It's Not All About Size: On the Role of Data Properties in Pedestrian Detection

Amir Rasouli(✉), Iuliia Kotseruba, and John K. Tsotsos

Department of Electrical Engineering and Computer Science and Center for Vision Research, York University, Toronto, ON M3J 1P3, Canada
{aras,yulia_k,tsotsos}@eecs.yorku.ca

Abstract. Pedestrian detection is central in applications such as autonomous driving. The performance of algorithms tailored to solve this problem has been extensively evaluated on benchmark datasets, such as Caltech, which do not adequately represent the diversity of traffic scenes. Consequently, the true performance of algorithms and their limitations in practice remain understudied.

To this end, we conduct an empirical study using 7 classical and state-of-the-art algorithms on the recently proposed JAAD dataset augmented with 16 additional labels for pedestrian attributes. Using this data we show that the relative performance of the algorithms varies depending on the properties of the training data.

We analyze the contribution of weather conditions and pedestrian attributes to performance changes and examine the major sources of detection errors. Finally, we show that the diversity of the training data leads to better generalizability of the algorithms across different datasets even with a smaller number of samples.

Keywords: Pedestrian detection · Data properties · Pedestrian attributes · Benchmark dataset · Evaluation framework · Autonomous driving

1 Introduction

1.1 Pedestrian Detection

With the rise of autonomous driving systems, visual perception algorithms are facing a new dilemma, that is the ability to detect and recognize objects in highly varying scenes. Among typical objects present in traffic scenes, pedestrians are particularly challenging for identification because they assume different poses, have high variability of appearance and can be easily confused with other objects with similar properties [25].

In the past decades numerous pedestrian detection algorithms [1,7,8,25,31] have been proposed, the majority of which have been tested on the publicly available datasets such as Caltech [5] and KITTI [11]. Although these datasets contain

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11129, pp. 210–225, 2019.
https://doi.org/10.1007/978-3-030-11009-3_12

(a) Localization errors (b) False positives (c) False negatives

Fig. 1. Different sources of detection errors due to the variability in the appearance of the pedestrians and scenes: (a) shows localization errors due to the presence of bags, backpack and umbrellas which are commonly associated with pedestrians observed in the scenes; (b) false positives caused by various environmental factors such as reflections on wet surfaces, over-exposure as well as the presence of objects resembling pedestrians; and (c) false negatives due to variation in shape, e.g. children who have different aspect ratio compared to adults, and appearance, e.g. pedestrians wearing hooded jackets, holding umbrellas or carrying bulky backpacks.

an adequately large amount of data for evaluating the performance of pedestrian detection algorithms, they lack sufficient variability in scene properties such as different lighting conditions and pedestrians' appearance corresponding to different weather conditions.

Given the dynamic nature of driving and the fact that autonomous vehicles should be able to handle a wide range of conditions robustly (see examples in Fig. 1), there is a need to examine the performance of pedestrian detection algorithms and measure their limitations under various visual conditions.

A number of past studies have investigated the role of data properties, such as deformation and occlusion [17], ground truth annotation [23,30], and scale [18] in pedestrian detection algorithms. What is missing, however, is determining the effects of visual appearances due to pedestrian attributes and environmental conditions.

The newly proposed detection datasets collected under various conditions, such as CityPersons [32] and JAAD [20], provide the opportunity to further investigate the role of data properties in the performance of pedestrian detection algorithms. To this end, we analyze the performance of state-of-the-art pedestrian detection algorithms using the publicly available JAAD dataset for which we annotated all pedestrian samples with information regarding their appearance, such as clothing, accessories, objects being carried and pose.

Using the newly annotated dataset together with available properties of JAAD, we show performance variation in detection algorithms as a result of the changes in train/test data. In particular, we investigate the influence of weather conditions under which the data is collected and attributes that impact

appearance and visibility of pedestrians. We also examine the effect of data diversity on generalizability by cross-evaluating the state-of-the-art pedestrian detection algorithms on the JAAD and Caltech datasets. As part of our contribution, we release an evaluation framework for training and testing pedestrian detection algorithms using common benchmarks and evaluation metrics.

2 Related Works

Pedestrian detection is a well-studied field. Over the years, a wide range of algorithms have been developed, ranging from models based on hand-crafted features [7,14,31] to modern convolutional neural networks [1,8,29], and hybrid algorithms benefitting from a combination of both of these techniques [15,28].

The modern pedestrian detection algorithms use various techniques to overcome the challenges of identifying pedestrians in the wild. For example, Tian *et al.* [24] propose a part-based detection algorithm to deal with occlusion. The model consists of a number of part detectors, combinations of which determine the existence of a pedestrian in a given location. In [25], the authors use semantic information of the scene in the form of pedestrian attributes, e.g. carrying a backpack, and scene attributes such as trees or vehicles to distinguish the pedestrians from the background.

