

CityPersons: A Diverse Dataset for Pedestrian Detection

Shanshan Zhang^{1,2}, Rodrigo Benenson², Bernt Schiele²

¹School of Computer Science and Engineering, Nanjing University of Science and Technology, China

²Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

shanshan.zhang@njust.edu.cn, firstname.lastname@mpi-inf.mpg.de

Abstract

Convnets have enabled significant progress in pedestrian detection recently, but there are still open questions regarding suitable architectures and training data. We revisit CNN design and point out key adaptations, enabling plain FasterRCNN to obtain state-of-the-art results on the Caltech dataset.

To achieve further improvement from more and better data, we introduce CityPersons, a new set of person annotations on top of the Cityscapes dataset. The diversity of CityPersons allows us for the first time to train one single CNN model that generalizes well over multiple benchmarks. Moreover, with additional training with CityPersons, we obtain top results using FasterRCNN on Caltech, improving especially for more difficult cases (heavy occlusion and small scale) and providing higher localization quality.

1. Introduction

Pedestrian detection is a popular topic in computer vision community, with wide applications in surveillance, driving assistance, mobile robotics, etc. During the last decade, several benchmarks have been created for this task [7, 8, 12]. These benchmarks have enabled great progress in this area [2].

While existing benchmarks have enabled progress, it is unclear how well this progress translate in open world performance. We think it is time to give emphasis not only to intra-dataset performance, but also across-datasets.

Lately, a wave of convolutional neural network (convnet) variants have taken the Caltech benchmark top ranks [31, 3, 21, 4, 27]. Many of these are custom architectures derived from the FasterRCNN [14, 13, 26] general object detector. We show here that a properly adapted FasterRCNN can match the detection quality of such custom architectures. However since convnets are high capacity models, it is unclear if such model will benefit from more data.

To move forward the field of pedestrian detection, we

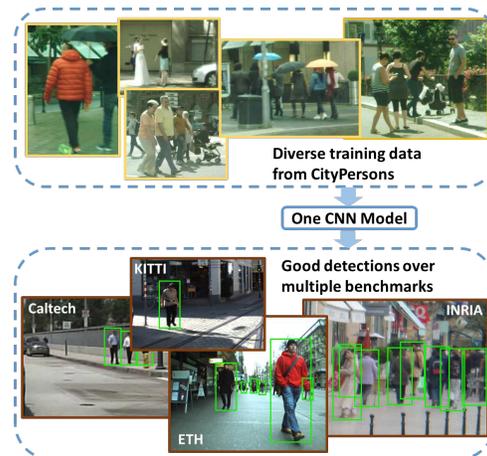


Figure 1: The diversity of the newly introduced CityPersons annotations allows to train one convnet model that generalizes well over multiple benchmarks.

introduce “CityPersons”, a new set of annotations on top of Cityscapes [5]. These are high quality annotations, that provide a rich diverse dataset, and enable new experiments both for training better models, and as new test benchmark.

In summary, our main contributions are:

1. We introduce CityPersons, a new set of high quality bounding box annotations for pedestrian detection on the Cityscapes dataset (train, validation, and test sets). The train/val. annotations will be public, and an online benchmark will be setup.
2. We report new state-of-art results for FasterRCNN on Caltech and KITTI dataset, thanks to properly adapting the model for pedestrian detection and using CityPersons pre-training. We show in particular improved results for more difficult detection cases (small and occluded), and overall higher localization precision.
3. Using CityPersons, we obtain the best reported across-dataset generalization results for pedestrian detection.
4. We show preliminary results exploiting the additional Cityscapes annotations. Using semantic labelling as addi-

tional supervision, we obtain promising improvements for detecting small persons.

Section 1.1 covers the related work, section 2 discusses how to adapt FasterRCNN for best detection quality, section 3 describes our annotation process, some statistics of the new data and baseline experiments. Finally, section 4 explores different ways to use CityPersons to improve person detection quality.

1.1. Related work

In this paper, we investigate convnets, datasets and semantic labels for pedestrian detection, so we discuss related works for these three aspects.

Convnets for pedestrian detection. Convolutional neural networks (convnets) have achieved great success in classification and detection on the ImageNet [20], Pascal, and MS COCO datasets [15]. FasterRCNN [14, 13, 26] has become the de-facto standard detector architecture. Many variants work try to extend it [25, 3, 18], but few improve results with a simpler architecture. A notable exception is SSD [23], which obtains comparable results with a simpler architecture.

