ST CrossingPose: A Spatial-Temporal Graph Convolutional Network for Skeleton-Based Pedestrian Crossing Intention Prediction

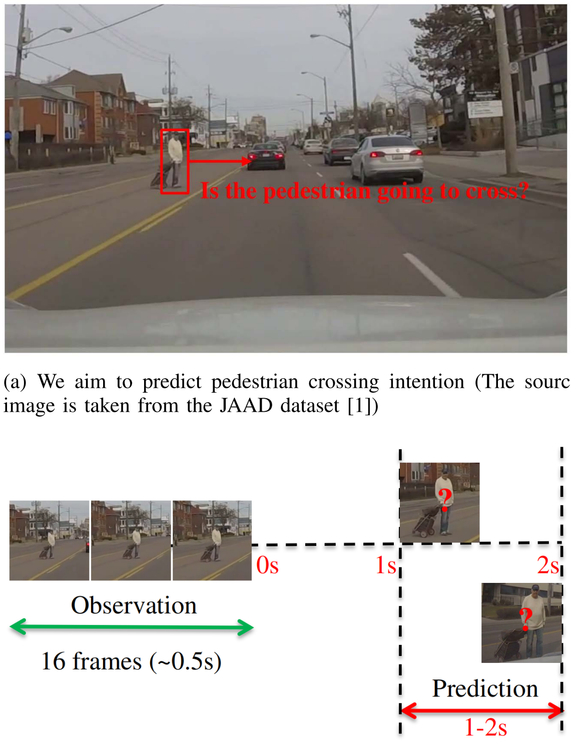
Xingchen Zhang, *Member, IEEE*, Panagiotis Angeloudis, and Yiannis Demiris, *Senior Member, IEEE*

***Abstract*—Pedestrian crossing intention prediction is crucial for the safety of pedestrians in the context of both autonomous and conventional vehicles and has attracted widespread interest recently. Various methods have been proposed to perform pedestrian crossing intention prediction, among which the skeleton-based methods have been very popular in recent years. However, most existing studies utilize manually designed features to handle skeleton data, limiting the performance of these methods. To solve this issue, we propose to predict pedestrian crossing intention based on spatial-temporal graph convolutional networks using skeleton data (ST CrossingPose). The proposed method can learn both spatial and temporal patterns from skeleton data, thus having a good feature representation ability. Extensive experiments on a public dataset demonstrate that the proposed method achieves very competitive performance in predicting crossing intention while maintaining a fast inference speed. We also analyze the effect of several factors, e.g., size of pedestrians, time to event, and occlusion, on the proposed method.**

***Index Terms*—Pedestrian crossing intention, human pose, human skeleton, graph convolutional networks, intelligent vehicle.**

I. Introduction

ROAD safety has been one of the main public health issues. In 2017 alone, 25,300 people lost their lives on EU roads, with human errors involved in around 95% of all road traffic accidents.[1] In 2020, 18,800 people were killed in EU's road accidents.[2] Autonomous vehicles (AVs) are expected to reduce these figures drastically and improve road safety.



(a) We aim to predict pedestrian crossing intention (The source image is taken from the JAAD dataset [1])

(b) Problem formulation

Fig. 1. Given 16 observation frames, our aim is to predict whether the pedestrian will cross the street or not in the future one to two seconds.

Safety is crucial in the context of AVs, especially for less structured environments where interaction between AVs and pedestrians is possible, such as pedestrian junctions and mixed traffic environments. In these cases, more fundamental research on safety aspects is needed since even minor contact between humans and vehicles poses severe dangers to unprotected humans. Pedestrian crossing intention prediction (Fig. 1(a)) is crucial for safe and smooth AV operation. However, although human drivers can generally make accurate inferences about the pedestrians' crossing intention in hundreds of milliseconds after seeing them, it is challenging for AVs to achieve this. Not being able to understand the crossing intention of pedestrians is a key factor in reducing the safety level of AVs, which can lead to traffic accidents or erratic behaviors towards pedestrians. Furthermore, if AVs cannot predict pedestrian intention, they might stop (possibly suddenly) for every pedestrian on the road. This unpredictable

Manuscript received 15 June 2021; revised 31 October 2021; accepted 22 April 2022. This work was supported in part by the InnovateUK DRISK Project (D-RISK: Validating the autonomous vehicle safety-case using simulation and adaptive test generation), in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant 101025274, and in part by a Royal Academy of Engineering Chair in Emerging Technologies. The Associate Editor for this article was P. Ye. *(Corresponding author: Xingchen Zhang.)*

Xingchen Zhang and Yiannis Demiris are with the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: xingchen.zhang@imperial.ac.uk; y.demiris@imperial.ac.uk).

Panagiotis Angeloudis is with the Transport Systems and Logistics Laboratory, Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: p.angeloudis@imperial.ac.uk).

Digital Object Identifier 10.1109/TITS.2022.3177367

[1] https://www.europarl.europa.eu/news/en/headlines/economy/20190110STO23102/self-driving-cars-in-the-eu-from-science-fiction-to-reality

[2] https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1767

behavior for other road users may not only result in rear-end accidents but also annoy other drivers and passengers. Therefore, pedestrian crossing intention prediction is crucial and has attracted a lot of interest in the computer vision and robotics communities [2]–[5]. It is worth mentioning that pedestrian crossing intention prediction is not only beneficial for AVs but also very relevant to conventional vehicles via the increasing adoption of advanced driving assistance systems.

Various methods have been proposed to predict or recognize the intention of pedestrians. The most common methods are based on the trajectory prediction of pedestrians [6]. However, merely relying on pedestrian trajectory is subject to error in many cases where pedestrians initiate walking suddenly or change their direction abruptly. Another kind of method developed recently first predicts future frames using the latest computer vision techniques and then recognizes the crossing action based on these predicted frames [7], [8]. These methods are essentially similar to trajectory-based ones, i.e., rely on motion features, so they suffer from similar problems. Alternatively, other information, such as body poses [9] or skeletons[3] and head orientation [10], have also been utilized to predict pedestrian crossing intention because such information is a good predictor of their situational awareness and future actions. In addition, the visual relationship between pedestrians and surroundings can also be used to predict intention [11].

Among these methods, the skeleton-based ones have been popular recently because skeletons convey important information about human actions. For example, a skeleton reveals information about the walking pattern, and existing studies [9], [19] have demonstrated that the walking pattern of pedestrians can be used to determine the crossing intention. To the best of our knowledge, Furuhashi et al. [12] presented the first pedestrian crossing intention estimation method based on poses. Subsequently, several researchers have developed skeleton-based methods to predict pedestrian crossing intention [9], [13]. A summary of representative skeleton-based pedestrian crossing recognition and prediction studies is presented in Table I.

However, existing skeleton-based methods have some shortcomings. First, hand-crafted features designed based on skeleton information (e.g., angles and distances between keypoints), which may not be very effective, are used in many methods [9], [14], [15], limiting their performance. Second, some methods [16]–[19] only recognize crossing action in the current frame or predict crossing intention in the next frame. They do not perform prediction for a longer time horizon. Finally, in some methods [20], skeletons are used together with other information for intention prediction. However, the inference speed of these methods is usually not very fast due to a large number of model parameters.