In [29] the authors use bootstrapping techniques to mine hard negative samples to minimize confusions caused by background while detecting pedestrians. The proposed algorithm uses features learned by a region proposal network (RPN) to train a cascaded boosted forest for the final hard negative mining and classification. In a more recent approach, Brazil *et al.* [1] show that jointly training a Faster R-CNN network and semantic segmentation network on pedestrian bounding boxes can improve the overall detection results.

As the performance of state-of-the-art pedestrian detection algorithms on benchmark datasets began to saturate (e.g. 7–9% miss rate reported on Caltech [5]), attention has shifted towards the effects of data properties on detection performance. A recent study on generic object recognition tasks shows that order of magnitude increase in the size of training samples can enhance performance even in the presence of up to 20% error in ground truth annotation [22].

As for pedestrian detection algorithms, the effect of occlusion and sample size [17], the balance between negative and positive samples [12], and the cleanness of ground truth annotations [23] have been investigated. Zhang *et al.* [30], for example, demonstrate that the percentage of miss-classification and localization error varies significantly depending on the algorithm. Through experimental evaluations, the authors show that simply by improving the quality of ground truth annotations, localization errors can be significantly reduced resulting in the overall performance boost of more than 7% miss rate in state-of-the-art pedestrian detection algorithms.

2.1 Datasets

There are a number of publicly available pedestrian detection datasets among which some, namely the Caltech [5] and KITTI [11] datasets, are widely used for evaluating the performance of pedestrian detection algorithms. These datasets, although large in scale, lack the diversity in data properties such as weather conditions, geographical locations, pedestrian attributes, etc. For example, Caltech contains 10 h of driving footage collected under sunny and clear weather conditions in streets of Los Angles. Likewise, KITTI is collected under similar weather conditions in streets of Karlsruhe in Germany.

Recently, we have witnessed the emergence of more diverse pedestrian detection datasets. For instance, CityPersons [32] is a pedestrian detection dataset comprised of data collected in various cities across Germany, in different seasons and under different weather conditions. Another pedestrian detection dataset, JAAD [20] is a set of high resolution image sequences collected in different countries and contains video footage recorded under clear and extreme weather conditions such as heavy rain.

The recently proposed datasets provide a variety of scenery and pedestrian samples suitable for studying the limitations of pedestrian detection algorithms under different conditions. Examples of errors caused by the changes data properties are illustrated in Fig. 1.

Despite the introduction of diverse pedestrian detection datasets, there are very few attempts on quantifying the effect of data properties on pedestrian detection algorithms. To this end, in this paper we analyze the effect of data properties in two ways: their impact on the performance of state-of-the-art, and generalizability of the algorithms across different datasets. More specifically, the contributions of this paper are as follows:

1. We introduce a large dataset of pedestrian attributes by annotating the pedestrian samples from the JAAD dataset [20] to study the effect of pedestrian appearance changes on detection algorithms.
2. We examine the performance of state-of-the-art pedestrian detection algorithms with respect to dataset properties and highlight changes in their behavior with respect to different training and testing samples.
3. We perform a cross-evaluation of the state-of-the-art algorithms on the JAAD and Caltech datasets to measure the generalizability of algorithms and datasets based on different properties of the data.
4. We propose a software framework for experimentation and benchmarking classical and state-of-the-art pedestrian detection algorithms using publicly available pedestrian datasets.

3 The Attribute Dataset

There are a number of existing pedestrian attribute datasets that provide fine-grained attributes (e.g. RAP [13], PETA [4]). These datasets primarily cater to applications such as surveillance and identification tasks, and, as a result,

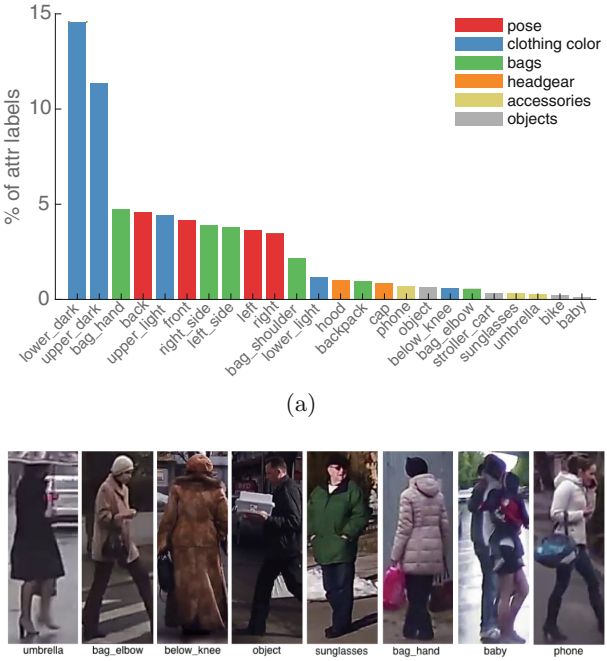

Fig. 2. (a) Types and frequency of new attribute labels in the JAAD dataset color-coded based on the attribute type (e.g. pose, clothing color, accessories); (b) Samples of pedestrians with select attribute labels shown. (Color figure online)

often contain indoor scenes or are recorded using on-site security cameras. Such characteristics make these datasets unsuitable for analyzing pedestrian detection algorithms for applications such as autonomous driving.

