Generalizable Pedestrian Detection: The Elephant In The Room

Irtiza Hasan [1], Shengcai Liao[1,†], Jinpeng Li[1], Saad Ullah Akram[2], Ling Shao[1]
Inception Institute of Artificial Intelligence (IIAI)[1], Aalto University, Finland[2]
{irtiza.hasan,shengcai.liao,jinpeng.li,ling.shao}@inceptioniai.org, saad.akram@aalto.fi

Abstract

Pedestrian detection is used in many vision based applications ranging from video surveillance to autonomous driving. Despite achieving high performance, it is still largely unknown how well existing detectors generalize to unseen data. This is important because a practical detector should be ready to use in various scenarios in applications. To this end, we conduct a comprehensive study in this paper, using a general principle of direct cross-dataset evaluation. Through this study, we find that existing state-of-the-art pedestrian detectors, though perform quite well when trained and tested on the same dataset, generalize poorly in cross dataset evaluation. We demonstrate that there are two reasons for this trend. Firstly, their designs (e.g. anchor settings) may be biased towards popular benchmarks in the traditional single-dataset training and test pipeline, but as a result largely limit their generalization capability. Secondly, the training source is generally not dense in pedestrians and diverse in scenarios. Under direct cross-dataset evaluation, surprisingly, we find that a general purpose object detector, without pedestrian-tailored adaptation in design, generalizes much better compared to existing state-of-the-art pedestrian detectors. Furthermore, we illustrate that diverse and dense datasets, collected by crawling the web, serve to be an efficient source of pre-training for pedestrian detection. Accordingly, we propose a progressive training pipeline and find that it works well for autonomous-driving oriented pedestrian detection. Consequently, the study conducted in this paper suggests that more emphasis should be put on cross-dataset evaluation for the future design of generalizable pedestrian detectors. Code and models can be accessed at https://github.com/hasanirtiza/Pedestron.

1. Introduction

Pedestrian detection is one of the longest standing problems in computer vision. Numerous real-world applications, such as, autonomous driving [9, 17], video surveillance [16], action recognition [48] and tracking [21] rely on accurate pedestrian/person detection. Recently, convolutional neural network (CNNs) based approaches have shown considerable progress in the field of pedestrian detection, where on certain benchmarks, the progress is within striking distance of a human baseline as shown in Fig. 1 left.

However, some current pedestrian detection methods show signs of over-fitting to source datasets, especially in the case of autonomous driving. As shown in Fig. 1 right, current pedestrian detectors, do not generalize well to other (target) pedestrian detection datasets, even when trained on a relatively large scale dataset which is reasonably closer to the target domain. This problem prevents pedestrian detection from scaling up to real-world applications.

Despite being a key problem, generalizable pedestrian detection has not received much attention in the past. More importantly, reasons behind poor performances of pedestrian detectors in cross-dataset evaluation has not been properly investigated or discussed. In this paper, we argue that this is mainly due to the fact that the current state-of-the-art pedestrian detectors are tailored for target datasets and their overall design is biased towards target datasets, thus reducing their generalization. Secondly, the training source is generally not dense in pedestrians and diverse in scenarios. Since current state-of-the-art methods are based on deep learning, their performance depend heavily on the quantity and quality of data and there is some evidence that the performance on some computer vision tasks (e.g. image classification) keeps improving at least up-to billions of samples [29].

At present, all autonomous driving related datasets have at least three main limitations, 1) limited number of unique pedestrians, 2) low pedestrian density, i.e. the challenging occlusion samples are relatively rare, and 3) limited diversity as the datasets are captured by a small team primarily for dataset creation instead of curating them from more diverse sources (e.g. youtube, facebook, etc.).

In last couple of years, few large and diverse datasets, CrowdHuman [35], WiderPerson [51] and Wider Pedestrian

[†]Corresponding author.

