

Benchmark for Evaluating Pedestrian Action Prediction

Iuliia Kotseruba
York University

yulia@eecs.yorku.ca

Amir Rasouli
York University

aras@eecs.yorku.ca

John K. Tsotsos
York University

tsotsos@eecs.yorku.ca

Abstract

Pedestrian action prediction has been a topic of active research in recent years resulting in many new algorithmic solutions. However, measuring the overall progress towards solving this problem is difficult due to the lack of publicly available benchmarks and common training and evaluation procedures. To this end, we introduce a benchmark based on two public datasets for pedestrian behavior understanding. Using the proposed evaluation procedures, we rank a number of baseline and state-of-the-art models and analyze their performance with respect to various properties of the data. Based on these findings we propose a new model for pedestrian crossing action prediction that uses attention mechanisms to effectively combine implicit and explicit features and demonstrate new state-of-the-art results. The code for models and evaluation is available at <https://github.com/ykotseruba/PedestrianActionBenchmark>.

1. Introduction

In the context of assistive and autonomous driving systems, the ability to predict pedestrian actions, crossing in particular, is deemed important for robust and reliable future planning. During the past decade, the topic of pedestrian action prediction has attracted significant interest in computer vision and robotics communities. As a result of compelling efforts in this field, many approaches have been proposed which report promising results on pedestrian crossing prediction.

To further advance research on pedestrian action prediction, it is important to compare the performance of the algorithms which requires common evaluation guidelines. As of now, such comparisons are difficult if not impossible due to the lack of agreed-upon evaluation criteria even if the performance of the models is assessed using the same datasets [47]. For instance, some works make continuous prediction even after the crossing even already began, use inconsistent data splits and sampling procedures. To this end, we propose a new benchmark for pedestrian cross-

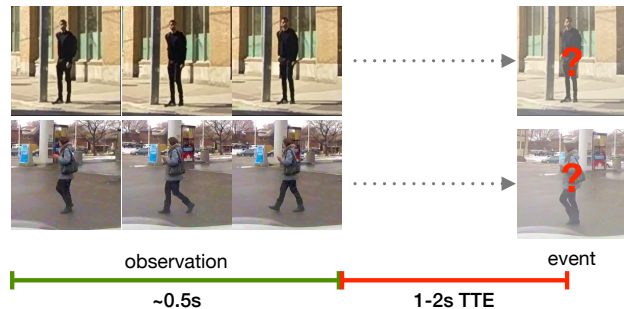


Figure 1: Crossing prediction task: given a set of observation frames predict action some time in the future to give the driver or autonomous vehicle sufficient time to react to pedestrian behavior.

ing prediction based on two publicly available datasets for pedestrian behavior understanding: Joint Attention for Autonomous Driving (JAAD) [44] and Pedestrian Intention Estimation (PIE) [42]. We evaluate the performance of baseline and state-of-the-art (SOTA) crossing prediction algorithms based on evaluation protocol established according to pedestrian behavioral studies. Besides ranking the algorithms by their performance in the same conditions, we examine the effect of different data properties across two datasets. Finally, based on the evaluation results and analysis of the data, we propose a novel method for crossing action prediction which achieves SOTA performance on benchmark datasets. As part of our effort for supporting future research on pedestrian action prediction, we will make the benchmarking code for all the algorithms public.

2. Related work

2.1. Behavior prediction

Behavior prediction is a topic of interest in many domains such as assistive robotics [35, 4, 30, 64], surveillance [27, 25, 17, 33], sports forecasting [29, 39, 10] and autonomous driving systems [46, 42, 32, 62]. Behavior prediction can be either implicit in the form of future trajectories or poses [4, 61, 25, 42, 39] or explicit in terms of predicting future actions or events [35, 46, 27, 17, 10].

Action prediction. Action prediction has a wide range of applications including human-object [23, 16, 1, 63] and human-human [17, 11, 48] interaction, accident prediction [60, 54, 55] and road traffic understanding [8, 18, 46, 28]. These methods are inherently similar to approaches used in action and activity recognition algorithms. Some commonly used learning architectures include 2D/3D convolutional networks [8, 18, 23], recurrent networks [55, 16, 63] or a combination of both [1] (see [41] for more details).

Pedestrian crossing prediction is a sub-problem within action prediction where the objective is to forecast whether a given pedestrian will cross the road at some point in the future. Crossing prediction algorithms rely on a variety of data modalities and architectures. For example, in one of the early works [44] static representations of traffic scene and pedestrian looking and walking actions encoded by CNNs are used to predict crossing action in a given frame. More recent approaches take advantage of sequence analysis techniques. In [49, 36], the authors use DenseNet-based architectures to make prediction based on pedestrian image sequences (along with pose features as in [36]). The method of [46] uses a multimodal approach and inputs different modalities, e.g. scenes, trajectories, according to a different level of complexity, into different levels of a stacked GRU network. Graph-based models are also used to model the connection between different body parts [5] or interactions between different agents [28] as part of the prediction system. Another group of algorithms relies on generative models to predict future scene representations, which in turn are fed into a classifier to predict whether a crossing event, by any pedestrians, is taking place [18, 9].

2.2. Crossing prediction evaluation

Comparing pedestrian crossing algorithms is difficult due to differences in their evaluation criteria. Such discrepancies exist even when the same datasets are used. Table 1 lists the recent models that were trained and tested on two publicly available datasets JAAD and PIE and the corresponding evaluation protocols in terms of observation endpoint and length as well as prediction horizon.

