HAL open science

Assessing Cross-dataset Generalization of Pedestrian Crossing Predictors

Joseph Gesnouin, Steve Pechberti, Bogdan Stanciulescu, Fabien Moutarde

To cite this version:

Joseph Gesnouin, Steve Pechberti, Bogdan Stanciulescu, Fabien Moutarde. Assessing Cross-dataset Generalization of Pedestrian Crossing Predictors. 33rd IEEE Intelligent Vehicles Symposium, Jun 2022, Aachen, Germany. hal-03682452

HAL Id: hal-03682452

https://minesparis-psl.hal.science/hal-03682452v1

Submitted on 31 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Assessing Cross-dataset Generalization of Pedestrian Crossing Predictors

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

***Abstract*—Pedestrian crossing prediction has been a topic of active research, resulting in many new algorithmic solutions. While measuring the overall progress of those solutions over time tends to be more and more established due to the new publicly available benchmark and standardized evaluation procedures, knowing how well existing predictors react to unseen data remains an unanswered question. This evaluation is imperative as serviceable crossing behavior predictors should be set to work in various scenarios without compromising pedestrian safety due to misprediction. To this end, we conduct a study based on direct cross-dataset evaluation. Our experiments show that current state-of-the-art pedestrian behavior predictors generalize poorly in cross-dataset evaluation scenarios, regardless of their robustness during a direct training-test set evaluation setting. In the light of what we observe, we argue that the future of pedestrian crossing prediction, *e.g.* reliable and generalizable implementations, should not be about tailoring models, trained with very little available data, and tested in a classical train-test scenario with the will to infer anything about their behavior in real life. It should be about evaluating models in a cross-dataset setting while considering their uncertainty estimates under domain shift.**

***Index Terms*—Pedestrian Intention Prediction, Uncertainty Estimation, Cross-dataset Evaluation**

I. Introduction

The topic of pedestrian discrete behavior prediction is deemed essential for robust and reliable planning leading to the deployment of autonomous vehicles. While the domain has attracted significant interest in computer vision and robotics communities for the past decade, the field of research has suffered for a long time from the lack of common evaluation protocols and standardized benchmarks, making the task of comparing performance between approaches complex if not impossible to achieve. To compensate for such problems, a standardized benchmark [1] to evaluate pedestrian behavior prediction for three datasets was recently proposed to advance research further. While this brought a breath of fresh air to the field of pedestrian behavior prediction, we believe that current evaluation protocols do not adequately represent the applicability of existing pedestrian prediction models for real-world scenarios. Comparable studies have previously been conducted in computer vision, questioning whether recent progress on the ImageNet [2] benchmark continues to represent meaningful generalization [3] and identifying various sources of bias and noise [4], [5]. However, going beyond accuracy to evaluate a model for a high-risk application with limited amount of training data, such as pedestrian crossing prediction, has never been properly investigated. In this work, we assess how pedestrian action prediction approaches react to small domain shifts and evaluate their generalization capability outside a standard train-test evaluation protocol. We show that all the current pedestrian behavior predictors show signs of over-fitting when evaluated during a direct training-test sets evaluation setting on those standardized benchmarks.

Fig. 1. Examples of crossing and non-crossing pedestrians from $JAAD$ and PIE datasets. The conditions under which pedestrians act from one scenario to another can differ drastically concerning input format and domain shift: pedestrian size, pedestrian positioning in the scene, illumination conditions, occlusion...

This problem leads to two major drawbacks for the field:

- The training source being generally not dense in variety of scenarios nor in the number of examples, the results of state-of-the-art approaches on each dataset might just come from noise: this noise effect should probably be further aggravated since the existing approaches are based on deep learning, depending heavily on the quantity and quality of data where the performance of approaches scales up with the amount of training data.
- It prevents pedestrian behavior predictors from scaling up to real-world applications, as they are not applicable in various scenarios with small domains shifts.

The above examples recap the general motivation of this work, encouraging us to rethink the evaluation methodology to rank current top-scoring behavior predictors from the perspective of uncertainty evaluation to small domain shifts. We argue that:

- The only empirical evaluation of models in a direct train-test sets evaluation offered by the original work