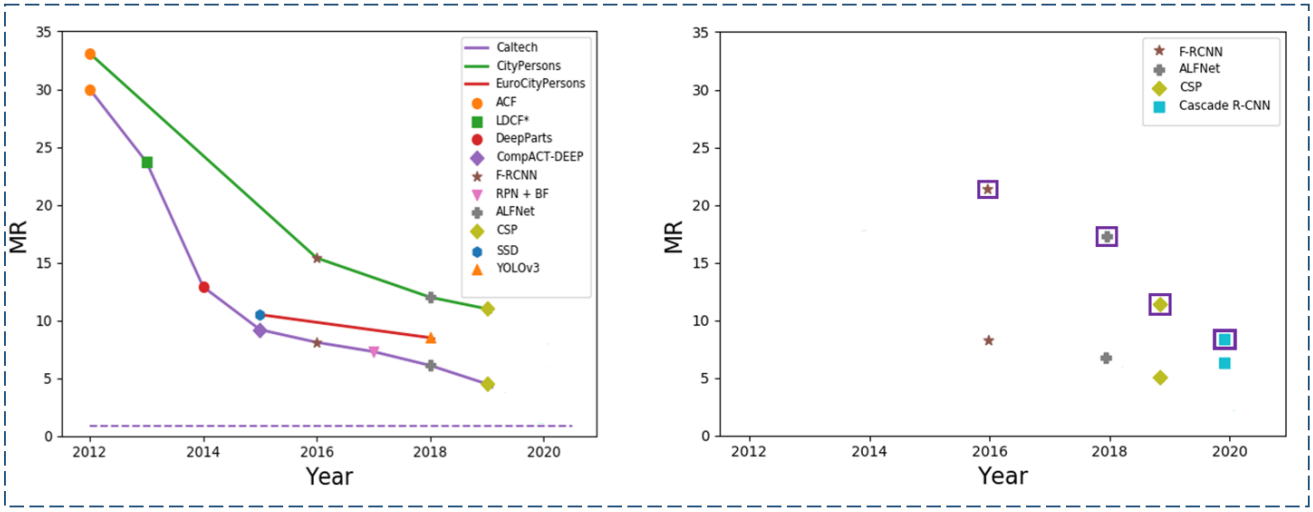

Figure 1: Left: Pedestrian detection performance over the years for *Caltech*, *CityPersons* and *EuroCityPersons* on the reasonable subset. *EuroCityPersons* was released in 2018 but we include results of few older models on it as well. Dotted line marks the human performance on *Caltech*. Right: We show comparison between traditional single-dataset train and test evaluation on Caltech [12] *vs.* cross-dataset evaluation for three pedestrian detectors and one general object detector (Cascade R-CNN). Methods enclosed with bounding boxes are trained on CityPersons [50] and evaluated on Caltech [12], while others are trained on Caltech.

[1], have been collected by crawling the web and through surveillance cameras. These datasets address the above mentioned limitations but as they are from a much broader domain, they do not sufficiently cover autonomous driving scenarios. Nevertheless, they can still be very valuable for learning a more general and robust model of pedestrians. As these datasets contain more person per image, they are likely to contain more human poses, appearances and occlusion scenarios, which is beneficial for autonomous driving scenarios, provided current pedestrian detectors have the innate ability to digest large-scale data.

In this paper, we demonstrate that the existing pedestrian detection methods fare poorly compared to general object detectors when provided with larger and more diverse datasets, and that the state-of-the-art general detectors when carefully trained can significantly out-perform pedestrian-specific detection methods on pedestrian detection task, without any pedestrian-specific adaptation on the target data (see Fig. 1 right). We also propose a progressive training pipeline for better utilization of general pedestrian datasets for improving the pedestrian detection performance in case of autonomous driving. We show that by progressively fine-tuning the models from the largest (but farthest away from the target domain) to smallest (but closest to the target domain) dataset, we can achieve large gains in performance in terms of MR^{-2} on reasonable subset of Caltech (3.7%) and CityPerson (1.5%) without fine-tuning on target domain. These improvement hold true for models from all pedestrian detection families that we tested such as Cascade R-CNN [8], Faster RCNN [34] and embedded vision based backbones such as MobileNet [20]

The rest of the paper is organized as follows. Section 2 reviews the relevant literature. We introduce datasets and evaluation protocol in Sec. 3. We benchmark our baseline in Sec. 4. We test the generalization capabilities of the pedestrian specific and general object detectors in Sec. 5. Finally, conclude the paper in Section 6.

2. Related Work

Pedestrian detection. Before the emergence of CNNs, a common way to address this problem was to exhaustively operate in a sliding window manner over all possible locations and scales, inspired from Viola and Jones [39]. Dalal and Triggs in their landmark pedestrian detection work [10] proposed Histogram of Oriented Gradients (HOG) feature descriptor for representing pedestrians. Dollar et al [11], proposed ACF, where the key idea was to use features across multiple channels. Similarly, [50, 32], used filtered channel features and low-level visual features along with spatial pooling respectively for pedestrian detection. However, the use of engineered features meant very limited generalization ability and limited performance.

In recent years, Convolutional Neural Networks (CNNs) have become the dominant paradigm in generic object detection [34, 18, 38, 24]. The same trend is also true for

the pedestrian detection [2, 19, 7]. Some of the pioneer works for CNN based pedestrian detection [19, 49] used R-CNN framework [15], which is still the most popular framework. RPN+BF [47] was the first work to use Region Proposal Network (RPN); it used boosted forest for improving pedestrian detection performance. This work also pointed out some problems in the underlying classification branch of Faster RCNN [34], namely that the resolution of the feature maps and class-imbalance. However, RPN+BF despite achieving good performances had a shortcoming of not being optimized end-to-end. After the initial works, Faster RCNN [34] became most popular framework with wide range of literature deploying it for pedestrian detection [52, 50, 6, 5, 30, 45].

Some of the recent state-of-the-art pedestrian detectors include ALF [26], CSP [27] and MGAN [33]. ALF [26] is based on Single Shot MultiBox Detector (SSD) [24], it stacks together multiple predictors to learn a better detection from default anchor boxes. MGAN [33] uses the segmentation mask of the visible region of a pedestrian to guide the network attention and improve performance on occluded pedestrians. Similarly, other methods such as [44, 28], used temporal information and bird-eye view map respectively to address occluded pedestrian detection. CSP is an anchor-less fully convolutional detector, which utilizes concatenated feature maps for predicting pedestrians.