To solve these issues, in this paper, we propose to predict pedestrian crossing intention using spatial-temporal graph convolution networks (ST-GCNs), inspired by ST-GCN [22] that achieves good action recognition performance. The proposed method takes as input a sequence (16 frames) of 2D skeletons and learns high-level features leveraging both spatial and temporal information to predict pedestrian crossing intention in a future time window (one to two seconds), as shown in Fig. 1(b).

TABLE I

MAIN EXISTING SKELETON-BASED PEDESTRIAN CROSSING RECOGNITION AND/OR PREDICTION METHODS

Reference	Joints	2D/3D	Network	Recognition	Prediction	Year
[14]	18	2D	RBF-SVM	Yes	Yes	2017
[15]	9	2D	Random Forest	Yes	Yes	2018
[9]	9	2D	Random Forest	Yes	Yes	2019
[16]	14	2D	GCN	Yes	No	2019
[5]	9	2D	RNN	Yes	Yes	2020
[17]	14	2D	CNN	Yes	No	2020
[18]	18	2D	GRU	Yes	No	2020
[19]	9	2D	Neural networks	Yes	No	2020
[21]	14	3D	NN, GRU, Encoder	Yes	Yes	2020

Note that although some studies have used a common dataset, i.e., the Joint Attention for Autonomous Driving (JAAD) dataset [1], for intention prediction, the preprocessing of the dataset and the problem formulation are not uniform between the attempted approaches [5], [9], [16]. Consequently, it is very difficult to compare the performance of those methods fairly. Fortunately, an excellent benchmark was recently released by Kotseruba *et al.* [20], which makes it possible to compare algorithm performance under the same standard. In this work, we adopt the same protocol as this benchmark using the JAAD dataset to ensure a fair comparison.

In summary, the main contributions of this paper are:

- We propose to predict pedestrian crossing intention based on spatial-temporal graph convolutional networks using 2D skeleton data. The proposed method can efficiently process skeleton data to extract spatial-temporal features. To the best of our knowledge, this is the first work to explore the application of spatial-temporal graph convolution networks in the pedestrian crossing intention prediction task using only skeleton data.
- Extensive experiments on a public benchmark dataset demonstrate that the proposed method achieves very competitive performance with a fast inference speed.
- A number of experiments have been carried out to investigate the effect of different factors, including loss function, time to event, pedestrian size, and occlusion, on the performance of the proposed method.

II. RELATED WORK

A. Human Activity Recognition and Prediction

Human activity recognition and prediction can be achieved using various modalities of input data, such as depth, appearance, and skeleton. Among human activity recognition methods, one of the most popular ones in recent years is the spatial-temporal graph-based method using skeleton data. For example, Yan *et al.* [22] proposed one of the first ST-GCN-based human activity recognition methods to recognize different human activities from skeleton data. Since then, many methods have been proposed following this way [23]–[25]. For example, Cheng *et al.* [24] proposed a shift graph convolutional network, which is composed of spatial and temporal shift graph convolution. Zhang *et al.* [25] proposed a context-aware graph convolution to utilize context information in addition to localized graph convolution. The main

[3] We do not distinguish between skeleton and pose in this paper.

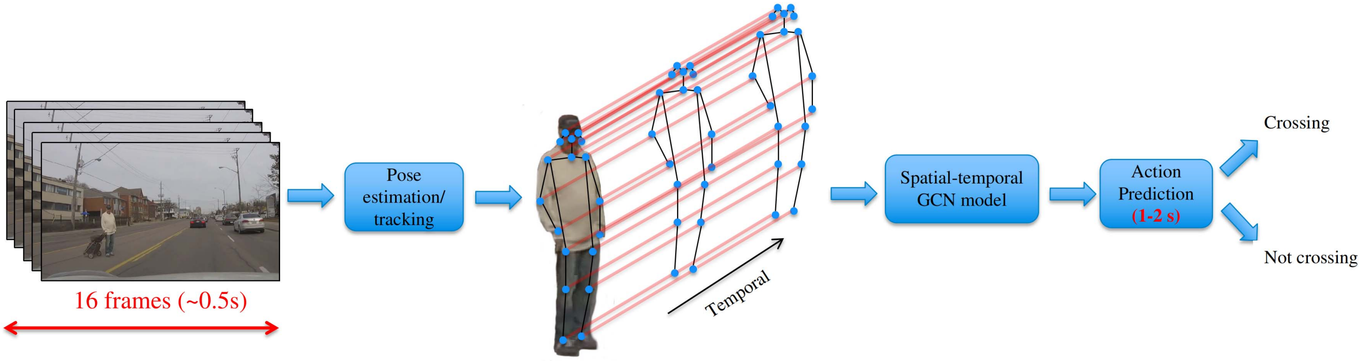

16 frames (~0.5s)
Pose estimation/ tracking
Temporal
Spatial-temporal GCN model
Action Prediction (1-2 s)
Crossing
Not crossing

Fig. 2. The Overall framework of the proposed method. Firstly, a sequence of skeleton data (16 frames) is extracted using pose estimation methods or captured using motion capture devices. The skeleton data is then constructed as a spatial-temporal graph (black lines illustrate spatial edges, and red lines illustrate temporal edges). A ST-GCN model is then applied to the spatial-temporal graph to extract features that encode both spatial and temporal information. Finally, binary classification is performed based on the features to predict whether the pedestrian will cross or not in the future one to two seconds.

advantage of ST-GCN-based methods is that ST-GCN can extract effective features considering both spatial and temporal information. Because of its advantages, ST-GCN has also been applied to other tasks [26].

B. Pedestrian Crossing Intention Prediction

Pedestrian crossing intention prediction can be achieved using various methods, such as trajectory-based methods [6], future frame prediction-based methods [7], [8], and context-based methods [4], [11]. In addition, pedestrian crossing intention prediction methods based on skeleton information have attracted much attention in recent years. For example, Fang *et al.* [14] proposed to determine whether a pedestrian is going to cross or not (C/NC) by analyzing the skeleton in several frames. This method was then extended to naturalistic driving conditions [15] to predict the crossing intention of both pedestrians and cyclists [9]. Later, Wang *et al.* [19] extended C/NC recognition to C/NC/LONG recognition. A major problem of these methods is that they use hand-crafted features to encode skeleton data, which may not be robust and effective. To solve this problem, Cadena *et al.* [16] proposed the first GCN-based crossing intention recognition method using skeleton data. However, only spatial information was utilized in that method, and it focused on action recognition rather than intention prediction. Recently, Zhang *et al.* [2] proposed to predict the crossing intention of pedestrians at intersections' red-light using pose data. However, they neglected moving pedestrians. To solve these issues, we propose to predict pedestrian crossing intention based on a spatial-temporal graph convolutional network using skeleton information, inspired by Yan *et al.* [22].

III. Method

A. Problem Formulation

As pointed in [20], although many works have been conducted regarding pedestrian crossing recognition and prediction, different datasets or different problem settings are used. Therefore, it is not easy to compare the performance of those methods. To ensure a fair performance comparison, we follow the problem formulation proposed in [20], where the crossing intention prediction is formulated as a binary classification problem. Specifically, we aim to predict whether a pedestrian is going to cross the street or not in the future 1 to 2 seconds based on 16 observation frames. Mathematically, we aim to predict the crossing intention $A \in \{0, 1\}$, given the skeleton of a pedestrian in 16 consecutive frames, i.e., $P_{obs,i} = \{p_i^{t-15}, \ldots, p_i^t\}$, where p_i consists of 2D coordinates of 18 joints provided by OpenPose [27], 0 indicates not-crossing and 1 indicates crossing. The frames per second (FPS) of the JAAD dataset used in [20] is 30. Therefore, the observation period is around 0.5 seconds, and the prediction horizon is 30 to 60 frames.