Initial attempts to apply convnets for pedestrian detection, used existing detectors (mainly decision forests over hand-crafted features [2, 33]) outputs and re-scored them with a convnet classifier (plus bounding box regression) [16, 28, 1, 27]. Better results are shown when using the reverse configuration: detections resulted from a convnet are re-scored with decision forests classifier (trained over convnet features) [4, 31]. Recently good results are presented by customized pure convnet architectures such as MS-CNN [3] and SA-FasterRCNN [21].

In this paper we show that a properly adapted plain FasterRCNN matches state-of-the-art detection quality without needing add-ons.

Pedestrian datasets. In the last decade several datasets have been created for pedestrian detection training and evaluation. INRIA [7], ETH [11], TudBrussels [29], and Daimler [10] represent early efforts to collect pedestrian datasets. These datasets have been superseded by larger and richer datasets such as the popular Caltech-USA [9] and KITTI [12]. Both datasets were recorded by driving through large cities and provide annotated frames on video sequences.

Despite the large number of frames, both datasets suffer from low-density. With an average of ~ 1 person per image, occlusions cases are severely under-represented. Another weakness of both dataset, is that each was recorded in a single city. Thus the diversity in pedestrian and background appearances is limited.

Building upon the strengths of the Cityscapes data [5], our new annotations provide high quality bounding boxes, with larger portions of occluded persons, and the diversity of 27 different cities. Such diversity enables models trained on CityPersons to better generalize to other test sets.

Semantic labels for pedestrian detection. In section 4.3 we will explore using the semantic labels from Cityscapes to train a pedestrian detector with better context modelling. The idea of using semantic labels to improve detections is at least a decade old [30], and two recent incarnations are [17, 6]. We will use the semantic probability maps computed from a semantic labeller network as additional input channels (next to RGB channels) for the pedestrian detection convnet (see section 4.3).

2. A convnet for pedestrian detection

Before delving into our new annotations (in section 3), we first build a strong reference detector, as a tool for our experiments in sections 3.4 and 4. We aim at finding a straightforward architecture that provides good performance on the Caltech-USA dataset [9].

Training, testing (MR^O , MR^N). We train our Caltech models using the improved $10\times$ annotations from [32], which are of higher quality than the original annotations (less false positives, higher recall, improved ignore regions, and better aligned bounding boxes). For evaluation we follow the standard Caltech evaluation [9]; log miss-rate (MR) is averaged over the FPPI (false positives per image) range of $[10^{-2}, 10^0]$ FPPI. Following [32], we evaluate both on the “original annotations” (MR^O) and new annotations (MR^N); and indicate specifically which test set is being used each time. Unless otherwise specified, the evaluation is done on the “reasonable” setup [9].

FasterRCNN. The FasterRCNN detector obtains competitive performance on general object detection. After re-training with default parameters it will under-perform on the pedestrian detection task (as reported in [31]). The reason why vanilla FasterRCNN underperforms on the Caltech dataset is that it fails to handle small scale objects ($50 \sim 70$ pixels), which are dominant on this dataset. To better handle small persons, we propose five modifications (M_i) that bring the MR^O (miss-rate) from 20.98 down to 10.27 (lower is better, see table 1). As of writing, the best reported results on Caltech is 9.6 MR^O , and our plain FasterRCNN ranks third with less than a point difference. We train FasterRCNN with VGG16 convolutional layers, initialized via ImageNet classification pre-training [26].

M_1 Quantized RPN scales. The default scales of the RPN (region proposal network in FasterRCNN) are sparse and

Detector aspect	MR ^O	ΔMR
FasterRCNN-vanilla	20.98	-
+ quantized rpn scales	18.11	+ 2.87
+ input up-scaling	14.37	+ 3.74
+ Adam solver	12.70	+ 1.67
+ ignore region handling	11.37	+ 1.33
+ finer feature stride	10.27	+ 1.10
FasterRCNN-ours	10.27	+ 10.71

Table 1: Step by step improvements on Caltech from vanilla FasterRCNN to our adapted version, we gain 10.71 MR points in total.

assume a uniform distribution of object scales. However, when we look at the training data on Caltech, we find much more small scale people than large ones. Our intuition is to let the network generate more proposals for small sizes, so as to better handle them. We split the full scale range in 10 quantile bins (equal amount of samples per bin), and use the resulting 11 endpoints as RPN scales to generate proposals.

M₂ Input up-scaling. Simply up-sampling the input images by 2x, provides a significant gain of 3.74 MR^O percent points (pp). We attribute this to a better match with the ImageNet pre-training appearance distribution. Using larger up-sampling factors does not show further improvement.