Pedestrian detection benchmarks. Over the years, several datasets for pedestrian detection have been created such as Town Center [3], USC [43], Daimler-DB [31], INRIA [10], ETH [13], and TUDBrussels [42]. All of the aforementioned datasets were typically collected for surveillance application. None of these datasets were created with the aim of providing large-scale images for the autonomous driving systems. However, in the last decade several datasets have been proposed from the context of autonomous driving such as KITTI [14], Caltech [12], CityPersons [50] and ECP [4]. Typically these datasets are captured by a vehicle-mounted camera navigating through crowded scenarios. These datsets have been used by several methods with Caltech [12] and CityPersons [50] being the most established benchmarks in this domain. However, Caltech [12] and CityPersons [50] datasets are monotonous in nature and they lack diverse scenarios (contain only street view images). Recently, ECP [4] dataset which is an order of magnitude larger than CityPersons [50] has been porposed. ECP [4] is much bigger and diverse as it contains images from all seasons, under both day and night times, in several different countries. However, despite its large scale, ECP [4] provides a limited diversity (in terms of scene and background) and density (number of people per frame is less than 10). Therefore, in this paper we argue that despite some recent large scale datasets, the ability of pedestrian detectors to generalize has been constrained by lack of diversity and density. Moreover, benchmarks such as WiderPerson [51] Wider Pedestrian [1] and CrowdHuman [35], which contain web crawled images provide a much larger diversity and density. This enables detectors to learn a more robust representation of pedestrians with increased generalization ability.

Cross-dataset evaluation. Previously, other works [4, 35, 50] have investigated the role of diverse and dense datasets in the performance of pedestrian detectors. Broadly, these works focused on the aspect that how much pre-training on a large-scale dataset helps in the performance of a pedestrian detector and used cross-dataset evaluation for this task. However, in this work we adopt cross-dataset evaluation to test the generalization abilities of several state-of-the-art pedestrian detectors. Under this principal, we illustrate that current state-of-the-art detectors lack in generalization, whereas, a general object detector generalizes better to unseen domains and through a progressive training pipeline, significantly surpasses current pedestrian detectors. Moreover, we include more recent pedestrian detection benchmarks in our evaluation setup.

3. Experiments

3.1. Experimental Settings

Datasets. We thoroughly evaluate and compare against state-of-the-art on three large-scale pedestrian detection benchmarks. These benchmarks are recorded from the context of autonomous driving, we refer to them as *autonomous driving* datasets. The **Caltech** [12] dataset has around 13K persons extracted from 10 hours of video recorded by a vehicle in Los Angeles, USA. All experiments on Caltech [12] are conducted using new annotations provided by [49]. **CityPersons** [50] is a more diverse dataset compared to Caltech as it is recorded in 27 different cities of Germany and neighboring countries. CityPersons dataset has roughly 31k annotated bounding boxes and its training, validation and testing sets contain 2,975, 500, 1,575 images, respectively. Finally, **EuroCity Persons** (ECP) [4] is a new pedestrian detection dataset, which surpasses Caltech and CityPersons in terms of diversity and difficulty. It is recorded in 31 different cities across 12 countries in Europe. It has images for both day and night-time (thus referred to as ECP day-time and ECP night-time). Total annotated bounding-boxes are over 200K. As mentioned in ECP [4], for the sake of comparison with other approaches, all experiment and comparisons are done on the day-time ECP. We report results on the validation set of ECP [4] unless stated otherwise. Evaluation server is available for the test set and frequency submissions are limited. Finally, in our experiments we also include two non-traffic related recent datasets namely, **CrowdHuman** [35] and

Table 1: Datasets statistics. ‡ Fixed aspect-ratio for bounding boxes.

	Caltech ‡	CityPersons ‡	ECP	CrowdHuman	Wider Pedestrian
images	42,782	2,975	21,795	15,000	90,000
persons	13,674	19,238	201,323	339,565	287,131
persons/image	0.32	6.47	9.2	22.64	3.2
unique persons	1,273	19,238	201,323	339,565	287,131

Table 2: Experimental settings.

Setting	Height	Visibility
Reasonable	[50, inf]	[0.65, inf]
Small	[50, 75]	[0.65, inf]
Heavy	[50, inf]	[0.2, 0.65]
Heavy*	[50, inf]	[0.0, 0.65]
All	[20, inf]	[0.2, inf]

Wider Pedestrian[1] [1]. Collectively we refer to Caltech, CityPersons and ECP as ***autonomous driving*** datasets and CrowdHuman, Wider Pedestrian as ***web-crawled*** datasets. Details of the datasets are presented in Table 1.