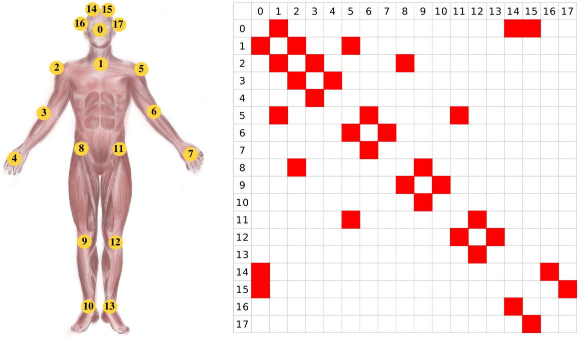
B. Overall Framework

The overall framework of the proposed method is shown in Fig. 2. A sequence of skeletons is first obtained. For each pedestrian, we construct a spatial-temporal graph. The joints in skeletons are used as nodes, and the natural connections between the joints are used as spatial edges. For the same node at different time steps, we connect them using temporal edges. We denote a graph as $G(V, E)$, where V denotes nodes and E denotes the spatial edges between these nodes. The adjacency matrix $\mathbf{A}$ denotes the connections between joints. G is a set of spatial graphs at different time steps t, i.e., G^t. Correspondingly, $\mathbf{A}$ is a set of $\mathbf{A}^t$.

Spatial-temporal graph convolutional networks [22], [23] are usually adopted for processing spatial-temporal graphs because they are able to extract both spatial and temporal features. In our framework, a spatial-temporal GCN model is applied to the spatial-temporal graph to extract features that encode both spatial and temporal information. Binary classification is then performed to predict pedestrian crossing intention based on these features.

C. Skeleton Information Used in This Work

The input skeletons can either be obtained using motion capture devices or pose estimation methods. In this work, we use the skeleton data of the JAAD dataset [1] provided in the benchmark [20] mentioned above. Specifically, 2D skeleton data with 18 joints generated by OpenPose [27] is used. The



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0		🟥													🟥	🟥		
1	🟥		🟥			🟥												
2		🟥		🟥					🟥									
3			🟥		🟥													
4				🟥														
5		🟥					🟥					🟥						
6						🟥		🟥										
7							🟥											
8			🟥							🟥								
9									🟥		🟥							
10										🟥								
11						🟥							🟥					
12												🟥		🟥				
13													🟥					
14	🟥																🟥	
15	🟥																	🟥
16															🟥			
17																🟥		

(a) 18 joints utilized in this work

(b) Adjacency matrix (red means those two joints are connected)

Fig. 3. The 18 joints utilized in this work and the corresponding adjacency matrix that is based on the natural connection between joints.

18 joints and the corresponding adjacency matrix showing the natural connections between joints are shown in Fig. 3. These joints include: 0-Nose, 1-Neck, 2-Right Shoulder, 3-Right Elbow, 4-Right Wrist, 5-Left Shoulder, 6-Left Elbow, 7-Left Wrist, 8-Right Hip, 9-Right Knee, 10-Right Ankle, 11-Left Hip, 12-Left Knee, 13-Left Ankle, 14-Right Eye, 15-Left Eye, 16-Right Ear, 17-Left Ear. Note that some studies employ 3D poses in pedestrian action recognition and prediction [28], [29]. However, in this study, to ensure a fair performance comparison with the methods in the benchmark [20], we utilize 2D poses to show the effectiveness of spatial-temporal GCNs in pedestrian crossing intention prediction.

D. Model Architecture

Our model is based on the work of Yan *et al.* [22]. The architecture of our model is shown in Fig. 4. Specifically, batch normalization is first applied to the input. The output is then passed to several ST-GCN units to extract spatial-temporal features. Each ST-GCN unit consists of a graph convolutional network (GCN) layer and a temporal convolution network (TCN) layer. Therefore, in each ST-GCN unit, we first perform spatial convolution and then perform temporal convolution. Then, global pooling is applied to the spatial-temporal features, followed by a fully connected layer to make a prediction. Note that the ResNet mechanism is adopted in each ST-GCN unit as suggested by Yan *et al.* [22].

1) GCN: We denote the input features of the GCN in one frame in unit l as $\mathbf{f}^l_{in}$, then the output features of the GCN in unit l is

$$\mathbf{f}^l_{out} = \mathbf{\Lambda}^{-\frac{1}{2}}\widehat{\mathbf{A}}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{f}^l_{in}\mathbf{W}^l, \tag{1}$$

where $\widehat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{\Lambda}$ is the diagonal matrix with node degrees of $\widehat{\mathbf{A}}$, $\mathbf{W}^l$ is the matrix of trainable parameters of GCN layer l. In addition, for the first GCN layer, the feature vector of each joint is a 2D vector, while for the rest of GCN layers, the dimension of the feature vector of each joint depends on the parameters of the previous ST-GCN unit. The spatial configuration partitioning proposed by Yan *et al.* [22] is utilized in our model. Therefore, the matrix $\widehat{\mathbf{A}}$ is divided into 3 matrices with the same dimension, i.e.,

$$\widehat{\mathbf{A}} = \sum_{i=1}^{3} \mathbf{A}_i, \tag{2}$$

where $\mathbf{A}_1$ is $\mathbf{I}$, $\mathbf{A}_2$ describes nodes that are closer to the skeleton gravity center than the root node, while $\mathbf{A}_3$ describes other nodes that are farther to the gravity center than the root node [22]. Consequently, Equation (1) becomes

$$\mathbf{f}^l_{out} = \sum_{i=1}^{3} \mathbf{\Lambda}_i^{-\frac{1}{2}}\mathbf{A}_i\mathbf{\Lambda}_i^{-\frac{1}{2}}\mathbf{f}^l_{in}\mathbf{W}^l_i, \tag{3}$$

where $\mathbf{\Lambda}_i$ is the degree matrix of $\mathbf{A}_i$.

2) TCN: After obtaining the spatial feature map $\mathbf{f}^l_{out}$, TCN is applied to obtain spatial-temporal features. TCN is implemented as a $K_t \times 1$ convolution along the temporal dimension [22], where K_t is the temporal kernel size. Concretely, TCN is first applied to one joint along the temporal dimension and then moved to the next joint until all joints are covered.

E. Loss Function

As mentioned previously, we formulate the pedestrian crossing intention prediction as a binary classification problem. However, the utilized dataset JAAD contains imbalanced samples. Specifically, there are more positive (crossing) samples than negative (not-crossing) samples. To handle this, we utilize the focal loss [30] as our loss function to guide the training of our model. The focal loss is defined by Lin *et al.* [30] as:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t), \tag{4}$$

where α_t is a weight parameter used to handle imbalanced data, γ is a focusing parameter, p_t is defined as:

$$p_t = \begin{cases} p & \text{if } y\text{=1} \\ 1 - p & \text{otherwise,} \end{cases} \tag{5}$$

where $y \in \{0, 1\}$ is the ground-truth class, $p \in [0, 1]$ is the probability estimated by the model for the positive class.