Tian *et al.* [25] introduced pedestrian attribute information for the Caltech dataset. The authors augmented the dataset with 9 attributes on 2.7K pedestrian samples. As was mentioned earlier, the Caltech dataset has insufficient variability of weather and scenery properties, hence the attributes lack diversity as well.

To investigate the effect of pedestrian attributes and data properties on detection algorithms, we utilized the publicly available JAAD dataset. The JAAD dataset is a naturalistic driving dataset which comprises videos gathered under different weather and road conditions and contains annotations for video properties, as well as some characteristics of pedestrians (e.g. their age and gender).

We further extended these annotations by adding 16 attributes for each of the 392K pedestrian samples, a total of 900K new attribute labels, summarized in Fig. 2(a)[1]. There are attributes for coarse pose (*left*, *right*, *back*, *front*), clothing

[1] The JAAD attributes are available at https://github.com/ykotseruba/JAAD_pedestrian.

color (*upper_dark* and *lower_dark*) and length (*below_knee* for long coats and skirts).

There are also several attributes for the presence and location of bags and their type: whether they are worn on the *left_side* or *right_side* relative to pedestrian's body and carried on the shoulder (*bag_shoulder*), elbow (*bag_elbow*), back (*backpack*) or held in the hand (*bag_hand*). In addition, we add labels for hooded clothing (*hood*) and caps (*cap*), accessories (e.g. *phone*, *sunglasses*) and various objects that pedestrians can hold in their hands (e.g. *object*, *baby*).

The attributes were selected based on their appropriateness for the driving tasks. For instance, pose of the pedestrian and color of their clothing affect visibility; long clothing obscures the shape and movement of the human body; caps, hoods, and sunglasses occlude pedestrian's face and may limit their view of the traffic scene as well; carrying large bags, backpacks or other objects may not only change appearance and shape of the pedestrian but limit their mobility; holding a phone does not change the pose significantly, but can be used to determine pedestrian's distraction [19], etc.

Clothing color and pose are the only attributes provided for all bounding boxes in the JAAD dataset and form the minimum attribute set. As can be seen from the bar plot in Fig. 2(a), most pedestrians in the dataset are wearing dark clothes, for instance, nearly 70% of pedestrians have both *upper_dark* and *lower_dark* attributes present.

Pose attributes, *left*, *right*, *back*, and *front*, are nearly equally distributed. Aside from clothing color and pose, the *bags* category is the most represented. In fact, nearly 50% of all pedestrians carry a bag or a backpack. In the following sections, we will consider the effect of the diversity and uneven distribution of attributes in the training data on detection.

4 Experimental Setup

4.1 Evaluation Framework

Our framework provides a unified API[2] for experimentation with 10 classical and state-of-the-art pedestrian detection algorithms including SPP+ [16], ACF+ [7], Faster-RCNN [21], CCF [28], Checkerboards [31], DeepPed [26], RPN+BF [29], LDCF+ [14], MS-CNN [2], and SDS-RCNN [1]. All algorithms in the API have training and testing code except SPP+ and DeepPed which only have test code as no official training code has been released by the authors.

The proposed framework is compatible with major publicly available pedestrian detection datasets including INRIA [3], ETH [10], TUD-Brussels [27], Daimler [9], Caltech [5], KITTI [11], CityPersons [32], and JAAD [20]. It allows the manipulation of these datasets in terms of scale, balancing training and testing samples, selection of ground truth, etc. The results can be evaluated using common metrics for pedestrian detection.

[2] The API is available at https://github.com/aras62/PBF.

The software is implemented in Matlab and is based on the code published by the authors of the corresponding algorithms. The training and testing functions are modified for the ease of use from a single API. Our framework uses the original and modified versions of the Piotr toolbox [6] and follows the Caltech benchmark standards [5].

4.2 Algorithms

For the experimental evaluations in this paper we chose three classical algorithms as baselines including ACF+ and its variation LDCF+ [7], and LDCF++ [14], and four state-of-the-art algorithms including RPN+BF [29], MS-CNN [2], SDS-RPN and SDS-RCNN [1] (the top performing algorithm as of ICCV 2017). From RPN+BF algorithm, we only report the results of its RPN component to highlight how the weak segmentation approach proposed in SDS-RPN would behave under different conditions.

The algorithms were trained on the subsets of the JAAD dataset using the default parameters proposed by the authors for the Caltech dataset. The only exception is that we modified the width of training and testing images to maintain the aspect ratio of the images in JAAD. For cross-evaluation with the Caltech dataset, we used the pre-trained models published by the authors of corresponding algorithms.