Evaluation protocol. Following the widely accepted protocol of Caltech [12], CityPersons [50] and ECP [4], the detection performance is evaluated using log average miss rate over False Positive Per Image (FPPI) over range $[10^{-2}, 10^{0}]$ denoted by (MR^{-2}). We evaluate and compare all methods using similar evaluation settings. We report numbers for different occlusion levels namely, **Reasonable**, **Small**, **Heavy**, **Heavy*** and **All**, unless stated otherwise, definition of each split is given in Table 2.

Cross-dataset evaluation. In cross-dataset evaluation, when written A→B, we train a model only on the training set of *A* and test it on the testing/validation set of *B*, this training and testing routine is consistent across all experiments.

Baseline. Since most of the top ranked methods on Caltech, CityPersons and ECP are direct extension of Faster/Mask R-CNN [34, 18] family, we also select recent **Cascade R-CNN [8]** (an extension of R-CNN family) as our **baseline**. In text, we interchangeably use baseline and Cascade RCNN, they both refer to exactly the same method Cascade R-CNN [8]. Cascade R-CNN contains multiple detection heads in a sequence, which progressively try to filter out harder and harder false positives. We tested several backbones with our baseline detector as shown in Table 3. HRNet [40] and ResNeXt [46] are two top performing backbones. We choose HRNet [40] as our backbone network. Better performance of HRNet [40] can be attributed to the fact that it retains feature maps at higher resolution, reducing the likelihood of important information being lost in repeated down-sampling and up-sampling, which is especially beneficial for pedestrian detection where the most difficult samples are very small.

[1]Wider Pedestrian has images from surveillance and autonomous driving scenarios. In our experiments, we used the data provided in 2019 challenge. Data can be accessed at : https://competitions.codalab.org/competitions/20132

4. Benchmarking

First, we present the benchmarking of our Cascade R-CNN [8] on three autonomous driving datasets. Table 4 presents benchmarking on Caltech [12] dataset, CityPersons [50] and on ECP [4] respectively. In the case of Caltech and CityPersons, our baseline (Cascade R-CNN [8]) without "bells and whistles" performs comparable to the existing state-of-the-art, which are tailored for pedestrian detection tasks. Its performance has a greater improvement compared to other methods with increasing dataset size. Its relative performance is the worst on the smallest dataset (Caltech) and the best on the largest dataset (EuroCityPersons).

5. Generalization Capabilities

As discussed in the previous sections, traditionally, pedestrian detectors have been evaluated using the classical within-dataset evaluation, i.e., they are trained and tested on the same dataset. We find that existing methods may over-fit on a single dataset, and so we suggest to put more emphasis on cross-dataset evaluation for a new way of benchmarking. Cross-dataset evaluation is an effective way of testing how well a given method adapts to unseen domain. Therefore, in this section we evaluated the robustness of each method using cross-dataset evaluation. [2]

5.1. Cross Dataset Evaluation of Existing State-of-the-Art

In this section we demonstrate that existing state-of-the art pedestrian detectors generalize worse than general object detector. We show that this is mainly due to the biases in the design of methods for the target set, even when other factors, such as backbone, are kept consistent.

To see how well state-of-the-art pedestrian detectors generalize to different datasets, we performed cross dataset

[2]Please see supplementary material for more qualitative and quantitative results.

Table 3: Evaluating generalization abilities of different backbones using our baseline detector.

Backbone	Training	Testing	Reasonable
HRNet	WiderPedestrian + CrowdHuman	CityPersons	**12.8**
ResNeXt	WiderPedestrian + CrowdHuman	CityPersons	**12.9**
Resnet-101	WiderPedestrian + CrowdHuman	CityPersons	15.8
ResNet-50	WiderPedestrian + CrowdHuman	CityPersons	16.0

Table 4: Benchmarking on autonomous driving datasets.

Method	Testing	Reasonable	Small	Heavy
ALFNet [26]	Caltech	6.1	7.9	51.0
Rep Loss [41]	Caltech	**5.0**	**5.2**	47.9
CSP [27]	Caltech	**5.0**	6.8	**46.6**
Cascade R-CNN [8]	Caltech	6.2	7.4	55.3
RepLoss [41]	CityPersons	13.2	-	-
ALFNet [26]	CityPersons	12.0	19.0	48.1
CSP [27]	CityPersons	**11.0**	16.0	39.4
Cascade R-CNN [8]	CityPersons	11.2	**14.0**	**37.1**
Faster R-CNN [4]	ECP	7.3	16.6	52.0
YOLOv3 [4]	ECP	8.5	17.8	37.0
SSD [4]	ECP	10.5	20.5	42.0
Cascade R-CNN [8]	ECP	**6.6**	**13.6**	**33.3**

evaluation of five state-of-the-art pedestrian detectors and our baseline (Cascade RCNN) on CityPersons [50] and Caltech [12] datasets. We evaluated recently proposed BGCNet [22], CSP [27], PRNet [37], ALFNet [26] and FRCNN [50](tailored for pedestrian detection). Furthermore, we added along with baseline, *Faster R-CNN [34]*, without "bells and whistles", but with a more recent backbone ResNext-101 [46] with FPN [23]. Moreover, we implemented a vanilla *FRCNN [50]* with VGG-16 [36] as a backbone and with no pedestrian specific adaptations proposed in [50] (namely quantized anchors, input scaling, finer feature stride, adam solver, ignore region handling, etc).