There are significant differences in how the observation frames are selected. About half of the works use only the frames preceding the crossing event for training and evaluation. For example, in [57] the last frame before the pedestrian starts to cross is used for predicting the behavior. In [34] observations from 2s before crossing to 0.5s after the crossing began. The rest of the models sample the entire trajectory. This is beneficial for training since there is more data available. In practice, however, the time before and immediately after the crossing point is the most crucial as the vehicle must decide whether to yield to the pedestrian or continue driving. Using the entire trajectory also affects the evaluation since prediction no longer applies if the action

Model	Year	Dataset	Observation endpoint	Observation length (s)	Prediction horizon
ATGC [44]	2017	JAAD	before event	0.3-0.5	next frame
Fang et al. [15]	2018	JAAD	all frames	0.46	next frame
Varytimidis et al. [57]	2018	JAAD	before event	0.03 (one frame)	next frame
CPN+GCN [5]	2019	JAAD	before event	10	next frame
Res-EnDec [18]	2019	JAAD	all frames	0.533	0.533
Neogi et al. [34]	2019	JAAD	before event	-	1.33
Pop et al. [38]	2019	JAAD	all frames	0.666	1.33
ST-DenseNet [49]	2019	JAAD	all frames	0.533	next frame
SF-GRU [46]	2019	PIE	before event	0.5	2
Chaabane et al. [9]	2020	JAAD	all frames	0.533	0.533
Kotseruba et al. [26]	2020	PIE	before event	0.5	0.3/0.5/1
Liu et al. [28]	2020	JAAD	before event	-	1/2/3
FuSSL-Net [36]	2020	JAAD	before event	0.533	0.533
VRUNet [40]	2020	JAAD	all frames	0.5/1	1

Table 1: Pedestrian action prediction models trained and evaluated on JAAD and PIE datasets. *Observation endpoint* indicates whether the entire pedestrian trajectory was sampled or only frames preceding the crossing action. *Observation length* is specified in seconds (‘-’ for value not specified). *Prediction horizon* indicates how early the action is predicted.

is already in progress. As experiments in previous works [46, 34, 28] and in Section 5.2 of this paper demonstrate, action prediction becomes progressively easier closer to the event (in terms of TTE) because the pedestrian is closer and more visual cues are observable. When the pedestrian is already crossing, most models achieve high prediction accuracy boosting their overall score and confounding their performance during the crucial moments before the crossing commences. To mitigate this, some authors subdivide samples into more fine-grained categories, e.g. *stopping*, *continuous crossing*, *starting* and *standing* in [34] and similar groups in [37], and evaluate the models on each subset.

The observation length also varies widely from a single frame used in [57] to 300 frames (10s) [5], although around 0.5s (or 15 – 16 frames at 30 fps) observation is more common. As for the prediction horizon, one-third of the models predict the action immediately following the observation and the rest predict the crossing action sometime in the future (0.5 – 3s) which is more desirable for practical scenarios as it gives the driver or an autonomous vehicle the time to respond to the unexpected events.

With the aim of resolving some of the inconsistencies pointed out above, our paper offers the following **contributions**: **1)** establishing common evaluation criteria motivated by the behavioral literature for the pedestrian crossing prediction task, **2)** implementations of baselines and state-of-the-art models, **3)** analysis of their performance with respect to various data properties, **4)** a new model that achieves SOTA on accurate crossing prediction.

3. Benchmark

Crossing action prediction task. We formulate the pedestrian action prediction as a binary classification problem where the objective is to predict whether the pedestrian will

start crossing the street at some time t given the observation of length m . We define the event at the time the pedestrian starts to cross or the last frame the pedestrian is observable in case no crossing takes place.

Datasets. For this benchmark we use two large public naturalistic datasets for studying pedestrian behavior in traffic: Joint Attention for Autonomous Driving (JAAD) [44] and Pedestrian Intention Estimation (PIE) [42] because they are freely available for unrestricted use by both academia and industry. JAAD contains 346 clips of pedestrians prior to crossing events filmed in several countries and under varying weather conditions. PIE contains 6 hours of continuous footage recorded in Toronto, Canada in clear weather.

Different from the JAAD dataset that focuses on pedestrians intending to cross, PIE provides annotations for all pedestrians sufficiently close to the road who may or may not attempt to cross in front of the ego-vehicle. As a result, both datasets have imbalances: JAAD behavioral data (JAAD_{beh}) is skewed towards pedestrians who are crossing (495) or are about to cross (191), whereas the complete JAAD dataset (JAAD_{all}) adds 2100 other visible pedestrians who are far away from the road and are not crossing. Like JAAD, PIE also contains more non-crossing pedestrians (1322 non-crossing vs 512 crossing), however they are more diverse in terms of appearance, type of behavior they exhibit, and location with respect to the curb.

Data sampling. For the JAAD dataset we use the same data split as in [45] which includes 323 clips (excluding ones recorded at lower resolution and clips with adverse weather and night conditions). The data is divided into 177, 117 and 29 clips for training, testing and validation respectively. The number of pedestrian tracks in data splits is 324, 48 and 276 in JAAD_{beh} and 1355, 202 and 1023 in JAAD_{all}. In the PIE dataset, we follow the data split defined in [42]: videos from *set01*, *set02* and *set06* are used for training, *set04* and *set05* for validation and *set03* for testing. The number of pedestrian tracks in PIE is 880, 243 and 719 in train, validation and test sets.

As illustrated in Figure 1, observation data for each pedestrian is sampled so that the last frame of observation is between 1 and 2s (or 30 – 60 frames) prior to the crossing event start (as provided in the dataset annotations). This time-to-event (TTE) is motivated by traffic studies [47]. Although pedestrians are unlikely to cross $< 2s$ TTE [43] some time should be allotted for emergency maneuver. Longer prediction horizon is impractical due to the unpredictable nature of most urban scenarios and human dynamics. The observation length for all models is fixed at 16 frames. The sample overlap is set to 0.6 for PIE and 0.8 for JAAD.