4.3 Data

The JAAD dataset contains HD quality images with dimensions of 1080×1920 pixels. To maximize the performance of the detection algorithms using default parameters tuned on Caltech, we resized all images to half-scale of 540×960. For evaluation and training, we selected samples with reasonable scale (bounding box height of 50 pixels or more) with partial occlusion (visibility of 75% or more).

For experimental evaluations, we divided JAAD into four different train/test subsets according to the property of the data in terms of weather conditions including *clear*, *cloudy*, *cloudy+clear* (*c+c*) and *mix*. As the names imply, *clear* and *cloudy* subsets only include training images collected under clear and cloudy skies with no rain/snow, and *mix* contains all weather conditions including clear and cloudy, and more extreme weather conditions such as rain/snow. It should be noted that we excluded the videos from the JAAD dataset that were collected under very poor visibility conditions such as nighttime and heavy rain.

The training images for each subset are generated by uniformly sampling 50% of the videos that are recorded under the given condition. Each training subset contains approximately 6.5K pedestrian samples. The remainder of the videos (which may include all weather conditions) are also uniformly sampled and divided into validation and test set.

4.4 Metrics

To report the performance of the algorithms, we use log-average miss rate over the precision range of $[10^{-2}, 10^{0}]$ (MR_2) and $[10^{-4}, 10^{0}]$ (MR_4) false positives

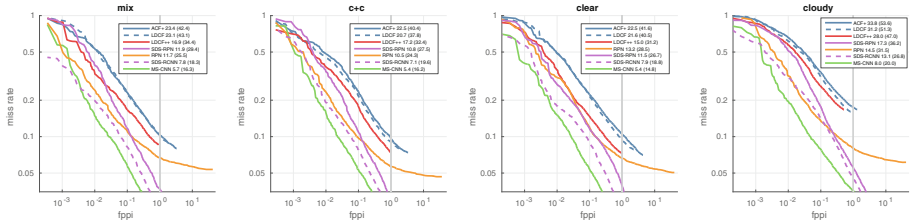

Fig. 3. ROC curves for all algorithms trained and tested on *mix*, *clear*, *cloudy* and *c+c* (clear and cloudy) datasets with detection threshold set to 0.5 IoU. Legends for each plot show the names of algorithms together with $MR_2(MR_4)$ measures. In each plot legend the algorithms are sorted by MR_2 in a descending order.

per image (FPPI) as in [29,30]. We also follow [30] and apply two oracle test cases to measure the contributions of background and localization errors. The localization oracle excludes all false positives that overlap with ground truth from evaluation thus reflecting the contribution of background error. The background oracle does not count false positives that do not overlap with ground truth hence showing the amount of localization error. All of our results are presented using the matching criterion of intersection over union (IoU) ≥ 0.5, unless otherwise stated.

5 Data Properties and Detection Accuracy

5.1 Weather

Weather conditions have multiple effects on the visibility of the pedestrians (e.g. due to rain) and their appearance (e.g. presence of sunglasses or umbrellas). In addition, the appearance of the scene itself may be altered by different lighting conditions, precipitation, reflections, sharp shadows, etc., leading to detection errors as illustrated in Fig. 1. In order to quantify these effects, we trained and tested all pedestrian detection algorithms on different subsets of JAAD dataset split by weather conditions as explained in Sect. 4.

We begin by reporting the ROC curves along with MR_2 and MR_4 metrics. As can be seen in Fig. 3, despite the changes in the overall performance of the algorithms, the rankings are the same across different subsets. The only exception is in the *clear* case where SDS-RPN outperforms RPN.

The main difference between SDS-RPN and regular RPN is that the former adds a weak segmentation component utilizing binary masks from bounding box annotations. It is apparent that using this technique is only effective under clear weather conditions which correspond to the properties of the Caltech dataset that this algorithm was originally tested on (see Table 2). Under different weather conditions, however, the weak segmentation results in a poorer performance compared to the regular RPN.

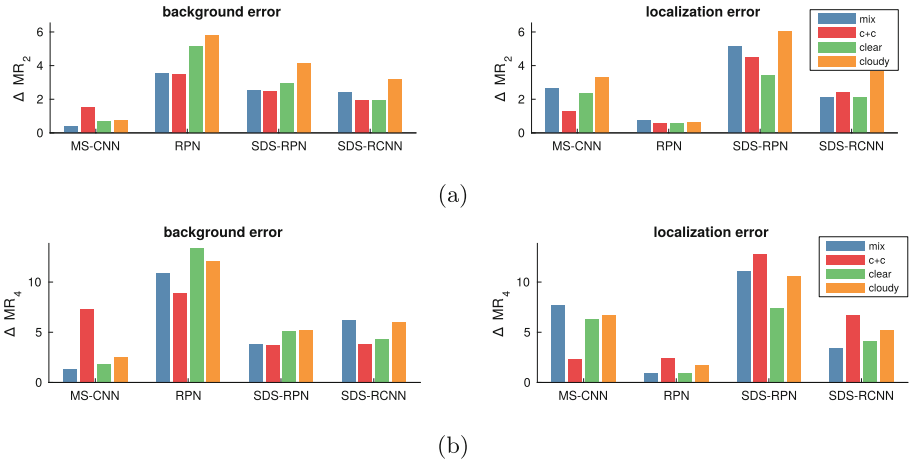

Fig. 4. The relative contribution of background and localization errors to the performance of the pedestrian detection algorithms. The errors are calculated as changes in (a) MR_2 and (b) MR_4 measures for algorithms trained and tested on different subsets of JAAD.

