EARLY PEDESTRIAN INTENT PREDICTION VIA FEATURES ESTIMATION

Nada Osman, Enrico Cancelli, Guglielmo Camporese, Pasquale Coscia and Lamberto Ballan

Department of Mathematics "Tullio Levi-Civita", University of Padova, Italy

{nadasalahmahmoud.osman, enrico.cancelli, guglielmo.camporese}@phd.unipd.it

{pasquale.coscia, lamberto.ballan}@unipd.it

ABSTRACT

Anticipating human motion is an essential requirement for autonomous vehicles and robots in order to primary guarantee people's safety. In urban scenarios, they interact with humans, the surrounding environment, and other vehicles relying on several cues to forecast crossing or not crossing intentions. For these reasons, this challenging task is often tackled using both visual and non-visual features to anticipate future actions from 2 s to 1 s earlier the event. Our work primarily aims to revise this standard evaluation protocol to forecast crossing events as early as possible. To this end, we conceive a solution upon an extensively used model for egocentric action anticipation (RU-LSTM), proposing to envision future features, or modalities, that can better infer human intentions using a properly attention-based fusion mechanism. We validate our model against JAAD and PIE datasets and demonstrate that an intent prediction model can benefit from these additional clues for anticipating pedestrians crossing events.

Index Terms— Pedestrian Intent Prediction, Action Anticipation, LSTM.

1. INTRODUCTION

Self-driving cars and autonomous robots have recently demonstrated to be capable of performing multiple tasks (*e.g.*, planning, control, manipulation, perception). Nevertheless, in crowded scenarios, such as streets, parks or airports, they must face critical decisions to avoid accidents and perform a smooth navigation. For example, assistive or delivery robots need to anticipate human motion to better plan their motion and be social compliant. In this regard, urban contexts represent relevant scenarios where predicting human intentions relies on both fast and correct scene perception. Furthermore, predicting future intentions as early as possible helps in better plan their interaction with the environment. Multiple cues are typically involved in this process, *e.g.*, relative speed, pedestrian pose and road signs [1, 2, 3, 4].

In our work, we primary focus on predicting pedestrians intentions to cross roads as early as possible given a fixed history of frames. To this end, we build on top of an action anticipation model an architecture that takes multiple input features and provides the probability that each monitored pedestrian is either crossing or not in the future. Using past motion along with the surrounding context is extremely useful when a pedestrian is walking on a sidewalk prior to crossing or standing due to low visibility in rainy or foggy days, for example. To improve our prediction accuracy, we conceive a *goal* module, whose aim is to predict future features to be fused to motion history (see Fig. 1).

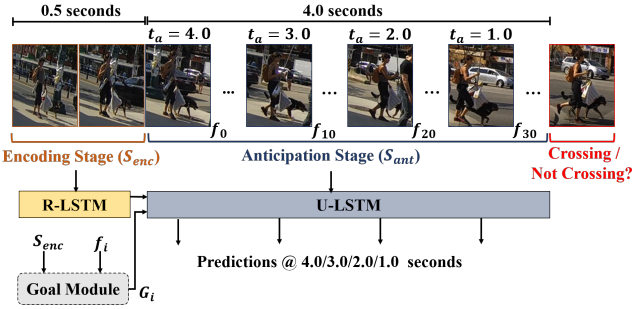

Fig. 1. Our model detects crossing events in two stages: an encoding stage (R-LSTM), processing the initial part of the sequence (0.5 s), and a decoding stage (U-LSTM), predicting the event at multiple anticipation times. We estimate future visual and non visual features, with an attention-based (goal) module, that are provided to the decoding stage. We consider a set of 4 anticipation times: 4.0, 3.0, 2.0, and 1.0 seconds.