IV. Experiments

A. Implementation Details

In this study, our model consists of three ST-GCN units. The model is trained in PyTorch using an RTX 3090 GPU. We trained the model using the Adam optimizer [31] with a learning rate of 0.000005 for 1000 epochs and selected the model according to its performance on the validation set. The batch size is set to 256. We chose the hyper-parameters based on a series of experiments. In the focal loss, α for not-crossing samples is set as 0.75, γ is set as 5.

B. Datasets and Evaluation Metrics

We choose the benchmark proposed by Kotseruba *et al.* [20] for performance comparison. Specifically, the JAAD dataset [1] is used, which is captured in real driving scenarios consisting of 346 video clips. In this dataset, many pedestrians (crossing and not-crossing) are included. The number of samples is shown in Table II. For completeness, we briefly introduce how these samples were generated in [20]. First, an event is defined. For pedestrians who will cross, the event is the moment that he or she

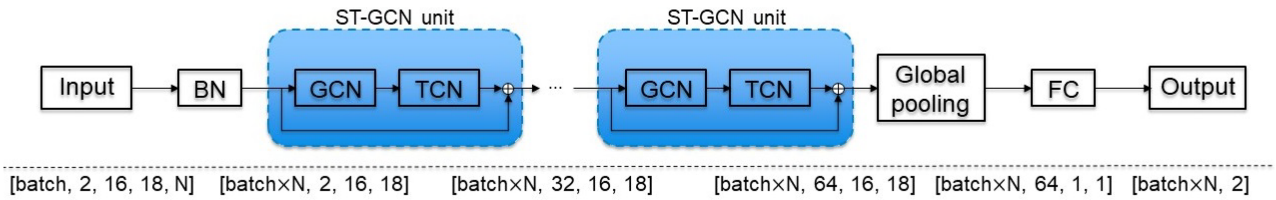
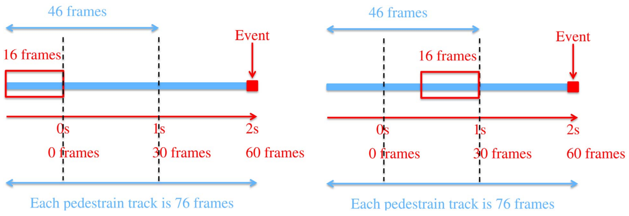


Fig. 4. The architecture of our model and the dimension of corresponding tensors.



(a) The first sample (TTE is 60) (b) The last sample (TTE is 30)

Fig. 5. Generation of samples using a sliding window technique.

TABLE II
THE NUMBER OF SAMPLES IN THE JAAD DATASET

Dataset	Type	Positive	Negative	Total
JAAD	Train	1760	374	2134
	Validation	176	66	242
	Test	1177	704	1881

starts to cross. For pedestrians who will not cross, the event is the last moment when the pedestrian is visible in the scene. For each pedestrian, a track consisting of 76 frames is generated. Then a sliding window technique is applied to generate more samples, as shown in Fig. 5. The time to event (TTE) is from 1 second (30 frames) to 2 seconds (60 frames). For more details, please refer to [20].

As proposed in the benchmark [20], five evaluation metrics are utilized to evaluate the performance of the proposed method, i.e., Accuracy, AUC, F1 Score, Precision, Recall. The definitions of these evaluation metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{8}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{9}$$

where TP is the number of true positive results, TN is true negative, FP is false positive, and FN is false negative. AUC is the area under the ROC curve.

C. Compared Methods

The 12 methods integrated in the benchmark [20] are chosen for comparison. Moreover, Yang *et al.* [3] recently reported their results on this benchmark. Therefore, we also compare with that method, resulting in 13 compared methods:

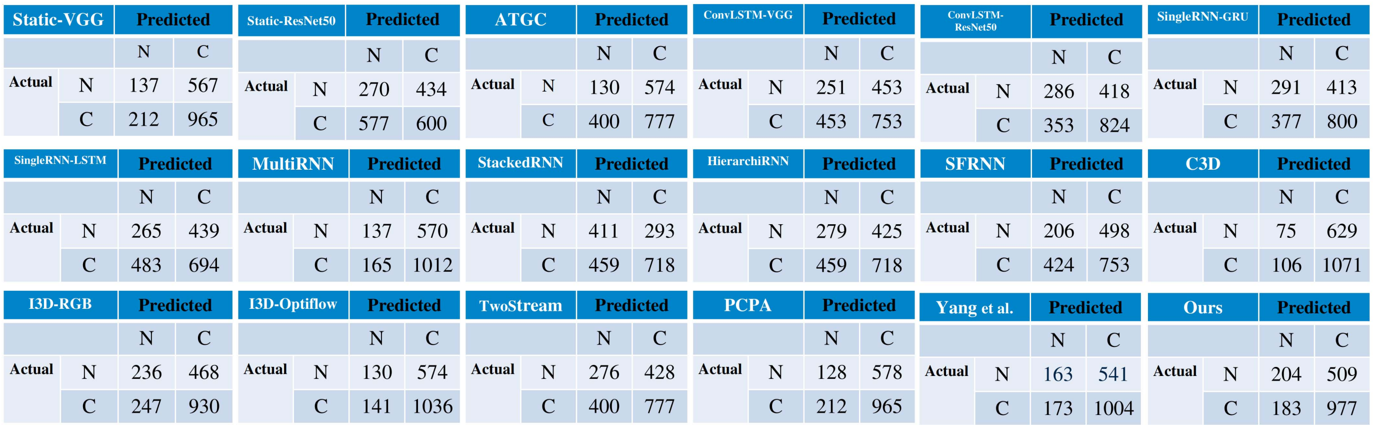
- **Static**: A method using only the last frame in the observation sequence to predict the action based on a fully connected layer. VGG16 [32] or ResNet50 [33] is used as the backend.
- **ATGC** [1]: A method using scene features, pedestrian gait, and head pose. Three CNN streams are utilized to process them, respectively. The fused feature is then fed into an SVM.
- **ConvLSTM** [34]: This method uses a pre-trained CNN to extract features, which are then processed by LSTM. The prediction is made via a fully connected layer based on the last hidden state.
- **SingleRNN** [35]: This method is based on recurrent neural networks (GRU or LSTM), which takes a single vector containing input features as input. A fully connected layer is used for action prediction.
- **StackedRNN** [36]: This method is based on a stack of RNN layers. Each RNN layer takes as input the hidden state of the RNN layer below.
- **MultiRNN** [37]: Several RNN streams are utilized to process different types of features. The hidden states of RNNs are concatenated and fed into a fully connected layer for prediction.
- **HierarchicalRNN** [38]: Several RNN streams are utilized to process different features. Another RNN is used to handle the concatenated hidden states of them. A fully connected layer is employed for prediction.
- **SFRNN** [39]: This is a modified version of StackedRNN. In this method, features are fused at each level gradually. Simpler features are fed at the top, while complex features are fed at the bottom layers.
- **C3D** [40]: This method receives a stack of RGB frames as input. A fully connected layer is used to handle the extracted features for prediction.
- **I3D** [41]: This method takes a stack of RGB frames as input. The prediction is generated through a fully connected layer.
- **TwoStream** [42]: Two CNN branches are utilized to process RGB images and optical flow, respectively. The final prediction is the average of the predictions made for each frame in the sequence.
- **PCPA** [20]: An attention-based method that uses the bounding box, pose, vehicle speed, and local context.
- **Yang *et al.*** [3]: A spatial-temporal method using different phenomena, including RGB sequences, segmentation masks, pose, and ego-vehicle speed.