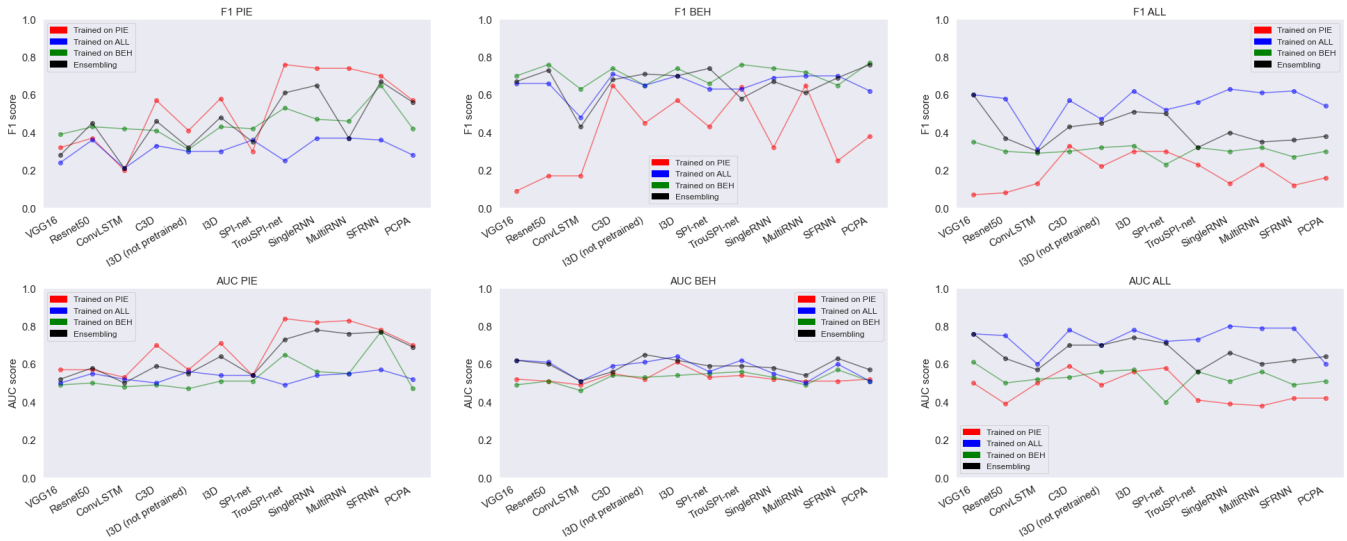

Fig. 2. Pedestrian crossing prediction performance for PIE, $JAAD_{behavior}$ and $JAAD_{all}$. We show a comparison between traditional single-dataset train and test evaluation on each dataset compared to cross-dataset evaluation for eleven methods representing the diversity of architectures and modalities usually used for pedestrian crossing prediction. Ensembling denotes the average output of instances of the same model trained on different training datasets and tested on a given test set.

introducing the method is not sufficient to effectively conclude anything about its applicability in a real-world scenario. The result is often statistically non-significant during a cross-dataset evaluation scenario and leads to an ever-changing state-of-the-art.

- It would be more interesting to compare each method by evaluating how trustworthy are their uncertainty estimates under different domain shifts.

II. Related Work

A. Pedestrian Crossing Prediction

Pedestrian crossing prediction formulates the prediction task as a binary classification problem where the objective is to determine if a pedestrian i will start crossing the street given the context observed up to some time t. The prediction can rely on multiple sources of information, including visual features of the pedestrians and their surroundings, pedestrian kinematics, spatial positioning of the pedestrian based on 2D bounding box locations, optical flow and ego-vehicle speed. Early works [6], [7], formulated the problem as a static image classification problem with 2D Convolutions [8], [9], using only the last frame in the observation sequence to predict crossing behaviors. More successful approaches were designed to take into account temporal coherence in short-term motions of visual features of the pedestrians by using ConvLSTMs [10], [11], 3D Convolutions [12], [13], [14], or Spatio-Temporal DenseNet [15]. Approaches trying to minimize the inference time of their models by avoiding the usage of RGB images were explored: [16] proposes a transformer using only spatial positioning of the pedestrian based on 2D bounding box locations. Crossing prediction based on kinematics only was also explored with various available learning architectures to monitor temporal evolution of skeletal joints such as convolutions [17], [18], [19], recurrent cells [20], [21] or graph-based models [22]. More recently, approaches combining multiple sources of information emerged. By combining several of these perception modalities in order to obtain a multi-modal representation of the scene, one obtain approaches that are often very discriminative and powerful for action prediction. However, this is at the expense of the inference's speed of the model and it highly depends on the quality of the fusion or co-learning algorithm. Therefore, multi-modal approaches differ by the way they merge the available sources, e.g. scenes, trajectories, poses and ego-vehicle speed, and the learning architecture used to infer a crossing prediction, e.g. RNN-based models [23], [24], [25], [1], [26] or Transformer-based models [27], [28].

B. Cross-dataset evaluation

In its first year of existence, proposed approaches evaluated on the benchmarks [1] constantly report higher classification scores [19], [16], [26], [29], [30], [31], giving the impression of clear improvements in pedestrian intention prediction. Usually, a new algorithm is proposed and the implicit hypothesis towards the proposed contribution is made such that it yields an improved performance over the existing state-of-the-art. To confirm such hypothesis, an empirical evaluation of the given contribution is realized in a direct train-test sets evaluation and the quality of the model is evaluated by regular classification metrics: newly proposed methods are then claimed as the new state-of-the-art as soon as they outperform previous ones even by a small margin. However, the ranking of the methods for a given task is currently only as good as the quality of the data used for comparison purposes, and the results obtained by one method on a given dataset do not always reflect its robustness in real-world applications.

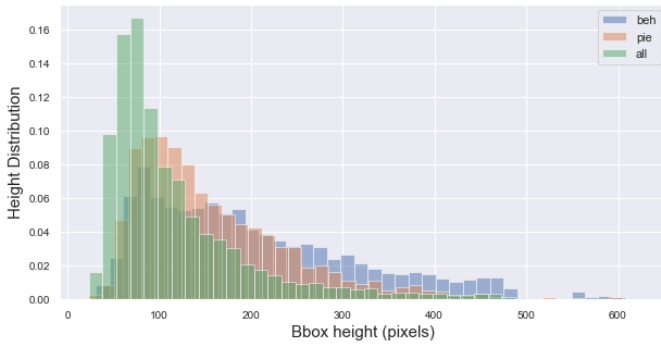

Fig. 3. Distribution of pedestrian bounding box height in pixel for PIE, $JAAD_{behavior}$ and $JAAD_{all}$.