We present results for Caltech and CityPersons in Table 5, respectively. We also report results when training is done on target dataset for readability purpose. For our results presented in Table 5 (Fourth column, CityPersons→Caltech), we trained each detector on CityPersons and tested on Caltech. Similarly, in the last column of the Table 5, all detectors were trained on the Caltech and evaluated on CityPersons benchmark. As expected, all methods suffer a performance drop when trained on CityPersons and tested on Caltech. Particularly, BCGNet [22], CSP [27], ALFNet [26] and FRCNN [50] degraded by more than 100 % (in comparison with fifth column, Caltech→Caltech). Whereas in the case of Cascade R-CNN [8], performance remained comparable to the model trained and tested on target set. Since, CityPersons is a relatively diverse and dense dataset in comparison with Caltech, this performance deterioration cannot be linked to dataset scale and crowd density. This illustrates better generalization ability of general object detectors over state-of-the-art pedestrian detectors. Moreover, it is noteworthy that BGCNet [22] like the Cascade R-CNN [8], also uses HRNet [40] as a backbone, making it directly comparably to the Cascade R-CNN [8].

Importantly, pedestrian specific FRCNN [50] performs worse in cross dataset (fourth column only), compared with its direct variant vanilla FRCNN. The only difference between between the two being pedestrian specific adaptations for the target set, highlighting the bias in the design of tailored pedestrian detectors.

Similarly, standard Faster R-CNN [34], though performs worse than FRCNN [50] when trained and tested on the target dataset, it performs better than FRCNN [50] when it is evaluated on Caltech without any training on Caltech.

It is noteworthy that Faster R-CNN [34] outperforms state-of-the-art pedestrian detectors (except for BGCNet [22]) as well in cross dataset evaluation, presented in Table 5. We again attribute this to the *bias* present in the design of current state-of-the-art pedestrian detectors, which are tailored for specific datasets and therefore limit their generalization ability. Moreover, a significant performance drop for all methods (though ranking is preserved except for vanilla FRCNN), including Cascade R-CNN [8], can be seen in Table 5, last column. However, this performance drop is attributed to lack of diversity and density of the Caltech dataset. Caltech dataset has less annotations than CityPersons and number of people per frame is less than 1 as reported in Table 1. However, still it is important to highlight, even when trained on a limited dataset, usually general object detectors are better at generalization than state-of-the-art pedestrian detectors. Interestingly, Faster R-CNN's [34] error is nearly twice as high as that of BGCNet [22] in within-dataset evaluation, whereas it outperforms in BGCNet [22] in cross-dataset evaluation.

As discussed previously, most pedestrian detection methods are extensions of general object detectors (FRCNN, SSD, etc.). However, they adapt to the task of pedestrian detection. These adaptations are often too specific to the dataset or detector/backbones (e.g. anchor settings [50, 26], finer stride [50], additional annotations [52, 33], constraining aspect-ratios and fixed body-line annotation [27, 22] etc.). These adaptations usually limit the generalization as shown in Table 5, and also discussed in [25], that task specific configurations of anchors limits general-

Table 5: Cross dataset evaluation on Caltech and CityPersons. A→B refers to training on *A* and testing on *B*.

Method	Bakcbone	CityPersons→CityPersons	CityPersons→Caltech	Caltech→Caltech	Caltech→CityPersons
FRCNN [50]	VGG-16	15.4	21.1	8.7	46.9
Vanilla FRCNN [50]	VGG-16	24.1	17.6	12.2	52.4
ALFNET [26]	ResNet-50	12.0	17.8	6.1	47.3
CSP [27]	ResNet-50	11.0	12.1	5.0	43.7
PRNet [37]	ResNet-50	10.8	10.7	-	-
BGCNet [22]	HRNet	**8.8**	10.2	**4.1**	41.4
Faster R-CNN [34]	ResNext-101	16.4	11.8	9.7	40.8
Cascade R-CNN [8]	HRNet	11.2	**8.8**	6.2	**36.5**

ization.

5.2. Autonomous Driving Datasets for Generalization

We illustrate that even when training dataset is as large as ECP and testing set is as small as Caltech, general object detection methods are better at learning a generic representation for pedestrians compared to existing pedestrian detectors (such as CSP[27]). Moreover, large scale dense autonomous driving datasets provide better generalization abilities.

As illustrated in Section 5.1, cross dataset evaluation provides insights on the generalization abilities of different methods. However, another vital factor in generalization is dataset itself. A diverse dataset should capture the true essence of real world without bias [4], detector trained on such dataset should be able to learn a generic representation that should handle subtle shifts in domain robustly. Deviating from previous studies [4, 35, 51] on the role of dataset in generalization, we perform a line by line comparison between state-of-the-art pedestrian detector and a general object detector when trained and tested on different datasets. In order to provide level playing field, we replace ResNet-50 in CSP [27] with a more powerful and recent backbone HRNet [40]. HRNet's effectiveness can be observed in Table 6, second row, where an improvement of 1.6% (11.0 *vs.* 9.4) in MR^{-2}an be seen.