M₃ Finer feature stride. Most pedestrians in Caltech have height × width = 80 × 40. The default VGG16 has a feature stride of 16 pixels. Having such a coarse stride compared to the object width reduces the chances of having a high score over persons, and forces the network to handle large displacement relative to the object appearance. Removing the fourth max-pooling layer from VGG16 reduces the stride to 8 pixels; helping the detector to handle small objects.

M₄ Ignore region handling. The vanilla FasterRCNN code does not cope with ignore regions (areas where the annotator cannot tell if a person is present or absent, and person groups where individuals cannot be told apart). Simply treating these regions as background introduces confusing samples, and has a negative impact on the detector quality. By ensuring that during training the RPN proposals avoid sampling the ignore regions, we observe a 1.33 MR pp improvement.

M₅ Solver. Switching from the standard Caffe SGD solver to the Adam solver [19], provides a consistent gain in our experiments.

We show the step-by-step improvements in table 1. M₁ and M₂ are key, while each of the other modifications add about ∼ 1 MR pp. All together these modifications adapt the vanilla FasterRCNN to the task of pedestrian detection.

Other architectures. We also explored other architectures such as SSD [23] or MS-CNN [3] but, even after adaptations, we did not manage to obtain improved results. Amongst all the variants reaching ∼ 10% MR our FasterRCNN is the simplest.

Conclusion. Once properly adapted, FasterRCNN obtains competitive performance for pedestrian detection on the Caltech dataset. This is the model we will use in all following experiments.

In section 3 we introduce a new dataset that will enable further improvements of detection performance.

3. CityPersons dataset

The Cityscapes dataset [5] was created for the task of semantic segmentation in urban street scenes. It consists of a large and diverse set of stereo video sequences recorded in streets from different cities in Germany and neighbouring countries. Fine pixel-level annotations of 30 visual classes are provided for 5 000 images from 27 cities. The fine annotations include instance labels for persons and vehicles. Additionally 20 000 images from 23 other cities are annotated with coarse semantic labels, without instance labels.

In this paper, we present the CityPersons dataset, built upon the Cityscapes data to provide a new dataset of interest for the pedestrian detection community. For each frame in the 5 000 fine-annotations subset, we have created high quality bounding box annotations for pedestrians (section 3.1). In section 3.2 we contrast CityPersons with previous datasets regarding: volume, diversity and occlusion. In section 4 we show how to use this new data to improve results on other datasets.

3.1. Bounding box annotations

The Cityscapes dataset already provides instance level segments for each human. These segments indicate the visible parts of humans. Simply using bounding boxes of these segments would raise three issues. I1) The box aspect ratio would be irregular, persons walking have varying width. It has been proposed to thus normalize aspect ratio for pedestrian annotations. I2) Even after normalizing aspect ratio, the boxes would not align amongst each other. They will be off in the horizontal axis due to being normalized based on the segment centre rather the object centre. They will be off in the vertical axis due to variable level of occlusion for each person. It has been shown that pedestrian detectors benefit from well aligned training samples [32], and conversely, training with misaligned samples will hamper results. I3) Existing datasets (INRIA, Caltech, KITTI) have defined bounding boxes covering the full object extent, not just the visible area. In order to train compatible, high quality models, we need to have annotations that align well the

full extent of the persons bodies (“amodal bounding box” [22]).

Fine-grained categories. In the Cityscapes dataset, humans are labelled as either person or rider. In this paper, we provide further fine-grained labels for persons. Based on the postures, we group all humans into four categories: pedestrian (walking, running or standing up), rider (riding bicycles or motorbikes), sitting person, and other person (with unusual postures, e.g. stretching).

Annotation protocol. For pedestrians and riders (cyclists, motorists), we follow the same protocol as used in [32], where the full body is annotated by drawing a line from the top of the head to the middle of two feet, and the bounding box is generated using a fixed aspect ratio (0.41). This protocol has been shown to provide accurate alignments. The visible bounding box for each instance is the tightest one fully covering the segment mask, and can be generated automatically from the segment. See an illustration in figure 2. The occlusion ratio can then be computed as $\frac{area(BB-vis)}{area(BB-full)}$.

As of other categories of persons, i.e. sitting and other persons, there is no uniform alignment to apply, so we only provide the segment bounding box for each of them without full body annotations.