In this work, we evaluate how pedestrian action prediction approaches react to small domain shifts by interchanging the training set of dataset A by the training set of dataset B and test it on the testing set of A. The given training routine is consistent across all experiments for all three datasets. This is referred throughout the paper as cross-dataset evaluation [32], [33], [34]. By adopting cross-dataset evaluation, we test the generalization abilities of several state-of-the-art pedestrian crossing predictors to distributional shift such as pedestrian size, as shown in Fig 3, pedestrian positioning in the scene, illumination conditions or occlusion as shown in Fig 1.

C. Uncertainty Metrics

In real-world scenarios, quantifying uncertainty is crucial as the input distributions are frequently shifted from the training distribution due to a number of causes such as sampling bias. Evaluating the generalization abilities of models by using cross-dataset evaluation and classification metrics only is not sufficient. In high-risk applications such as pedestrian behavior prediction, the idea that a model's predicted probabilities of outcomes reflect true probabilities of those outcomes is mandatory for high-level decisions (*i.e.*, vehicle planning module in crowded urban traffic environments). Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) are standard uncertainty[1] metrics in this context [35], [36], [37], [38]. Predictions are divided into M interval bins according to a given binning strategy, we then calculate the accuracy of each bin to estimate the predicted accuracy from finite data. Let B_m denote the set of sample indices for which prediction confidence is inside one interval bin. The accuracy of B_m is defined as

$$\text{acc}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\left(\hat{y}_i = y_i\right) \tag{1}$$

where $\hat{y}_i$ and y_i are respectively the predicted and true class labels for sample i. The average confidence within one interval bin B_m is defined as:

$$\text{conf}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{2}$$

where $\hat{p}_i$ is the model confidence for sample i. Throughout our experiments, the maximum softmax probability [39] is used as the confidence score. We therefore compare each model output pseudo-probabilities to its accuracy. We obtain the following metrics to rank methods based on their calibration:

Expected Calibration Error (ECE): takes a weighted average of the absolute difference in accuracy and confidence.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}\left(B_m\right) - \text{conf}\left(B_m\right)\right| \tag{3}$$

Maximum Calibration Error (MCE): measures the maximum discrepancy between accuracy and confidence.

$$MCE = \max_{m} \left|\text{acc}\left(B_m\right) - \text{conf}\left(B_m\right)\right| \tag{4}$$

Since the underlying binning approach has a significant impact on the accuracy and reliability of ECE and MCE, we use an adaptive binning strategy [40] instead of a uniform partition[2]: the number of samples in a bin is adaptive to the distribution of the samples in the confidence range.

III. Generalization Capabilities

A. Datasets and Implementation Details

For this evaluation, we use two large public naturalistic datasets for studying pedestrian behavior prediction: $JAAD$ [41] and PIE [42]. These datasets are typically obtained by a vehicle-mounted camera as it navigates through crowded urban traffic environments: $JAAD$ contains 346 clips and focuses on pedestrians intending to cross, PIE contains 6 hours of continuous footage and provides annotations for all pedestrians sufficiently close to the road regardless of their intent to cross in front of the ego-vehicle and provides more diverse behaviors of pedestrians. There are significant differences between JAAD and PIE dataset in terms of sensors: 3 different cameras are used in JAAD with narrow FOV while PIE continuous footage was recorded with with a single wide-angle lens camera. The $JAAD$ dataset is split into $JAAD_{behavior}$ and $JAAD_{all}$. $JAAD_{behavior}$ is biased towards pedestrians attempting to cross the street (402 crossing out of 648) and the smallest dataset available. $JAAD_{all}$ adds all visible pedestrians in $JAAD$, regardless of their position in the scene and contains more non-crossing pedestrians (490 crossing out of 2580). Similarly, PIE contains more non-crossing pedestrians (512 crossing out of 1842). All three datasets are heavily skewed towards one class. To compensate for such significant datasets shifts label-wise, we train all our models using class weights inversely proportional to the percentage of samples for each class. Following the existing evaluation procedures [1], we use the same data sampling method, the same splits and the same inputs sets for our experiments[3]. However, we disregard the ego-vehicle speed input for all our models as the sensor data used for the ego-vehicle speed is only available for PIE and could not be used for cross-dataset evaluation purposes.

[1]Because confidence is the additive inverse of uncertainty with regard to 1, the terms are often interchanged.

[2]https://github.com/yding5/AdaptiveBinning

[3]https://github.com/ykotseruba/PedestrianActionBenchmark

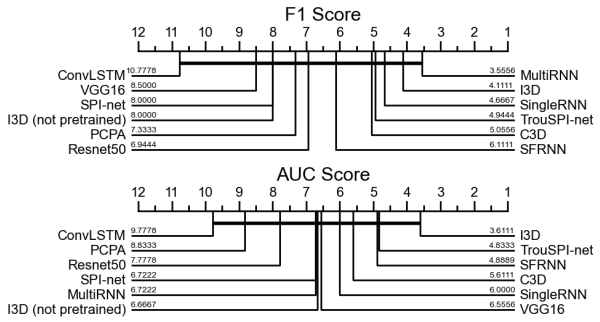

Fig. 4. Critical Difference Diagram [43]: first a Friedman test is performed to reject the null hypothesis, we then proceed with a post-hoc analysis based on the Wilcoxon-Holm method. We compare the robustness of classifiers over multiple training and testing sets shifts. We can see how each method ranks on average. A thick horizontal line groups a set of classifiers that are not significantly different ($\alpha = 0.1$).