Another observation is that the MS-CNN algorithm (which according to [1] is not among top five performing algorithms on Caltech) achieves the best performance by a large margin (up to 2% on *mix, clear* and *c+c* subsets and more than 5% on *cloudy*) compared to state-of-the-art SDS-RCNN.

To further understand the underlying factors impacting the performance of each algorithm, we report background and localization errors under different weather conditions. As depicted in Fig. 4, testing and training on the subsets of JAAD with different properties reveal inconsistencies in the performance of each detection algorithm as well as their relative performance compared to other algorithms. For example, in the case of *c+c*, MS-CNN reaches its highest background error while at the same time it achieves the lowest localization error compared to others.

For RPN-based models the same trend does not hold as they all perform poorly in terms of localization error, when trained and tested on *c+c*. Comparatively, MS-CNN has the lowest background error on the *mix, clear* and *cloudy* subsets and the second worst on *c+c*.

Likewise, on average, RPN performs best in terms of localization error, however, it is the worst in terms of background error. One interesting observation is the added benefit of the weak segmentation component to RPN (in SDS-RPN) which helps improve the background error but at the price of reducing its localization accuracy.

Table 1. The performance of pedestrian detection algorithms in the presence of individual attributes. The results are reported as MR_4 metric. The top performing algorithms for each attribute are highlighted in bold.

Algorithms	Attributes										
	female	*male*	*pose_back*	*pose_front*	*pose_left*	*pose_right*	*child*	*backpack*	*bag*	*cap_hood*	*umbrella*
ACF+	38.96	34.66	39.71	38.28	34.70	33.91	60.92	38.88	36.00	40.21	69.18
LDCF+	37.02	33.84	35.27	37.24	32.90	30.94	55.02	33.50	33.94	38.27	68.16
LDCF++	30.09	28.30	34.41	31.79	26.44	26.71	55.16	32.76	26.69	33.29	56.64
MS-CNN	**13.49**	**14.03**	17.77	**14.00**	15.20	**11.19**	45.37	16.01	**10.77**	**14.08**	31.06
RPN	21.99	25.79	28.03	26.82	22.72	21.34	53.59	24.59	19.48	28.97	37.35
SDS-RPN	24.31	22.57	26.58	23.67	21.51	22.74	52.54	19.50	20.12	24.61	31.68
SDS-RCNN	14.30	15.77	**17.72**	15.29	**14.46**	13.60	**43.14**	**15.85**	12.25	15.68	**25.57**

5.2 Pedestrian Attributes

In this section, we evaluate the contribution of select attributes (shown in Table 1) on the performance of detection algorithms trained and tested on the *mix* dataset.

Due to the fact that many attributes often appear together in various combinations, it is very hard to disentangle the effect of the individual attributes on the overall detection accuracy of each algorithm. However, major differences can be observed in the relative performances of the algorithms in the presence of certain attributes in the scene.

As one would expect, the performance of classical models is inferior compared to CNN-based algorithms, particularly with respect to some of the rarely occurring attributes such as *child* and *umbrella*. The performance of the state-of-the-art also varies on different attributes. For example, MS-CNN, which shows the highest results on the *mix* subset of JAAD, underperforms compared to SDS-RCNN on select attributes such as *umbrella, backpack, child, pose-back*.

To investigate the common causes of error for MS-CNN and SDS-RCNN we group false positive (FP) and false negative (FN) detections at 0.1 FPPI by the object present in the bounding box as shown in Fig. 5.

With respect to FP, SDS-RCNN and MS-CNN differ greatly not only in the relative contributions of background and localization errors but also in terms of the objects they commonly confuse with pedestrians. Aside from annotation errors, MS-CNN is much more distracted by elongated objects often found in the street scenes, such as tree trunks, hydrants and parking meters.

Many of the localization errors for both MS-CNN and SDS-RCNN are caused by not being able to distinguish pedestrians in groups of 2 or more, particularly when children are also present (attribute *group child* in Fig. 5b). SDS-RCNN also has a higher tendency to place bounding boxes on body parts of the pedestrians or objects they carry (e.g. bags) than MS-CNN. Finally, for both MS-CNN and SDS-RCNN, partially occluded pedestrians, groups of pedestrians and children stand out as main sources of false negative detections.

Note that despite individual sensitivities to certain attributes, both MS-CNN and SDS-RCNN have trouble detecting children and pedestrians with infrequently occurring attributes such as backpacks, umbrellas, hooded clothing, etc.