Previous works have mainly addressed human action anticipation tasks in both third-person [5, 6, 7, 8, 9] and first-person videos [10, 11, 12, 13, 14]. In this regard, predicting if a pedestrian will cross or not is still a challenging problem. Prior works on intent crossing prediction only focus on a time window of 0.3-0.5 s before the event to extract context features and classify the event [1]. More recent works extend this anticipation time to different values with different observation lengths, in addition to developing more advanced models with multiple input features [2, 3, 4]. In [15], an evaluation benchmark is proposed to tackle this problem, also adopted in [16, 17, 18, 19]. This benchmark focuses on predicting future intentions between 1.0 to 2.0 s earlier the event and uses overlapping windows of 0.5 s as motion history. An averaging operation is then employed to obtain the final pre-

diction. By contrast, our work aims to extend this anticipation time from $1.0\ s$ to $4.0\ s$ and use an adaptive observation window. To extract predictions at fixed anticipation times, we do not employ an averaging operator yet use pedestrian records as single samples. We mainly focus on two largely adopted benchmarks, namely Joint Attention for Autonomous Driving (JAAD) [1] and Pedestrian Intention Estimation (PIE) [20], which include a large set of annotations.

Our contributions can be summarized as follows: (1) We extend the standard evaluation protocol to predict pedestrian crossing intentions as early as possible. (2) We build on top of a state-of-the-art action anticipation model by introducing a *goal* module to envision future motion in multiple feature spaces. We fuse these features with the motion history using an attention-based mechanism to predict crossing events. (3) We demonstrate, experimenting on multiple benchmarks, that pedestrian intents can be foreseen several seconds in advance, thus improving human safety and social awareness.

2. METHOD

Pedestrian intent prediction relies on past motion to detect whether a pedestrian will cross or not the street, and it can be treated as a binary classification task at different anticipation times. More specifically, let r be a set of RGB frames of a pedestrian starting from t_0 to the crossing event at time T, and let $T-t_a$ be the anticipation time, where t_a represents the remaining time from the current observation until the event. Our task is to observe r from t_0 until $T-t_a$ and predict the crossing/not crossing action at T. We aim to predict crossing intentions at different anticipation times, from very early ($t_a \uparrow$) to very close the event ($t_a \downarrow$).

RU-LSTM. As backbone, we consider RU-LSTM [11], which processes the observed video frames in two stages: an encoding stage of S_{enc} steps and an anticipation stage of S_{ant} steps, with α as time interval between subsequent time steps. This model uses an LSTM-based encoder-decoder, referred as to *rolling*-LSTM (R-LSTM) and *unrolling*-LSTM (U-LSTM), respectively. During the decoding stage, it simultaneously produces predictions at various anticipation times. RU-LSTM uses multi-modal features, where different modalities are fused using an attention-based mechanism[1](MATT).

Goal Estimation. Since RU-LSTM cannot use any information after $T-t_a$ during the evaluation phase, it repeats the observed frame at $T-t_a$ for each step in the unrolling stage. Nevertheless, this model may benefit from having a glimpse into the future. For this reason, we define a *goal* module that predicts features that might be extracted at T and fuse this information with the features at $T-t_a$ to enhance its unrolling capability. Let i be the index of the frame at time $T-t_a$; then, for each modality m, we use both encoding sequence (f^m_{enc}) and frame features (f^m_i) to predict future features at T, as follows:

$$G^m_i = D^m_{goal}(LSTM^m_{goal}(f^m_{enc}) \oplus f^m_i), \tag{1}$$

where $\oplus$ denotes the concatenation operation, $LSTM^m_{goal}$ represents the encoding process related to the goal prediction of modality m, and D^m_{goal} is a feed-forward neural network.

Goal Fusion. Using an MLP-based network, we define an attention mechanism (see Fig. 2) to predict weights to be assigned to *goal* features at each unrolling stage, denoted by W^m_i in Eq. 2. W^m_i inherits its length l_i from the corresponding unrolling stage, where $l_i = \frac{t_a}{\alpha}$ is equal to the number of time steps in t_a seconds, with an interval of α seconds; for example, in Fig. 2, $l_0 = 5$, $l_1 = 4$, and the length keeps decreasing reaching the event, where $l_4 = 1$ at $t_a = \alpha$ and $t = T - \alpha$. Concretely,

$$W^m_i = Softmax(D^m_{GATT}(h^m_i \oplus c^m_i)) \tag{2}$$

where D^m_{GATT} is a feed-forward neural network representing our goal attention mechanism (GATT) which uses hidden (h^m_i) and cell (c^m_i) vectors of R-LSTM at time t_a. Our new input to the U-LSTM is then a weighted average of G^m_i and f^m_i, defined as $I^m_i = W^m_i \times G^m_i + (1 - W^m_i) \times f^m_i$.

