EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes

Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M. Gavrila

Abstract—Big data has had a great share in the success of deep learning in computer vision. Recent works suggest that there is significant further potential to increase object detection performance by utilizing even bigger datasets. In this paper, we introduce the EuroCity Persons dataset, which provides a large number of highly diverse, accurate and detailed annotations of pedestrians, cyclists and other riders in urban traffic scenes. The images for this dataset were collected on-board a moving vehicle in 31 cities of 12 European countries. With over 238,200 person instances manually labeled in over 47,300 images, EuroCity Persons is nearly one *order* of magnitude larger than datasets used previously for person detection in traffic scenes. The dataset furthermore contains a large number of person orientation annotations (over 211,200). We optimize four state-of-the-art deep learning approaches (Faster R-CNN, R-FCN, SSD and YOLOv3) to serve as baselines for the new object detection benchmark. In experiments with previous datasets we analyze the generalization capabilities of these detectors when trained with the new dataset. We furthermore study the effect of the training set size, the dataset diversity (day- versus night-time, geographical region), the dataset detail (i.e., availability of object orientation information) and the annotation quality on the detector performance. Finally, we analyze error sources and discuss the road ahead.

Index Terms—Object detection, benchmarking

1 Introduction

Person detection in images is a key task in a number of important application domains, such as intelligent vehicles, surveillance, and robotics. Despite two decades of steady progress, it is still an open research problem. The wide variation in person appearance, arising from articulated pose, clothing, background and visibility conditions (time of day, weather), makes person detection particularly challenging. It therefore often features as canonical task to assess the performance of generic object detectors.

In this paper, we focus on the application setting of detecting persons in urban traffic scenes, as observed from cameras on-board a moving vehicle. Detection performance has improved to the point that pedestrian and cyclist detection is incorporated in active safety systems of various premium vehicles on the market. Still, such systems are deployed in the context of driver assistance, meaning that a correct detection performance of about 90 percent is acceptable, as long as the false alarm rate is essentially zero. With the advent of fully self-driving vehicles, performance needs to be significantly upped, as a driver is no longer available. A recent paper [1] argues that current pedestrian detection performance lags that of an attentive human by an order of magnitude. How can this performance gap be closed?

Datasets play a crucial role in today's computer vision research [6]. Corresponding benchmarks reveal strengths and weaknesses of existing approaches and are instrumental in guiding research forward. Still, [7] argues that even larger datasets are needed. Experiments on their 300 million images dataset show that the classification performance further increases logarithmically with the size of the training dataset. Deep learning has also been very successful in the context of object detection [8], [9] and [10]. More data could prove useful for object detection as well [4].

During the last two decades an extensive amount of research has been spent on pedestrian detection [1], [2], [11], [12], [13], [14]. For several years, progress in this domain was monitored on benchmarks like Caltech [2] and KITTI [3]. However, these datasets have come into age since. The recording conditions back then (i.e., image resolution and quality) do not reflect the current state of the art anymore. The comparatively small size of the training data (i.e., several thousands samples) furthermore makes these benchmarks prone to dataset bias and to over-fitting [15]. Recently, CityPersons [4] was released with higher resolution images and a larger quantity of training data ($\approx 35{,}000$ person samples). Although these data additions are helpful, [4] conclude that more training data is necessary for the recent high-capacity deep learning architectures. Data diversity is another important aspect. The before-mentioned datasets were captured in few countries ($1-3$), and in daylight and dry weather conditions only; this hampers generalization to real world applications.

- *M. Braun and S. Krebs are with the Environment Perception Group, Daimler AG, Ulm 89081, Germany, and also with the Intelligent Vehicles Group, TU Delft, Delft 2628 CD, Netherlands. E-mail: {markus.ma.braun, sebastian.krebs}@daimler.com.*
- *F. Flohr is with the Environment Perception Group, Daimler AG, Ulm 89081, Germany. E-mail: fabian.flohr@daimler.com.*
- *D. M. Gavrila is with the Intelligent Vehicles Group at TU Delft, Delft 2628 CD, Netherlands. E-mail: d.m.gavrila@tudelft.nl.*

Manuscript received 13 May 2018; revised 24 Oct. 2018; accepted 16 Jan. 2019. Date of publication 4 Feb. 2019; date of current version 11 July 2019. (Corresponding author: Dariu Gavrila.)
Recommended for acceptance by S. Lazebnik.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2019.2897684

Fig. 1. The EuroCity Persons dataset was recorded in 31 cities of 12 European countries: Croatia (Zagreb), Czech Republic (Brno, Prague), France (Lyon, Marseille, Montpellier, Toulouse), Germany (Berlin, Dresden, Hamburg, Köln, Leipzig, Nürnberg, Potsdam, Stuttgart, Ulm and Würzburg), Hungary (Budapest), Italy (Bologna, Firenze, Milano, Pisa, Roma and Torino), The Netherlands (Amsterdam), Poland (Szczecin), Slovak Republic (Bratislava), Slovania (Ljubljana), Spain (Barcelona) and Switzerland (Basel, Zürich). The map itself was compiled from 500 randomly sampled pedestrian bounding boxes from our dataset.

To address these limitations we introduce a new dataset for vision-based person detection coined EuroCity Persons. The images for this dataset were collected on-board a moving vehicle in 31 cities of 12 European countries, see Fig. 1 and Table 5. With over 238,200 person instances manually labeled in over 47,300 images, EuroCity Persons is nearly one order of magnitude larger than person datasets used previously for benchmarking, in terms of manual annotations. Due to its comparatively large geographic coverage, its recordings during both day and night-time, and during all four seasons (light/short summer to thick/long winter clothing) it provides a new level of data diversity. EuroCity Persons furthermore offers detailed annotations; besides bounding box information, it includes tags for occlusion/truncation and annotates body orientation (the latter has relevance for object tracking and path prediction). Finally, thanks to the implemented quality control procedures, annotations are overall accurate.

By means of an experimental study using EuroCity Persons, we address a number of questions: how much do recent deep learning methods improve by an increased amount of training data? How well does this dataset generalize to existing datasets? What is the day- and night-time performance? Is there a geographical bias? How does annotation quality affect object detection performance? Does multi-tasking (orientation estimation) help object detection?

2 Related Work

2.1 Datasets

A number of early datasets focus on pedestrian classification (e.g., Daimler-CB [16], CVC [17], and NICTA [18]) and detection (e.g., Daimler-DB [11], INRIA [19], ETH [20], and TUD-Brussels [21]). See [11] for an overview. Currently, KITTI [3] and the Caltech [2] are the established pedestrian detection benchmarks. The latter has been extended by [1] with corrected annotations. The Tsinghua-Daimler Cyclist (TDC) dataset [5] focuses on cyclists and other riders. In [22] a multi-spectral dataset for pedestrian detection is introduced, combining RGB and infrared modalities.

The Cityscapes dataset [23] was recorded in 50 cities during three seasons. Similar to earlier scene labeling challenges like Pascal VOC [24] and Microsoft COCO [25], it provides pixel-wise segmentations for a number of semantic object classes. The CityPersons dataset [4] extends part of the Cityscapes dataset by bounding-box labels for the full extent of pedestrians. This enables occlusion analysis as the segmentation masks cover the visible areas only.

See Table 1 for an overview of the main person detection benchmarks in vehicle context. In terms of the annotation quantity and data diversity, CityPersons [4] and Tsinghua-Daimler Cyclist [5] had, so far, the most to offer for the pedestrian and the riders class. Although Caltech [2] lists a large number of pedestrian annotations, only an unspecified subset of these annotations were done manually, the remainder was obtained by interpolation (we estimate the number of manual annotations to be an order of magnitude smaller). In total there are about 2,300 unique persons in this dataset. Training and evaluation on Caltech is typically performed on

TABLE 1
Comparison of Person Detection Benchmarks in Vehicle Context

	Caltech [2]	KITTI [3]	CityPersons [4]	TDC [5]	**EuroCity Persons**
# countries	1	1	3	1	**12**
# cities	1	1	27	1	**31**
# seasons	1	1	3	1	**4**
# images (day / night)	**249,884** / -	14,999 / -	5,000 / -	14,674 / -	40,217 / **7,118**
# pedestrians (day / night)	**289,395**[a] / -	~9,400[b] / -	31,514 / -	8,919 / -	183,004 / **35,309**
# riders (day / night)	- / -	~3,300[b] / -	3,502 / -	**23,442** / -	18,216 / **1,564**
# ignore regions (day / night)	57,226[a] / -	~22,600[b] / -	13,172 / -	- / -	**75,673** / **20,032**
# orientations (day / night)	- / -	~12,700[b] / -	- / -	- / -	**176,879** / **34,393**
resolution	640 × 480	1240 × 376	**2048×1024**	**2048×1024**	1920 × 1024
weather	dry	dry	dry	dry	**dry, wet**
train-val-test split (%)	50-0-50	50-0-50	60-10-30	71-8-21	60-10-30

[a]*Only an unspecified subset of these annotations were done manually, the remainder was obtained by interpolation (we estimate the number of manual annotations to be an order of magnitude smaller).*
[b]*Number estimated on the basis of the average number of pedestrians per image, since the test set is private and the authors did not report the actual number.*

TABLE 2
Overview of Recent Deep Learning Detection Methods

	two stage methods			one stage methods		
	Fast R-CNN [9]	**Faster R-CNN** [10]	**R-FCN**[31]	YOLO [32]	**YOLOv3** [33]	**SSD** [34]
region proposals	external	RPN	RPN	gridbased	anchor boxes	default boxes
hard example mining	implicit	implicit	explicit	none	none	explicit
used feature maps	last	last	last	last	several	several

Methods evaluated in this work are bold-faced.

a subset of the dataset, using every 30th frame. Cyclist and other riders annotations are missing in the Caltech dataset, and orientation annotations are missing in both Caltech and CityPersons datasets. KITTI, Caltech and TDC datasets have been collected in one city only. CityPersons was recorded in 27 different cities but, apart of Strasbourg and Zurich, it covers only Germany and recordings were not made throughout all seasons. Very recently, the Berkeley Deep Drive dataset (BDD) [26] was made available, which in total provides 100,000 images recorded in a vehicle context. A white paper describing the dataset was announced.

Other person datasets relate to attribute recognition [27], [28], [29]. Notable for its sheer size is furthermore the recent Open Images V4 dataset [30], containing 15.4M bounding boxes on 1.9M images for 600 different categories.

2.2 Methods

Deformable Part Models (DPM) using Histograms of Oriented Gradients (HOG) features [35], [36], [37], and Decision Forests using ICF features [38], [39], [40], [41] were until a few years ago the established pedestrian detection methods [12]. Successes of deep learning for image classification (e.g., AlexNet [42]) also lead to its incorporation in object detection. By training deep convolutional neural networks (CNN) like GoogleNet [43], VGG [44] and ResNet [45] on the ImageNet dataset [6] for classification, models learn to extract powerful features from raw pixels, which can be used effectively for other tasks like object detection [46].

