TU Delft

Delft University of Technology

Visual Detection and Pose Estimation of Vulnerable Road Users for Automated Driving

Braun, M.

DOI
10.4233/uuid:983cffe4-f7ac-47bc-9fdc-8b671008c23c

Publication date
2022

Document Version
Final published version

Citation (APA)
Braun, M. (2022). *Visual Detection and Pose Estimation of Vulnerable Road Users for Automated Driving*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:983cffe4-f7ac-47bc-9fdc-8b671008c23c

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

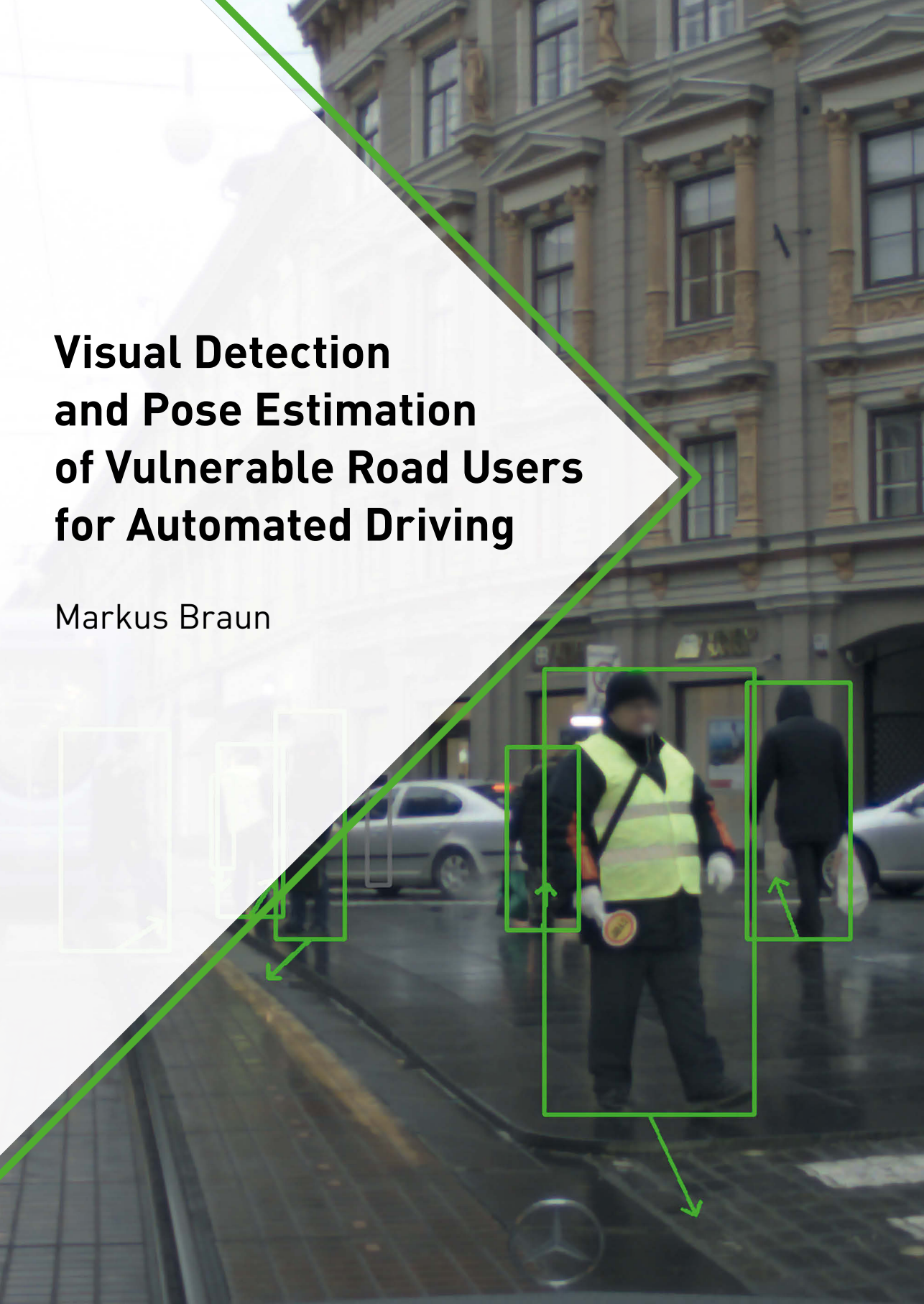
Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Visual Detection and Pose Estimation of Vulnerable Road Users for Automated Driving

Markus Braun



VISUAL DETECTION AND POSE ESTIMATION OF VULNERABLE ROAD USERS FOR AUTOMATED DRIVING

VISUAL DETECTION AND POSE ESTIMATION OF VULNERABLE ROAD USERS FOR AUTOMATED DRIVING

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op vrijdag 9 december 2022 om 12:30 uur

door

Markus BRAUN

Master of Sciene in Informatics,
Karlsruhe Institute of Technology, Karlsruhe, Duitsland,
geboren te Heilbronn, Duitsland.