Finally, predictions from each modality are combined with the modality attention (MATT) mechanism, as proposed in [11], to provide the final prediction.

3. EXPERIMENTS

We rely on two popular datasets for our evaluation, namely JAAD and PIE. JAAD dataset contains 2786 pedestrian samples, annotated with bounding boxes and is split into two subsets (JAAD_{beh} and JAAD_{all}). PIE dataset contains 1842 annotated pedestrian samples with bounding boxes and behavioral tags, in addition to speed, GPS coordinates, and heading direction of the ego-vehicle.

Features. Three main modalities for both datasets are employed: bounding box coordinates, pose features and visual local context features.The PIE dataset also includes ego-vehicle speed as an additional modality.

Baselines. To evaluate the performance of the proposed method for early predictions, we develop an LSTM-based model using RU-LSTM without the unrolling stage (R-LSTM). Furthermore, we compare our model, at different anticipation times, against PCPA [15] which uses fixed and overlapped observations of $0.5\ s$ in the range $[2-1]$ s earlier the event. For our purposes, this model is modified to output earlier predictions in the range $[4-1]$ s. We also consider CAPformer [17] and TrouSPI-Net [19] models.

Evaluation Metrics. As classification metrics we use accuracy, AUC and F1 score.

[1]We refer the reader to [11] for a full description of this model.

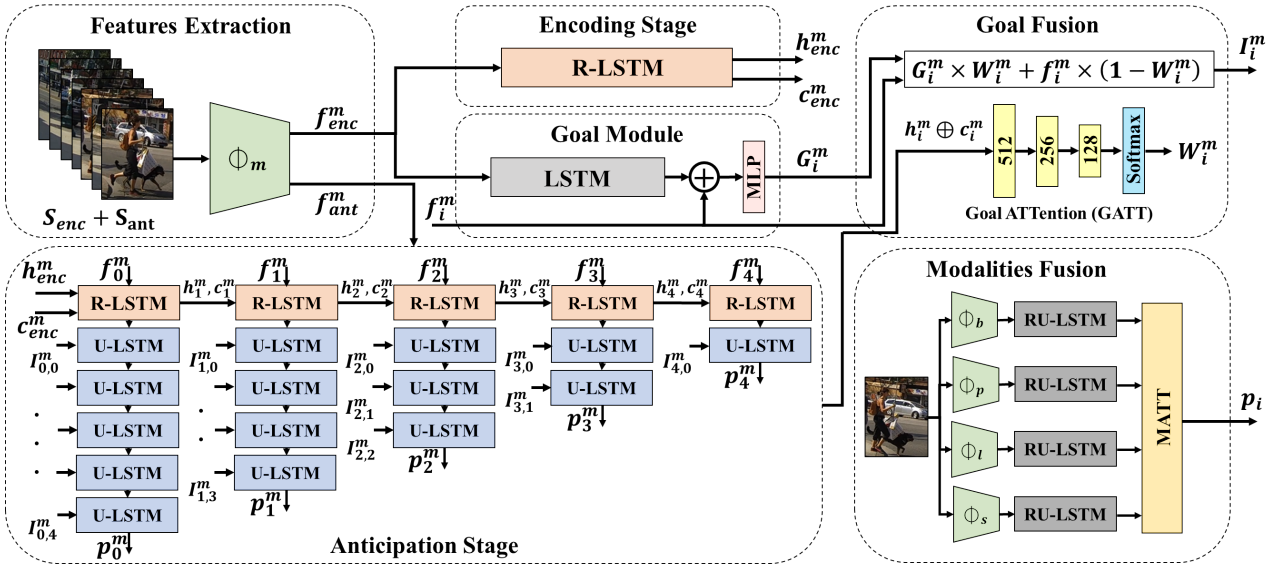

Fig. 2. Our architecture is composed of six sequential steps: **1) Features Extraction**: where a function Φ_m extracts different types of features from raw images; **2) Encoding stage**: encodes the motion history with length S_{enc}; **3) Goal Module**: takes the input features of the encoding part, along with the features at the i^{th} anticipation step (f^m_i) and predicts goal features associated to the modality m (G^m_i); **4) Goal Fusion**: fuses predicted goal features G^m_i and extracted frame features f^m_i, and produces a new input representation for the next step; **5) Anticipation Stage**: takes I^m_i as input and predicts the crossing probability at each time step (p^m_i); **6) Modalities Fusion**: represents our final stage taking predictions provided by each modality and applies an attention-based fusion mechanism to produce the final prediction.