A comparison of selected detection methods building up on feature maps of CNNs is shown in Table 2. They can be clustered into two stage methods [9], [10], [31], that use a proposal stage and a downstream classification stage, and one stage methods that go without the proposal stage [32], [33], [34]. The R-CNN methods [8], [9], [10] are the basis for most current two stage methods. R-CNN [8] and its extension Fast R-CNN [9] depend on proposals for possible object locations from an external input. R-CNN uses a CNN to classify each proposal separately. Fast R-CNN optimizes the runtime by executing the CNN on a complete image to share the calculated features. For every (mapped) region proposal, features are pooled and used for separate classification and bounding box regression by fully connected layers.

The relation between proposal recall and the overall detection performance is shown in [47] for a lot of different proposal methods like selective search [48], MCG [49] and BING [50]. Proposal methods based on depth data [51], [52] increase the detection performance of Fast R-CNN as the proposal recall is larger.

Faster R-CNN [10] does without external proposals by implementing a region proposal network (RPN). Thus, the two stages are combined in a single network jointly trainable end-to-end. Inside the RPN anchor-boxes of varying scales, positions, and aspect ratios are convolutionally classified as fore- or background. Foreground anchors are then used as proposals for feature pooling. Regardless of the scale of an anchor-box only features are pooled from the last layer. Hereby the spatial support of the features can be a lot larger or even smaller than the objects to be detected. The problem of varying object sizes in pedestrian detection is tackled in the extensions [53], [54], [55], [56]. In SDP [55] features are pooled from different layers in dependence of the proposal size. MS-CNN [54] directly appends proposal networks on feature maps of different scales.

A great part of the computational complexity of Fast R-CNN and Faster R-CNN depends on the number of proposals. The minibatches during training consist of a sampled subset, which is usually several orders of magnitude smaller than the total amount of proposals. [9] and [35] argue that the selection of background samples slightly overlapping with positive samples can be seen as a heuristic hard negative mining. R-FCN [31] does not use fully connected layers and thus does not have to resort to limiting the number of proposals by sampling. Instead it uses convolutional layers to generate scoring maps. Final detection is performed by pooling from these scoring maps without any further calculations dependent on trainable weights. As all proposals are classified, online hard example mining [57] is applicable.

One stage detection methods like YOLO [32], its extensions YOLOv2 [58], YOLOv3 [33], and others [34], [59] go without a distinct proposal stage. In YOLO the final downsampled feature map is divided into grid cells. For each grid cell fully connected layers are trained to detect objects that are centered within this cell using the complete image as spatial support. This approach has weaknesses for small objects and object groups, that cluster within a single cell. That is why YOLOv2 [58] adopts the anchor boxes of Faster R-CNN. Scales and aspect ratios of these boxes are set by calculating dimension clusters using k-means clustering. Features are stacked from different layers to further support the detection of varying object sizes, still the boxes themselves are anchored in a single layer. In YOLOv3 [33] three different layers with three different strides are used to predict classes and precise positions for the anchor boxes. Furthermore, they propose the Darknet-53 network architecture specialised for fast object detection, combining ideas of other CNNs [43], [44], [45].

SSD [34] detects objects based on default boxes. These default boxes are similar to anchor boxes, but they are applied on different feature layers at different resolutions. Hereby the receptive field sizes are approximately proportional to the sizes of the default boxes. In the SSD512 variant, seven layers are used for prediction which means a finer discretization of the output space than with YOLOv3. Unlike

the YOLO methods not all negative boxes or gridcells are used in backpropagation. Hard negative mining is applied to select the boxes with the highest confidence loss similar to R-FCN. [59] introduces a recurrent neural network based on a VGG-16 architecture that improves the localisation accuracy of one stage methods. This is achieved by applying a recurrent rolling convolution on several feature layers.

Generative adversarial networks (GAN) [60] are also used for pedestrian detection. In [61] a Fast R-CNN architecture is extended by a generator branch that adds super resolved features after region proposal pooling to improve the detection performance for small objects. The adversarial branch is trained to discriminate super resolved features of small objects from real features of large scale objects. In [62] inspired by GANs a discriminator is trained to select realistic looking images rendered by a game engine. An extension of Faster R-CNN coined RPN+ is then trained on this data to improve the detection performance for unusual pedestrians.

Deep learning has also been used for estimating orientations of common objects in traffic scenarios on datasets [3] that provide orientation ground-truth. In [63] and [64], orientation estimation is handled as a multi-class classification problem. [65] introduced the Biternion Net, which regresses continuous orientation angles. The Biternion representation is adapted in the Pose-RCNN [52] approach, such that orientation estimation is trained jointly with detection in a Fast R-CNN architecture. In [51] a L1 loss is used instead of the Biternion-based Von-Mises loss for estimating continuous orientation angles.

2.3 Performance Analysis

In [2], 16 different detection methods are evaluated on the Caltech dataset. Small sizes and occlusion are identified as major challenges for pedestrian detectors. The "reasonable" test set typically used for evaluation contains pedestrians larger than 50 *px* with no partial occlusion. In [12] more than 40 detectors are evaluated on the Caltech dataset to analyze the main cause for improvement during the last 10 years. Deep models are examined as one of several possible causes. Still, they are outclassed by the design of better features as the main driver of performance improvement. In [13] also deep models on the Caltech dataset are analyzed. False positives which are touching ground-truth samples are considered as localization error. The remaining false positives are considered as confusion of background and foreground. Hereby, the authors find that confusion is the most frequent reason for false positives. Discriminating false positives by localization and confusion errors is also done in [1]. The authors focus on the boosted decision forests-based methods RotatedFilters [66] and Checkerboards [41]. In addition to categorizing false positives as localization or classification errors, they automatically analyze the effect of contrast, size and blurring on the detection score. Furthermore, they manually cluster false positives and false negatives at a fixed false positives per image by qualitative failure reasons. In contrast, [14] applies an automatic failure analysis for ACF [67] on Caltech and KITTI. They assign failure reasons to false negatives, such as truncation, occlusion, small objects heights, unusual aspect ratios, and localization in one study. As more than one of the sources could qualify as failure reason a certain prioritization provides the primary reason.

Methods [53], [54], [68] building upon the work of Fast/Faster R-CNN are the top-performing methods on the Caltech dataset [1]. [68] uses decision forests for classification instead of fully connected layers but the performance depends on the feature layers of the CNNs. Regarding the KITTI benchmark, the top performing non-anonymous submissions all rely on deep CNNs [54], [55], [56], [59], [69]. Apart from [59] all of these are two stage methods building upon the work of Fast/Faster R-CNN.

[70] evaluates R-FCN, SSD and Faster R-CNN on the generic object detection benchmark MS COCO [25]. By varying the feature extractor, the image resolution and other parameters various speed/accuracy trade-offs are examined.

2.4 Main Contributions

Our contributions are threefold:

- We introduce the EuroCity Persons dataset, which provides a large number of highly diverse, accurate and detailed annotations of persons (pedestrians, cyclists, and other riders) in urban traffic scenes across Europe. It also contains night-time scenes. Annotations extend beyond bounding boxes and include overall body orientations and a variety of object- and image-related tags. See Section 3.
- We optimize four deep learning approaches (Faster R-CNN [10], R-FCN [31], SSD [34] and YOLOv3 [33]) to serve as baselines for the new person detection benchmark. We prove the generalization capabilities of detectors trained with the new dataset and thereby its usefulness. See Sections 4.1 and 4.2.
- We provide insights regarding to the effect of several dataset characteristics on detector performance: the training set size, the dataset bias (day- versus night-time, geographical region), the dataset detail (i.e., availability of object orientation information) and the annotation quality. We analyze error sources and discuss the road ahead. See Sections 4.3 and 5.

3 Benchmark

3.1 Dataset Collection

We collected the images of the EuroCity Persons dataset from a moving vehicle in 31 cities of 12 European countries. Recordings were made with a state-of-the-art automotive-grade two megapixel camera (1920 x 1024) with rolling shutter at a frame rate of 20 Hz. The camera, mounted behind the windshield, originally yielded 16 bit color images; this high dynamic-range was important for capturing scenes with strong illumination variation (e.g., night-time, low-standing sun shining directly into the camera). Images were debayered and rectified afterwards. For the purpose of EuroCity Persons benchmark, and for allowing comparisons with existing datasets, the original 16-bit color images were converted to 8-bit by means of a logarithmic compression curve with a parameter setting different for day and night.

We collected 53 hours of image data in total, for an average of 1.7 hours per city. To limit selection bias [15], we extracted every 80th frame for our detection benchmark without further filtering. This means that a substantial fraction of the person annotations in the dataset are unique,

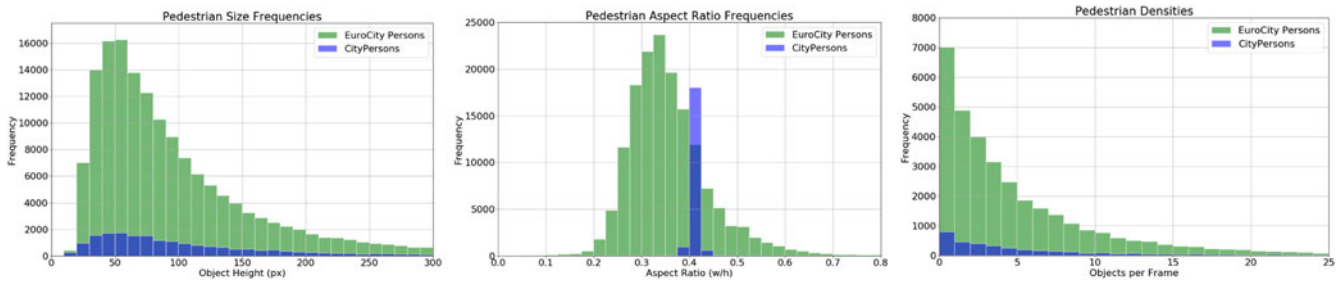


Fig. 2. Statistics of EuroCity Persons and CityPersons for pedestrians of the training and validation datasets (height, aspect ratio and density).

although especially at traffic lights and in slow moving traffic, same persons might appear in different annotations. Even so, due to sparse sampling at every four seconds, image resolutions and body poses will differ.

3.2 Dataset Annotation

We annotated pedestrians and riders; the latter were further distinguished by their ride-vehicle type: bicycle, buggy, motorbike, scooter, tricycle, wheelchair.