Some methods have different variants. For example, the Static method can use VGG16 or ResNet50 as the backend. In this paper, we also compare these variants with our method. For more details of these methods, please refer to the corresponding references and [20]. There are also other crossing intention prediction algorithms. However, because very few of them provided their code and different experiment

TABLE III

PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS. THE RESULTS OF [3] ARE EXTRACTED FROM THE CORRESPONDING PAPER, AND THE RESULTS OF OTHER METHODS ARE EXTRACTED FROM THE PEDESTRIAN CROSSING BENCHMARK [20]. THE BEST VALUES ARE MARKED IN BOLD, AND THE SECOND-BEST VALUES ARE UNDERLINED

Method	Model/Variants	Accuracy	AUC	F1 Score	Precision	Recall
Static	VGG16	0.59	0.52	0.71	0.63	0.82
Static	ResNet50	0.46	0.45	0.54	0.58	0.51
ATGC [1]	AlexNet	0.48	0.41	0.62	0.58	0.66
ConvLSTM [34]	VGG16	0.53	0.49	0.64	0.64	0.64
ConvLSTM [34]	ResNet50	0.59	0.55	0.69	0.68	0.70
SingleRNN [35]	GRU	0.58	0.54	0.67	0.67	0.68
SingleRNN [35]	LSTM	0.51	0.48	0.61	0.63	0.59
MultiRNN [37]	GRU	0.61	0.50	0.74	0.64	0.86
StackedRNN [36]	GRU	0.60	**0.60**	0.66	**0.73**	0.61
HierarchicalRNN [38]	GRU	0.53	0.50	0.63	0.64	0.61
SFRNN [39]	GRU	0.51	0.45	0.63	0.61	0.64
C3D [40]	RGB	0.61	0.51	**0.75**	0.63	**0.91**
I3D [41]	RGB	0.62	0.56	0.73	0.68	0.79
I3D [41]	Optical flow	0.62	0.51	**0.75**	0.65	0.88
TwoStream [42]	VGG16	0.56	0.52	0.66	0.66	0.66
PCPA [20]	GRU	0.58	0.50	0.71	0.63	0.82
Yang et al. [3]	VGG+GRU	0.62	0.54	0.74	0.65	0.85
Ours (with focal loss, $\alpha = 0.75, \gamma = 5$)	STGCN	**0.63**	0.56	0.74	0.66	0.83



Static-VGG		Predicted	
		N	C
Actual	N	137	567
	C	212	965

Static-ResNet50		Predicted	
		N	C
Actual	N	270	434
	C	577	600

ATGC		Predicted	
		N	C
Actual	N	130	574
	C	400	777

ConvLSTM-VGG		Predicted	
		N	C
Actual	N	251	453
	C	453	753

ConvLSTM-ResNet50		Predicted	
		N	C
Actual	N	286	418
	C	353	824

SingleRNN-GRU		Predicted	
		N	C
Actual	N	291	413
	C	377	800

SingleRNN-LSTM		Predicted	
		N	C
Actual	N	265	439
	C	483	694

MultiRNN		Predicted	
		N	C
Actual	N	137	570
	C	165	1012

StackedRNN		Predicted	
		N	C
Actual	N	411	293
	C	459	718

HierarchiRNN		Predicted	
		N	C
Actual	N	279	425
	C	459	718

SFRNN		Predicted	
		N	C
Actual	N	206	498
	C	424	753

C3D		Predicted	
		N	C
Actual	N	75	629
	C	106	1071

I3D-RGB		Predicted	
		N	C
Actual	N	236	468
	C	247	930

I3D-Optiflow		Predicted	
		N	C
Actual	N	130	574
	C	141	1036

TwoStream		Predicted	
		N	C
Actual	N	276	428
	C	400	777

PCPA		Predicted	
		N	C
Actual	N	128	578
	C	212	965

Yang et al.		Predicted	
		N	C
Actual	N	163	541
	C	173	1004

Ours		Predicted	
		N	C
Actual	N	204	509
	C	183	977

Fig. 6. Confusion matrices of all compared methods.

configurations were utilized, we do not compare our method with them.

D. Results

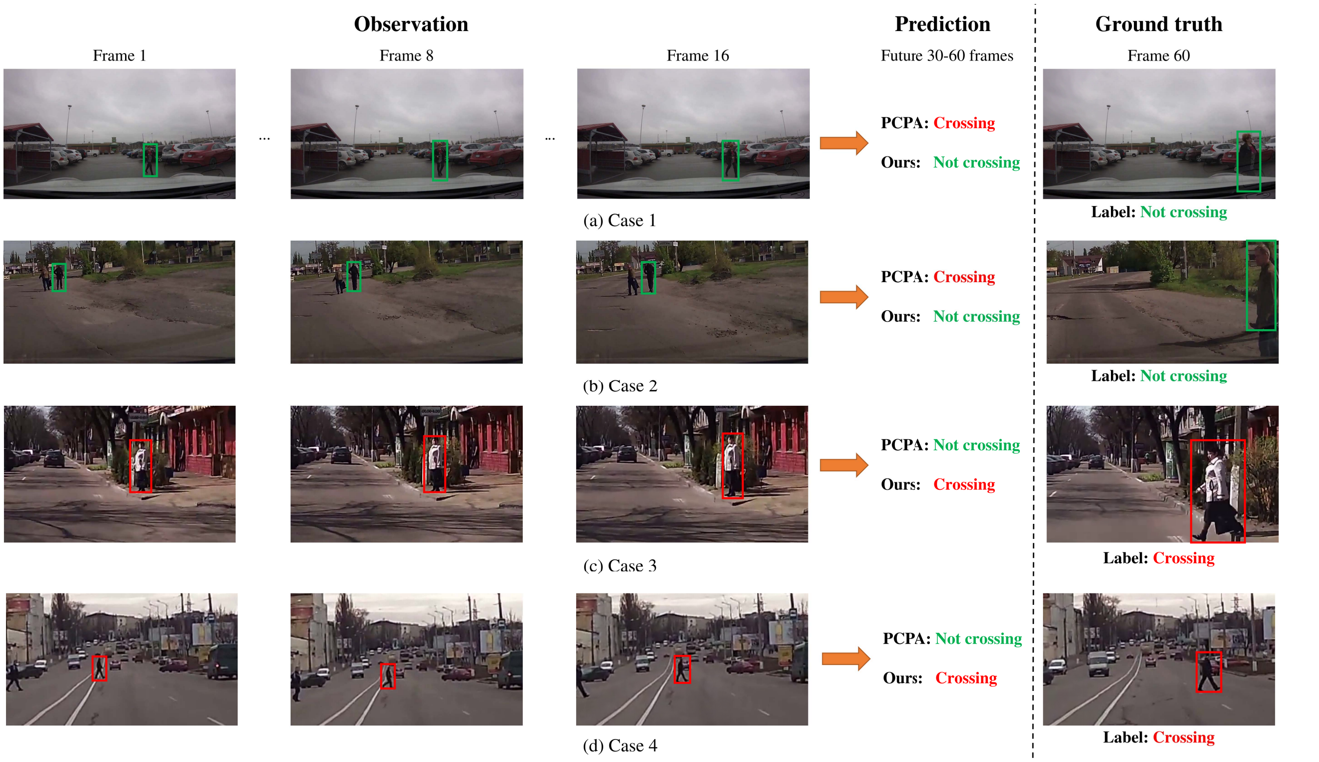
1) Quantitative Results: As illustrated in Table III, the proposed method obtains very competitive results on the JAAD dataset. Specifically, our ST CrossingPose obtains the best results in terms of Acc and the second-best results in terms of AUC and F1 Score. Our performance of precision and recall is also very competitive. It is worth pointing out that only skeleton data is used in the proposed ST CrossingPose, while several other methods (e.g., PCPA and Yang *et al.* [3]) utilize different additional types of information, such as bounding boxes and context information. This clearly demonstrates the effectiveness of the proposed method in predicting pedestrian crossing intention with spatial-temporal graph neural networks.