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotie commissie bestaat uit:

Rector Magnificus,	voorzitter
Prof. dr. D.M. Gavrila,	Technische Universiteit Delft, promotor
Dr. J.F.P. Kooij,	Technische Universiteit Delft, copromotor
Onafhankelijke leden:	
Prof. dr. S. Nedevschi	Technical University of Cluj-Napoca
Prof. dr. B. Leibe	RWTH Aachen University
Dr. J.C. van Gemert	Technische Universiteit Delft
Prof. dr. ir. M. Wisse	Technische Universiteit Delft
Prof. dr. ir. H. Hellendoorn	Technische Universiteit Delft

Keywords: Person detection, Human pose estimation, Benchmarking, Intelligent vehicles, Automated driving

Cover art: NEUEFORM corporate designers, Hermann Schmidt.

Copyright © 2022 by M. Braun

ISBN 978-94-6384-397-3

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Science does not increase the infinite beauty of nature
but it may contribute to preserving it for our children and all generations to come.

Markus Braun

Contents

SUMMARY

This thesis addresses the topic of visual person detection and pose estimation. While these tasks are relevant for a broad range of applications, this thesis focuses on the domain of intelligent vehicles in urban traffic scenes. This domain is particularly interesting due to specific challenges related to visual perception from a moving vehicle. Accident statistics show that a great proportion of traffic fatalities affect vulnerable road users such as pedestrians and riders. This motivates the interest in reproducing or even surpassing the capabilities of an attentive human driver for driver assistance systems and fully automated driving to improve safety. Deep learning contributed to narrowing the performance gap between computer vision methods and human visual perception. Especially the capability of convolutional neural networks to learn powerful features is helpful for person detection and pose estimation. Throughout this thesis new deep learning methods for these tasks will be presented. The thesis not only focuses on methodical extensions but also on the creation of new datasets for training, evaluation, and benchmarking in the intelligent vehicles domain.

First, a novel approach for joint object detection and orientation estimation with a single deep convolutional neural network is presented. The orientation estimation is implemented by extending an existing convolutional network architecture with several carefully designed layers and an appropriate loss function. The network depends on external proposals for object candidate regions, whose accuracy is crucial for the overall performance. Therefore, two proposal methods are introduced that make use of 3D sensor data - precisely stereo as well as lidar data. The KITTI dataset, which is commonly used for object detection benchmarking in the automotive domain, serves for training and evaluation. The experiments on the KITTI dataset show that by combining proposals of both sensor modalities, high recall can be achieved while keeping the number of proposals low. Furthermore, the method for joint detection and orientation estimation is competitive with other state of the art approaches. It outperforms the state of the art for a test scenario of the bicycle class.

Big data has had a great share in the success of deep learning in computer vision. Still, the number of pedestrians and riders in the KITTI dataset is rather limited and previous works suggest that there is significant further potential to increase object detection performance by utilizing bigger datasets. Regarding benchmarking, small datasets are prone to dataset bias and overfitting.

Therefore, the second part of this thesis introduces 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 onboard a moving vehicle in 31 cities of 12 European countries. With over 238200 person instances manually labeled in over 47300 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 211200). Four state of the art deep learning approaches are thoroughly optimized to serve as baselines for the new object detection benchmark. In experiments with previous datasets, the generalization capabilities of these detectors when trained with the new dataset are analyzed. Furthermore, this thesis studies the effect of the training set size, the dataset diversity (day- vs. night-time, geographical region), the dataset detail (i.e., availability of object orientation information), and the annotation quality on the detector performance.

The qualitative and quantitative analysis of error sources for the best-performing detector reveals methodical weaknesses in dense traffic scenes. For these, the commonly used (greedy) implementation of non-maximum suppression, which is needed in the post-processing of the analyzed deep learning methods, poses a tradeoff between recall and precision.