Location. All objects were annotated with tight bounding boxes of the complete extent of the entity. If an object is partly occluded, its full extent was estimated (this is useful for later processing steps such as tracking) and the level of occlusion was annotated. We discriminated between no occlusion, low occlusion (10-40 percent), moderate occlusion (40-80 percent), and strong occlusion (larger than 80 percent). Similar annotations were performed with respect to the level of object truncation at the image border (here, full object extent was not estimated). For riders, we labeled the riding person and its ride-vehicle with two separate bounding boxes, and annotated the ride-vehicle type. Riderless-vehicles of the same type in close proximity were captured by one class-specific group box (e.g., several bicycles on a rack).

In [1] and [4] one vertical line is drawn and automatically converted into a rectangular box of a fixed aspect ratio. Because of the diverse pedestrian aspect ratios (see Fig. 2 middle) and to be comparable with the KITTI dataset, we remained with the classical bounding-box convention of labeling the outermost object parts. For every sampled frame, all visible persons were annotated; otherwise, missed annotations could lead to the flawed generation of background samples during training and bootstrapping. Also persons in non-upright poses (e.g., sitting, lying) were annotated or persons behind glass. These cases were tagged separately.

A person is annotated with a rectangular (class-specific) ignore region if a person is smaller than 20 *px*, if there are doubts that an object really belongs to the appropriate class, and if instances of a group can not be discriminated properly. In the latter case, several instances may be grouped inside a single ignore region.

Orientation. The overall object orientation is an important cue for the prediction of future motion of persons in traffic scenes. We provide this information for all persons larger than 40 *px* (including those riding).

Additional Tags. Person depictions (e.g., large poster) and reflections (e.g., in store windows) were annotated as a separate object class. Additional events were tagged at the image level, such a lens flare, motion blur, and rain drops or a wiper in front of the camera.

All annotations were manually performed; no automated support was used, as it might introduce an undesirable bias towards certain algorithms during benchmarking. We placed reasonably high demands on accuracy. The amount of missed and hallucinated objects were each to lie within 1 percent of the annotated number. Annotators were asked to be accurate within two pixels for bounding box sides (apart from ignore regions) and within 20 degrees for orientation. Annotations were double checked by a quality validation team that was disjoint from the annotation team. If needed, several feedback iterations were run between the teams to achieve a consolidated outcome. Experiments regarding annotation quality are listed in Section 4.3.

3.3 Data Subsets

We define various data subsets on the overall EuroCity Persons dataset. First, we distinguish a day-time and a night-time data subset, each with its own separate training, validation and test set. Three overlapping data subsets are furthermore defined, considering the ground-truth annotations, similar to [3], [4], [5]:

- *Reasonable:* Persons with a bounding box height greater than 40 *px* which are occluded/truncated less than 40 percent
- *Small:* Persons with a height between 30 *px* and 60 *px* which are occluded/truncated less than 40 percent
- *Occluded:* Persons with a bounding box height greater than 40 *px* which are occluded between 40 and 80 percent

These data subsets can be used in test cases to selectively evaluate properties of person detection methods for various sizes or degrees of occlusion.

Each city recording lasted on average 1.7 hours. In order to increase the chances that certain time-dependent environmental conditions (e.g., a rain shower, particular type of road infrastructure or buildings) were well represented across training, validation and test set, for each city the recordings are separated into chunks with a duration of at least 20 minutes. The recorded images of each chunk were split into training, evaluation, and test by 60, 10, and 30 percent respectively, as illustrated in Fig. 3. During halts

Fig. 3. The applied test, val, and train split visualized for one city. Assuming a recording length of one hour for this city, the whole session is divided into three equidistant 20 minute subsets. Each subset is then split into train, validation, and test by a 60, 10, 30 percent distribution.

due to traffic lights or jams people could appear in several consecutive frames. To facilitate that the test, validation and training sets are disjunct in terms of people we only splitted sequences at points in time where the recording vehicle had a speed larger than 7 $\mathrm{km/h}$. By placing furthermore the validation set intermittently with the training and test, it was all but avoided that the latter two would contain the same physical person.

3.4 Dataset Characteristics

See Table 1 and Fig. 2 for some statistics on the new EuroCity Persons dataset. Seasonality, weather, time of day and, to some degree, geographical location, all influence clothing and thus person appearance. These factors also influence the person density observed, which, as shown in Fig. 2 (right) varies a lot, not only per frame but also per city. For example, the lowest average number of pedestrians per city (1.8) occurred in Leipzig likely due to the rainy weather during recording. Very crowded scenarios have been collected in Lyon with on average 9.5 pedestrians per image. These imply challenging occlusions and overlapping objects that complicate non-maximum suppression (these difficult scenarios are missing in KITTI and Caltech, where on average there is about one pedestrian per frame). Geographical location also influences the background (i.e., vehicles, road furniture, buildings). The time-of-the-day has furthermore a significant impact on scene appearance. Recordings at nighttime suffer from low contrast, color loss and motion blur.

By driving through a large part of Europe, during all four seasons, in most weather conditions (apart from heavy rain or snowfall), and during day and night, we recorded very diverse backgrounds and person appearances, see Table 5.

3.5 Evaluation Metrics

To evaluate detection performance, we plot the miss-rate (mr) against the number of false positives per image ($fppi$) in log-log plots

$$mr(c) = \frac{fn(c)}{tp(c) + fn(c)}, \tag{1}$$

$$fppi(c) = \frac{fp(c)}{\#img}, \tag{2}$$

where $tp(c)$ is the number of true positives, $fp(c)$ is the number of false positives, and $fn(c)$ is the number of false negatives, all for a given confidence value c such that only detections are taken into account with a confidence value greater or equal than c. As commonly applied in object detection evaluation [2], [3], [4], [24] the confidence threshold c is used as a control variable. By decreasing c, more detections are taken into account for evaluation resulting in more possible true or false positives, and possible less false negatives. We define the log average miss-rate ($LAMR$) as

$$LAMR = \exp\left(\frac{1}{9}\sum_{f} \log\left(mr(\arg\max_{fppi(c)\leq f} fppi(c))\right)\right), \tag{3}$$

where the 9 $fppi$ reference points f are equally spaced in the log space, such that $f \in \{10^{-2}, 10^{-1.75}, \ldots, 10^{0}\}$. For each $fppi$ reference point the corresponding mr value is used. In the absence of a miss-rate value for a given f the highest existent $fppi$ value is used as new reference point, which is enforced by $mr(\mathrm{argmax}_{fppi(c)\leq f} fppi(c))$. This definition enables $LAMR$ to be applied as a single detection performance indicator at image level. At each image the set of all detections is compared to the ground-truth annotations by utilizing a greedy matching algorithm. An object is considered as detected (true positive) if the Intersection over Union (*IoU*) of the detection and ground-truth bounding box exceeds a pre-defined threshold. Due to the high non-rigidness of pedestrians we follow the common choice of an *IoU* threshold of 0.5. Since no multiple matches are allowed for one ground-truth annotation, in the case of multiple matches the detection with the largest score is selected, whereas all other matching detections are considered false positives. After the matching is performed, all non matched ground-truth annotations and detections, count as false negatives and false positives, respectively. In addition, to allow a comparison with results from other work [3], [5] we also utilize the Average Precision (AP), which is defined as

$$AP = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}} \max_{re(c)\geq r} pr(c), \tag{4}$$

with the recall $re(c) = tp(c)/(tp(c) + fn(c))$, and precision $pr(c) = tp(c)/(tp(c) + fp(c))$, both for a given confidence threshold c.

For the evaluation of joint object detection and pose estimation we use the average orientation similarity (AOS) [3]

$$AOS = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}} \max_{\tilde{r}:\tilde{r}\geq r} s(\tilde{r}), \tag{5}$$

where s is the orientation similarity given by

$$s(r) = \frac{1}{|\mathcal{D}(r)|}\sum_{i\in\mathcal{D}(r)} \frac{1 + \cos\Delta_{\theta}^{(i)}}{2}\delta_i. \tag{6}$$

$\mathcal{D}(r)$ denotes the set of all object detections at recall r and $\Delta_{\theta}^{(i)}$ is the difference between the estimated and the ground-truth angle. δ_i is set to 1, if detection i has been assigned to a ground truth bounding box ($IoU > 0.5$) else it is set to zero, to penalize multiple detections which explain a single object. Thus, the upper bound of the AOS is given by the AP score.

As in [3], [4], neighboring classes and ignore regions are used during evaluation. Neighboring classes involve entities that are semantically similar, for example bicycle and moped riders. Some applications might require their precise distinction (*enforce*) whereas others might not (*ignore*). In the latter case, during matching correct/false detections are not credited/penalized. If not stated otherwise, neighboring classes are ignored in the evaluation. In addition to ignored neighboring classes all persons annotations with the tags *behind glass* or *sitting-lying* are treated as ignore regions. Further, as mentioned in Section 3.2, ignore regions are used for cases where no precise bounding box annotation is possible (either because the objects are too small or because there are too many objects in close proximity which renders the instance based labeling infeasible). Since there is no precise information about the number or the location of objects in the ignore region, all unmatched detections which share an intersection

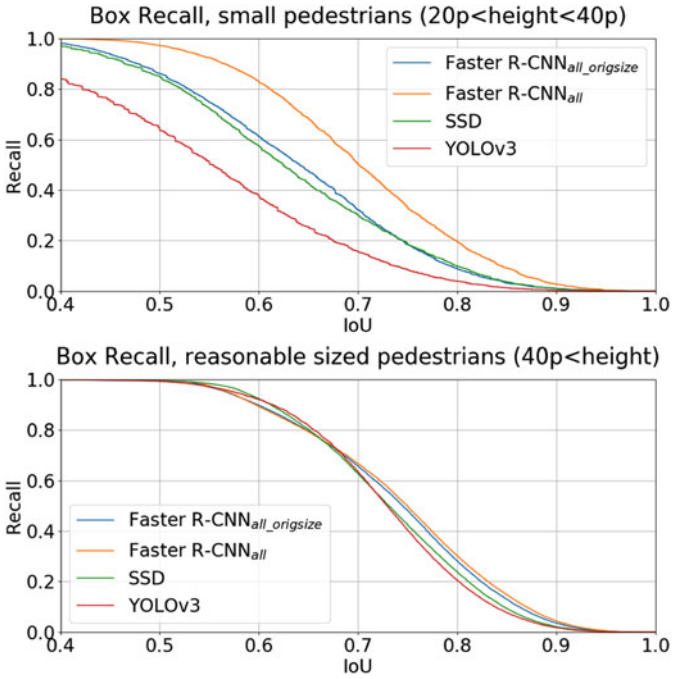


Fig. 4. Recall versus *IoU* for small pedestrians (top) and pedestrians of the "reasonable" test case (down) for the optimized anchor-boxes of Faster R-CNN and YOLOv3 and the SSD default boxes.

of more than 0.5 with these regions are not considered as false positives.