Because JAAD is an imbalanced dataset, we also show the confusion matrix of each method in Fig. 6 to better compare the quantitative results.[4] As can be seen, although some methods (Static-ResNet50, ConvLSTM, SingleRNN, StackedRNN, HierarchiRNN, SFCNN, TwoStream) can correctly predict more not-crossing samples than our method, their ability to correctly predict crossing samples is not good. By contrast, some algorithms (MultiRNN, C3D, I3D-Optiflow, Yang *et al.* [3]) can successfully predict more crossing samples than our method, but they cannot predict not-crossing samples very well. A very competitive method is I3D-RGB, which has relatively good performance on both crossing and not-crossing samples. However, the proposed ST CrossingPose has better Acc, F1, and recall than I3D-RGB. In summary, the proposed ST CrossingPose achieves a better balance in predicting crossing and not-crossing samples.

2) Qualitative Results: We compare our method qualitatively with a state-of-the-art method (PCPA) proposed in the above-mentioned benchmarking study [20].[5] PCPA employs various information, including bounding box, local context, pose, and vehicle speed, to make the prediction. Figure 7

[4] The confusion matrices of compared methods (except the method of Yang *et al.* [3]) were obtained according to Table III.

[5] The qualitative results of PCPA shown in this section were produced by us using the code provided at https://github.com/ykotseruba/PedestrianActionBenchmark



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Fig. 7. Qualitative results. (a) and (b) are not-crossing examples, while (c) and (d) are crossing examples. As can be seen, our method successfully predicts the crossing intention in these cases, while the PCPA method fails.

TABLE IV

EFFECTIVENESS OF FOCAL LOSS. THE BEST VALUES ARE MARKED IN BOLD

Variants	Acc	AUC	F1	Precision	Recall
V1	0.52	0.53	0.57	**0.66**	0.50
V2 (negative weight 0.82)	0.55	0.54	0.62	**0.66**	0.59
V3 (negative weight 0.75)	0.61	0.54	0.72	0.65	0.80
V4 (negative weight 0.7)	0.61	0.53	0.73	0.65	0.84
V5 ($\alpha = 0.82, \gamma = 0.5$)	0.55	0.54	0.62	**0.66**	0.59
V6 ($\alpha = 0.82, \gamma = 2$)	0.56	0.54	0.63	**0.66**	0.60
V7 ($\alpha = 0.82, \gamma = 5$)	0.55	0.54	0.62	**0.66**	0.58
V8 ($\alpha = 0.70, \gamma = 0.5$)	0.61	0.53	0.73	0.64	0.85
V9 ($\alpha = 0.70, \gamma = 2$)	**0.63**	0.55	**0.74**	0.65	**0.86**
V10 ($\alpha = 0.70, \gamma = 5$)	**0.63**	0.55	**0.74**	**0.66**	**0.86**
V11 ($\alpha = 0.75, \gamma = 0.5$)	0.62	0.55	0.73	**0.66**	0.82
V12 ($\alpha = 0.75, \gamma = 2$)	0.62	0.55	0.73	0.65	0.83
Ours ($\alpha = 0.75, \gamma = 5$)	**0.63**	**0.56**	**0.74**	**0.66**	0.83

TABLE V

THE EFFECT OF NUMBER OF ST-GCN UNITS. THE BEST VALUES ARE MARKED IN BOLD. $\alpha = 0.75$ AND $\gamma = 5$ ARE USED IN THESE EXPERIMENTS. THE TOTAL RUNNING TIME ON THE TEST SET (1881 SAMPLES) IS ALSO SHOWN FOR COMPARISON

Units	Output dimension	Acc	AUC	F1	P	R	Running time (s)
1	32	0.60	0.52	0.73	0.64	**0.85**	**1.80**
2	32,64	0.59	0.52	0.72	0.64	0.81	1.84
3	32,64,64	**0.63**	**0.56**	**0.74**	**0.66**	0.83	1.86
4	32,64,64,64	0.61	0.54	0.73	0.65	0.82	1.90
5	32,64,64,64,128	0.62	0.54	0.73	0.65	0.84	2.08

shows the qualitative results of PCPA and the proposed method in four cases. As can be seen, in these cases, the proposed method correctly predicts the crossing intention of pedestrians, while PCPA has difficulties. This is due to the effectiveness of the proposed method in extracting both spatial and temporal features using ST-GCNs.

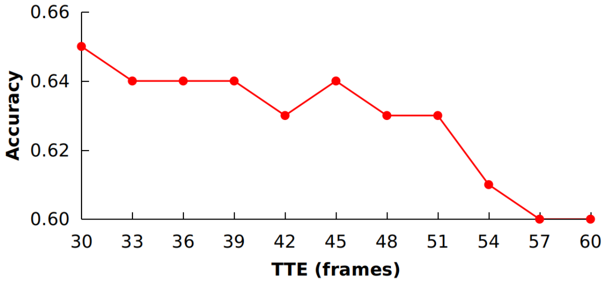


Fig. 8. The relationship between accuracy and TTE.

E. Ablation Studies

1) Effectiveness of Focal Loss: To show the effectiveness of focal loss, we trained several variants of our model. We first used the cross-entropy loss during training. We denote this variant as V1. Then, we used class weights, which are inversely proportional to the percentage of samples (the weight for not-crossing samples is 0.82), in the cross-entropy loss function as suggested in [20]. We denote this variant as V2. We also chose other weights, i.e., 0.75 (V3) and 0.7 (V4). Moreover, we chose different combinations of α and γ in the focal loss (Equation (4)), giving another nine variants. The performance comparison of these variants is shown in Table IV. As can be seen, the usage of class weights significantly improves the performance of our model. Moreover, the usage of focal loss further improves our method to have better performance. In addition, the value of α and γ used in Equation (4) has a significant impact on the results. To have a good balance between crossing and not-crossing samples, we selected $\alpha = 0.75$ and $\gamma = 5$.