	4 s			3 s			2 s			1 s			[2-1] s		
	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
JAAD$_{beh}$															
PCPA [15]	-	-	-	-	-	-	-	-	-	-	-	-	0.58	0.50	0.71
PCPA† [15]	0.54	0.51	0.62	0.47	0.46	0.54	0.49	0.45	0.61	0.45	0.52	0.63	0.56	0.50	0.67
R-LSTM	0.67	0.64	0.74	0.70	0.64	0.78	0.66	**0.62**	0.75	0.65	0.60	0.75	0.65	0.59	0.74
RU-LSTM	**0.72**	**0.67**	**0.79**	**0.72**	**0.64**	**0.81**	**0.69**	**0.62**	**0.78**	**0.70**	**0.63**	**0.79**	**0.69**	**0.62**	**0.78**
JAAD$_{all}$															
PCPA [15]	-	-	-	-	-	-	-	-	-	-	-	-	0.85	**0.86**	**0.68**
PCPA† [15]	0.75	0.75	0.50	0.74	0.76	0.53	0.72	0.75	0.52	0.76	**0.79**	0.55	0.80	0.79	0.57
R-LSTM	0.83	0.74	0.54	0.86	0.73	0.58	**0.85**	0.73	0.57	**0.87**	0.77	**0.62**	**0.86**	0.76	0.60
RU-LSTM	**0.84**	**0.76**	**0.57**	**0.87**	**0.78**	**0.64**	**0.85**	**0.76**	**0.59**	0.86	0.78	**0.62**	**0.86**	0.78	0.62
PIE															
PCPA [15]	-	-	-	-	-	-	-	-	-	-	-	-	**0.87**	**0.86**	**0.77**
PCPA† [15]	0.76	0.75	0.62	0.77	0.76	0.63	0.83	**0.84**	0.73	0.86	**0.85**	0.77	0.86	**0.86**	**0.77**
R-LSTM	0.75	0.64	0.48	0.75	0.66	0.50	0.76	0.66	0.51	0.76	0.67	0.52	0.76	0.67	0.52
RU-LSTM	**0.77**	**0.76**	**0.63**	**0.80**	**0.79**	**0.68**	**0.85**	0.82	**0.74**	**0.88**	**0.85**	**0.79**	**0.87**	0.84	**0.77**

Table 1. Comparison among multipe intent prediction models for both JAAD and PIE datasets. We consider four anticipation times and the standard evaluation protocol averaging predictions within the range [2-1] s. PCPA† [15] denotes our retrained PCPA model (using the original code and configurations provided in the official GitHub repository).

Implementation Details. Our LSTMs have 512 hidden layers, with $S_{enc} = 6$ and $S_{ant} = 40$. The time interval is 3 frames ($\alpha = 0.1$ s), while our models output 40 predictions from $t_a = 4$ s until $t_a = 0.1$ s. For the sake of clarity, we consider only predictions at $t_a \in \{4, 3, 2, 1\}$ s. For each anticipation time (t_a), we discard videos that do not satisfy the minimum length requirement, *i.e.*, videos whose length is less than t_a. Therefore, we consider 4 s as the earliest anticipation time, where longer predictions would lead to discarding too many examples (more than 50% of data from the JAAD