Features. Observation consists of several explicit features annotated in the JAAD and PIE datasets: bounding box coordinates, ego-vehicle speed (sensor data in PIE

dataset, and qualitative labels in JAAD), and various attributes for pedestrians and the scene.

In addition, we use pose information generated by OpenPose [6] (18 body joint coordinates concatenated into a 36D feature vector) and optical flow for the entire scene calculated using FlowNet2 [21].

Several types of visual features are generated: *scene* - the entire frame, *local box* - frame cropped to the size of the bounding box, *local context* - image area 1.5 the size of the bounding box which gives a wider view of the scene around the pedestrian, and *local surround* - *local context* where area inside the pedestrian bounding box is grayed out.

Metrics. We report the results using standard binary classification metrics: accuracy, ROC curve, AUC, precision, recall and F1 defined as $F1 = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$.

4. Baselines and state-of-the-art models

We included several models covered in Section 2.2 that provided implementations in addition to a number of baselines adopted from the activity recognition domain. Approaches proposed in [18, 9] predict crossing action for the scene, not individual pedestrians, and thus were not included in the benchmark.

Static. A baseline model that uses VGG16 [53] and ResNet50 [19] backends and an fully-connected (fc) layer to predict the action using only the last frame in the observation sequence.

Are They Going to Cross (ATGC) [44]. An architecture consisting of three CNN streams that encode scene features, pedestrian gait and head pose into shared representation for each frame which is then fed into SVM. The final prediction is computed as the mean over all observation frames.

ConvLSTM [51]. A model that takes as input a stack of images, pre-processes them with pre-trained CNN and feeds the extracted features into convolutional LSTM. The last hidden state is input into fc layer for prediction.

SingleRNN (RNN) [26]. The input features are concatenated into a single vector and fed into a recurrent network (GRU [13] or LSTM [20]) followed by an fc layer for action prediction.

Stacked RNN [59]. A stack of RNN layers where each RNN in the stack receives hidden state of the RNN from the layer below.

Multi-stream RNN (MultiRNN) [3] is comprised of separate RNN streams independently processing each feature type. At the end, the hidden states of the RNNs are concatenated and input into an fc layer for prediction.

Hierarchical RNN [14]. An architecture where separate RNN branches independently process each feature type. The corresponding hidden states are concatenated and fed into another RNN layer followed by an fc layer for classification.

Model name	ModelVariants	PIE					JAAD _{beh}					JAAD _{all}				
		Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
Static	VGG16	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	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 [44]	AlexNet	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	0.8
ConvLSTM [51]	VGG16	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	0.54	0.46	0.26	0.23	0.29	0.59	0.55	0.69	0.68	0.7	0.63	0.58	0.33	0.25	0.49
SingleRNN [26]	GRU	0.83	0.77	0.67	0.7	0.64	0.58	0.54	0.67	0.67	0.68	0.65	0.59	0.34	0.26	0.49
	LSTM	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	0.7
MultiRNN [3]	GRU	0.83	0.8	0.71	0.69	0.73	0.61	0.5	0.74	0.64	0.86	0.79	0.79	0.58	0.45	0.79
StackedRNN [59]	GRU	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 [14]	GRU	0.82	0.77	0.67	0.68	0.66	0.53	0.5	0.63	0.64	0.61	0.8	0.79	0.59	0.47	0.79
SFRNN [46]	GRU	0.82	0.79	0.69	0.67	0.7	0.51	0.45	0.63	0.61	0.64	0.84	0.84	0.65	0.54	0.84
C3D [56]	RGB	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 [7]	RGB	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	0.81	0.83	0.72	0.60	0.9	0.62	0.51	0.75	0.65	0.88	0.84	0.8	0.63	0.55	0.73
TwoStream [52]	VGG16	0.64	0.54	0.32	0.33	0.31	0.56	0.52	0.66	0.66	0.66	0.6	0.69	0.43	0.29	0.83

Table 2: Evaluation results for baseline and state-of-the-art models and their variants on PIE and JAAD datasets. JAAD_{beh} is a subset of the dataset with behavioral annotations (i.e. only pedestrians that intend to cross) and JAAD_{all} includes all detected pedestrians. Dashed lines separate different types of architectures: static, recurrent, 3D convolutional and two-stream.

Stacked with multilevel Fusion RNN (SFRNN) [46]. A modification of the Stacked RNN, where the features are processed hierarchically and are gradually fused at each level with more complex features fed at the bottom layers and simpler features at the top.

Convolutional 3D (C3D) [56] network takes as input a stack of RGB frames and generates final prediction by feeding the features to an fc layer.

Inflated 3D (I3D) [7] network takes as input a stack of RGB frames and generates final prediction using an fc layer.

TwoStream [52] uses two CNN branches where the RGB images and optical flow are processed in parallel and combined to generate the final prediction. The prediction is made for each frame in the sequence and averaged.

5. Evaluation

5.1. Training

For models that were previously trained on PIE and JAAD we used the hyper-parameters specified in the corresponding papers. For the remaining models we performed hyper-parameter search by varying model architecture (e.g. backbone, number of hidden units), feature type (optical flow vs rgb), learning rate and number of epochs and selected the best model to report the results. To offset significant dataset bias we applied class weights inversely proportional to the percentage of samples of each class.