Apart from real persons, we also ask the annotators to search over the whole image for areas containing fake humans, for instance, people on posters, statue, mannequin, people’s reflection in mirror or window, etc., and mark them as ignore regions.

Annotation tool. Since we already have the segment mask for each instance, we can do the annotations in a more efficient way than from scratch. To this end, we develop a new annotation tool to avoid searching for persons over the images by exploiting the available instance segments. This tool pops out one person segment at a time and asks the annotator to recognize the fine-grained category first and then do the full body annotation for pedestrians and riders. Thanks to the high-quality of segmentation annotations, using such a tool also reduces the risk of missing persons, especially at crowded scenes. But the ignore region annotations have to be done by searching over the whole images.

3.2. Statistics

Volume. We show the number of bounding box annotations provided by us in table 2. In a total of 5 000 images, we have ~35k person and ~13k ignore region annotations. And we notice the density of persons are consistent across train/validation/test subsets. Please note we use the same split as Cityscapes.



Figure 2: Illustration of bounding box annotations for pedestrians. For each person, the top of the head and middle of the feet is drawn by the annotator. An aligned bounding box is automatically generated using the fixed aspect ratio (0.41). The bounding box covering the segmentation mask is used to estimate the visible part.

	Train	Val.	Test	Sum
# cities	18	3	6	27
# images	2 975	500	1 575	5 000
# persons	19 654	3 938	11 424	35 016
# ignore regions	6 768	1 631	4 773	13 172

Table 2: Statistics of bounding box annotations on CityPersons dataset.

Diversity. We compare the diversity of Caltech, KITTI and CityPersons in table 3. Since KITTI test set annotations are not publicly available, we only consider the training subset for a fair comparison.

The CityPersons training subset was recorded across 18 different cities, three different seasons, and various weather conditions. While the Caltech and KITTI datasets are only recorded in one city at one season each.

In terms of density, we have on average ~7 persons per image. This number is much higher than that on the Caltech and KITTI datasets, where each image only contains ~1 person on average.

Also, the number of identical persons is another important evidence of diversity. On our CityPersons dataset, the number of unique persons amounts up to ~20 000. In contrast, the Caltech and KITTI dataset only contains ~1 300 and ~6 000 unique pedestrians respectively. Note KITTI and CityPersons frames are sampled very sparsely, so each person is considered as unique.

CityPersons also provides fine-grained labels for persons. As shown in figure 3, pedestrians are the majority (83%). Although riders and sitting persons only occupy 10% and 5% respectively, the absolute numbers are still considerable, as we have a large pool of ~35k persons.

Occlusion. The Cityscapes data was collected by driving through the centre of some highly populated cities, e.g. Frankfurt and Hamburg. We notice that on some im-

	Caltech	KITTI	CityPersons
# country	1	1	3
# city	1	1	18
# season	1	1	3
# person/image	1.4	0.8	7.0
# unique person	1273	6336	19654

Table 3: Comparison of diversity on different datasets (training subset only).

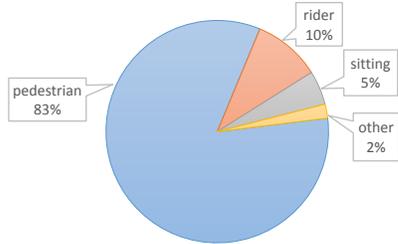


Figure 3: Fine-grained person categories on CityPersons.

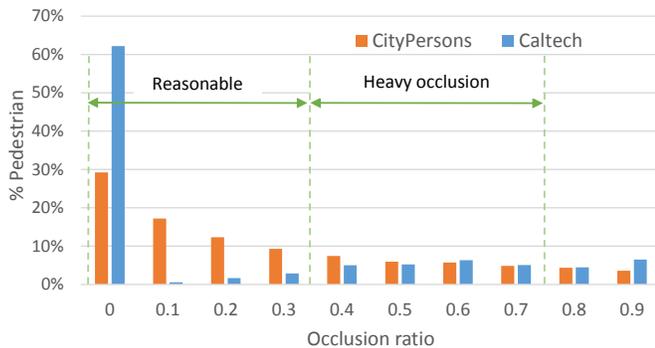


Figure 4: Comparison of occlusion distributions on CityPersons and Caltech datasets. CityPersons contains more occlusions in the reasonable subset than Caltech.