3.6 Benchmarking

The EuroCity Persons dataset, including its annotations for the training and validation sets, is made freely available to academic and non-profit organizations for non-commercial, scientific use. The test set annotations are withheld. An evaluation server is made available for researchers to test their detections, following the metrics discussed in previous Section. Results are tallied online, either by name or anonymous. The frequency of submissions is limited.

4 Experiments

All the baseline and generalization experiments (Sections 4.1 and 4.2) involved the day-time EuroCity Persons dataset and the pedestrian class, for comparison purposes with earlier works. This also holds in part for the data aspects experiments (Section 4.3), unless stated otherwise.

4.1 Baselines

As the top ranking methods on KITTI and Caltech use deep convolutional neural networks, we select our baselines among these methods. Many recent pedestrian detection methods [54], [55], [56], [61], [62], [69], [71] are extensions of Fast/Faster R-CNN and profit from the basic concepts of these methods. Therefore, *Faster R-CNN* is evaluated as prominent representative of the two stage methods. As shown in [4], it can reach top performance for pedestrian detection if it is properly optimized. The one stage methods often trade faster inference against a lower detection accuracy. YOLO [32] is one of the first methods within this group. We evaluate its latest extension *YOLOv3* [33], as in comparison with its predecessors, its design is promising regarding the detection of smaller objects. Within both groups we also select methods with explicit hard example mining, namely *R-FCN* [31] and *SSD* [34].

Faster R-CNN, R-FCN and SSD are trained with the Caffe framework [72] using VGG-16 [44] as base architecture (as done for pedestrian detection in [4], [59], [61], [62], [71]; using ResNet as base architecture for Faster R-CNN did not improve our experimental results, see supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2897684). YOLOv3 is trained with the Darknet framework [73] and Darknet-53 [33] as base architecture. The base architectures are pre-trained on ImageNet [6].

Adaptations and Training. We optimize the box recall for all methods as it is important for the overall detection performance. For Faster R-CNN and R-FCN we apply improvements from [4] adapting the scales and aspect ratios of the anchor-boxes, reducing the feature stride by removing the last max pooling layer and upscaling the input image during training and testing. SSD and YOLOv3 can in practice not be trained on upscaled images because of higher memory demands and the limitations of the used graphics cards. Still, we optimize the default boxes of SSD and the anchor boxes of YoloV3 resulting in similar recalls for all methods for the "reasonable" test case as shown in Fig. 4. For Faster R-CNN

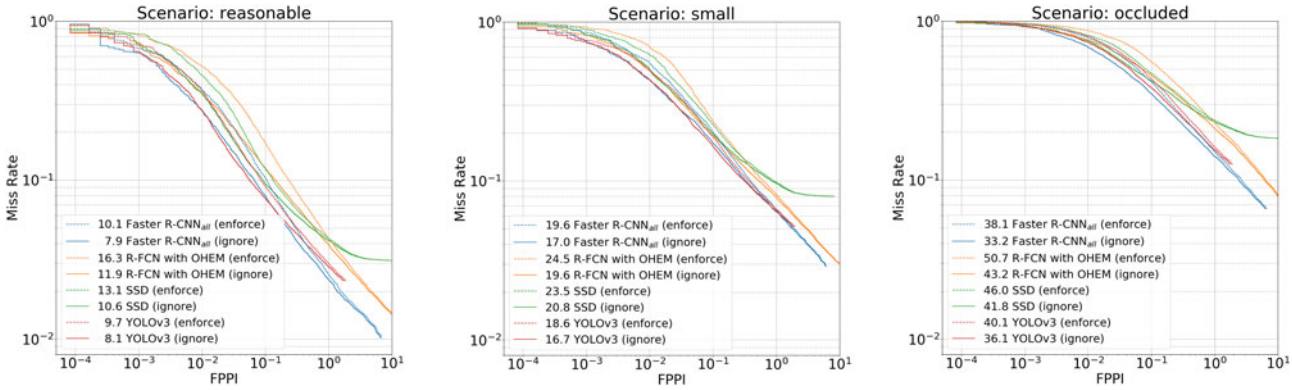


Fig. 5. Miss-rate curves on the EuroCity Persons test set for our selected methods for the "reasonable" (left), "small" (middle) and "occluded" (right) test case. The required *IoU* for a detection to be matched with a ground-truth sample is 0.5. For every method, the curves are shown for enforcing or ignoring precise class label with respect to neighboring classes.

TABLE 3
Training Settings of the Faster R-CNN Method, Differing in the Heights and Degree of Occlusion of the Samples Used for Training and in the Upscaling Factor used by Bilinear Interpolation (between Brackets)

	height	occlusion	upscaling
Faster R-CNN$_{small}$	$[20, \infty]$	$[0, 40]$	yes (1.3)
Faster R-CNN$_{reasonable}$	$[40, \infty]$	$[0, 40]$	yes (1.3)
Faster R-CNN$_{occluded}$	$[40, \infty]$	$[0, 80]$	yes (1.3)
Faster R-CNN$_{all}$	$[20, \infty]$	$[0, 80]$	yes (1.3)
Faster R-CNN$_{all_origsize}$	$[20, \infty]$	$[0, 80]$	no
Faster R-CNN$_{baseline}$	$[20, \infty]$	$[0, 40]$	no

TABLE 4
Log Average Miss-Rate ($LAMR$) on the Test Set of the EuroCity Persons Benchmark for Different Settings of the Optimized Methods

	Test Case		
	reasonable	small	occluded
Faster R-CNN$_{small}$	**7.2**	**16.4**	51.3
Faster R-CNN$_{reasonable}$	7.3	24.7	50.0
Faster R-CNN$_{occluded}$	7.8	25.1	33.3
Faster R-CNN$_{all}$	7.9	17.0	**33.2**
Faster R-CNN$_{all_origsize}$	9.2	23.1	34.5
Faster R-CNN$_{baseline}$	9.3	22.5	54.4
YOLOv3	8.1	16.7	36.1
SSD	10.6	20.8	41.8
R-FCN OHEM	11.9	19.6	43.2
R-FCN NoOHEM	12.0	19.4	44.9

and R-FCN we implement an ignore region handling similar to [4]. Furthermore, we filter training samples according to different test cases to train several Faster R-CNN models as summarized in Table 3. For all experiments with R-FCN, SSD and YOLOv3 we filter samples that are more than 80 percent occluded or smaller than 20 *px* in height. SGD is used as back-propagation algorithm on the training dataset with a step-wise reduced learning rate. The model to be evaluated on the test dataset is selected on the validation dataset. See supplemental material, available online for further details.

Results. See Table 4 for the quantitative results obtained with the methods considered. Variants of the two stage method Faster R-CNN perform overall best on the three test cases. Faster R-CNN$_{small}$ performs best on the corresponding "small" test case, and interestingly, also slightly better on the "reasonable" test case. Faster R-CNN$_{all}$, that is trained with pedestrians of all sizes and of occlusions up to 80 percent, performs best overall. It also performs slightly better than Faster R-CNN$_{occluded}$ on the "occluded" test case. The Faster R-CNN variants ($all\_origsize, baseline$) that are trained and tested with the original image resolution perform slightly worse for the "reasonable" and "occluded" test cases than the other Faster R-CNN variants. Still, they run 66 percent faster during training and testing. As could be expected by the lower box recall shown in Fig. 4, there is a considerable performance difference for small sized pedestrians. Interestingly, both one stage detectors YOLOv3 and SSD perform better than R-FCN at least on the "reasonable" and "occluded" test cases. One of the main differences between Faster R-CNN and R-FCN is the use of the bootstrapping method OHEM. OHEM proves useful when comparing results for the two R-FCN variants with enabled and disabled OHEM for the "occluded" test case.

See Table 5 for some illustrations of typical results with Faster R-CNN$_{all}$ (we include night-time and rider results, not part of this section).

Failure Analysis. We now analyze the detection errors of our best-performer on Faster R-CNN$_{all}$ qualitatively and quantitatively. Table 6 illustrates false positives and false negatives of this method at a false positive per image rate of 0.3 for the "reasonable" test case, clustered by main error source. As can be seen, clothes, depictions and reflections are main sources for confusion with real pedestrians and thus for false positives (our evaluation policy is strict and we count these wrong due to application considerations; note, however, that depictions and reflections are annotated in our dataset, thus a more lenient policy to ignore false positives of these type is readily implemented).

Certain pedestrian poses and aspect ratios can lead to multiple detections for the same pedestrian as shown in the *Multi-detections* category. Non-maximum suppression (NMS) is used by Faster R-CNN and other deep learning methods to suppress multiple detections. We use an *IoU* threshold of 0.5 which is not sufficient to suppress detections that have very diverse aspects. On the other hand, a higher *IoU* threshold would result in more false negatives. These already occur for an *IoU* threshold of 0.5 as shown in the *NMS repressing* category. In these instances, pedestrians are occluded less than 40 percent and thus have to be detected in the "reasonable" test case. Because of the high *IoU* between pedestrians not all of them can be detected because of the greedy NMS. Thus, NMS is an important part of many deep learning methods that is usually not trained but has a great influence on detection performance.

Small and occluded pedestrians are a further common source for false negatives as already shown by the "small" and "occluded" test cases. In traffic scenarios usually only the lower part of a pedestrian is occluded due to parked cars or other obstacles. In our qualitative analysis we have false negatives where the head is occluded. These are particularly challenging for pedestrian detection methods, as these cases are quite rare in the training dataset. Further challenges are rare poses or pedestrians leaning on bicycles as shown in the *Others* group.

For the quantitative analysis of false positives we build upon the ideas of oracle tests as in [1]. There, false positives touching ground-truth samples are regarded as localization error. Non-touching false positives are regarded as confusion of fore- and background. We analyze false positives types for a finely discretized range of false positive per image (*fppi*), see Fig. 6. In this study, we further subdivide the localization errors in four groups: multiple detections ($IoU > 0.5$ with ground-truth samples, as we penalize multiple assignments), and detections touching matched ground truth samples, non-matched ground truth samples, and ignore regions, respectively. In this context an ignore region may either be an ignore region annotation or an object that has not to be detected in the "reasonable" test case. We also subdivide the fore- and background confusions into three groups: detections that can be matched with depictions and reflections, and other background, further subdivided whether smaller than 80 px in height or not.

TABLE 5
Qualitative Detection Results of Faster R-CNN$_{all}$ at *fppi* of 0.3 (Green: Pedestrians, Blue: Riders)

True Positives

Samples are recorded during dry weather (first row), rainy weather and wintertime (second row), and during dusk and night (last two rows).