5.2. Discussion of the results

Table 2 summarizes the evaluation results for all models and their variants. We split models into 4 groups: static, recurrent, 3D convolution and two-stream. For the following analysis we selected the top-performing models from each category: Static (with ResNet-50 backbone), Multi-

RNN and I3D (trained on optical flow features). Due to low performance of the TwoStream model it was not used in the analysis.

Performance of different architectures. Even though action prediction is a spatio-temporal problem it is still possible to make predictions based on a single image as illustrated by the fact that the simplest Static model outperforms many temporal models on JAAD dataset and few on PIE. This could be partially attributed to the fact that for JAAD only pedestrians with clear intention to cross were selected whereas PIE is much more diverse and contains many samples of pedestrians with and without intention to cross. As a result, samples in PIE are more diverse in terms of appearance, trajectories, poses and surroundings which are more difficult to capture with Static model. Given that JAAD is also smaller than PIE, it might have sufficient data for learning temporal patterns.

Among recurrent models, ConvLSTM is by far the weakest model in the group and is also inferior to the Static model. Other recurrent models incorporate visual information with explicit features such as poses, trajectories, and ego-vehicle speed. All these models show very similar performance across all datasets suggesting that the choice of the particular architecture (GRU vs LSTM) or how features are combined (e.g. parallel, stacked or hierarchical) does not have a significant effect.

3D convolutional networks similar to ConvLSTM rely only on visual information. C3D and I3D process visual information more effectively and achieve results comparable to recurrent models that use additional data. In the case of I3D, we found that using optical flow instead of RGB images led to better results.

The TwoStream model generalized poorly to the task of crossing action prediction despite following best practices

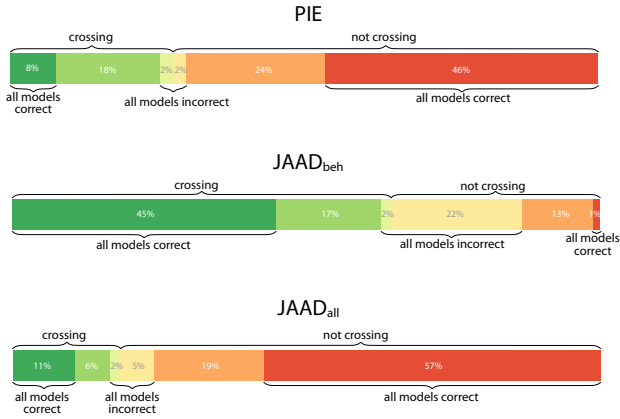


Figure 2: Proportions of crossing and non-crossing samples in PIE and JAAD that are correctly classified by all/some/none of top models from each category: Static, MultiRNN and I3D.

[58]. Perhaps, the reason is that it requires larger and more diverse training data.

Model agreement. Figure 2 shows which crossing and non-crossing pedestrian samples were correctly classified by all, some or none of the models (referred to as *easy*, *medium* and *difficult* samples respectively). We looked at the distribution of the various data properties of samples, however, no strong correlations were found with model agreement. For instance, the distributions of various TTEs, occlusions and bounding box sizes in easy, medium and difficult sets were very similar.

Likewise, there were no strong correlations between model agreement and semantic properties of the scene. For instance, it is reasonable to expect that analyzing pedestrian behavior at unsignalized intersections or behavior of jaywalking pedestrians may be more difficult than at signalized intersections where traffic signals and signs define the behavior of the users. In fact, in JAAD, jaywalking pedestrians (away from intersections) were somewhat harder to classify, whereas in PIE the presence of signal or zebra at the intersection did not correlate with the difficulty of the sample.

Other factors, such as how far the pedestrians were located with respect to the curb, which way they were facing and whether they observed the traffic, were also similarly distributed across easy, medium and difficult sets.

Performance gap between JAAD_{beh} and JAAD_{all}. As mentioned in Section 3, JAAD_{all} extends the set of samples with behavioral annotations (JAAD_{beh}) by adding extra pedestrians detected on the sidewalks thus increasing the number of non-crossing samples by more than 15 times. Having more training data clearly boosts the performance of all models by 10 – 20% in all metrics on JAAD_{all} but also



Figure 3: Examples of crossing and non-crossing pedestrians from the intersection of the hard sets from JAAD_{beh} and JAAD_{all} misclassified by all models.

has undesirable effects on hard samples in JAAD_{all} compared to JAAD_{beh}. For instance, in the set of non-crossing samples shared by the two datasets, there are 3% misclassified by all models in JAAD_{beh} and 3.8% in JAAD_{all}.

Given that misclassification of the crossing samples is potentially more dangerous, it is concerning that the proportion of hard crossing samples increased significantly from 3% in JAAD_{beh} to 10% in JAAD_{all}. At the same time, the number of *easy* crossing samples correctly classified by all models decreased from 69% in JAAD_{beh} to 40% in JAAD_{all}.

A similar conclusion can be made from analyzing the intersection of the hard sets of samples in JAAD_{beh} and JAAD_{all}. The overlap comprises about 30% of the union of the two sets and consists of crossing and non-crossing samples which remain difficult for all models trained on the larger JAAD_{all} dataset. Models trained on JAAD_{all} are significantly better at classifying non-crossing pedestrians compared to models trained on JAAD_{beh} (*easy* set comprises 70% of non-crossing pedestrians in JAAD_{all} and only 1% in JAAD_{beh}). However, the bulk of the improvement is from correctly classifying the additional non-crossing pedestrians in JAAD_{all} that are fairly uniform in their properties (far away from the road and moving alongside the traffic) which helps to distinguish them. At the same time, there is still a significant portion of pedestrians near the road without intention to cross that are incorrectly classified comprising 40% of the overlapping hard sets from JAAD_{beh} and JAAD_{all}.