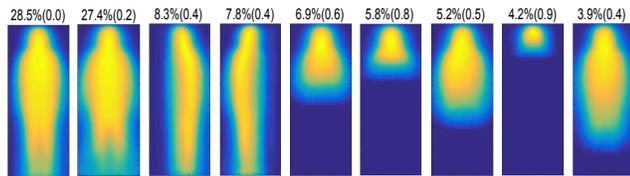


Figure 5: Top 9 of quantized 11 occlusion patterns of pedestrians on CityPersons dataset. Two numbers on top indicate percentage and average occlusion ratio of samples clustered into each pattern.

ages, there are ~100 people walking on the street, highly occluded by each other. Such a high occlusion is rarely seen in previous datasets. In figure 4, we compare the distribution of pedestrians at different occlusion levels for Caltech and CityPersons. We notice that on Caltech

there are more than 60% fully visible pedestrians, while on CityPersons there are less than 30%. This indicates we have two times more occlusions than Caltech, which makes CityPersons a more interesting ground for occlusion handling. Moreover, on the reasonable subset (≤ 0.35 occlusion) the community typically use, Caltech is dominated by fully visible pedestrians, while CityPersons has more occlusion cases.

In order to understand which kinds of occlusions we have on CityPersons, we quantize all persons into 11 patterns and show the top 9 of them in figure 5 (the last two patterns are not shown as they are of less than 1% and thus noisy). For visualization, we resize each full body bounding box to a fixed size, and then overlay the segmentation mask. For each pattern, the bright area shows the visible part and the two numbers on top indicate the percentage and average occlusion ratio of corresponding pattern. The first two patterns (55.9%) roughly cover the “reasonable” subset; the third and fourth patterns correspond to occlusions from either left or right side. Apart from that, we still have about 30% pedestrians distributed in various patterns, some of which have a very high occlusion ratio (>0.9). Such distributed occlusion patterns increase the diversity of the data and hence makes the dataset a more challenging test base.

3.3. Benchmarking

With the publication of this paper, we release the data on the Cityscapes website¹, where train/validation annotations can be downloaded, and an online evaluation server is available to compute numbers over the held-out test annotations.

We follow the same evaluation protocol as used for Caltech [9], by allowing evaluation on different subsets. In this paper, MR stands for log-average miss rate on the “reasonable” setup (scale [50, ∞], occlusion ratio [0, 0.35]) unless otherwise specified. While evaluating pedestrian detection performance, cyclists/sitting persons/other persons/ignore regions are not considered, which means detections matching with those areas are not counted as mistakes.

3.4. Baseline experiments

To understand the difficulties of pedestrian detection on the CityPersons dataset, we train and evaluate three different detectors. ACF [8] and Checkerboards [33] are representatives from the Integral Channel Features detector (ICF) family, while FasterRCNN [26] acts as the state-of-the-art detector. We set up the FasterRCNN detector by following the practices we learned from Caltech experiments (section 2). Since CityPersons images are ~7 times larger than Caltech, we are only able to use an upsampling factor of 1.3 to fit in 12GB of GPU memory.

We re-train each detector using the CityPersons training set and then evaluate on the validation set. Note that

¹<https://www.cityscapes-dataset.com/>

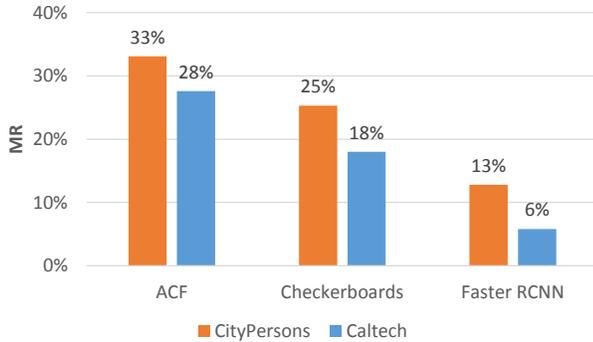


Figure 6: Comparison of baseline detectors on Caltech test and `CityPersons` val. set (reasonable). Numbers are MR^N on Caltech and MR on `CityPersons` (lower is better). Ranking of methods between two datasets is stable. For all methods, `CityPersons` is more difficult to solve than Caltech.

all the `CityPersons` numbers reported in this paper are on the validation set. Consistent with the reasonable evaluation protocol, we only use the reasonable subset of pedestrians for training; cyclists/sitting persons/other persons/ignore regions are avoided for negative sampling.

In figure 6, we show the comparison of the above three detectors on `CityPersons` and Caltech. FasterRCNN outperforms ICF detectors by a large margin, which indicates the adaptation of FasterRCNN on Caltech is also transferable to `CityPersons`. Moreover, we find the ranking of three detectors on `CityPersons` is consistent with that on Caltech, but the performance on `CityPersons` dataset is lower for all three detectors. This comparison shows that `CityPersons` is a more challenging dataset, thus more interesting for future research in this area.