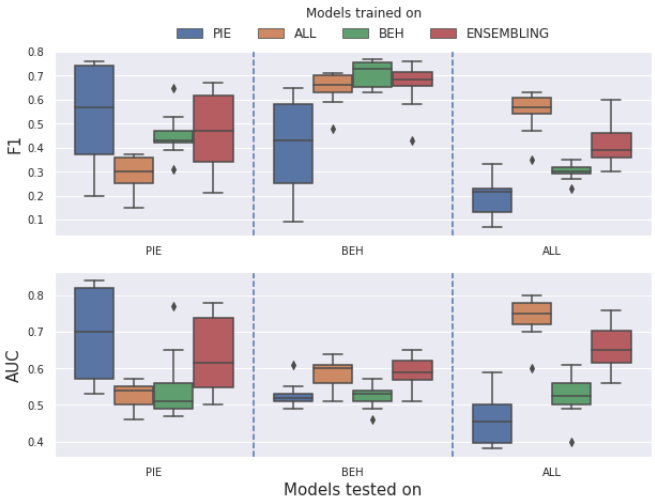

Fig. 5. Distribution of the performance of the eleven selected approaches when evaluated in a direct train-test scenario and when evaluated in cross-dataset scenarios.

The observation length for all models is fixed at 16 frames. In order to combine different models trained on different data sets, the sample overlap is set to 0.8 for both PIE and $JAAD$ trainings. We report the results using standard binary classification metrics: AUC and F1 Score and standard confidence calibration metrics: adaptive ECE and MCE.

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

We select a subset of methods from the pedestrian crossing prediction literature, and more broadly, action recognition literature for their prevalence, practical applicability and diversity in terms of architectures and input modalities. These include:

- **VGG16** [8] and **Resnet50** [9] : two baseline static models that use only the last frame in the observation sequence to predict the crossing behavior of a pedestrian.
- **ConvLSTM** [10]: A model using a stack of images as input, pre-process those images with pre-trained CNN and apply ConvLSTM on those features.
- **Convolutional-3D** (*C3D*) [12] and **Inflated-3D** (*I3D*) [13]: two models pretrained on Sports1M [44] using a stack of images as input and applying 3D convolutions to extract features.
- **SPI-net** [18] and **TrouSPI-net** [19]: two multi-modal models relying on pedestrians' pose kinematics extracted by OpenPose [45], relative euclidean distance of key-points and evolution of the pedestrian spatial positioning. Poses sequences are converted into 2D image-like spatio-temporal representations and self-spatio-temporal attention is applied via CNN-based models for multiple time resolutions. Each remaining feature is independently processed via either U-GRUs [46] or feed forward neural network and fused by either applying temporal and modality attention or sent to a *fc* layer to predict crossing behaviors.
- **SingleRNN** [23], **Multi-stream RNN** (*MultiRNN*) [24] and **Stacked with multilevel Fusion RNN** (*SFRNN*) [47]: Three multi-modal models relying on RGB Images extracted by VGG16 [8], pose kinematics extracted by OpenPose [45] and evolution of the pedestrian spatial positioning. Input features are either concatenated into a single vector and sent to a recurrent network followed by a *fc* layer for crossing prediction, either processed independently by GRUs [48] and the hidden state of GRUs are then concatenated and sent into a *fc* layer for crossing prediction or either processed by GRUs [48] and fused gradually at different levels of processing and complexity.
- **Pedestrian Crossing Prediction with Attention** (*PCPA*) [1]: A multi-modal model relying on RGB images extracted by C3D [12], pose kinematics extracted by OpenPose [45] and evolution of the pedestrian spatial positioning. Non-images features are independently encoded by GRUs [48] and each is fed to a temporal attention block. 3D Convoluted features are flattened and fed into a *fc* layer. Modality attention is then applied to all the branches to fuse them into a single representation by weighted summation of the information from individual modalities.

C. Cross-dataset Evaluation Results

We present the coarse results of our cross-dataset evaluation on Fig 2. For readability purposes, the corresponding critical difference diagram is reported on Fig 4 and the average distribution of performance of the selected approaches is reported on Fig 5. The results of the average prediction given by the three models trained on each training set for one given test set is reported on Table I. As expected, all methods, regardless of their architecture or input modalities, suffer a consequent performance drop when trained on PIE and tested on $JAAD$ and vice versa. Fig 5 shows that however robust the individual classifier is, there is a general trend for classifiers to decline when exposed to a different test set than the expected one. This is consistent towards all our experiments with the exception of $JAAD_{behavior}$. $JAAD_{all}$