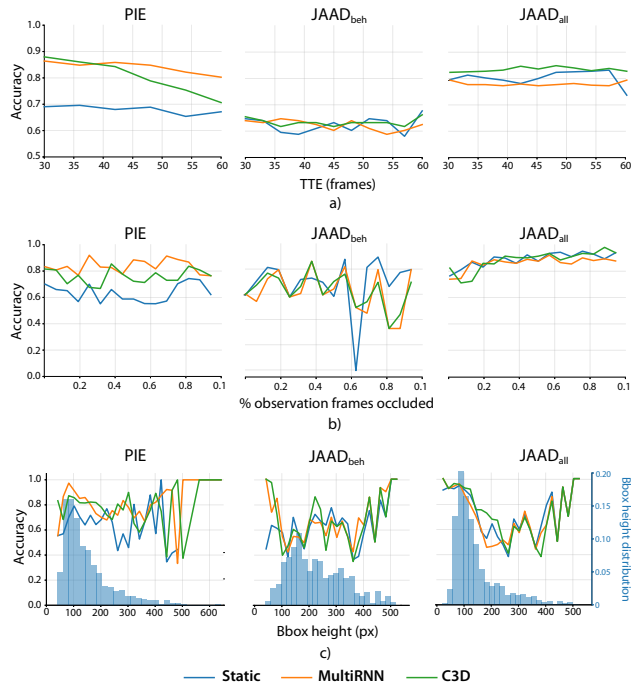


Figure 4: Effects of a) TTE, b) occlusion and c) bounding box height on accuracy of models in PIE, JAAD_{beh} and JAAD_{all} datasets. In c) the distribution of the bounding boxes in the datasets is shown in blue with the corresponding right axis.

Figure 3 shows examples of hard crossing and non-crossing samples. Crossing samples that are difficult for all models typically have low contrast, occlusion and some are crowded, whereas non-crossing samples predominantly consist of pedestrians who walk alongside or in front of the ego-vehicle.

Together these findings suggest that the addition of irrelevant pedestrians increases the overall volume of the training data and boosts the results but does not improve the performance of the models on the most important samples. Overall, the bulk of performance improvement on JAAD_{all} is due to a large number of correctly classified non-crossing pedestrians on the sidewalks at the expense of reduced classification rates of pedestrians close to the road. Although we cannot verify whether the same holds true for the PIE dataset, it might be a factor to consider for future data collection and annotation efforts.

Effect of time-to-event. Even though samples with different TTEs are uniformly distributed in both datasets there is a correlation between TTE and accuracy as shown in Figure 4a. It is especially evident in the PIE dataset where the accuracy of all models gradually decreases by up to 20% as the TTE increases from 30 to 60 frames. A similar trend is also visible in JAAD_{beh} but is not as promi-

nent or consistent across models. A possible explanation is that in JAAD_{beh} the ego-vehicle moves slower than in PIE, as a result, many samples appear reasonably large even at higher TTEs (see pedestrian bounding box height distribution in JAAD_{beh} shown in Figure 4c). In JAAD_{all} this trend is absent due to a large number of non-crossing pedestrian samples on the sidewalks who are easily distinguished from the pedestrians who interact with the ego-vehicle.

Effect of occlusion. The vast majority of the observations in all three datasets are unoccluded (82% in PIE, 85% in JAAD_{beh} and 75% in JAAD_{all}). In only 1 – 2% of samples more than half of the observation frames are fully or partially occluded. As expected, the accuracy of the models fluctuates rather unpredictably on the occluded samples where training and test data are scarce (Figure 4b).

Behavioral research shows that some parts of the pedestrians or scene may be more informative than others for predicting the likelihood of crossing [50, 12]. Likely, computational models are similarly affected by occlusion (e.g. some hard crossing samples are occluded in Figure 3), however deeper investigation is not possible since annotations do not provide specific locations of occluded elements.

Effect of scale. Figure 4c shows the accuracy of the models overlaid on top of the distribution of the bounding boxes heights in the datasets. In PIE, most of the train and test data is concentrated in [80, 120] px range. Note that there is a significant drop-off in performance on smaller bounding boxes with height less than 80 px likely due to insufficient visual information available (especially for 3D convolution models). At the same time, near-perfect performance is achieved by some models on the large bounding boxes > 500 px in the PIE dataset and > 400 px in JAAD_{all} respectively. These are the pedestrian samples that are close to the ego-vehicle and the ego-vehicle itself is typically stationary or moving very slowly creating less noise in the observation.

In JAAD_{beh}, pedestrians of different scales are more uniformly distributed in the train and test data and fluctuation in performance is likely caused by other factors. In JAAD_{all} the addition of other non-crossing pedestrians visible in the scene increases the proportion of samples in the range of [20, 150] px [44] which improves accuracy within that range for all models compared to JAAD_{beh}. The accuracy in the range between [150, 200] px is fluctuating similarly in both datasets.