We begin by using the largest dataset in terms of diversity (more countries and cities included) and pedestrian density from the context of autonomous driving, ECP, for training and evaluate both Cascade RCNN and CSP on CityPersons (Table 6 third and fourth row respectively). It can be seen that Cascade RCNN adapts better on CityPersons, compared to CSP (Reasonable setting), provided the same backbone. ECP is large scale dataset and intuitively one would expect CSP to outperform Cascade RCNN, since in within-dataset evaluation, CSP is better by significant margin (nearly 2% MR^{-2} points).

Furthermore, we swapped our training and testing set, and evaluated on ECP [4]. Cascade RCNN adapted better than CSP, even when the training source is not diverse. Besides *Reasonable* setting, the difference between the performances are at least 5 % MR^{-2} points (across small scale pedestrians, its 10.5 % MR^{-2}). Lastly, we fixed the smallest dataset Caltech as our testing set and used both ECP and CityPersons as our training source. Last four rows of Table 6, illustrates the robustness of a Cascade RCNN across all settings. Importantly, when trained on a dense and diverse dataset ECP Cascade RCNN has more ability to learn a better representation than CSP across all settings.

5.3. Diverse General Person Detection Datasets for Generalization

We investigated how well diverse and dense datasets improve generalization. We conclude, in the case of small autonomous driving datasets, such as Caltech [12], training on diverse and dense sources, which may be further away from the target domain can also benefit. However, in the case of large scale target sets, training on sources close to target domains are more effective. General object detection methods, such as cascade RCNN tend to benefit more from diverse and dense datasets than a pedestrian detector such as CSP.

Table 7, presents results of pre-training of Cascade R-CNN [8] and CSP [27] (HRNet [40] as a backbone) on CrowdHuman [35] and Wider Pedestrian [1] datasets. These two datasets are different from autonomous driving datasets, as CrowdHuman [35] contain web-crawled images of persons in different scenarios and Wider Pedestrian [1] contains images from surveillance cameras and street view images (not just street view images, making them both diverse and dense). Since the autonomous driving datasets, lack in density and diversity [1], CrowdHuman [35] and Wider Pedestrian [1] are a suitable choice for pre-training, since average person per frame and crowd density is much larger in CrowdHuman [35] and Wider Pedestrian [1] combines street view images and surveillance cameras based images, adding a different form of diversity. In Table 7, it can be observed that training on CrowdHuman [35] and Wider Pedestrian [1] can reduce nearly half of the error on Caltech dataset for Cascade RCNN, outperforming previous state-of-the-art, that are trained only on Caltech. Performance improvement is also consistent in CSP [27], though the margin of improvement is less than that of a general

Table 6: Cross dataset evaluation of (Casc. R-CNN and CSP) on Autonomous driving benchmarks. Both detectors are trained with HRNet as a backbone.

Method	Training	Testing	Reasonable	Small	Heavy
Casc. RCNN	CityPersons	CityPersons	11.2	14.0	37.0
CSP	CityPersons	CityPersons	**9.4**	**11.4**	**36.7**
Casc. RCNN	ECP	CityPersons	10.9	**11.4**	40.9
CSP	ECP	CityPersons	11.5	16.6	38.2
Casc. RCNN	ECP	ECP	**6.9**	**12.6**	**33.1**
CSP	ECP	ECP	19.4	50.4	57.3
Casc. RCNN	CityPersons	ECP	17.4	40.5	49.3
CSP	CityPersons	ECP	19.6	51.0	56.4
Casc. RCNN	CityPersons	Caltech	8.8	9.8	**28.8**
CSP	CityPersons	Caltech	10.1	13.3	34.4
Casc. RCNN	ECP	Caltech	**8.1**	**9.6**	29.9
CSP	ECP	Caltech	10.4	13.7	31.3

Table 7: Benchmarking with CrowdHuman and Wider Pedestrian dataset.

Method	Training	Testing	Reasonable	Small	Heavy
Casc. RCNN	CrowdHuman	Caltech	3.4	11.2	32.3
CSP	CrowdHuman	Caltech	4.8	5.7	31.9
Casc. RCNN	CrowdHuman	CityPersons	15.1	21.4	49.8
CSP	CrowdHuman	CityPersons	11.8	18.3	44.8
Casc. RCNN	CrowdHuman	ECP	17.9	36.5	56.9
CSP	CrowdHuman	ECP	19.8	48.9	60.1
Casc. RCNN	Wider Pedestrian	Caltech	3.2	10.8	31.7
CSP	Wider Pedestrian	Caltech	3.4	3.0	29.5
Casc. RCNN	Wider Pedestrian	CityPersons	16.0	21.6	57.4
CSP	Wider Pedestrian	CityPersons	17.0	22.4	58.2
Casc. RCNN	Wider Pedestrian	ECP	16.1	32.8	58.0
CSP	Wider Pedestrian	ECP	24.1	62.6	76.7