To understand the impact of having a larger amount of training data, we show how performance grows as training data increases in figure 7. We can see performance keeps improving with more data. Therefore, it is of great importance to provide CNNs with a large amount of data.

Considering the trade off between speed and quality, we use an alternative model of our FasterRCNN by switching off input image upsampling for the analysis experiments shown in figure 7 and section 4.3. This model is about 2x faster at both training and test time, but only drops the performance by ~2 pp (from 13% MR to 15% MR).

Conclusion. The `CityPersons` dataset can serve as a large and diverse database for training a powerful model, as well as a more challenging test base for future research on pedestrian detection.

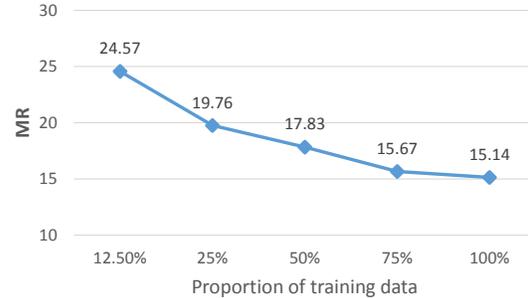


Figure 7: Quality as function of training volume. Faster-RCNN model trained/evaluated on `CityPersons` train/val. set (MR: lower is better).

4. Improve quality using `CityPersons`

Having the `CityPersons` dataset at hand, we now proceed to illustrate three different ways it enables to improve pedestrian detection results (§4.1, §4.2, §4.3). As we will see, `CityPersons` is particularly effective at improving results for small scale pedestrians, occluded ones, and providing higher localization accuracy.

4.1. Generalization across datasets

Commonly, a detector is trained on the training set of the target benchmark. As such, one needs to train multiple detectors for different benchmarks. Ideally, one would wish to train one detector that is able to perform well on multiple benchmarks. Since the `CityPersons` dataset is large and diverse, we wonder if it can allow us to train a detector with good generalization capabilities.

To see how well `CityPersons` data generalizes across different datasets, we train models on Caltech, KITTI and `CityPersons` datasets, and then apply each of them on six different test sets: Caltech, KITTI, `CityPersons`, INRIA, ETH and Tud-Brussels. For KITTI, we split the public training data into training and validation subsets (2:1) by random sampling. Table 4 shows comparisons of two detectors: ACF [8] and FasterRCNN [26].

We observe:

(1) Overall, when trained with the same data Faster-RCNN generalizes better across datasets than ACF. (Note that FasterRCNN benefits from ImageNet pre-training, while ACF does not.)

(2) For both detectors, the mean MR across test sets is significantly better for models trained with `CityPersons` training data. `CityPersons` generalizes better than Caltech and KITTI.

These experiments confirm the generalization ability of `CityPersons` dataset, that we attribute to the size and diversity of the Cityscapes data, and to the quality of the bounding boxes annotations.

Train	Caltech	KITTI	CityPersons
Caltech	27.63	63.15	51.28
KITTI	49.99	32.06	46.74
CityPersons	72.89	94.28	33.10
INRIA	63.39	67.49	50.23
ETH	78.64	89.94	56.30
Tud-Brussels	63.22	69.25	67.21
mean MR	59.29	69.36	50.81

(a) ACF

Train	Caltech	KITTI	CityPersons
Caltech	10.27	46.86	21.18
KITTI	10.50	8.37	8.67
CityPersons	46.91	51.21	12.81
INRIA	11.47	27.53	10.44
ETH	57.85	49.00	35.64
Tud-Brussels	42.89	45.28	36.98
mean MR	29.98	38.04	20.95

(b) FasterRCNN

Table 4: Generalization ability of two different methods, trained and tested over different datasets. All numbers are MR on reasonable subset. Bold indicates the best results obtained via generalization across datasets (different train and test).

4.2. Better pre-training improves quality

In table 4, we find the CityPersons data acts as very good source of training data for different datasets, assuming we are blind to the target domain. Furthermore, when we have some training data from the target domain, we show CityPersons data can be also used as effective external training data, which helps to further boost performance.

First, we consider Caltech as the target domain, and compare the quality of two models. One is trained on Caltech data only; and the other is first trained on CityPersons, and then finetuned on Caltech (CityPersons→Caltech). From table 5, we can see the additional training with CityPersons data improves the performance in the following three aspects.