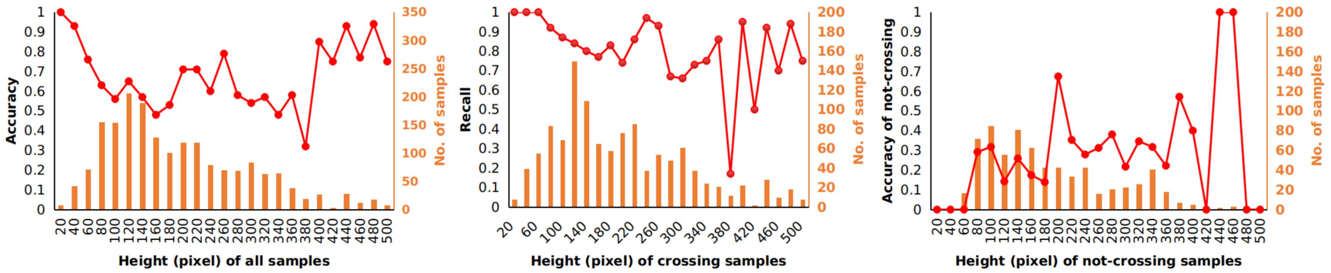


Fig. 9. The relationship between results and bounding box height (in pixels). Left: all samples. Middle: crossing examples. Right: not-crossing samples.

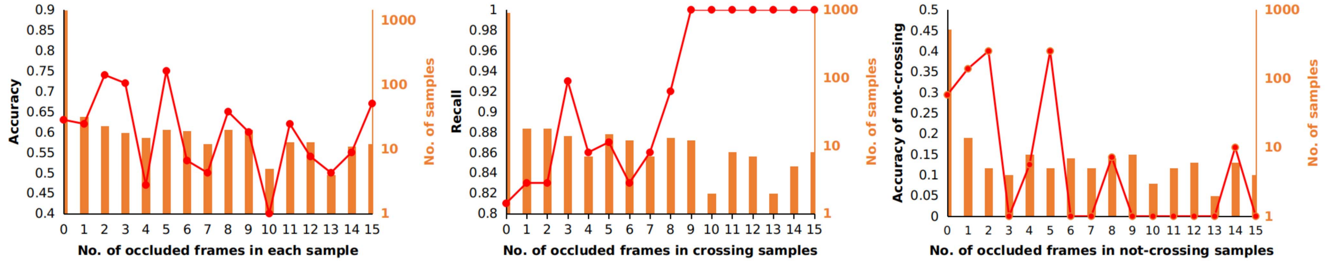


Fig. 10. The relationship between results and the number of occluded frames. Left: all samples. Middle: crossing examples. Right: not-crossing samples.

2) Effect of the Number of ST-GCN Units: We trained several variants with different numbers of ST-GCN units. The results are presented in Table V. As can be seen, the number of ST-GCNs has a significant impact on the prediction performance. However, more units do not necessarily ensure better performance. In our study, the model shows the best performance when three ST-GCN units are used. As a result, we adopt three ST-GCN units.

F. Analysis of the Performance w.r.t. Properties of the Data

1) Effect of TTE: The relationship between accuracy and TTE is shown in Fig. 8. Generally speaking, the accuracy decreases as TTE increases from 30 to 60. This is as expected because when the TTE is longer, there is a higher possibility that the pedestrian shows different motion patterns from the observation frames. Another possible explanation is that the pedestrian is usually larger when the TTE is shorter. Therefore, better skeleton data can be extracted.

2) Effect of Pedestrian Size: The size (bounding box height) of pedestrians can implicitly reveal the distance between pedestrians and the ego vehicle. The relationship between results and the average bounding box height (in pixels) of each sample is shown in Fig. 9. From the figure, we can see that there is no apparent relationship between the results and the size of pedestrians. However, in general, when the bounding box height is greater than 400, the model shows good overall performance. This is because better skeletons can be extracted from the images in these cases. Therefore, our model has better input data.

3) Effect of Occlusion: In the JAAD dataset, some pedestrians are partially (>25% of the pedestrian is occluded) or fully (>75% of the pedestrian is occluded) occluded [1]. We also investigated the relationship between prediction results and occlusion, as shown in Fig. 10. From the left figure in Fig. 10, we can see that the accuracy of our model fluctuates with the occluded samples. We also show the accuracy on crossing and not-crossing samples in the middle and right figure in Fig. 10. As can be seen, for crossing samples, our model shows good performance in general. Although some frames are occluded, our model consistently obtains recall values larger than 0.8. By contrast, for not-crossing samples, the performance is significantly affected by occlusion. Specifically, when the number of occluded frames is 3, 6, 7, 9, 10, 11, 12, 13, and 15, our model cannot correctly predict any not-crossing samples.

TABLE VI

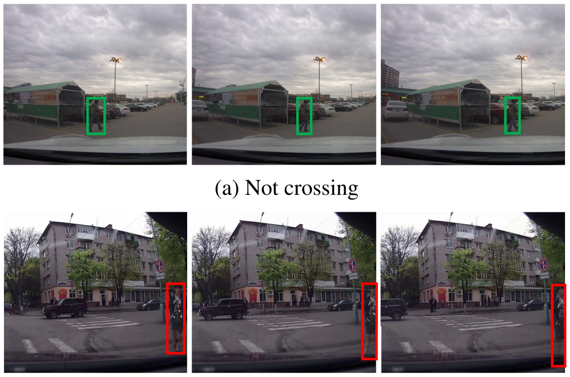
THE EFFECT OF OBSERVATION LENGTH ON PERFORMANCE

Observation frames	Acc	AUC	F1 Score	Precision	Recall
4	0.64	0.56	0.75	0.68	0.85
8	0.63	0.56	0.74	0.67	0.83
12	0.64	0.57	0.75	0.67	0.84
16	0.63	0.56	0.74	0.66	0.83
20	0.61	0.54	0.73	0.65	0.81
24	0.60	0.53	0.72	0.65	0.81

4) Effect of Observation Length: We have shown the results of using 16 observation frames as per the benchmark [20]. However, it is interesting to investigate the effect of observation length on performance. To this end, we chose different observation frames and retrained the proposed method. Specifically, we chose five different observation lengths, i.e., 4, 8, 12, 20, and 24. For each case, we adopted the above-mentioned sliding window technique to generate samples. All other settings were kept the same. The results are shown in Table VI. As can be seen, our model works with different observation lengths, showing its effectiveness. Generally speaking, better results are obtained when a shorter observation length is used. This is because when a shorter observation length is chosen, more samples can be generated from the JAAD dataset using the sliding window technique. Therefore, the model is trained with more data.

G. Failure Cases

Figure 11 presents two failure cases of our method. In the first case, the pedestrian is walking along the road. However, because the car is turning left, the relative motion of the


(a) Not crossing

(b) Crossing

Fig. 11. Failure cases.

pedestrian appears like a crossing. Therefore, our method erroneously predicts that the pedestrian will cross. In the second case, the pedestrian is partially occluded by the car, or a part of her body is outside the camera view in most observation frames. Therefore, it is difficult to detect all 18 joints of the skeleton. Consequently, our method erroneously predicts that this pedestrian will not cross.