object detector. On CityPersons [50], training on CrowdHuman [35] does not improve the performance for CSP [27] or Cascade RCNN, since, CityPersons [50] is a relatively challenging dataset compared to Caltech [12] (in terms of density and diversity), and requires training on sources closer to the domain. This trend can also be seen in the case of ECP [4], where for Cascade RCNN and CSP [27], the performance is lower when trained on CrowdHuman [35], compared to training on CityPersons [50] as in Table 6. Interestingly, in the case of Wider Pedestrian [1] (bottom half Table 7), besides CityPersons [50], the relative improvements in the case of Wider Pedestrian [1] is relatively larger for general object detector. The potential reason is that compared with CrowdHuman [35], Wider Pedestrian [1] is large scale and closer to the target domain. Since it contains images essentially for two views (street view and surveillance), where as CrowdHuman [35] contains web-crawled person images appearing in different poses and scenes.

5.4. Progressive Training Pipeline

We conducted experiments to show that performance can be significantly improved through progressive fine-tuning, where starting from a general diverse dataset (farther from target domain), and subsequently fine-tuning on dataset closer to the target domain.

We conduct additional experiments on the importance of progressive training. To be consistent, we do not fine-tune on the target set and for training we only use the training subset of each respective dataset. $A \rightarrow B$ refers to pre-training on dataset A and fine-tuning on B. Whereas, $A + B$ refers to simply merging the two datasets together and training the model on merged larger set. For our results presented in Table 8, we used CityPersons [50] and Caltech [12] as our testing sets. It can be seen, in Table 8, first two rows, that progressive training pipeline significantly improves the performance of Cascade RCNN. Particularly, pre-training on Wider Pedestrian [1] and fine-tuning

Table 8: Investigating the effect on performance when CrowdHuman, Wider Pedestrian and ECP are merged and Cascade R-CNN [8] is trained only on the merged dataset.

Method	Training	Testing	Reasonable	Small	Heavy
Casc. RCNN	CrowdHuman → ECP	CP	10.3	12.6	40.7
Casc. RCNN	Wider Pedestrian → ECP	CP	**9.7**	**11.8**	**37.7**
Casc. RCNN	Wider Pedestrian + CrowdHuman + ECP	CP	10.9	12.7	43.1
Casc. RCNN	Wider Pedestrian + CrowdHuman → ECP	CP	9.7	12.1	39.8
Casc. RCNN	CrowdHuman → ECP	Caltech	2.9	11.4	30.8
Casc. RCNN	Wider Pedestrian → ECP	Caltech	**2.5**	**9.9**	**31.0**

on ECP [4] brings Cascade RCNN on par with other state-of-the-art approaches on CityPersons [50], without training on the CityPersons [50]. Similarly, in the case of Caltech [12] as well, progressive training, outperformed previously established state-of-the-art on Caltech [12] dataset. Noteworthy is the fact that performance on Caltech [12] is within a close vicinity of a human-baseline (0.88).

Finally, concatenating all datasets (Table 8, third and fourth row), leads to improvement in performance, but it is still slightly worse than the progressive training that we have used, where we fine-tune on the autonomous driving benchmark. The results illustrate that this strategy enables us to significantly improve the performances of state-of-the-art without fine-tuning on the actual target set. This illustrates the generalization capability of the proposed approaches can be enhanced by progressive training strategy, without exposure to the target set, Cascade R-CNN [8] is on par with top performer on CityPersons and best performing on Caltech [12].

5.5. Application Oriented Models

In many pedestrian detection applications, such as autonomous driving and cameras mounted on drones to localize persons, the size and computational cost of models is constrained. We experiment with a small and light-weight model MobileNet [20] v2, which is designed for mobile and embedded vision applications, to investigate if with progressive training pipeline, even with a light backbone, the performance improvements hold true.

Table 9 show results on CityPersons [50] using MobileNet [20] as a backbone network architecture into Cascade R-CNN [8]. First row of Table 9, is for reference when, MobileNet [20] trained and evaluated on CityPersons [50]. Intuitively, MobileNet [20] performs worse than the HRNet [40]. However, in the case of MobileNet [20] as well, we see pre-training on CrowdHuman [35] and fine-tuning on ECP [4] improves the performance of the MobileNet [20]. Furthermore, we replaced CrowdHuman [35] with Wider Pedestrian [1] as the initial source of pre-training. Improvement over the Cascade R-CNN [8] (1st row) can be observed (3rd row), where with Wider Pedestrian [1] pre-

Table 9: Investigating the performance of embedded vision model, when pre-trained on diverse and dense datasets.

Training	Testing	Reasonable	Small	Heavy
CP	CP	12.0	15.3	47.8
ECP	CP	19.1	19.3	51.3
CrowdHuman→ECP	CP	11.9	15.7	48.9
Wider Pedestrian→ECP	CP	**11.4**	**14.6**	**43.4**

training and fine-tuning on ECP [4], a performance gain of 0.6% MR^{-2}an be seen. This is consistent with our previous finding reported in Table 7, Wider Pedestrian [1] is a better source of pre-training than CrowdHuman [35], since it has images of autonomous driving scenes as well, making it more closer to the target domain than CrowdHuman [35]. Interestingly, in the case of CrowdHuman [35] and Wider Pedestrian [1], even with a light-weight architecture, Cascade R-CNN [8] with MobileNet [20], is comparable state-of-the-art pedestrian detector CSP [27] (ResNet-50).