(1) CityPersons data improves overall performance. When evaluated on the reasonable setup, the CityPersons→Caltech model obtains ~1 pp gain.

(2) CityPersons data improves more for harder cases, e.g. smaller scale, heavy occlusion. We notice the gap for heavy occlusion is large (~9 pp), due to more occluded training samples on the CityPersons dataset. Similar trend is also found for smaller scale persons ([30,80]).

(3) CityPersons data helps to produce better-aligned

O/N	Setup	Scale range	IoU	Caltech	CityPersons →Caltech	Δ MR
MR ^O	Reasonable	[50, ∞]	0.5	10.3	9.2	+ 1.1
MR ^O	Smaller	[30, 80]	0.5	52.0	48.5	+ 3.5
MR ^O	Heavy occl.	[50, ∞]	0.5	68.3	57.7	+ 8.6
MR ^N	Reasonable	[50, ∞]	0.5	5.8	5.1	+ 0.7
MR ^N	Reasonable	[50, ∞]	0.75	30.6	25.8	+ 4.8

Table 5: Gains from additional CityPersons training at different evaluation setups on Caltech test set. MR^O and MR^N indicate numbers evaluated on original and new annotations [32]. CityPersons pre-training helps more for more difficult cases. See also table 6.

Setup	Scale range	IoU	KITTI	CityPersons →KITTI	Δ MR
Reasonable	[50, ∞]	0.5	8.4	5.9	+ 2.5
Reasonable	[50, ∞]	0.75	43.3	39.2	+ 4.1
Smaller	[30, 80]	0.5	37.8	27.1	+ 10.7

Table 6: Gains from additional CityPersons training at different evaluation setups on KITTI validation set. All numbers are MR (see §2). Here also, CityPersons pre-training helps more for more difficult cases. See also table 5.

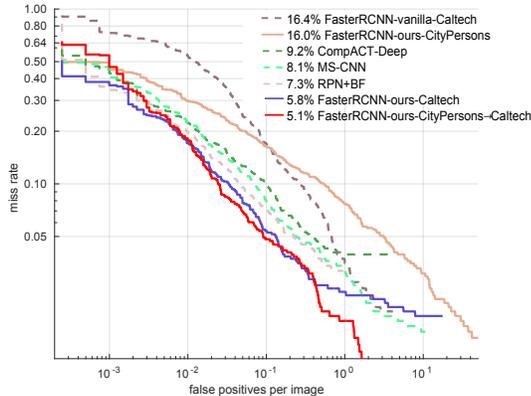
detections. The Caltech new annotations are well aligned, thus a good test base for alignment quality of detections. When we increase the IoU threshold for matching from 0.50 to 0.75, the gain from CityPersons data also grows from 1 pp to 5 pp. This gap indicates the high quality of CityPersons annotations are beneficial to produce better-aligned detections.

Compared with other state-of-the-art detectors, our best model using CityPersons for pre-training obtains 5.1% MR^N at IoU 0.50 evaluation, outperforming previous best reported results (7.3% MR^N) by 2.2 pp (figure 8a); this gap becomes even larger (~ 20 pp) when we use a stricter IoU of 0.75 (figure 8b). From the comparison, our FasterRCNN detector obtains state-of-the-art results on Caltech, and improves the localization quality significantly.

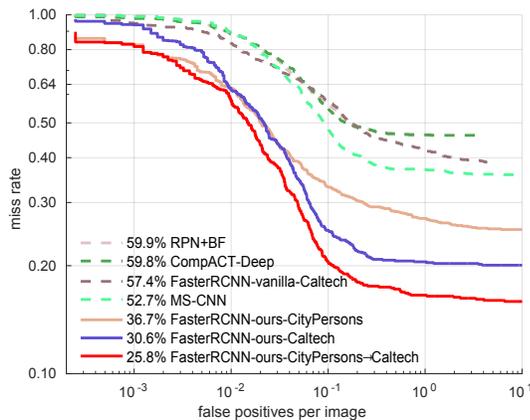
When we consider KITTI as the target domain, we also see improvements brought by additional training with CityPersons data. As shown in table 6, the gain on reasonable evaluation setup is 2.5 pp, while for smaller scale, the gap becomes more impressive (10.7 pp). The 4.1 pp gap at IoU 0.75 again verifies CityPersons data helps to produce better aligned detections.

