET Computer Vision*, 13(3):319–328, 2018.

[28] F. Piccoli, R. Balakrishnan, M. J. Perez, M. Sachdeo, C. Nunez, M. Tang, K. Andreasson, K. Bjurek, R. Dass Raj, E. Davidsson, C. Eriksson, V. Hagman, J. Sjoberg, Y. Li, L. Srikar Muppirisetty, and S. Roychowdhury. FuSSI-Net: Fusion of Spatio-temporal Skeletons for Intention Prediction Network. *arXiv e-prints*, page arXiv:2005.07796, May 2020.

[29] D. Pop, A. Rogozan, C. Chatelain, F. Nashashibi, and A. Bensrhair. Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction. *IEEE Access*, PP:1–1, 10 2019.

[30] Y. Raaj, H. Idrees, G. Hidalgo, and Y. Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] A. Ranga, F. Giruzzi, J. Bhanushali, E. Wirbel, P. Pérez, T.-H. Vu, and X. Perotton. Vrunet: Multi-task learning model for intent prediction of vulnerable road users. *Electronic Imaging*, 2020(16):109–1, 2020.

[32] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *ICCV*, 2019.

[33] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269, 2017.

[34] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017.

[35] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC*, 2019.

[36] R. Rozenberg, J. Gesnouin, and F. Moutarde. Asymmetrical bi-rnn for pedestrian trajectory encoding. *arXiv preprint arXiv:2106.04419*, 2021.

[37] K. Saleh, M. Hossny, and S. Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9704–9710. IEEE, 2019.

[38] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition*, pages 281–297. Springer, 2018.

[39] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[40] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 2015:802–810, 2015.

[41] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[42] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 568–576. Curran Associates, Inc., 2014.

[43] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, page arXiv:1409.1556, Sept. 2014.

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[45] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017.

[46] D. Varytimidis, F. Alonso-Fernandez, B. Duran, and C. Englund. Action and intention recognition of pedestrians in urban traffic. In *2018 14th International conference on signal-image technology & internet-based systems (SITIS)*, pages 676–682. IEEE, 2018.

[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[48] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Compressed video action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[49] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu. Pose Flow: Efficient Online Pose Tracking. *arXiv e-prints*, page arXiv:1802.00977, Feb. 2018.

[50] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[51] F. Yang, Y. Wu, S. Sakti, and S. Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019.

[52] Yong Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.

[53] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[54] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2718–2726, 2016.

[55] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba. Lookahead Optimizer: k steps forward, 1 step back. *arXiv e-prints*, page arXiv:1907.08610, July 2019.

[56] X. Zhang, C. Xu, X. Tian, and D. Tao. Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE transactions on neural networks and learning systems*, 31(8):3047–3060, 2019.