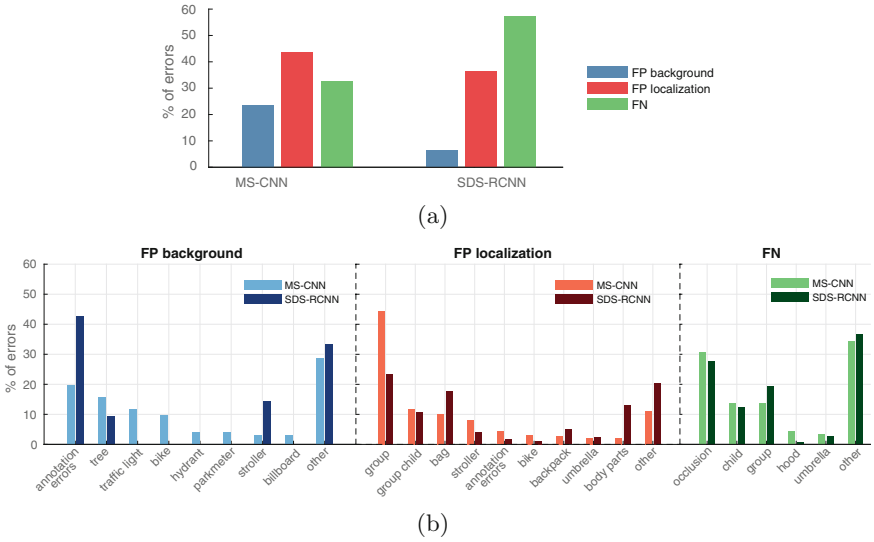

Fig. 5. Error analysis for MS-CNN and SDS-RCNN trained and tested on the *mix* reasonable subset of JAAD. Plot (a) shows the relative percentages of false positives (FP) and false negatives (FN) for each algorithm at 0.1 FPPI. FP is further split into localization and background errors depending on whether the detected bounding box overlaps with the ground truth or not. Plot (b) shows a detailed breakdown of false positive and false negative errors grouped by the corresponding attributes.

There is also evidence that algorithms may learn the appearance of common attributes such as bags instead of the pedestrian itself leading to poor localization.

The former issue may be addressed by increasing the variability of the training data either by explicitly ensuring the presence of certain hard attributes or implicitly, by gathering data under different weather conditions, which in turn affect the appearance of the pedestrians. On the other hand, explicitly learning the attributes may also help, as demonstrated by [25].

5.3 Generalizability Across Different Datasets

Here, our goal is to identify the link between the generalizability of the dataset and its properties, i.e. we want to measure whether using training data from a diverse dataset can improve the performance of detection algorithms on other datasets with more uniform properties.

For this purpose, we employed the widely used Caltech dataset [5] and JAAD. We evaluated the algorithms trained on Caltech using the test data from the *mix* subset of JAAD, and also the models trained on different subsets of JAAD using Caltech test set. All the tests are done on a reasonable set of pedestrians with the height of 50 pixels and above. The minimum allowable visibility is set to 75% on the Caltech test set to match the partial occlusion of the JAAD dataset.

Table 2. The performance of state-of-the-art pedestrian detection algorithms on the Caltech and JAAD *mix* datasets. The table shows the results for algorithms trained and tested on the same dataset. The performances on the Caltech test set are reported on both old (MR^O) and new (MR^N) annotations. The best results are highlighted with blue color.

	$C \rightarrow C$ $MR_2^N(MR_2^O)$	$mix \rightarrow mix$ MR_2
ACF+	26.27 (30.55)	23.36
LDCF+	23.07 (25.79)	23.07
LDCF++	13.66 (16.10)	16.90
RPN	11.71 (14.33)	11.71
MS-CNN	9.47 (11.21)	5.70
SDS-RPN	8.15 (9.27)	11.89
SDS-RCNN	6.58 (7.59)	7.78

Table 3. The performance of state-of-the-art pedestrian detection algorithms on the Caltech and different subsets of the JAAD dataset. The results show the performance of the algorithms trained on Caltech and tested on JAAD ($C \rightarrow mix$) and trained on different subsets of JAAD and tested on Caltech ($J \rightarrow C$). The performances on the Caltech test set are reported on both old (MR^O) and new (MR^N) annotations. The best and second best results are highlighted with blue and green color respectively.