6. Proposed model

Based on the results of the benchmark summarized in Table 2 where the top-performing models use RNN and 3D convolutions, we propose a new model for pedestrian crossing prediction with attention (PCPA). Our model uses a 3D convolutional branch for encoding visual information and individual RNN branches to process in parallel

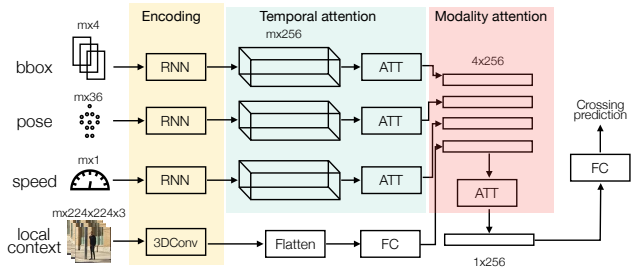


Figure 5: Diagram of the proposed model. The inputs to the model consist of bounding boxes, poses and ego-vehicle speed and visual features represented by local context around the detected pedestrian with observation length $m = 16$. Non-visual features are independently encoded by RNNs and each is fed into an attention block which acts as a temporal attention. 3D convolutional features are flattened and fed through a fully-connected (FC) layer to make their dimension equal to the size of the RNN outputs. Modality attention is applied to the outputs of all 4 branches and the re-weighted outputs are fed into the final FC layer to make crossing prediction.

explicit features such as pose, location, and ego-vehicle speed. Even though I3D performed better than C3D on all datasets, due to its larger size, I3D in combination with RNN branches showed a tendency to overfit especially on the JAAD dataset. Therefore for all experiments below, we use C3D to compute 3D convolutional features.

Given the past successful applications of attention in many vision tasks, we add attention mechanisms inspired by [31, 2]. In one instance it is applied to hidden states of the RNN branches (temporal attention) and again to the outputs of the branches (modality attention). The diagram of the model is shown in Figure 5 and the implementation details follow below.

The model predicts crossing action $A \in \{0, 1\}$ given observation consisting of the following features: 1) local visual context around the pedestrian i $C_{obs} = \{c_i^{t-m}, c_i^{t-m+1}, \dots, c_i^t\}$, where c is a cropped RGB image of the scene around the pedestrian, 2) trajectory $L_{obs} = \{l_i^{t-m}, l_i^{t-m+1}, \dots, l_i^t\}$, where l is a 2D bounding box defined by top-left and bottom-right points, 3) pose of the pedestrian $P_{obs} = \{p_i^{t-m}, p_i^{t-m+1}, \dots, p_i^t\}$, where p is a 36D vector of 2D coordinates of 18 joints detected by OpenPose, and 4) speed of the ego-vehicle $S_{obs} = \{s_i^{t-m}, s_i^{t-m+1}, \dots, s_i^t\}$. t is the time 30 – 60 frames ($\sim 0.5 - 1s$) before the event (crossing or non-crossing).

Each RNN encoder produces a vector $c = q(\{h_1, \dots, h_m\})$ of hidden states $h_i = f(x_i, h_{i-1})$, where f and q are nonlinear functions.

The attention weight vector α is of the same length m as observation and is computed as the score that takes into account the last hidden state of the observation with each

of the previous hidden states. In doing so, it functions as temporal attention that changes the importance of frames in the observation relative to the last seen frame. The weights are expressed as follows:

$$\alpha = \frac{\exp(s(h_m, \bar{h}_i))}{\sum_{i'=1}^T \exp(s(h_m, \bar{h}_{i'}))},$$

where h_m is the last hidden state of the encoder and $s(h_m, \bar{h}_i) = h_m^T W \bar{h}_i$ is the score function with W being a trainable weight matrix.

The last hidden state h_m is concatenated with the sum of all attention weighted hidden states $c_t = \sum_i \alpha_i \bar{h}_i$ (a compact representation of the observed context prior to the last observed frame) and fed into the feedforward network to produce the attention output vector a :

$$a = f(c_m, h_m) = \text{tanh}(W_c [c_m; h_m]).$$

We then apply the same attention mechanism to the weighted outputs of the RNN streams and the output of 3D convolutional branch (flattened and passed through the FC layer with sigmoid activation to reduce the dimension to the number of hidden states), i.e. the new vector c is defined as $c = q(\{\bar{C}, \bar{h}_L, \bar{h}_P, \bar{h}_S\})$, where \bar{h} is the attention weighted last hidden state of the RNN stream. The rest of the computation is the same. This mechanism functions as modality attention.

The output of the modality attention block is then passed to an FC layer for prediction of the crossing action A for the given observation.

Implementation. We use GRUs with 256 hidden units for encoding the non-visual features and C3D network with weights pre-trained on Sports-1M dataset [22] for encoding the visual features (*local context* as defined in Section 3). L2 regularization of 0.001 is added to the final FC prediction layer and a dropout of 0.5 is added after the attention block. The number of observation frames m is set to 16. The ego-vehicle speed feature is used only in the PIE dataset and omitted in JAAD.

Training. We train the model with Adam optimizer [24], binary crossentropy loss and batch size set to 8. We train for 60 epochs on PIE dataset with learning rate set to $5 \cdot 10^{-5}$ and for 80 epochs on JAAD dataset with learning rate $5 \cdot 10^{-7}$.

6.1. Results

The results of the final model with temporal and modality attention are presented in Table 3. Results are most improved on the PIE dataset, where accuracy is increased by 4% and F1 score by 6%. On JAAD_{all} our model performs on par with state-of-the-art across some metrics. Since the inclusion of attention mechanisms increases the number of

Model	Model Variants	PIE			JAAD _{beh}			JAAD _{all}		
		Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
MultiRNN	GRU	0.83	0.8	0.71	0.61	0.5	0.74	0.79	0.79	0.58
C3D	RGB	0.77	0.67	0.52	0.61	0.51	0.75	0.84	0.81	0.65
I3D	Optical flow	0.81	0.83	0.72	0.62	0.51	0.75	0.84	0.8	0.63
PCPA	No attention	0.83	0.83	0.73	0.52	0.48	0.63	0.78	0.8	0.58
	temp. attention	0.85	0.84	0.75	0.56	0.49	0.68	0.81	0.83	0.62
	mod. Attention	0.86	0.86	0.77	0.57	0.48	0.71	0.84	0.79	0.62
	Temp. + mod. attention	0.87	0.86	0.77	0.58	0.5	0.71	0.85	0.86	0.68

Table 3: Results for the proposed model PCPA trained with different types of attention mechanisms. Results for other models are shown for comparison above the dashed line.