	AUC (↑)			F1(↑)			ECE(↓)			MCE (↓)		
Method	**pie** (±0.02)	**beh** (±0.02)	**all** (±0.01)	**pie** (±0.03)	**beh** (±0.01)	**all** (±0.02)	**pie** (±0.01)	**beh** (±0.02)	**all** (±0.02)	**pie** (±0.02)	**beh** (±0.03)	**all** (±0.03)
VGG16 [8]	0.52	**0.62**	**0.76**	0.28	0.67	**0.60**	0.07	0.06	0.20	0.24	**0.13**	0.25
Resnet [9]	0.58	0.60	0.63	0.45	0.68	0.37	0.09	0.05	**0.04**	0.37	**0.15**	0.44
ConvLSTM [10]	0.50	0.51	0.57	0.21	0.43	0.30	0.09	0.14	0.10	0.22	0.25	0.41
C3D [12]	0.59	0.56	0.70	0.46	0.73	0.43	0.17	0.08	**0.03**	0.43	**0.12**	**0.11**
I3D [13]	0.64	**0.62**	**0.74**	0.48	0.71	0.51	**0.05**	0.08	0.09	**0.13**	**0.15**	**0.16**
PCPA [1]	0.69	0.57	0.64	0.56	0.67	0.38	0.12	0.04	0.12	0.36	**0.13**	0.28
SingleRNN [23]	**0.78**	0.58	0.66	**0.65**	0.69	0.40	0.09	**0.02**	0.09	0.16	**0.15**	**0.14**
MultiRNN [24]	**0.76**	0.54	0.60	**0.64**	**0.74**	0.35	**0.06**	0.08	0.19	**0.13**	**0.17**	0.378
SFRNN [47]	**0.77**	**0.63**	0.62	**0.67**	0.58	0.36	0.07	0.08	0.11	**0.11**	0.29	**0.16**
Spi-Net [18]	0.54	0.59	0.71	0.35	0.61	0.50	0.10	0.07	0.22	0.30	**0.15**	0.33
TrouSPI-net [19]	0.73	0.59	0.56	0.61	**0.76**	0.32	0.07	0.05	0.24	**0.12**	**0.13**	0.41

TABLE I

AVERAGE PREDICTION GIVEN BY THREE INDIVIDUAL MODELS AND THEIR RESPECTIVE OUTPUTS FOR EACH TEST SET (ENSEMBLING), EACH INDIVIDUAL MODEL IS EITHER TRAINED ON PIE, $JAAD_{behavior}$ OR $JAAD_{all}$. IN ADDITION TO CLASSIFICATION METRICS (WE USE ARROWS TO INDICATE WHICH DIRECTION IS BETTER), WE COMPARE MODELS WITH PREDICTIVE UNCERTAINTY METRICS SUCH AS EXPECTED CALIBRATION ERROR (ECE) AND MAXIMUM CALIBRATION ERROR (MCE). DASHED LINES SEPARATE DIFFERENT TYPES OF ARCHITECTURES.

being an extension to the set of samples with behavioral annotations, $JAAD_{all}$ "generalizes" well on $JAAD_{behavior}$ but unsurprisingly, the converse is far from true. Even when trained on a relatively diverse dataset (PIE) and inferred on a smaller one in comparison ($JAAD_{behavior}$), selected methods barely show signs of generalization. More alarming, some methods even under-performed a random binary guess based on class distribution when exposed to a different testing set than the expected one. While the task, standardized inputs and observation length are the same across all three datasets, none of the tested models reaches a satisfactory level of generalization across any other testing set. When it comes to compare performance towards small domain-shift at the granular level of individuals methods, the critical difference diagram reported on Fig 4, shows that none of the selected methods arise as a clear winner when it comes to cross-dataset ranking. More importantly, the obtained ranks of each method when evaluated under cross-dataset evaluation is far from the one we usually consider when developing pedestrian crossing behavior predictors: some general methods such as I3D or C3D are on par with multi-modal methods specifically designed to tackle the problem of pedestrian crossing prediction, while part of this could be due to the removal of ego-vehicle data which is an important source of information exploited by many multi-modal approaches. This confirms the importance of rethinking the evaluation methodology of our approaches. The ensembling provided in Table I, is the closest plausible approximation of the selected models' robustness for real-world application as it integrates all available conditions and training instances while removing the sampling biases of each specific training set. It shows that the only empirical evaluation of models in a direct train-test sets evaluation is not sufficient to effectively conclude anything about its applicability in a real-world scenario. This also demonstrates that the use of classification metrics alone is not representative of the overall capacity of the models. For two given models which are equivalent with respect to classification metrics (AUC or F1 score), their calibration (ECE and MCE) can differ drastically. This supports our argument that the usage of uncertainty metrics should complement the metrics conventionally used in order to obtain a comprehensive view of the robustness of existing approaches.

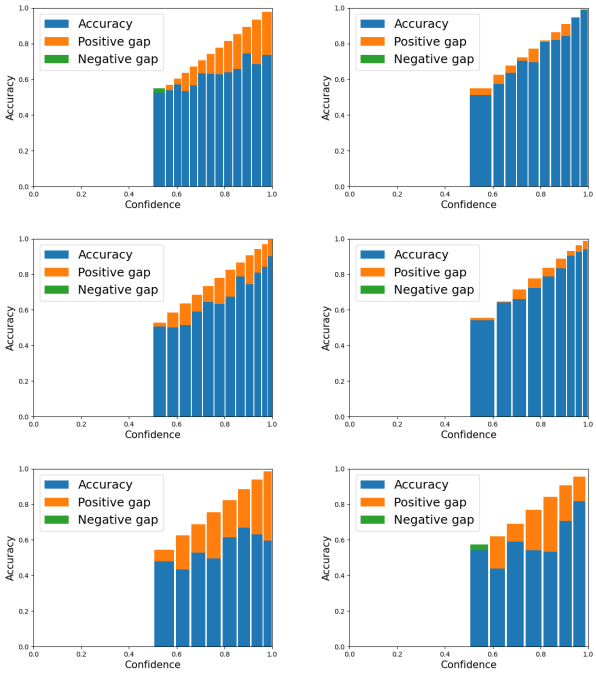

Fig. 6. Reliability Diagrams between I3D [13] randomly initialized (left) and pre-trained on Sports1M [44](right) on PIE, $JAAD_{all}$ and $JAAD_{behavior}$ datasets. If the model is perfectly calibrated, then the diagram plots the identity function. Any deviation from a perfect diagonal represents miscalibration: the model is either overconfident (orange) or subconfident (green).