Fig. 6 shows that localization errors account for about 60 percent of all errors at a high *fppi* of 6, decreasing to about 40 percent for a low *fppi* rate of 4×10^{-3}. The share of false positives touching ground-truth samples remains approximately the same for the entire *fppi* range. Of these touched ground-truth samples, an increasing proportion is non-matched, for decreasing *fppi*. The share of false positives touching ignore regions is similar for a large *fppi* range but decreases somewhat for *fppi* below 10^{-2}. Possible objects inside these ignore regions seem to lead to erroneous detections in their surroundings. In terms of classification errors, depictions and reflections are among the hardest error sources to take care off: at decreasing *fppi* the share of this error type increases. Also the share of larger other-background objects increases with decreasing *fppi*.

Computational Efficiency. Processing rates for the R-FCN, Faster R-CNN, SSD and YOLOv3 on non-upscaled test images were 1.2 *fps*, 1.7 *fps*, 2.4 *fps* and 3.8 *fps*, respectively, on a Intel(R) Core(TM) i7-5960X CPU 3.00 GHz processor and a NVidia GeForce GTX TITAN X with 12.2 GB memory. There are several possibilities to optimize the runtime, such as replacing the VGG base architecture by a GoogLeNet model [43] and upgrading to the latest GPU processor; this was outside the scope of this study.

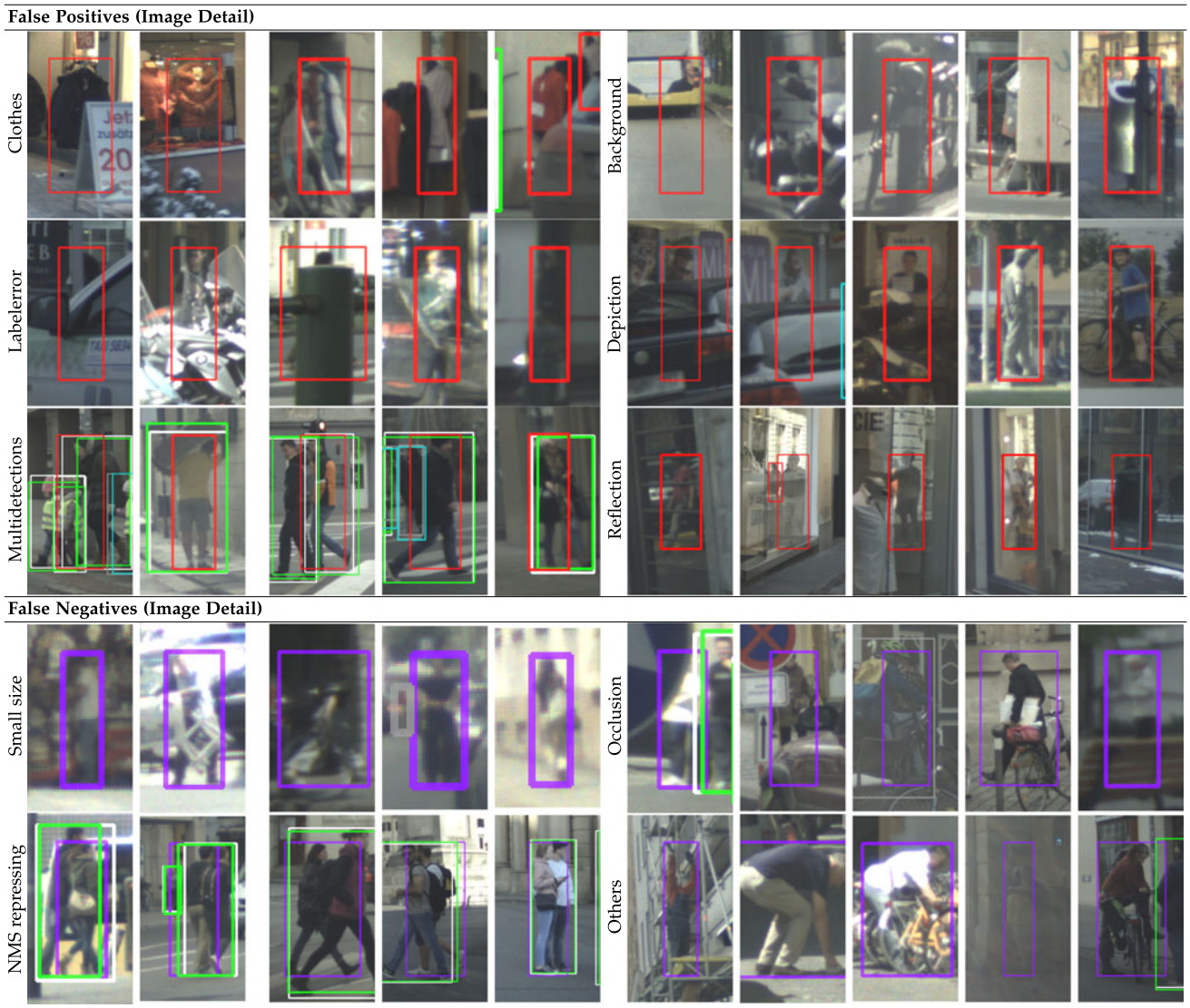
For our remaining experiments we focus on Faster R-CNN as best performing method. Results for other methods are shown when they lead to additional insights.

4.2 Generalization Capabilities

A dataset with a reduced bias should better capture the true world, and result in superior generalization capabilities of the detectors which are trained on this dataset. KITTI, CityPersons (CP) and EuroCity Persons (ECP) all involve traffic-related datasets but contain differences. KITTI and ECP, for example, differ in camera types used for recording. Even for a casual observer the images of these datasets look differently regarding colors and style. The CP and ECP datasets have been recorded with similar cameras. Still, they differ regarding the annotation bias, as the aspect ratios of all bounding boxes provided by CP are the same, unlike ECP (cf. Section 3.2). The Open Images V4 dataset (OP), on the other hand, contains iconic images of persons; this "generic" setting is quite different to the traffic setting of KITTI, CP and ECP (an obvious difference is the much larger person sizes in OP).

Here, we want to examine how the various datasets generalize with respect to the traffic-related ("target") datasets KITTI, CP, and ECP. For this, we consider various training

TABLE 6
Qualitative Detection Results for Faster R-CNN$_{all}$ at 0.3 $fppi$ (Green: True Positives, Red: False Positives, Purple: False Negatives, White: Ground Truth)



sets (in isolation and with pre-training) and measure the performance of a reference model (i.e., the optimized Faster R-CNN baseline) on a target evaluation set.

The OP dataset contains 3.2M individually labeled persons from 736,433 images. Labeled groups of persons are used as ignore regions in our experiments. To compensate for the large person sizes we downscale the OP images by a factor of 2 (OP512) or by a factor of 4 (OP256). We split the official KITTI training dataset into two equally sized, disjunct subsets to obtain our KITTI training and validation datasets, as in [51].

All models derived from the individual KITTI, CP, OP, and ECP datasets are initialized with ImageNet [6]. Pre-training a model with a source dataset means selecting its best performing version during training based on evaluation on the validation set of the pre-training dataset. The training strategy and all hyper-parameters for fine-tuning are kept the same to ensure the changes in performance can be traced back to the model used for initialization.

The results of our generalization experiments are shown in Tables 7, 8 and 9. A first observation is that if no pre-training is used (rows 1, and 6-9 of Tables 7, 8, and 9), then the best performance on the target evaluation dataset is obtained when training with the target training dataset (row 1 of respective tables). The second best performance in that case is achieved by training with the ECP training set (for KITTI and CP as targets, see Tables 7 and 8). Training with the OP-only training set gives notably bad results, despite its large size.

A second observation is that pre-training with very large training sets (ECP, OP) allows to surpass the performances significantly of using solely the target training sets, for the target test sets of smaller size (KITTI and CP). Pre-training results in an improvement of about 6, 9, and 12 percentage points in average precision for the "easy", "moderate" and "hard" KITTI validation datasets, respectively, when compared to using the original KITTI training data set. Similarly, pre-training results in an improvement of 3, 9, and 6 percentage points in $LAMR$ for the "reasonable", "small", and "occluded" CP validation datasets, respectively, when compared to using the original CP training data set. Pre-training

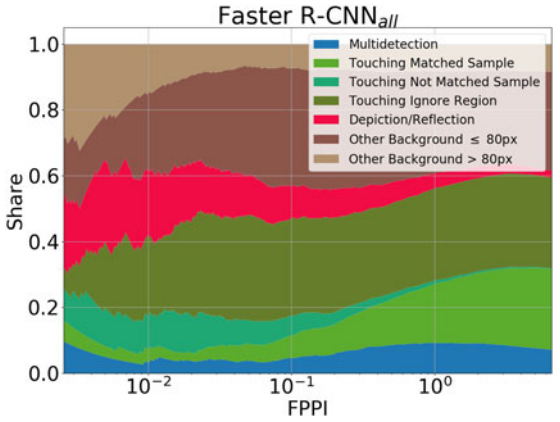


Fig. 6. The contribution of various sources to the number of false positives of Faster R-CNN$_{all}$, depending on $fppi$.

TABLE 8
Log Average Miss-Rate ($LAMR$) on the CityPersons (CP) Validation Set for Different Training Settings of Faster R-CNN

	CityPersons Validation Set		
Training Data	reasonable	small	occluded
CP	17.2	38.9	52.0
ECP→CP	15.0	**30.0**	**45.8**
KITTI→CP	17.0	39.3	51.7
OP256→CP	15.6	30.9	47.3
OP512→CP	**14.7**	32.3	48.0
ECP	25.5	43.8	62.6
KITTI	57.7	81.4	88.1
OP256	55.5	67.8	88.8
OP512	48.2	66.6	85.3

$A \rightarrow B$ denotes pre-training on A and finetuning on B.

with ECP is especially valuable for the hard or occluded cases, involving improvements of about 10 percentage points in $LAMR$ or average precision.

Pre-training with OP and with ECP do similarly well for the easier test case of CP (see "reasonable" column of Table 8). That OP is competitive with ECP in this case should perhaps not come as a big surprise, given this test case involves comparatively large and un-occluded pedestrians, where the OP dataset has some similarity with the target dataset. Yet size is not all that matters. Despite being one order of magnitude larger in size than ECP, when it comes to the harder test cases (see "moderate/small" and "hard/occluded" columns of Tables 7 and 8), pre-training with ECP outperforms pre-training with OP significantly. For the KITTI validation set (Table 7), we see an improvement of at least 1.3 and 2.9 in average precision for the "moderate" and "hard" test cases. For the CP validation set (Table 8) this improvement is at least 0.9 and 1.5 in $LAMR$.