	JAAD$_{beh}$				
	4 s	3 s	2 s	1 s	[2-1] s
	Bounding Box				
Without Goal	0.55	0.59	0.63	<u>0.64</u>	0.58
Interpolation	**0.61**	0.56	0.61	**0.65**	**0.64**
Concatenation	**0.61**	**0.64**	<u>0.63</u>	0.63	<u>0.63</u>
Attention	**0.61**	<u>0.63</u>	**0.65**	**0.65**	**0.64**
	Pose				
Without Goal	0.51	0.60	<u>0.65</u>	<u>0.65</u>	<u>0.65</u>
Interpolation	0.59	<u>0.64</u>	0.63	0.64	0.63
Concatenation	<u>0.60</u>	**0.67**	**0.66**	**0.66**	**0.66**
Attention	**0.62**	**0.67**	0.57	<u>0.65</u>	<u>0.65</u>
	Local Context				
Without Goal	0.68	0.66	0.65	0.67	<u>0.66</u>
Interpolation	0.73	<u>0.67</u>	0.63	0.66	0.61
Concatenation	0.65	**0.70**	**0.68**	<u>0.68</u>	**0.68**
Attention	**0.69**	<u>0.67</u>	<u>0.66</u>	**0.69**	<u>0.66</u>

Table 2. Ablation study reporting the accuracy metric using different fusion methods for each modality on JAAD$_{beh}$. Underlined numbers refer to the 2^{nd} best performing model.

	JAAD$_{beh}$			JAAD$_{all}$		
	Acc	AUC	F1	Acc	AUC	F1
PCPA [15]	0.58	0.50	0.71	0.85	**0.86**	**0.68**
CAPformer [17]	-	0.55	0.76	-	0.82	0.63
TrouSPI-Net [19]	0.64	0.56	0.76	0.82	0.77	0.58
RU-LSTM	0.69	0.62	0.78	**0.86**	0.78	0.62
G-RULSTM	**0.72**	**0.65**	**0.80**	**0.86**	0.80	0.63
Imp.	**3%**	**3%**	**2%**	-	**2%**	**1%**

Table 3. Comparison between our architectures and state-of-the-art models within the $[2-1]$ s range on JAAD dataset.

dataset). For a fair comparison, in the range [2-1] s, we use the same evaluation protocol proposed in [15], averaging the predictions with a step of 0.1 s for JAAD and 0.2 s for PIE.

Results. Firstly, we evaluate our models on different anticipation times and, eventually, measure the impact of the *goal* module. Table 1 reports our results for different anticipation times (from 4 s to 1 s). Depending on the dataset, two main trends emerge when approaching the event. On JAAD$_{beh}$ dataset, all metrics remain stable approaching the event confirming our intuition to forecast features in advance. This behavior may also be related to the limited samples of this dataset which is quite unbalanced yet on JAAD$_{all}$ this trend is confirmed since no remarkable drops in performance can be observed. Among the used models, RU-LSTM performs the best in both cases. By contrast, on PIE dataset our metrics increase when the anticipation time gets close to the event. It is worth noting that RU-LSTM systematically improves performance metrics compared to the other considered models for the different anticipation times. In the $[2-1]$ s range, RU-LSTM outperforms the considered models on JAAD$_{beh}$ dataset, while is on par with PCPA [15] on JAAD$_{all}$ and PIE datasets. It is also worth mentioning that the standard protocol used in [15] increases the number of training samples considering overlapping windows. By contrast, our proposed protocol dumps this overlapping technique leading to a more realistic evaluation of this task yet largely reducing the number of training samples. This could also explain the limited performance of RU-LSTM in the $[2-1]$ s range, where PCPA uses overlapping windows, while it outperforms at fixed anticipation times all the models using the same number of training samples.

Table 2 reports an ablation study considering three different fusion techniques for estimating future features: **1) Interpolation**, where I_i^m is obtained as a linear combination of f_i^m and G_i^m, based on the time distance d_j from the current step to the frame corresponding to the event, *i.e.*, $I_i^m = \frac{(1-d_j)}{l_i} \times G_i^m + \frac{d_j}{l_i} \times f_i^m$; **2) Concatenation** followed by an MLP layer; **3) Attention**, as defined in Sec. 2. We observe that our goal module, with any considered fusion technique, improves prediction metrics over the baseline, for all modalities and anticipation times. Furthermore, depending on the considered modality, a different fusion techniques could be considered. For example, for bounding box coordinates, representing pedestrian's motion within the image, a linear relationship over time is reported. By contrast, both pose and local context features show a different trend. In this case, concatenation and attention perform better. We select the attention-based technique for all modalities to adapt to both their linear and non-linear relationships.