D. Role of pre-training in uncertainty calibration

Table I illustrates that generic baseline methods (*i.e.* VGG16, C3D, I3D) pre-trained on well diverse and dense datasets further away from the target domain, benefit in terms of generalization and uncertainty calibration as they are on par with the methods specifically designed to tackle the problem of pedestrian crossing prediction, which was not the case in a simple train-test evaluation setting.

To better isolate the effects of pre-training with larger datasets we consider two I3D [13] but trained with different

	AUC (↑)			F1 (↑)			ECE (↓)			MCE (↓)		
Method	**pie**	**beh**	**all**	**pie**	**beh**	**all**	**pie**	**beh**	**all**	**pie**	**beh**	**all**
Non-pretrained	0.55 (±0.04)	0.50 (±0.02)	0.69 (± 0.05)	0.34 (±0.09)	0.65 (±0.08)	0.54 (±0.06)	0.205 (±0.062)	0.184 (± 0.089)	0.111 (± 0.069)	0.290 (±0.039)	0.338 (±0.191)	0.162 (±0.075)
Ens-Nonpretrained	0.59 (±0.06)	0.58 (±0.05)	0.64 (±0.04)	0.37 (±0.10)	0.55 (±0.10)	0.39 (±0.05)	0.065 (±0.022)	0.131 (±0.058)	0.091 (±0.026)	0.248 (±0.122)	0.644 (±0.152)	0.280 (±0.172)
Deterministic	0.72 (±0.01)	0.56 (±0.03)	0.76 (±0.03)	0.60 (±0.02)	0.74 (±0.02)	0.61 (±0.02)	0.026 (±0.007)	0.143 (± 0.020)	0.054 (±0.010)	0.063 (±0.007)	0.239 (± 0.030)	0.118 (±0.020)
Ens-Deterministic	0.64 (±0.01)	0.62 (±0.01)	0.73 (± 0.1)	0.49 (±0.02)	0.70 (±0.01)	0.51 (±0.01)	0.053 (±0.003)	0.080 (±0.001)	0.097 (±0.016)	0.120 (±0.013)	0.138 (±0.024)	0.172 (±0.022)
MC Dropout	0.73 (±0.01)	0.55 (±0.01)	**0.78** (±0.01)	0.61 (±0.01)	0.67 (±0.01)	0.60 (± 0.01)	0.064 (± 0.003)	**0.063** (±0.005)	0.040 (±0.002)	0.106 (±0.004)	**0.134** (±0.012)	**0.059** (±0.009)
Ens-MC Dropout	0.61 (±0.01)	0.61 (±0.01)	0.73 (±0.01)	0.42 (±0.01)	0.49 (±0.02)	0.53 (±0.01)	0.053 (±0.002)	0.053 (±0.003)	0.129 (±0.002)	0.096 (±0.005)	0.120 (±0.013)	0.181 (±0.007)
TempScaling	0.70 (±0.02)	**0.58** (±0.02)	0.76 (±0.01)	0.57 (±0.03)	0.72 (±0.01)	0.61 (± 0.02)	**0.020** (±0.005)	0.070 (± 0.020)	0.037 (±0.08)	**0.050** (±0.010)	0.300 (±0.130)	0.146 (±0.035)
Ens-TempScaling	0.61 (± 0.01)	**0.66** (±0.01)	**0.75** (±0.01)	0.41 (±0.02)	0.69 (±0.01)	**0.56** (±0.01)	0.058 (±0.004)	0.054 (±0.008)	0.142 (±0.015)	0.127 (±0.012)	0.237 (±0.068)	0.215 (±0.016)
LL Dropout	0.71 (±0.01)	0.54 (±0.003)	**0.78** (±0.001)	0.59 (±0.01)	0.74 (± 0.001)	**0.62** (±0.003)	**0.020** (± 0.003)	0.155 (±0.003)	0.061 (± 0.001)	0.063 (±0.012)	0.247 (±0.01)	0.107 (±0.016)
Ens-LL Dropout	0.65 (± 0.002)	0.62 (±0.003)	0.73 (±0.002)	0.49 (±0.003)	0.70 (±0.001)	0.50 (±0.003)	0.052 (±0.003)	0.081 (±0.002)	0.098 (±0.002)	**0.105** (±0.002)	0.145 (±0.023)	0.176 (±0.005)
LL SVI	**0.74** (±0.01)	0.53 (±0.01)	0.77 (± 0.003)	**0.62** (±0.01)	**0.76** (±0.01)	0.57 (±0.004)	**0.021** (±0.002)	0.162 (±0.004)	**0.026** (±0.003)	0.059 (±0.009)	0.214 (±0.006)	**0.054** (±0.009)
Ens-LL SVI	**0.68** (±0.003)	0.61 (±0.003)	0.69 (±0.003)	**0.55** (±0.003)	**0.73** (±0.002)	0.43 (±0.004)	**0.045** (±0.003)	**0.036** (±0.003)	**0.046** (±0.005)	0.146 (±0.027)	**0.075** (±0.009)	**0.133** (±0.016)