Pre-training with ECP furthermore strongly outperforms pre-training with KITTI or CP across the board. For the KITTI validation set, we see an improvement of 2.8, 3.6 and 5.6 in average precision for the "easy", "moderate" and "hard" test cases versus pre-training with CP (rows 2 and 3 in Table 7). For the CP validation set we see an improvement of 2.0, 9.3 and 5.9 in $LAMR$ (rows 2 and 3 in Table 8). Note that the $LAMR$ listed in [4] for training and testing on CP was 12.8 rather than 17.2 listed here. The difference arises from a difference in the "reasonable" test case settings used. If we use the exact same settings as in [4], we arrive at an even better $LAMR$ of 12.2, which is improved by ECP pre-training to 10.2.

We also tested the benefit of pre-training with ECP on the official KITTI test set by submitting to the evaluation server on the KITTI website. Our pre-trained model on ECP achieved an average precision of 74.3 for the moderate setting. At the moment of our submission this results in rank 6. The Faster R-CNN model trained with KITTI data alone achieved an average precision of 63.5 resulting in rank 32.

A third observation is that when considering the ECP dataset as target, pre-training on the other datasets only helps marginally, if at all (see Table 9).

4.3 Dataset Aspects

What aspects make a dataset worthwhile and facilitate that it generalizes well? We argue that these aspects are diversity, quantity, accuracy, and detail. We now examine these in turn for the ECP dataset. Faster R-CNN$_{baseline}$ is used as training setting without upscaling images because of computational considerations.

Quantity. [7] shows a logarithmic relation between the amount of training data and the performance of deep learning methods. We validate this relation on our benchmark. Therefore we train our baseline methods on different sized subsets which are randomly sampled from all cities. The detection results for our baseline methods with the use of

TABLE 7
Average Precision on the KITTI Validation Set for Different Training Settings of Faster R-CNN

	KITTI Validation Set		
Training Data	easy	moderate	hard
KITTI	80.8	72.3	62.6
ECP→KITTI	**86.4**	**81.1**	**74.1**
CP→KITTI	83.6	77.5	68.5
OP256→KITTI	84.9	79.8	71.2
OP512→KITTI	85.2	78.7	69.3
ECP	73.9	68.7	61.4
CP	69.8	65.2	58.6
OP256	67.7	60.0	51.5
OP512	72.7	65.9	55.7

$A \rightarrow B$ denotes pre-training on A and finetuning on B.

TABLE 9
Log Average Miss-Rate ($LAMR$) on the EuroCity Persons (ECP) Test Set for Different Training Settings of Faster R-CNN

	EuroCity Persons Test Set		
Training Data	reasonable	small	occluded
ECP	**7.2**	16.4	33.2
CP→ECP	**7.2**	16.8	32.2
KITTI→ECP	7.4	16.5	32.9
OP256→ECP	7.4	**15.8**	31.6
OP512→ECP	**7.2**	16.3	**31.4**
CP	30.7	48.4	68.6
KITTI	65.3	82.8	92.3
OP256	66.8	77.3	93.2
OP512	51.9	74.4	90.9

$A \rightarrow B$ denotes pre-training on A and finetuning on B.

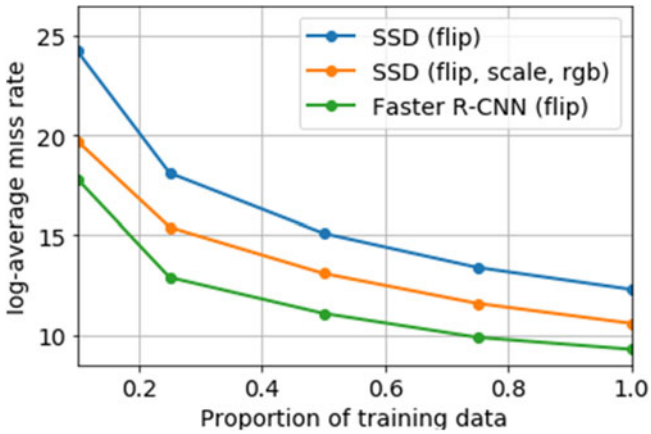


Fig. 7. Detection performance ($LAMR$) of Faster R-CNN and SSD as a function of training set size.

different augmentation modes in dependence of the dataset proportion are shown in Fig. 7. As image augmentations the images may be flipped or scaled in size. The *rgb* augmentation randomly shifts the colors of an image independently for the three color channels. We observe that logarithmic relation between training set size and detection performance also holds on our benchmark for Faster R-CNN and SSD.

Diversity. We wish to investigate whether overall geographical region introduces a dataset bias which influences person detection performance. For this, we constructed two datasets that are similar in terms of other influencing factors (i.e., season, weather, time of day, person density):

- *Central West Europe (WE):* Basel, Dresden, Köln, Nürnberg, Stuttgart, Ulm, Würzburg
- *Central East Europe (EE):* Bratislava, Budapest, Ljubljana, Prague, Zagreb

We split these datasets into subsets for training, validation and testing as described in Section 3, such that the number of pedestrians in each training dataset is 15,000. [74] shows that resampling of a dataset can be applied to evaluate the significance of benchmark results. We permutate the train-val-test blocks and vary the block length (between 10 and 30 minutes) resulting in 20 different dataset combinations for training, validation and testing. For every dataset combination one model is trained per region and evaluated on the corresponding test datasets of the two regions. The mean performances over all different dataset combinations and the standard deviations for these are shown in Table 10. In the case of a non existent dataset bias the difference between the output of both models comes from a distribution with zero median. This is used as the null-hypothesis

TABLE 10
Effect of Geographical Bias on Detection Performance ($LAMR$) for the "reasonable" Test Case: Central West Europe (WE) versus Central East Europe (EE)

	Test Set			
Training Set	WE (mean)	WE (std)	EE (mean)	EE (std)
WE	12.7	1.3	11.0	0.7
EE	14.4	2.3	**9.0**	0.4
WE&EE	**12.2**	1.1	9.6	0.7

Datasets compiled to provide otherwise similar conditions. Results involve averages over different dataset splits.

for the Wilcoxon signed-rank test [74]. For the same test set the 20 results for the model trained on the same location and the model trained on the other location are paired. We calculate the respective p-value, which is the probability of observing the test results given the null-hypothesis is true. For the WE and EE test sets, these values are 0.0098 and 0.0020, respectively. Hence, with a confidence interval of 99 percent, the null-hypothesis (the non-existence of a regional bias) can be rejected for both regions.

Another diversity factor is the time of day. Table 11 shows detection results for the day-time, night-time and combined datasets. As the night-time dataset is only 20 percent of day data (Table 1) we reduce for this experiment the number of training samples used for the day-time and combined models accordingly. Table 11 shows that training on day-time and testing on night-time gives significantly worse results than training and testing on the same time-of-day. Overall results are worse than those of other experiments due to the comparatively small training sets used.

Detail. The importance of additional annotations for ignore regions, for riders, and for orientations is now examined. Table 12 shows results for a model trained without ignore region handling compared to our baseline method. In accordance with earlier findings [4], we observe that detection performance deteriorates when not using ignore regions during training. For the "reasonable" and "small" test cases the $LAMR$ drops by about two points.

We extended the baseline detection method by an orientation estimation layer as in [52] (Two variants for the orientation loss is considered: L1 and Biternion loss). Hence, the network performs multi-tasking: classification, bounding box and orientation regression. As body orientation correlates with the aspect ratio we assume that the bounding box regression task and hereby the detection performance could also benefit from learning all three tasks jointly in one network. In contrast to [9] which shows that training multiple tasks together can improve the overall result, the detection

Fig. 8. Qualitative results for orientation estimation. Left and middle image show correct estimations. Right image contains a rare failure case (left person has orientation offset of about 180 degrees).

TABLE 11
Effect of Day- versus Night-Time Condition on Detection Performance ($LAMR$) for the "Reasonable" Test Case

	Test Set	
Training Set	Night	Day
Night	**18.4**	21.4
Day	33.3	**14.3**
Day and Night	22.7	14.5

Datasets compiled to provide otherwise similar conditions.

results decrease slightly for the multitask network with the Biternion loss as shown in Table 12. Fig. 9 shows the orientation estimation error as a function of object size (distance). The Biternion loss is superior to the L1 loss as it does not suffer from the periodicity of an orientation angle. Using the aggregated AOS metric from Section 3.5 for the "reasonable" test case we get a score of 85.9 for the L1 loss and 86.7 for the Biternion loss. See Fig. 8 for typical results with the Biternion loss.

The evaluation protocol described in Section 3.5 ignores detected neighboring classes. For pedestrians this means that riders are not considered as false positives. If these neighboring classes are instead counted as false positives, detection performance decreases as expected: the $LAMR$ for our baseline method increases from 9.3 to 11.0, as shown in Table 13. By adding riders as an additional class, one observes that the pedestrian detection performance improves for the protocol which requires pedestrians to be classified as such (10.3 versus 11.0). There is only a slight difference in performance when the network is trained to regress a bounding box for the rider alone or for the rider including the ride type. The absolute detection performance for pedestrians and riders is quite similar although there are 10 times more pedestrians than riders in our training dataset.

Accuracy. Here we evaluate to what degree our annotation accuracy requirements from Section 3.2 were actually met in practice in the final EuroCity Persons annotations.

To estimate the amount of missed annotations, we compare these with the object detector output. At a *fppi* of 0.3 for Faster R-CNN$_{all}$ on the "reasonable" test case we manually count 230 missed annotation larger than 32 *px*. However, the miss-rate for Faster R-CNN$_{all}$ at this *fppi* is about 10 percent for the small test scenario and about 30 percent for the occluded test scenario. Using the more conservative 30 percent figure, we estimate that, in fact, there are additional 99 missed annotations for pedestrians larger than 32 px, bringing the total missed annotation to 329. As there are about 48,000 pedestrians in the test dataset, this corresponds to 0.7 percent missed annotations, which lies within the 1 percent quality requirement of Section 3.2.

TABLE 12
Log Average Miss-Rate ($LAMR$) of the Detail Study

	Test Case	
Training Scenario	reasonable	small
Baseline	**9.3**	**22.5**
NoIgnoreHandling	10.8	24.5
Orientation L1	**9.3**	22.7
Orientation Bit	10.1	24.0

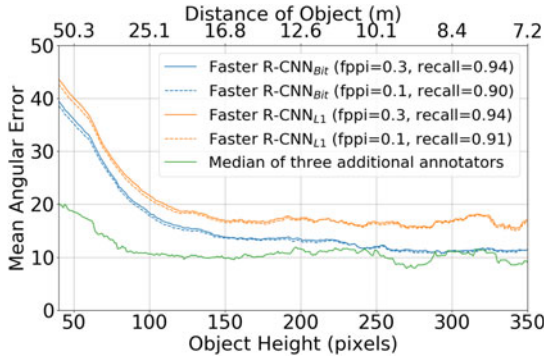


Fig. 9. Person orientation estimation quality versus object size (distance).