	$C \rightarrow mix$ MR_2	$J \rightarrow C$ $MR_2^N(MR_2^O)$			
		mix	c+c	cloudy	clear
ACF+	77.94	46.97 (53.63)	49.52 (55.06)	70.79 (74.06)	49.99 (55.23)
LDCF+	54.82	43.61 (49.93)	44.89 (50.85)	59.18 (64.11)	47.29 (52.54)
LDCF++	47.94	37.66 (46.04)	40.41 (48.54)	54.86 (60.72)	44.77 (51.93)
RPN	40.15	27.80 (41.19)	25.74 (38.18)	34.67 (47.34)	28.75 (40.05)
MS-CNN	35.09	22.87 (34.83)	26.30 (38.11)	31.55 (46.35)	29.49 (41.64)
SDS-RPN	43.40	24.24 (30.84)	26.64 (33.61)	35.62 (42.90)	30.85 (38.52)
SDS-RCNN	25.45	21.47 (27.73)	25.29 (32.69)	35.20 (42.35)	23.81 (31.75)

Given that a large portion of the original bounding box annotations in the Caltech dataset are poorly localized, following the advice of [30], we report the results on both the original and newly clean Caltech test set. We denote the miss rate results as MR^O and MR^N for old and new annotations respectively. All detections are calculated on $IoU \geq 0.5$. The results of the evaluations of the algorithms trained and tested on the same dataset are summarized in Table 2 and the results of cross-evaluation between algorithms trained and tested on Caltech and subsets of JAAD are shown in Table 3.

The first observation is that the performance of algorithms on a uniform dataset compared to a diverse one varies significantly. SDS-RCNN algorithm that

achieves state-of-the-art performance on Caltech is the second best in JAAD and its counterpart, SDS-RPN, which has the second-best performance on Caltech, performs worse compared to the regular RPN algorithm. MS-CNN, on the other hand, performs best on the *mix* subset, even though on Caltech it is the third best in our evaluation and not even in top five in the latest benchmarks [1].

As was mentioned earlier, the Caltech dataset contains images collected during daylight under the clear sky. Surprisingly, we observe that the *clear* subset of JAAD that has similar properties does not generalize best to Caltech. Besides having the second-best performance on SDS-RCNN models, it ranks third in other cases. In fact, we can see that diversifying the data by training on *c+c* and further adding extreme weather conditions such as rainy and snowy samples achieves the best results on the Caltech dataset.

Partly, such performance improvement is owing to better localization. For instance, MS-CNN and SDS-RCNN on average have IoUs of 0.73 and 0.75 respectively when trained on JAAD *clear* and 0.74 and 0.76 when trained on JAAD *mix*. The same models trained on Caltech, however, have an average IoU of 0.73.

It should be noted that the CNN-based models in the table are trained on Caltech10x [31] which contains over 45K images with more than 16K training samples. The diverse *mix* dataset contains less than 7K samples, yet generalizes better on Caltech than vice versa.

6 Discussion

In this paper, we conducted a series of experiments to investigate the effect of dataset diversity on the performance of pedestrian detection algorithms (see some qualitative examples in Fig. 6). Using the newly proposed JAAD dataset, we showed that the performance measures reported on the classical benchmark

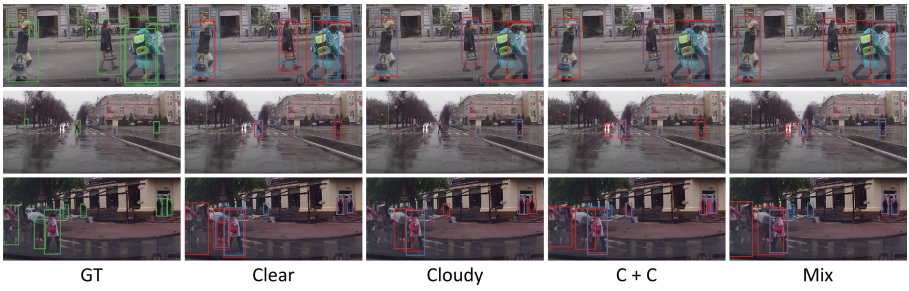

Fig. 6. Examples of the performance of state-of-the-art pedestrian detection algorithms on samples with different weather conditions and pedestrian attributes. From left to right, the ground truth (GT) and the results of algorithms trained on different subsets of the JAAD dataset are shown. Colors green, red and blue correspond to the ground truth, MS-CNN and SDS-RCNN respectively. The results show that the behaviors of both detection algorithms are affected based on the changes in the training data, but in different and somewhat unpredictable ways. For instance, in the example in the second row, SDS-RCNN performs better when trained on the *mix* subset whereas MS-CNN does so when trained on the *clear* subset. (Color figure online)

datasets, such as Caltech, do not necessarily reflect the true potential of detection algorithms in dealing with a wider range of environmental conditions. For instance, MS-CNN which does not even rank top five in the recent state-of-the-art benchmarks, outperforms the current top ranking algorithm, SDS-RCNN, by a significant margin on all subsets of the JAAD dataset.

We showed that the changes in relative performance can be attributed to different properties of the datasets, e.g. depending on what types of weather conditions are represented in the training data. For example, SDS-RPN outperforms the classical RPN on the Caltech dataset owing to the use of a weak segmentation technique, however, it shows inferior results on the JAAD dataset under all weather conditions except clear (which is the most similar to Caltech).