In Table 3, we compare our model to state-of-the-art architectures. RU-LSTM does not contain the goal module, while G-RULSM uses future features estimation. We observe that G-RULSTM outperforms state-the-art-models for JAAD$_{beh}$, in addition to a noticeable improvement over RU-LSTM. For JAAD$_{all}$, goal-boosted G-RULSTM outperforms our baseline for AUC and F1 metrics, which are more robust metrics for unbalanced datasets. Our proposed model suffers from a hard reduction of training samples, compared to [15] which limits its performance on larger datasets, *e.g.*, JAAD$_{all}$. Nevertheless, our model is on par with CAPformer [17] and outperforms TrouSPI-Net [19] for both subsets.

4. CONCLUSION

In this work, we revise the standard evaluation protocol used to measure the performance of pedestrian intent prediction models. We demonstrate that crossing events can be predicted several seconds in advance with no (or negligible) impact on the performance. To validate our intuition, we build upon an action anticipation model a *goal* module to forecast future features and improve its prediction metrics. This information can increase prediction accuracy up to 3% compared to models that do not envision future features.

Acknowledgements. This work was supported by the MUR PRIN-17 PREVUE project. We also acknowledge UniPD the HPC resources of UniPD (DM and CAPRI clusters).

5. REFERENCES

[1] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. of IEEE/CVF International Conference on Computer Vision Workshops*, 2017.

[2] Satyajit Neogi, Michael Hoy, Kang Dang, Hang Yu, and Justin Dauwels, "Context model for pedestrian intention prediction using factored latent-dynamic conditional random fields.," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6821–6832, 2020.

[3] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos, "Do they want to cross? understanding pedestrian intention for behavior prediction.," in *Proc. of IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[4] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[5] Yazan Abu Farha, Alexander Richard, and Juergen Gall, "When will you do what? - anticipating temporal occurrences of activities," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Panna Felsen, Pulkit Agrawal, and Jitendra Malik, "What will happen next? forecasting player moves in sports videos," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[7] Jiyang Gao, Zhenheng Yang, and Ram Nevatia, "Red: Reinforced encoder-decoder networks for action anticipation," in *Proc. of the British Machine Vision Conference (BMVC)*, 2017.

[8] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese, "A hierarchical representation for future action prediction," in *Proc. of European Conference on Computer Vision (ECCV)*, 2014.

[9] Kuo-Hao Zeng, William B. Shen, De-An Huang, Min Sun, and Juan Carlos Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[10] Chenyou Fan, Jangwon Lee, and Michael S. Ryoo, "Forecasting hands and objects in future frames," in *Proc. of the European Conference on Computer Vision Workshops*, 2018.

[11] Antonino Furnari and Giovanni M. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, pp. 4021–4036, 2021.

[12] Nicholas Rhinehart and Kris M. Kitani, "First-person activity forecasting with online inverse reinforcement learning," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[13] Guglielmo Camporese, Pasquale Coscia, Antonino Furnari, Giovanni M. Farinella, and Lamberto Ballan, "Knowledge distillation for action anticipation via label smoothing," in *Proc. of the IAPR International Conference on Pattern Recognition (ICPR)*, 2021.

[14] Nada Osman, Guglielmo Camporese, Pasquale Coscia, and Lamberto Ballan, "SlowFast Rolling-Unrolling LSTMs for action anticipation in egocentric videos," in *Proc. of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[15] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[16] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith Redmill, and Ümit Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention.," *arXiv preprint arXiv:2104.05485*, 2021.

[17] Javier Lorenzo, Ignacio Parra Alonso, Rubén Izquierdo, Augusto Luis Ballardini, Álvaro Hernández Saz, David Fernández Llorca, and Miguel Ángel Sotelo, "CAPformer: pedestrian crossing action prediction using transformer.," *Sensors*, vol. 21, no. 17, 2021.

[18] Joseph Gesnouin, Steve Pechberti, Bogdan Stanciulcscu, and Fabien Moutarde, "Rnn-based pedestrian crossing prediction using activity and pose-related features," in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[19] Joseph Gesnouin, Steve Pechberti, Bogdan Stanciulcscu, and Fabien Moutarde, "TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and u-grus for skeletal pedestrian crossing prediction," in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[20] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.