To determine the inter-annotator agreement and thus obtain an indication about achieved accuracy with respect to bounding box localization and orientation annotation, a random subset of 1,000 not occluded pedestrians was labeled again by three different persons. We analyze the average deviation between the median value of the three annotators and the corresponding Eurocity Persons annotation, in dependence of the object size. Fig. 10 shows that the average deviation of the bounding box extents stays below 1.4 *px* for objects up to 200 *px* high (interestingly, upper/lower box side more accurate than left/right side). Fig. 9 shows that in terms of orientation angle, the average deviation starts at 20 degrees for object sizes of 40 *px* and reduces to about 10 degrees for object sizes larger than 100 *px*. We note that this lies within the requirements of Section 3.2 as well.

We now artificially disturb the annotation quality of the training dataset in the following experiments, see Table 14. First, we randomly delete bounding boxes of instances and groups to simulate the effect of missed objects during annotation ("delete"). Second, we move bounding boxes by four pixels up or down and left or right ("jitter"). Third, we add (erroneous) ground-truth boxes to simulate the effect of hallucinating objects during annotation ("hallucination"). For this, a selected ground-truth bounding box itself is not changed but an additional, identically sized bounding box of the pedestrian class is placed at a random location in the image. Lastly, we introduce hallucinations that are more likely to resemble pedestrians, by running a SSD model of an early training stage on the training dataset (after 80,000 iterations). The 11,000 highest scoring false positives of these detections (corresponds to 10 percent of all pedestrians in the training dataset) are handled as regular groundtruth boxes

TABLE 13
Effect of Multi-Class Handling (Pedestrian versus Riders) on Detection Performance ($LAMR$) for the "Reasonable" Test Case

	Test			
	pedestrians		riders	
Training	ignore	enforce	ignore	enforce
Baseline (pedestrians)	9.3	11.0	-	-
+Riders only	**9.2**	**10.3**	**8.9**	**11.0**
+Riders with ride-vehicle	**9.2**	10.4	10.7	12.1

The "enforce" ("ignore") settings involves (not) penalizing samples of the other class for being categorized as the respective class. The first row (baseline) involves a single class, the second and third row involve two classes.

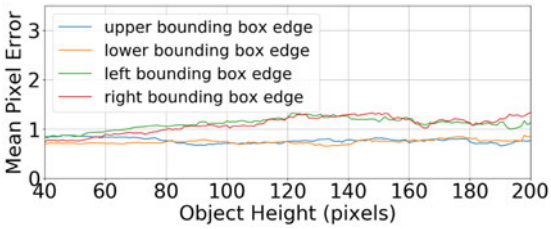


Fig. 10. Mean pixel error between median of three additional annotators and the ECP dataset annotations, in dependence of object height p (averaged over the interval [p-20, p+20]).

and added to the training dataset for the "false positives" experiment. We examine different levels of disturbances by manipulating different amounts of bounding boxes. The effects for disturbances that are even worse than in our very first pilot study are also evaluated. The probability for a bounding box to be disturbed is given in the Table 14.

The detection performance of Faster R-CNN suffers from deleting and disturbing the bounding box locations. Deleting 25 percent of the bounding boxes results in a miss-rate of 11.3. Note that with 75 percent of the training samples a $LAMR$ of 10.0 is achieved (see Fig. 7). Pedestrians without bounding box labels may be used as background samples during training which results in the confusion of pedestrians and background during testing. This effect is even stronger when OHEM is applied as seen when comparing R-FCN results with and without OHEM. Placing hallucinations at random locations only slightly influences the overall detection performance. Adding 10 percent hallucinations that more resemble pedestrians ("false positives") result in a more significant drop in performance of 3.3 points.

TABLE 14
Perturbation Analysis of Annotation, Effects on Performance

Method	Disturbance	Prob.	LAMR	Δ
Faster R-CNN	none	-	9.3	-
Faster R-CNN	delete	10%	9.9	+0.6
Faster R-CNN	delete	25%	11.3	+2.0
Faster R-CNN	false positives	10%	12.6	+3.3
Faster R-CNN	hallucination	20%	9.3	0.0
Faster R-CNN	hallucination	50%	9.8	+0.5
Faster R-CNN	jitter	10%	9.5	+0.2
Faster R-CNN	jitter	20%	9.7	+0.4
Faster R-CNN	jitter	50%	12.3	+3.0
R-FCN OHEM	none	-	11.9	-
R-FCN OHEM	delete	25%	14.9	+3.0
R-FCN NoOHEM	none	-	12.0	-
R-FCN NoOHEM	delete	25%	13.7	+1.7

5 Discussion

A main outcome from the EuroCity Persons experiments is that data still remains a driving factor for the person detection performance in traffic scenes: Even at training data sizes that are about one order of magnitude larger than existing ones (cf. Table 1), the considered state-of-the-art deep learning methods (Faster R-CNN and SSD) do not saturate in detection performance.

The fact that saturation does not occur can be attributed to the diversity of the data. The ECP dataset covers a large geographical region, day and night, and different weather conditions. This quality is reflected in its generalization capability across datasets. As was shown in Section 4.2, pre-training on ECP and fine-tuning (post-training) on a smaller target dataset (KITTI, CP) yields significantly better results than training solely on the target dataset. Pre-training on ECP also leads to better results than pre-training with other datasets on these target datasets. Conversely, pre-training with other datasets helps only marginally, if at all, when evaluating on the ECP test dataset. A "generic" dataset like Open Images V4 was shown to be beneficial for pre-training of the smaller traffic-related datasets (KITTI, CP), when ECP is not used. It could not outright replace the training sets of the latter.

The ECP dataset allowed us to analyze some biases in more detail. Foremost, experiments suggest that there is indeed a bias derived from large geographical region. We compiled datasets for central West Europe versus central East Europe, where other factors influencing performance were held similar. We found that the existence of a bias is statistically significant with a confidence interval of 99 percent.

Comparing day- and night-time detection performance, one observes from Table 11 that at equal training set sizes, night-time performance is worse (a $LAMR$ of four points higher). This difference is enlarged when the entire day- and night-time training sets of ECP are used as the former is an order of magnitude larger. See Fig. 11. The drop in recall for pedestrians closer than 8 m could be due to the headlights of the recording vehicle. These could result in very bright spots for the lower body of pedestrians and complicate detection. Our dataset provides the possibilities to further research in this direction and compare differences between day and night recordings.

The way annotations are performed proves to be important as well. As in [4] we show that a correct ignore region handling has an impact on detection performance. In our case it boosts performance by 1.5 points (see Table 12). This is a larger difference than that between the performances using 75 and 100 percent of the training data in Fig. 7. We go beyond [4] to show that it is beneficial to train specific detectors for classes that otherwise might be confused with the target class. In our experiments, the jointly trained detection models for riders and pedestrians achieve a lower miss rate for the pedestrian class, than models trained for pedestrians-only, when the precise class is enforced. In the evaluation protocol of [4] this case is not considered as riders are always handled as ignore regions.

It is interesting to put the current traffic-related person detection performance in context. When viewed in historic context, the best-performer on an early benchmark [11] was a method based on HOG features and SVM classifier. When comparing its performance with that of the best-performer in this paper, the R-CNN, one observes that performance has improved by an order of magnitude over the past decade, in terms of the reduction of the number of false positives at given correct detection rate, albeit dealing with two different datasets of urban traffic (Fig. 8 in [11] versus Fig. 5 here).

State-of-the-art detection performance (e.g., correct detection around 90 percent at $0.1 - 0.3$ $fppi$) is sometimes cited as evidence that performance is far away from practical use for an on-board vehicle application. This is incorrect, as can be readily inferred from the fact that there are already several

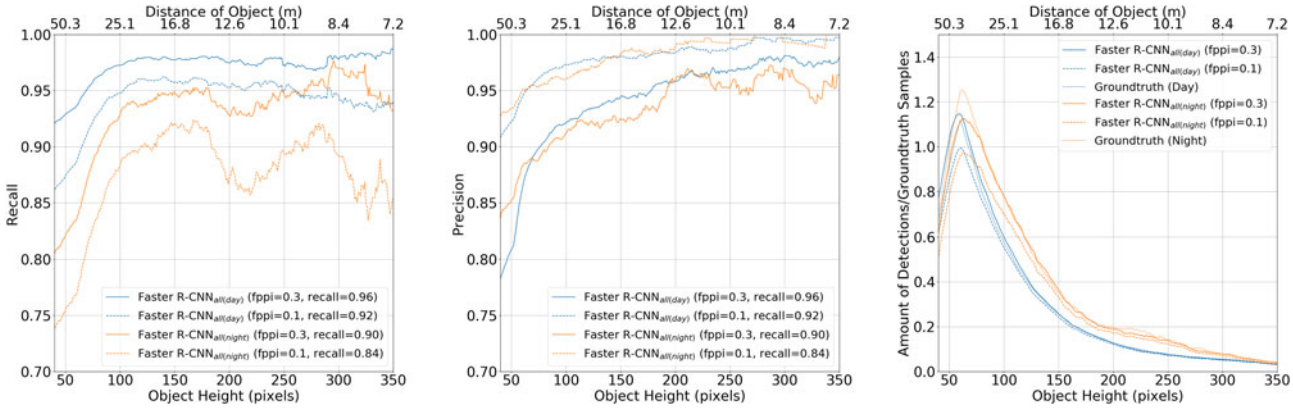


Fig. 11. Recall (left), precision (middle) and the associated per-image detection and ground-truth sample counts (right) versus object height at two operating points for the Faster R-CNN variant at day- and night-time (each trained and tested separately on upscaled day- and night-time images of EuroCity Persons reasonable). To calculate the distance of an object (upper x-axis) the camera calibration is used and a fixed object height of 1.7 m is assumed. For smoothing reasons, the recall and precision for object height p in pixels (*px*) is computed within the height range [p-20 *px*, p + 20 *px*].

vision-based person detection systems on-board production vehicles on the market. A number of factors improve performance in the vehicle application. First, other than we assume in this study, not all errors are equal in the vehicle application. Errors increasingly matter when they involve objects close to the vehicle. The detectors improve their performance with decreasing distance (increasing object size). See Fig. 11, the detection rate increases to 97 percent at a distance of 25 m (object height 100 *px*). Second, some false positives can be eliminated, when taking advantage of known scene geometry constraints (e.g., pedestrians or riders should be on the ground plane, their heights should be physically plausible when accounting for perspective mapping). Third, many false positives arise by an accidental overlaying of structures at different depths, and are not consistent over time when observed from a moving camera. Tracking can suppress such false positives ([11] shows a reduction by up to 37 percent). Last but not least, active safety systems for pedestrians and cyclists involve additional sensors for detecting obstacles in front of the vehicle: a second camera (stereo vision), radar or LiDAR. Thus vehicle actuation (braking, steering) does not solely rely on monocular object detection. It should be finally noted that current commercial systems are in the context of driver assistance, meaning that a correct detection performance of about 90 percent is acceptable, as long as the false alarm rate is essentially zero.