Similar fluctuations in the performance of detection algorithms can be seen with respect to pedestrian attributes. Particularly, rarely occurring attributes such as *child, backpack* and *umbrella* are associated with the highest miss rate for all algorithms. On the other hand, some of the most frequently occurring attributes such as hand bags are shown to be frequently localized instead of the pedestrians.

The diversity of training data also leads to the better generalization of pedestrian detection algorithms across different datasets. Our empirical results suggest that mixing samples with different properties can improve the performance of algorithms even on a more uniform dataset. For example, the MS-CNN algorithm trained on the *mix* subset of JAAD had 7% and 3% lower miss rates on Caltech compared to the models trained on the *clear* and *c+c* subsets respectively.

A carefully selected dataset can also reduce the need for a large volume of training data. For example, the models trained on the *mix* subset of JAAD using only 7K training samples performed better on the Caltech dataset compared to models that were trained on more than 16K training samples from Caltech and tested on the JAAD *mix*.

In conclusion, our study shows that the selection of benchmark datasets for the evaluation of pedestrian detection algorithms for practical applications such as autonomous driving should be revisited to properly assess their performance and limitations under different conditions, and to better reflect the nature of generalizability that is desired.

Using larger datasets certainly benefits the training of the algorithms as does balancing the data with respect to underrepresented weather conditions and pedestrian categories. On the other hand, overrepresented attributes in the data can cause detection errors which should be taken into account when designing pedestrian detection algorithms.

Acknowledgement. This research was supported by several sources, via grants to the senior author, for which the authors are grateful: Air Force Office of Scientific Research USA (FA9550-18-1-0054), the Canada Research Chairs Program (950-231659), and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05352), and the NSERC Canadian Field Robotics Network (NETGP-417354-11).

References

1. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: ICCV, pp. 4950–4959 (2017)
2. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 354–370. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_22
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893 (2005)
4. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: The ACM International Conference on Multimedia, pp. 789–792 (2014)
5. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: CVPR, pp. 304–311 (2009)
6. Dollár, P.: Piotr's Computer Vision Matlab Toolbox (PMT). https://github.com/pdollar/toolbox
7. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. PAMI **36**(8), 1532–1545 (2014)
8. Du, X., El-Khamy, M., Lee, J., Davis, L.: Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection. In: Winter Conference on Applications of Computer Vision (WACV), pp. 953–961 (2017)
9. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. PAMI **31**(12), 2179–2195 (2009)
10. Ess, A., Leibe, B., Schindler, K., van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR, pp. 1–8 (2008)
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. (IJRR) **32**(11), 1231–1237 (2013)
12. Jung, S.I., Hong, K.S.: Deep network aided by guiding network for pedestrian detection. Pattern Recognit. Lett. **90**, 43–49 (2017)
13. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A richly annotated dataset for pedestrian attribute recognition. arXiv:1603.07054 (2016)
14. Ohn-Bar, E., Trivedi, M.M.: To boost or not to boost? on the limits of boosted trees for object detection. In: ICPR, pp. 3350–3355 (2016)
15. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV, pp. 2056–2063 (2013)
16. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 546–561. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_36
17. Rajaram, R.N., Ohn-Bar, E., Trivedi, M.M.: An exploration of why and when pedestrian detection fails. In: International Conference on Intelligent Transportation Systems (ITSC), pp. 2335–2340 (2015)
18. Rajaram, R.N., Ohn-Bar, E., Trivedi, M.M.: Looking at pedestrians at different scales: a multiresolution approach and evaluations. IEEE Trans. Intell. Transp. Syst. **17**(12), 3565–3576 (2016)
19. Rangesh, A., Ohn-Bar, E., Yuen, K., Trivedi, M.M.: Pedestrians and their phones-Detecting phone-based activities of pedestrians for autonomous vehicles. In: International Conference on Intelligent Transportation Systems (ITSC), pp. 1882–1887 (2016)

20. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In: ICCV, pp. 206–213 (2017)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 91–99 (2015)
22. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV, pp. 843–852 (2017)
23. Taiana, M., Nascimento, J., Bernardino, A.: On the purity of training and testing data for learning: the case of pedestrian detection. Neurocomputing **150**, 214–226 (2015)
24. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: ICCV, pp. 1904–1912 (2015)
25. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: CVPR, pp. 5079–5087 (2015)
26. Tomè, D., Monti, F., Baroffio, L., Bondi, L., Tagliasacchi, M., Tubaro, S.: Deep convolutional neural networks for pedestrian detection. Signal Process.: Image Commun. **47**, 482–489 (2016)
27. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 794–801 (2009)
28. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: ICCV, pp. 82–90 (2015)
29. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: ECCV, pp. 443–457 (2016)
30. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR, pp. 1259–1267 (2016)
31. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR, pp. 1751–1760 (2015)
32. Zhang, S., Benenson, R., Schiele, B.: CityPersons: a diverse dataset for pedestrian detection. In: CVPR, pp. 3213–3221 (2017)