TABLE II

AVERAGE PEDESTRIAN CROSSING PREDICTION PERFORMANCE FOR PIE, $JAAD_{behavior}$ AND $JAAD_{all}$ (5 RUNS). DASHED LINES SEPARATE EACH PROBABILISTIC DEEP LEARNING BASELINE. EACH BASELINE IS TESTED TWICE: FIRST, IN A CLASSICAL TRAIN-TEST EVALUATION PROTOCOL AND THEN TESTED BY ENSEMBLING ALL THREE MODELS TRAINED ON EACH TRAINING SET TO EVALUATE ITS ROBUSTNESS TO SMALL DOMAIN SHIFT. WE HIGHLIGHT THE HIGHEST SCORES FOR EACH METRIC AND FOR BOTH EVALUATION PROTOCOLS: TRAIN-TEST OR ENSEMBLING.

configuration: the first one being randomly initialized and the second one being pre-trained on Sports1M [44]. We assess their performance on the same datasets and report our findings in Fig 6. We show that pre-trained models significantly outperform randomly initialized models across all three datasets in terms of calibration. As far as robustness aspects towards small domain shifts are concerned, this may become an important factor to consider when designing pedestrian crossing behavior approaches for real-world scenarios. The training source being generally not dense in variety of conditions nor in the number of examples, the results provided on each dataset might just come from noise on testing sets. Pre-training well-established models on diverse and dense datasets further away from the target domain before fine-tunning to our target task might prove efficient and mandatory for the next step of pedestrian crossing behavior prediction: generalization and vehicle implementation.

IV. IMPROVING UNCERTAINTY CALIBRATION

For the very same approach, there is a significant discrepancy between traditional train-test and cross-dataset evaluation results. This calls into question the reliability of current methods in regard to their capacity to generalize. In addition, we have shown that the standard classification metrics are not sufficient to reliably evaluate an approach since the use of uncertainty metrics raises additional issues that are not reflected otherwise. We are confident that the future breakthroughs in the area will not occur by outperforming current state-of-the-art by a small margin on conventionally used evaluation protocols as they currently fail to provide the big picture of pedestrian crossing behavior prediction.

As we encourage the community to change the direction in which we are taking the research field, we investigate how additional baselines from the probabilistic deep learning literature improve the generalization ability of pedestrian behavior predictors towards small domain shifts. We believe that those methods could prove useful for the next generation of predictors and present our results with the intention that they will serve as a baseline for future work addressing our prescriptions.

A. Baselines from the probabilistic deep learning literature

Below, we present the selected methods from the probabilistic deep learning literature applied on top of an I3D [13] model:

- **Non-pretrained and Deterministic**: Maximum softmax probability [39] of N networks trained independently on each dataset using either random initialization or pre-trained weights from Sports1M [44]. (We set $N = 5$ for each method below.)
- **Monte-Carlo Dropout** (*MC Dropout*): Dropout activated at test time as an approximate bayesian inference in deep Gaussian processes [49].
- **Temperature Scaling**[4] (*TempScaling*): Post-hoc calibration of softmax probability by temperature scaling using a validation set [36].
- **Last Layer Dropout** (*LL Dropout*): Bayesian inference for the parameters of the last layer only: Dropout activated at test time on the activations before the last layer.
- **Last Layer Stochastic Variational Bayesian Inference** (*LL SVI*): Mean field stochastic variational inference on the last layer using Flipout [50].
- **Ensembling** (*Ens*): Average prediction of three networks trained independently on each training set using pre-trained weights [51]. Similarly to Table I, we use ensembling as a plausible approximation of one model's robustness for real-world scenarios.

B. Discussion

We present the results obtained by probabilistic methods for both evaluation protocols: train-test and ensembling on Table II. This allows us to report the effect of dataset shift on accuracy and calibration for the probabilistic deep learning methods. Naturally, we would like to obtain a model, that is well-calibrated on the training and testing distributions of each dataset and remains calibrated with ensembling. We observe that, similarly to the deterministic methods, the quality of predictions consistently degrades with dataset shift regardless of the selected probabilistic method for both PIE and $JAAD_{all}$. However, overall robustness degrades more

[4]https://github.com/gpleiss/temperature_scaling

significantly for some methods. For instance, TempScaling, *e.g.* post-hoc calibration of softmax probability, seems to be one of the best train-test probabilistic methods in regards to expected calibration error (ECE) when evaluated in a standard train-test procedure but falls behind when evaluated under dataset shift. In fact, when evaluated under dataset shift, all the methods except Non-pretrained ones outperform TempScaling in regards to ECE. Similarly, we report that better calibration and accuracy on each test set does not correlate with better calibration under ensembling: the average ECE of the methods when evaluated with classical train-test scenario is [0.166, 0.074, 0.056, 0.042, 0.079, 0.070] and the average ECE of the same methods under dataset shift are [0.096, 0.077, 0.078, 0.085, 0.077, 0.042]. Interestingly, most of the selected probabilistic methods perform better on average than the deterministic I3D under train-test evaluation protocols but fail to generalize when exposed to dataset shift. The exception to the rule is LL SVI, which looks very promising in terms of generalization to small domain shift. As our experiments required pre-trained weights from I3D, we could not replace each convolutional layer with mean-field variational Flipout layers, we only changed the last layer of the given model to obtain a variational bayesian inference for a quick baseline. Nevertheless, we believe that this could be a future research to consider. We should explore the effects of transferring initially learned features on large bases further away from the target task and explore how probabilistic methods react to transfer-learning and domain-shift.