This brings us to the human baseline. A visual inspection shows that the remaining errors are indeed "hard", even for a human, see Table 6. A recent paper [1] finds that current single-frame pedestrian detection performance lags that of an attentive human by an order of magnitude. Thus there is a potential for a substantial further performance improvement; an improvement which would be important with the advent of fully self-driving vehicles.

More data remains part of the solution on how to improve performance. Our study shows that performance still improves with increasing training set size with a decent gradient (i.e., Fig. 7). A further doubling of the current training size (110,000 pedestrians) is projected to yield a reduction of the $LAMR$ from 9.3 to about 7.3 points. More training data is especially helpful for persons in non-standard poses, in rainy or night-time conditions, or under partial occlusions. The found relations between annotation quality and quantity on one hand and detection performance on the other (i.e., Table 14), together with a price tag for annotations at various quality levels can help optimizing the requirement specification for dataset annotation.

In terms of vision methods, better solutions are needed to provide accurate localization in the presence of multiple persons and significant occlusion. Recent detection methods like R-FCN or Faster R-CNN have profited from incorporating the proposal generation in an end-to-end learning strategy. Still, the proposal boxes are classified independently of each other resulting in multiple detections for the same object in particular if the proposals share similar image locations. In general, there is no loss enforcing a one to one matching between detections and ground-truth samples. The task of suppressing multiple detections for the same object is usually solved by the decoupled non-maximum suppression. Interestingly, most top performing methods of the common generic object detection benchmarks depend on a simple greedy non-maximum suppression [75]. This NMS poses a problem for overlapping objects e.g., in pedestrian groups. When selecting the *IoU* threshold there is a tradeoff between recall and precision as shown in Table 6. In [76] a neural network is trained to rescore detections, which renders a further NMS stage unnecessary. Still, the neural network solely relies on bounding box locations and confidence scores as input. To further improve the performance [76] proposes to incorporate image features in future works. Doing so the network could be informed about how many objects are present. As it is already a neural network architecture it can easily be integrated into existing detection networks. Thus, the three steps proposal generation, classification/bounding box regression, and NMS would finally be combined in a true end-to-end approach.

A number of methodical avenues could improve classification performance. In Fig. 11 and in our baseline experiments we show that small objects are still very challenging despite the great amount of small sized pedestrians present in our training dataset. Approximately 75 percent of the false positives at 0.3 *fppi* analyzed in Fig. 6 are smaller than

80 pixels. Recently, methods have been published that are tuned for the detection of smaller objects like MS-CNN. Such methods have to be analysed in detail to find still remaining weaknesses and further possibilities for improvement. We show quantitatively in Fig. 6 and qualitatively in Table 6 that depictions, reflections and clothes are often confused with real pedestrians. These confusions result in high scoring false positives also for sizes larger than 80 *px*. That necessitates the design of appropriate multi-task deep nets that more effectively incorporate global scene context. When training a detection network jointly for pedestrians and riders we have already shown that confusions between the two person classes can be reduced. Utilizing the already annotated reflections and depictions as additional classes during training could improve the discrimination performance as well. An ensemble of specialized deep learning models could take advantage of known bias (particular location and digital maps, weather, time of day). Such an approach could even switch on a per frame basis between sub-models, e.g., when there is a sudden change in lighting. For example lenseflares might occur from one frame to another when the vehicle turns into the direction of the sun.

As person detection is being perfected, the focus of research will likely shift to tracking and motion prediction. Motion prediction based on point kinematics is often not accurate because of abrupt changes in person motion. Systems like [77] come into play which take into account additional pose information. In preparation for this, we included in this benchmark the orientation estimation of the overall body, and showed that the latter can be jointly trained with the detection task at minimal performance loss.

6 Conclusions

We have created the new EuroCity Persons dataset, which takes annotations of persons in urban traffic scenes to a new level in terms of quantity, diversity and detail. We optimized four state-of-the-art deep learning approaches (Faster R-CNN, R-FCN, SSD and YOLOv3) to serve as baselines for the new person detection benchmark; we found a variant of Faster R-CNN to perform overall best, with a log-average-miss-rate of 7.9, 17.0 and 33.2 on the "reasonable", "small" and "occluded" test cases, respectively.

The experiments show that data is still a driving factor for the person detection performance in urban traffic scenes: Even at the new training data sizes that are about one order of magnitude larger than previous ones, the considered deep learning methods do not saturate in detection performance. This can also be attributed to the diversity of the dataset. In experiments on transfer learning, we showed that detectors pre-trained with the new dataset and fine-tuned on a target dataset, yield superior performance than those trained on the target dataset only (improvements on KITTI and CityPersons by 6-12 and 2-9 points, respectively). Conversely, pre-training with other datasets helped only marginally, if at all, when evaluating on the ECP test dataset.

The experiments also showed that night-performance is a few percentage points lower than day-time performance. Experimental results furthermore indicate that a statistically significant bias exists on detection performance across large-scale regions in Europe, resulting in performance variations of the same order. Adding orientation estimation to object detection lowers the detection performance by a single percentage point for the Biternion loss.

System performance regarding person detection in urban traffic settings has improved by an order of magnitude over the past decade; it is now closing in on human performance. Future improvement will in part still come from additional data. More person training data is especially helpful in non-standard poses, in rainy or night-time conditions, or under partial occlusions. We provided some insights regarding the effect of annotation accuracy on performance that could be useful for future annotation efforts. The development of appropriate multi-task deep networks, which combine a holistic approach to scene understanding with specialized person detection, taking advantage of known bias (geolocation, time of day, weather condition) seem promising. We hope that the new EuroCity Persons benchmark will stimulate research towards finding the "perfect" person detector for traffic scenes, a detector that could save lives on-board intelligent vehicles.

References

[1] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–985, Apr. 2018.

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[4] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4457–4465.

[5] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A new benchmark for vison-based cyclist detection," in *Proc. IEEE Intell. Veh. Symp.*, 2016, pp. 1028–1033.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[7] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.

[8] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[9] R. B. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[11] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[12] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2014, pp. 613–627.

[13] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4073–4082.

[14] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "An exploration of why and when pedestrian detection fails," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, 2015, pp. 2335–2340.

[15] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1521–1528.

[16] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.

[17] D. Gerónimo, A. Sappa, A. López, and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," in *Proc. 5th Int. Conf. Comput. Vis. Syst.*, 2007, vol. 39.

[18] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 373–378.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[20] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[21] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 794–801.

[22] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[23] 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. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[24] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[26] Berkeley deep drive dataset, 2018. [Online]. Available: http://bdd-data.berkeley.edu/

[27] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1543–1550.

[28] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5482–5491.

[29] G. Sharma and F. Jurie, "Learning discriminative spatial representation for image classification," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[30] (2018). [Online]. Available: https://storage.googleapis.com/openimages/web/index.html

[31] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[33] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv: 1804.02767*, 2018.

[34] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[36] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3033–3040.

[37] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3198–3205.

[38] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.

[39] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 947–954.

[40] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3666–3673.

[41] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1751–1760.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. Int. Conf. Learn. Representations*, 2015.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[46] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?" *arXiv:1608.08614*, 2016.

[47] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, Apr. 2016.

[48] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[49] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 328–335.

[50] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300 fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.

[51] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[52] M. Braun, Q. Rao, Y. Wang, and F. Flohr, "Pose-RCNN: Joint object detection and pose estimation using 3D object proposals," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, 2016, pp. 1546–1551.

[53] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[54] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[55] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2129–2137.

[56] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu, "Scale-adaptive deconvolutional regression network for pedestrian detection," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 416–430.

[57] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

[58] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.

[59] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 752–760.

[60] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[61] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1951–1959.

[62] S. Huang and D. Ramanan, "Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4664–4673.

[63] S. Tulsiani and J. Malik, "Viewpoints and keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1510–1519.

[64] H. Su, C. R. Qi, Y. Li, and L. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2686–2694.

[65] L. Beyer, A. Hermans, and B. Leibe, "Biternion nets: Continous head pose regression from discrete training labels," in *Proc. German Conf. Pattern Recognit.*, 2015, pp. 157–168.

[66] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1259–1267, 2016.

[67] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[68] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–457.
[69] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 924–933.
[70] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3296–3297.
[71] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6034–6043.
[72] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[73] J. Redmon, "Darknet: Open source neural networks in C," 2013–2016. [Online]. Available: http://pjreddie.com/darknet/
[74] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[75] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5562–5570.
[76] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6469–6477.
[77] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 618–633.

Markus Braun received the MSc degree in computer science from the Karlsruhe Institute of Technology, Germany, in 2015. Since then he is working toward the PhD degree at TU Delft, Delft, The Netherlands. He is also currently with Daimler Research and Development with the Environment Perception Department, Ulm, Germany. His research interests include machine learning and video analysis for automated driving, with a focus on detection and pose estimation of vulnerable road users.

Sebastian Krebs received the MSc degree in computer science from the University of Ulm, Germany, in 2016. Since then he is working toward the PhD degree at TU Delft, Delft, The Netherlands. He is also currently with Daimler Research and Development with the Environment Perception Department, Ulm, Germany. His research interests include machine learning and video analysis for automated driving, with a focus on vulnerable road user tracking.

Fabian Flohr received the MSc degree in computer science from the Karlsruhe Institute of Technology, Germany, in 2012. He is currently working toward the PhD degree at the University of Amsterdam, The Netherlands. He is also currently with Daimler Research and Development, Ulm, Germany, where he has the technical lead for vulnerable road user sensing.

Dariu M. Gavrila received the PhD degree in computer science from the University of Maryland, College Park, in 1996. From 1997 until 2016, he was with Daimler R&D, Ulm, Germany, where he became a Distinguished scientist. He led the multi-year pedestrian detection research effort at Daimler, which was incorporated in the Mercedes-Benz S-, E-, and C-Class models (2013-2014). He was awarded the Outstanding Application Award 2014 from the IEEE Intelligent Transportation Systems Society, as part of a Daimler team. In 2003, he also became a part-time professor with the University of Amsterdam, in the area of intelligent perception systems. Over the past 20 years, he has focused on visual systems for detecting human presence and activity, with application to intelligent vehicles, smart surveillance, and social robotics. In 2016, he moved to TU Delft, where he since heads the Intelligent Vehicles group as a full-time professor.