3D Features	Acc	AUC	F1
<i>local box</i>	0.86	0.86	0.77
<i>local context</i>	0.87	0.86	0.77
<i>scene context</i>	0.88	0.85	0.78

Table 4: Results for the final PCPA model trained on the PIE dataset with different visual features.

parameters it is beneficial to have a larger and more varied training set. The performance is not as strong on JAAD_{beh}, which has the smallest amount of training data sampled with significant overlap.

Figure 6 shows samples from both datasets showcasing success and failure cases of the model. Some of the correctly classified samples are from the *hard* set described in Section 5.2, e.g. pedestrians facing away from the vehicle and samples with significant occlusion. The pedestrian having a conversation (bottom row) is misclassified by the model suggesting the necessity of modeling the interactions between agents in the scene.

Effect of attention mechanisms. Table 3 also shows the results for PCPA model trained with different types of attention. It can be seen that even when applied individually temporal and modality attention mechanisms improve the results across all metrics and additional boost is achieved when both mechanisms are combined. It should be noted that attention mechanisms increase the complexity of the model and number of parameters and thus require larger and more varied data for training to avoid overfitting.

Effect of visual context. Besides attention, we also experimented with various visual features defined in Section 3. The effect of training on different types of visual context was negligible in the JAAD dataset, likely due to the size and low variability of the data. The results of the experiments on PIE are shown in Table 4. Both contexts around the pedestrian and the entire scene provide superior results compared to using the appearance information only within the bounding box.

7. Conclusions

We proposed a novel evaluation protocol for benchmarking pedestrian action prediction algorithms. Using the new criteria we implemented and evaluated a number of baseline and state-of-the-art models. Using two publicly avail-

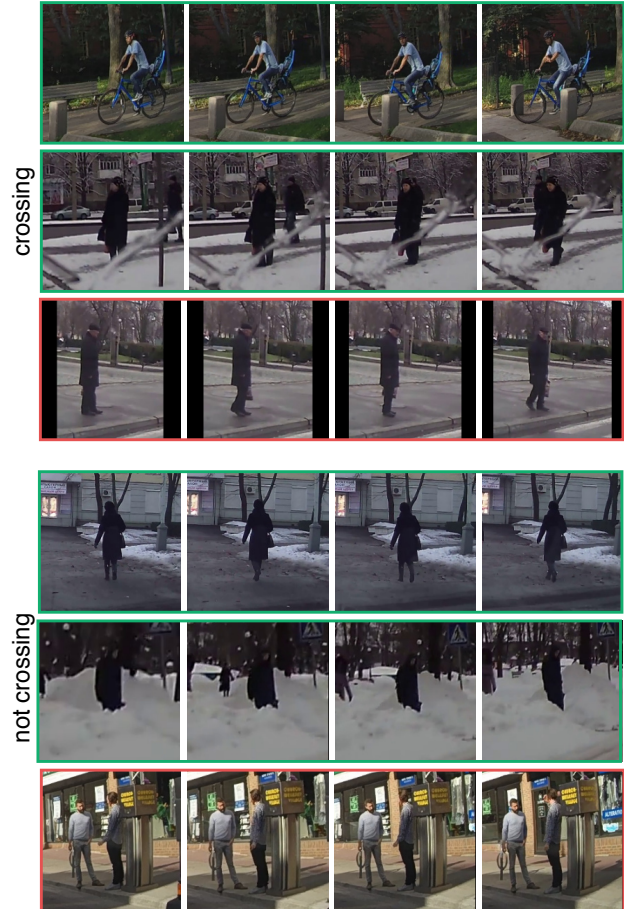


Figure 6: Examples of crossing and non-crossing pedestrians from JAAD and PIE. Samples correctly classified by the proposed model are shown with green outline and misclassified cases are highlighted with red.

able datasets, PIE and JAAD, we analyzed the performance of the static, recurrent, 3D convolutional and two-stream architectures with respect to various properties of the data such as time-to-event, occlusion, and scale. We found that it is difficult to attribute the difficulty of the samples to specific data properties. In addition, we did not observe the consistency in the model agreement.

Based on the benchmark results, we proposed a novel hybrid model combining recurrent and 3D convolutional approaches with temporal and modality attention mechanisms. The proposed model achieves state-of-the-art on both PIE and JAAD datasets.

Acknowledgements. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the NSERC Canadian Robotics Network (NCRN), the Air Force Office for Scientific Research (USA), and the Canada Research Chairs Program through grants to JKT.