4.3. Exploiting Cityscapes semantic labels

In this subsection, we explore how much improvement can be obtained for pedestrian detection by leveraging the



(a) IoU=0.50



(b) IoU=0.75

Figure 8: Comparison of state-of-the-art results on the Caltech test set (reasonable subset), MR^N .



(a) Original image

(b) Semantic map

Figure 9: Example of semantic map generated by an FCN-8s model trained on Cityscapes coarse annotations.

semantic labels available on the Cityscapes dataset.

We use an FCN-8s [24] model trained on Cityscapes coarse annotations to predict semantic labels. Note we cannot involve fine-annotation images in this semantic labelling training, otherwise our following detection training will suffer from overfitting. Although this model is only trained on coarse annotations, we can see the semantic segmentation mask provides a reasonable structure for the whole scene (figure 9). Then we concatenate semantic channels

Scale range	Baseline	+ Semantic	ΔMR
[50, ∞]	15.4	14.8	+ 0.6
[100, ∞]	7.9	8.0	+ 0.1
[75, 100]	7.2	6.7	+ 0.5
[50, 75]	25.6	22.6	+ 3.0

Table 7: Improvements from semantic channels in different scale ranges. Numbers are MR on the CityPersons val. set. Albeit there is small overall gain, adding semantic channels helps for the difficult case of small persons.

with RGB channels and feed them altogether into convnets, letting convnets to figure out the hidden complementarity.

For the reasonable evaluation setup, we get an overall improvement of ~ 0.6 pp from semantic channels. When we look at the fine-grained improvements for different scale ranges, we find that semantic channels help more for small persons, which is a hard case for our task (table 7).

As a preliminary trial, we already get some improvements from semantic labels, which encourage us to explore more effective ways of using semantic information.

5. Summary

In this paper, we first show that a properly adapted FasterRCNN can achieve state-of-the-art performance on Caltech. Aiming for further improvement from more and better data, we propose a new diverse dataset namely CityPersons by providing bounding box annotations for persons on top of Cityscapes dataset. CityPersons shows high contrast to previous datasets as it consists of images recorded across 27 cities, 3 seasons, various weather conditions and more common crowds.

Serving as training data, CityPersons shows strong generalization ability from across dataset experiments. Our FasterRCNN model trained on CityPersons obtains reasonable performance over six different benchmarks. Moreover, it further improves the detection performance with additional finetuning on the target data, especially for harder cases (small scale and heavy occlusion), and also enhance the localization quality.

On the other hand, CityPersons can also be used as a new test benchmark as there are more challenges, e.g. more occlusions and diverse environments. We will create a website for this benchmark and only allows for online evaluations by holding out the test set annotations.

Other than bounding box annotations for persons, there are additional information to leverage on CityPersons, e.g. fine semantic segmentations, other modalities of data (stereo, GPS), and un-annotated neighbouring frames. Our preliminary results of using semantic labels show promising complementarity. These rich data will motivate more efforts to solve the problem of pedestrian detection.

References

- [1] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson. Real-time pedestrian detection with deep network cascades. In *BMVC*, 2015. 2
- [2] R. Benenson, M. Omran, J. Hosang, , and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV, CVRSUAD workshop*, 2014. 1, 2
- [3] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016. 1, 2, 3
- [4] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015. 1, 2
- [5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3
- [6] A. D. Costea and S. Nedeveschi. Semantic channels for fast pedestrian detection. In *CVPR*, 2016. 2
- [7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2
- [8] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014. 1, 5, 6
- [9] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2012. 2, 5
- [10] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 2009. 2
- [11] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008. 2
- [12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 2
- [13] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 1, 2
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2
- [16] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015. 2
- [17] Q. Hu, P. Wang, C. Shen, A. van den Hengel, and F. Porikli. Pushing the limits of deep cnns for pedestrian detection. *ArXiv*, 2016. 2
- [18] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv*, 2015. 2
- [19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2
- [21] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *arXiv*, 2016. 1, 2
- [22] K. Li and J. Malik. Amodal instance segmentation. In *ECCV*, 2016. 4
- [23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2015. 2, 3
- [24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional models for semantic segmentation. In *CVPR*, 2015. 8
- [25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2
- [26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1, 2, 5, 6
- [27] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, 2015. 1, 2
- [28] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, 2015. 2
- [29] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009. 2
- [30] L. Wolf and S. M. Bileschi. A critical view of context. *IJCV*, 2006. 2
- [31] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, 2016. 1, 2
- [32] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *CVPR*, 2016. 2, 3, 4, 7
- [33] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015. 2, 5