H. Inference Time

The inference time of our model with different numbers of ST-GCN units is given in Table V. As can be seen, the inference time increases as more ST-GCN units are used because more ST-GCN units indicate a larger model. For the model with three ST-GCN units, it takes 1.86 seconds to predict the 1881 test samples of the JAAD dataset. Therefore, the average inference time for each sample is 0.99ms. For comparison, we also run the PCPA method [20] and the method of Yang *et al.* [3] using the same machine, which needs 32.21 seconds and 26.06 seconds for 1881 samples, respectively. Therefore, our method is more efficient. This is because we only utilize skeleton data and a spatial-temporal GCN with three ST-GCN units. Consequently, our model is less complicated than PCPA and the method of Yang *et al.* [3] that have multiple branches to process different kinds of source information. Note that we only countered the inference time after all required features (for example, skeletons and context information) of the JAAD dataset had been extracted.

V. Conclusion and Future Work

In this work, we proposed to predict pedestrian crossing intention using spatial-temporal graph neural networks based on pedestrian skeletons. Specifically, given the skeleton data of an observation sequence of 16 frames, the proposed ST CrossingPose can handle the skeleton data of pedestrians to extract both spatial and temporal features. Experiments on a public dataset, i.e., JAAD, demonstrate that the proposed method achieves very competitive prediction performance. In particular, although only skeleton data is used, the proposed ST CrossingPose outperforms several algorithms that utilize different kinds of information (such as bounding box and pose) on the JAAD dataset. This further demonstrates the effectiveness of the proposed method. Moreover, the proposed method can perform prediction efficiently. The proposed method is not only beneficial for autonomous vehicles but also very useful for conventional vehicles via the increasing adoption of advanced driving assistance technologies.

Note that the proposed method can also be extended to multiple pedestrians. A possible idea is to first extract the pose of multiple pedestrians in different frames and then construct a graph for each pedestrian. The proposed method can then be employed to predict their crossing intention. Moreover, although we do not explicitly use vehicle speed information in our model, vehicle movement is implicitly embedded in pedestrian poses in the observation frames and utilized by our spatial-temporal model. In our future work, we will explore combining context information with pedestrian pose to predict pedestrian crossing intention. Finally, we only employed 2D poses of pedestrians to perform crossing intention prediction. However, the real world is 3D. Therefore, it is worth investigating the application of 3D poses [28], [29] in pedestrian crossing intention prediction using spatial-temporal graph convolutional networks.

Acknowledgment

The authors would like to thank Iuliia Kotseruba from York University for insightful discussions on their benchmark, which was very helpful for this work.

References

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 206–213.

[2] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian crossing intention prediction at red-light using pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2331–2339, Mar. 2022.

[3] D. Yang, H. Zhang, E. Yurtsever, K. Redmill, and U. Ozguner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Trans. Intell. Vehicles*, early access, Mar. 28, 2022, doi: 10.1109/TIV.2022.3162719.

[4] B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, "Crossing or not? Context-based recognition of pedestrian crossing intention in the urban environment," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 2, 2021, doi: 10.1109/TITS.2021.3053031.

[5] F. Piccoli *et al.*, "FuSSI-net: Fusion of spatio-temporal skeletons for intention prediction network," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2020, pp. 68–72.

[6] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[7] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2097–2103.

[8] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2297–2306.

[9] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4773–4783, Nov. 2020.

[10] J. F. P. Kooij, M. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 618–633.

[11] B. Liu *et al.*, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3485–3492, Apr. 2020.

[12] R. Furuhashi and K. Yamada, "Estimation of street crossing intention from a pedestrian's posture on a sidewalk using multiple image frames," in *Proc. 1st Asian Conf. Pattern Recognit.*, Nov. 2011, pp. 17–21.

[13] J. Hariyono and K.-H. Jo, "Detection of pedestrian crossing road: A study on pedestrian pose recognition," *Neurocomputing*, vol. 234, pp. 144–153, Apr. 2017.

[14] Z. Fang, D. Vázquez, and A. M. López, "On-board detection of pedestrian intentions," *Sensors*, vol. 17, no. 10, p. 2193, Sep. 2017.

[15] Z. Fang and A. M. Lopez, "Is the pedestrian going to cross? Answering by 2D pose estimation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1271–1276.

[16] 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 *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 2000–2005.

[17] 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*, vol. 13, no. 12, p. 331, 2020.

[18] F. Li, S. Fan, P. Chen, and X. Li, "Pedestrian motion state estimation from 2D pose," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1682–1687.

[19] Z. Wang and N. Papanikolopoulos, "Estimating pedestrian crossing states based on single 2D body pose," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 2205–2210.

[20] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1258–1268.

[21] U.-H. Kim, D. Ka, H. Yeo, and J.-H. Kim, "A real-time predictive pedestrian collision warning service for cooperative intelligent transportation systems using 3D pose estimation," 2020, *arXiv:2009.10868*.

[22] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–10.

[23] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[24] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 183–192.

[25] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14333–14342.

[26] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3289–3299.

[27] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[28] R. Q. Mínguez, I. P. Alonso, D. Fernández-Llorca, and M. Á. Sotelo, "Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1803–1814, May 2019.

[29] V. Kress, S. Schreck, S. Zernetsch, K. Doll, and B. Sick, "Pose based action recognition of vulnerable road users using recurrent neural networks," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 2723–2730.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represenation*, 2015, pp. 1–15.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[35] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Do they want to cross? Understanding pedestrian intention for behavior prediction," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1688–1693.

[36] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[37] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4194–4202.

[38] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[39] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked RNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–13.

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[42] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Comput. Linguistics*, vol. 1, no. 4, pp. 568–576, 2014.

Xingchen Zhang (Member, IEEE) received the B.Sc. degree from the Huazhong University of Science and Technology in 2012, and the Ph.D. degree from the Queen Mary University of London in 2018. He is currently a Marie Skłodowska-Curie Individual Fellow at the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London. Prior to this, he was a Teaching Fellow and Research Associate at the Department of Electrical and Electronic Engineering, Imperial College London. His main research interests include human intention prediction, image fusion, and object tracking.

Panagiotis Angeloudis received the M.Eng. degree in civil and environmental engineering and the Ph.D. degree in transport operations from Imperial College London in 2005 and 2009, respectively. He is currently a Reader and the Director of the Transport Systems and Logistics Laboratory, part of the Centre for Transport Studies and the Department of Civil and Environmental Engineering, Imperial College London. He was recently appointed by the U.K. Department for Transport to the Expert Panel for Maritime 2050. His research interests include transport systems and networks operations, with a focus on the efficient and reliable movement of people and goods across land, sea, and water. He was a member of the U.K. Government Office of Science Future of Mobility Review Team.

Yiannis Demiris (Senior Member, IEEE) received the B.Sc. degree (Hons.) in artificial intelligence and computer science and the Ph.D. degree in intelligent robotics from the Department of Artificial Intelligence, The University of Edinburgh, Edinburgh, U.K., in 1994 and 1999, respectively. He is currently a Professor with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., where he is the Royal Academy of Engineering Chair in Emerging Technologies, and the Head of the Personal Robotics Laboratory. His current research interests include human–robot interaction, machine learning, user modeling, and assistive robotics. He is a fellow of IET and BCS. He was a recipient of the Rector's Award for Teaching Excellence in 2012 and the FoE Award for Excellence in Engineering Education in 2012.