References

- [1] Edoardo Alati, Lorenzo Mauro, Valsamis Ntouskos, and Fiora Pirri. Help by predicting what to do. In *IEEE International Conference on Image Processing (ICIP)*, 2019.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [3] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4194–4202, 2018.
- [4] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *IEEE international conference on robotics and automation (ICRA)*, 2018.
- [5] Pablo Rodrigo Gantier Cadena, Ming Yang, Ye qiang Qian, and Chunxiang Wang. Pedestrian Graph: Pedestrian Crossing Prediction Based on 2D Pose Estimation and Graph Convolutional Networks. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2000–2005, 2019.
- [6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.
- [7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.
- [8] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning (CORL)*, 2018.
- [9] Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, and Ross Beveridge. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2297–2306, 2020.
- [10] Junwen Chen, Wentao Bao, and Yu Kong. Group activity prediction with sequential relational anticipation model. In *European Conference on Computer Vision (ECCV)*, 2020.
- [11] Lei Chen, Jiwen Lu, Zhanjie Song, and Jie Zhou. Part-activated deep reinforcement learning for action prediction. In *European Conference on Computer Vision (ECCV)*, 2018.
- [12] Wenxiang Chen, Xiangling Zhuang, Zixin Cui, and Guojie Ma. Drivers recognition of pedestrian road-crossing intentions: Performance and process. *Transportation Research Part F: Traffic Psychology and Behaviour*, 64:552–564, 2019.
- [13] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
- [14] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.
- [15] Zhijie Fang and Antonio M López. Is the pedestrian going to cross? Answering by 2D pose estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1271–1276, 2018.
- [16] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [17] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [18] Pratik Gujjar and Richard Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *International Conference on Robotics and Automation (ICRA)*, pages 2097–2103, 2019.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [21] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017.
- [22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014.
- [23] Qihong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [25] Vineet Kosaraju, Amir Sadeghian, Roberto Martin-Martin, Ian Reid, Hamid Rezaatofghi, and Silvio Savarese. Socialbigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [26] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Do They Want to Cross? Understanding Pedestrian Intention for Behavior Prediction. In *IEEE Intelligent Vehicles Symposium (IV)*, 2020.
- [27] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G. Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [28] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Sheno, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

- [29] Chaochao Lu, Michael Hirsch, and Bernhard Scholkopf. Flexible spatio-temporal networks for video prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [30] Ren C Luo and Licong Mai. Human intention inference and on-line human hand motion prediction for human-robot collaboration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5958–5964, 2019.
- [31] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.
- [32] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [33] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *ECCV*, 2020.
- [34] 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*, 2019.
- [35] Chonhyon Park, Jan Ondřej, Max Gilbert, Kyle Freeman, and Carol O’Sullivan. Hi robot: Human intention-aware robot planning for safe and efficient navigation in crowds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3320–3326, 2016.
- [36] Francesco Piccoli, Rajarathnam Balakrishnan, Maria Jesus Perez, Moraldeepsingh Sachdeo, Carlos Nunez, Matthew Tang, Kajsa Andreasson, Kalle Bjurek, Ria Dass Raj, Ebba Davidsson, et al. FuSSI-Net: Fusion of Spatio-temporal Skeletons for Intention Prediction Network. *arXiv preprint arXiv:2005.07796*, 2020.
- [37] Dnuț Ovidiu Pop. Detection of pedestrian actions based on deep learning approach. *Studia UBB Informatica*, 2019.
- [38] Dănuț Ovidiu Pop, Alexandrina Rogozan, Clement Chate-lain, Fawzi Nashashibi, and Abdelaziz Benschraï. Multi-Task Deep Learning for Pedestrian Detection, Action Recognition and Time to Cross Prediction. *IEEE Access*, 7:149318–149327, 2019.
- [39] Mengshi Qi, Jie Qin, Yu Wu, and Yi Yang. Imitative non-autoregressive modeling for trajectory forecasting and imputation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [40] Adithya Ranga, Filippo Giruzzi, Jagdish Bhanushali, Emilie Wirbel, Patrick Pérez, Tuan-Hung Vu, and Xavier Peroton. VRUNet: Multi-Task Learning Model for Intent Prediction of Vulnerable Road Users. In *International Symposium on Electronic Imaging: Autonomous Vehicles and Machines*, pages 1–9, 2020.
- [41] Amir Rasouli. Deep learning for vision-based prediction: A survey. *arXiv:2007.00095*, 2020.
- [42] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 6262–6271, 2019.
- [43] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269. IEEE, 2017.
- [44] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017.
- [45] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. It’s not all about size: On the role of data properties in pedestrian detection. In *European Conference on Computer Vision (ECCV)*, 2018.
- [46] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian Action Anticipation using Contextual Feature Fusion in Stacked RNNs. 2019.
- [47] Amir Rasouli and John K Tsotsos. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):900–918, 2020.
- [48] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [49] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *International Conference on Robotics and Automation (ICRA)*, pages 9704–9710, 2019.
- [50] Sarah Schmidt and Berthold Faerber. Pedestrians at the kerb—recognising the action intentions of humans. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(4):300–310, 2009.
- [51] Xingjian Shi, Zhoung Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 802–810, 2015.
- [52] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 568–576, 2014.
- [53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [54] Mark Strickland, Georgios Fainekos, and Heni Ben Amor. Deep predictive models for collision risk assessment in autonomous driving. In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [55] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [56] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

- [57] Dimitrios Varytimidis, Fernando Alonso-Fernandez, Boris Duran, and Cristofer Englund. Action and intention recognition of pedestrians in urban traffic. In *International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 676–682, 2018.
- [58] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.
- [59] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, 2015.
- [60] Kuo-Hao Zeng, Shih-Han Chou, Fu-Hsiang Chan, Juan Carlos Niebles, and Min Sun. Agent-centric risk assessment: Accident anticipation and risky region localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [61] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [62] Zhishuai Zhang, Jiyang Gao, Junhua Mao, Yukai Liu, Dragomir Anguelov, and Congcong Li. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [63] He Zhao and Richard P. Wildes. On diverse asynchronous activity anticipation. In *European Conference on Computer Vision (ECCV)*, 2020.
- [64] Siyu Zhou, Mariano J Phielipp, Jorge A Sefair, Sara I Walker, and Heni Ben Amor. Clone swarms: Learning to predict and control multi-robot systems by imitation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4092–4099, 2019.