V. Conclusion

In this paper, we show that the classical train-test sets evaluation for pedestrian crossing prediction, *i.e.*, models being trained and tested on the same dataset, is not sufficient to efficiently compare nor conclude anything about their applicability in a real-world scenario: the benchmarks being either too small or too loose in variety of scenarios, it is easy for a given model to over-fit on a specific target dataset. In order to evaluate the generalization capacity of the approaches, we conduct a study based on direct cross-dataset evaluation for eleven methods representing the diversity of architectures and modalities used for pedestrian crossing prediction. We found a huge lack of generalization and robustness for all selected approaches. This led us to a ranking of existing approaches that is much more complex and less absolute than the standard one. We secondly discuss the importance of quantifying a model's uncertainty. Although this is currently completely disregarded, it is common sense to use it in our field of application. We discover two interesting properties: pre-training well-established models on diverse and dense datasets further away from the target domain before fine-tuning to our target task improves calibration and, two models with equivalent classification scores do not necessarily have equivalent calibration scores. This may prove interesting to consider when comparing their usefulness in real-world scenarios with inputs distribution frequently shifted from the training distribution. Finally, we enforce the importance of evaluating the robustness of pedestrian crossing behavior models by evaluating how trustworthy are their uncertainty estimates under domain shifts with cross-dataset evaluation. We encourage the community to consider those new protocols and metrics in order to reach the end-goal of pedestrian crossing behavior predictors: vehicle implementation. In order to build the foundation on which future work should be based on, and, in addition to the eleven deterministic baselines evaluated under domain shift, we report the results of multiple baselines from the probabilistic deep learning literature, designed to tackle the problem of improving model calibration. Given all of the above, we advise the community to change the direction in which we are taking the research field: with so little existing data, non-existent generalization of models, and inconclusive ranking of them, we need to agree to properly evaluate our approaches in order to minimize the noise of our productions and thus, make the research field more sustainable and representative of the real advances to come.

References

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

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord, "Are we done with imagenet?" *arXiv preprint arXiv:2006.07159*, 2020.

[4] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[5] P. Stock and M. Cisse, "Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 498–512.

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

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

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

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

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

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

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

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

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

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

[16] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is attention to bounding boxes all you need for pedestrian action prediction?" *arXiv preprint arXiv:2107.08031*, 2021.

[17] A. Ranga, F. Giruzzi, J. Bhanushali, E. Wirbel, P. Pérez, T.-H. Vu, and X. Perotton, "Vrunet: Multi-task learning model for intent prediction of vulnerable road users," *Electronic Imaging*, vol. 2020, no. 16, pp. 109–1, 2020.

[18] 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.

[19] J. Gesnouin, S. Pechberti, B. Stanciulcscu, and F. Moutarde, "Trouspinet: Spatio-temporal attention on parallel atrous convolutions and u-grus for skeletal pedestrian crossing prediction," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–07.

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

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

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

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

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

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

[26] D. Yang, H. Zhang, E. Yurtsever, K. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *arXiv preprint arXiv:2104.05485*, 2021.

[27] J. Lorenzo, I. P. Alonso, R. Izquierdo, A. L. Ballardini, Á. H. Saz, D. F. Llorca, and M. Á. Sotelo, "Capformer: Pedestrian crossing action prediction using transformer," *Sensors*, vol. 21, no. 17, p. 5694, 2021.

[28] J. Lorenzo, I. Parra, and M. Sotelo, "Intformer: Predicting pedestrian intention with the aid of the transformer architecture," *arXiv preprint arXiv:2105.08647*, 2021.

[29] S. Bapu Sridhar and A. Moosakhanian, "Pedestrian intent prediction using deep machine learning," 2020.

[30] A. Singh and U. Suddamalla, "Multi-input fusion for practical pedestrian intention prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2304–2311.

[31] D. Zhang, F. Shi, Y. Meng, Y. Xu, X. Xiao, and W. Li, "Pedestrian intention prediction via depth augmented scene restoration," in *2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI)*. IEEE, 2021, pp. 1–6.

[32] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 328–11 337.

[33] Y. Chen, P. Liu, M. Zhong, Z.-Y. Dou, D. Wang, X. Qiu, and X. Huang, "Cdevalsumm: An empirical study of cross-dataset evaluation for neural summarization systems," *arXiv preprint arXiv:2010.05139*, 2020.

[34] P. Guo, Z. Xue, L. R. Long, and S. Antani, "Cross-dataset evaluation of deep learning networks for uterine cervix segmentation," *Diagnostics*, vol. 10, no. 1, p. 44, 2020.

[35] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[36] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[37] J. Heo, H. B. Lee, S. Kim, J. Lee, K. J. Kim, E. Yang, and S. J. Hwang, "Uncertainty-aware attention for reliable interpretation and prediction," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 917–926.

[38] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, pp. 13 991–14 002, 2019.

[39] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[40] Y. Ding, J. Liu, J. Xiong, and Y. Shi, "Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 4–5.

[41] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

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

[43] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[44] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

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

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

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

[48] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[49] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[50] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, "Flipout: Efficient pseudo-independent weight perturbations on mini-batches," *arXiv preprint arXiv:1803.04386*, 2018.

[51] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.