6. Conclusions

Encouraged by the recent progress of pedestrian detectors on existing benchmarks from the context of autonomous driving, we assessed real-world performance of several state-of-the-art pedestrian detectors using standard cross-dataset evaluation. We conclude that current state-of-the-art pedestrian detectors, despite achieving impressive performances on several benchmarks, poorly handle even small domain shifts. This is due to the fact that the current state-of-the-art pedestrian detectors are tailored for target datasets and their overall design contains biasness towards target datasets, thus reducing their generalization. In contrast, general object detectors are more robust and generalize better to new datasets. We thoroughly investigated and verified that general object detectors due to generic design can benefit more from large-scale datasets diverse in scenes and dense in pedestrians. Besides, a progressive training pipeline is proposed which works well for autonomous-driving oriented pedestrian detection. In summary, our findings in this paper can serve as a stepping stone in developing new generalizable pedestrian detectors.

References

[1] Wider pedestrian 2019. https://competitions.codalab.org/competitions/20132. 2, 3, 4, 6, 7, 8

[2] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian detection with deep network cascades. 2015. 3

[3] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011*, pages 3457–3464. IEEE, 2011. 3

[4] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019. 3, 4, 5, 6, 7, 8

[5] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4950–4959, 2017. 3

[6] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016. 3

[7] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3361–3369, 2015. 3

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 4, 5, 6, 8

[9] Victor Campmany, Sergio Silva, Antonio Espinosa, Juan Carlos Moure, David Vázquez, and Antonio M López. Gpu-based pedestrian detection for autonomous driving. *Procedia Computer Science*, 80:2377–2381, 2016. 1

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005. 2, 3

[11] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 2

[12] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012. 2, 3, 4, 5, 6, 7, 8

[13] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 3

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[16] Hironori Hattori, Vishnu Naresh Boddeti, Kris M Kitani, and Takeo Kanade. Learning scene-specific pedestrian detectors without real data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3819–3827, 2015. 1

[17] Amal Hbaieb, Jihene Rezgui, and Lamia Chaari. Pedestrian detection for autonomous driving within cooperative communication system. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2019. 1

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 4

[19] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4073–4082, 2015. 3

[20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2, 8

[21] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Bridging the gap between detection and tracking: A unified approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3999–4009, 2019. 1

[22] Jinpeng Li, Shengcai Liao, Hangzhi Jiang, and Ling Shao. Box guided convolution for pedestrian detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1615–1624, 2020. 5, 6

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2, 3

[25] Wei Liu, Irtiza Hasan, and Shengcai Liao. Center and scale prediction: A box-free approach for pedestrian and face detection. *arXiv preprint arXiv:1904.02948*, 2019. 5

[26] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018. 3, 5, 6

[27] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 5, 6, 7, 8

[28] Yan Luo, Chongyang Zhang, Muming Zhao, Hao Zhou, and Jun Sun. Where, what, whether: Multi-modal learning meets pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14065–14073, 2020. 3

[29] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 1

[30] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3136, 2017. 3

[31] Stefan Munder and Dariu M Gavrila. An experimental study on pedestrian classification. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1863–1868, 2006. 3

[32] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *European Conference on Computer Vision*, pages 546–561. Springer, 2014. 2

[33] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection, 2019. 3, 5

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3, 4, 5, 6

[35] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1, 3, 6, 7, 8

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[37] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu Honggang Zhang, and Jun Guo. Progressive refinement network for occluded pedestrian detection. In *Proc. European Conference on Computer Vision*, volume 7, page 9, 2020. 5, 6

[38] Shuyang Sun, Jiangmiao Pang, Jianping Shi, Shuai Yi, and Wanli Ouyang. Fishnet: A versatile backbone for image, region, and pixel level prediction. In *Advances in Neural Information Processing Systems*, pages 760–770, 2018. 2

[39] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 2

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *arXiv preprint arXiv:1908.07919*, 2019. 4, 5, 6, 8

[41] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018. 5

[42] Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 794–801. IEEE, 2009. 3

[43] Bo Wu and Ram Nevatia. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 3

[44] Jialian Wu, Chunluan Zhou, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Temporal-context enhanced detection of heavily occluded pedestrians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13430–13439, 2020. 3

[45] Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang, and Junsong Yuan. Self-mimic learning for small-scale pedestrian detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2012–2020, 2020. 3

[46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4, 5

[47] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer, 2016. 3

[48] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1112–1121, 2020. 1

[49] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1267, 2016. 3

[50] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 2, 3, 4, 5, 6, 7, 8

[51] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 2019. 1, 3, 6

[52] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–151, 2018. 3, 5