As the robustness of detection and pose estimation is also important in dense groups of persons, the third part of the thesis focuses on improving both tasks for such scenarios. Learning the task of non-maximum suppression with a neural network architecture incorporating the head boxes of pedestrians as further attributes to discriminate persons in groups does not improve performance. Yet, the experiments reveal issues with ambiguities in detection and attribute estimation (e.g. head box estimation) for pedestrians that highly overlap each other. To solve this ambiguity for pairwise constellations of persons a new pose estimation method is proposed that relies on pairwise detections as input and jointly estimates the two poses of such pairs in a single forward pass within a deep convolutional neural network. As the availability of automotive datasets providing poses and a fair amount of crowded scenes is limited, the EuroCity Persons dataset is extended by additional images and pose annotations, which are made publicly available as the *EuroCity Persons Dense Pose* dataset. This dataset is the largest pose dataset recorded from a moving vehicle. The experiments on this dataset with the new method show improved performance for poses of pedestrian pairs in comparison with a state of the art method for human pose estimation in crowds.

The final chapter of the thesis draws conclusions from the content of the previous chapters of the thesis and discusses the required performance for automated driving. Furthermore, it reasons about efficiency aspects regarding the collection, annotation, and usage of data for deep learning and presents potential future work regarding methodical improvements and end-to-end training of the functional chain for automated driving including the integration of multiple sensors.

SAMENVATTING

Dit proefschrift betreft de detectie van personen en de schatting van lichaamshouding op basis van video beelden. Alhoewel het relevant is voor meerdere applicaties, is de focus van dit proefschrift zelfrijdende voertuigen in stedelijke verkeerssituaties. Dit domein in het bijzonder is interessant door de uitdagingen die voortkomen uit visuele perceptie vanuit een bewegend voertuig. Verkeersstatistieken geven aan dat 'vulnerable road users' - kwetsbare verkeersdeelnemers, zoals voetgangers en fietsers - betrokken zijn bij een groot deel dodelijke verkeersongelukken. Dit motiveert onderzoek naar het reproduceren of zelfs overtreffen van het menselijke rijvermogen in assistentiesystemen en zelfrijdende voertuigen om de veiligheid te waarborgen. *Deep learning* is medeverantwoordelijk voor het dichten van het gat tussen computer visie methodes en menselijke visuele perceptie. Vooral de krachtige *features* die convolutionele neurale netwerken kunnen leren zijn behulpzaam voor de detectie van personen en schatting van lichaamshouding. Nieuwe *deep learning* methodes voor deze taken zullen in dit proefschrift gepresenteerd worden. Dit proefschrift zal niet alleen methodologische uitbreidingen introduceren, maar ook de creatie van nieuwe datasets voor het trainen, evalueren, en het ijken van perceptie systemen in het zelfrijdende voertuigen-domein.

Eerst zal een nieuwe methode voor gezamelijke detectie en houding schatting met een enkel 'diep' convolutioneel neuraal netwerk gepresenteerd worden. Houding schatting is bereikt door een bestaande convolutioneel netwerk architectuur uit te breiden met een aantal zorgvuldig ontworpen *layers* en een toepasselijke *loss* functie. Het netwerk is afhankelijk van externe voorstellen voor object kandidaat regios, waarvan de nauwkeurigheid van groot belang is voor de uiteindelijke prestatie van het netwerk. Twee voorstel methodes die gebruik maken van *3D sensor data* - *stereo* en *lidar data* om precies te zijn - worden daarom geïntroduceerd. De KITTI-dataset, die vaak gebruikt wordt voor benchmarken in het zelfrijdende voertuigen-domein, is benut voor het trainen en evalueren. De experimenten op de KITTI-dataset geven aan dat, met het combineren van voorstellen gebaseerd op beide sensormodaliteiten, een hoge *recall* bereikt kan worden, alhoewel er weinig voorstellen nodig zijn voor deze methode. Verder is de methode voor gezamelijke detectie en houding schatting competitief met andere *state of the art* methodes. Deze methode presteert beter dan de *state of the art* voor fietsers in een *test scenario*.

Big data is een belangrijke factor in het succes van *deep learning* in computer visie. Desondanks is het aantal voetgangers, fietsers, en rijders in de KITTI-dataset nogal beperkt, de bestaande literatuur suggereert dat de prestatie van object detectie methoden verbeterd zou kunnen worden door het gebruik van grotere datasets. Met betrekking tot benchmarken zijn kleine datasets daarnaast gevoelig voor *dataset bias* en *overfitting*.

Daarom introduceert het tweede deel van dit proefschrift de EuroCity Persons dataset. Deze dataset heeft een groot aantal diverse, nauwkeurige, en gedetailleerde annotaties van voetgangers, fietsers, en andere rijders in stedelijke verkeerssituaties. De afbeeldingen van deze dataset zijn verzameld vanuit een rijdend voertuig in 31 steden in 12 Europese

landen. Met meer dan 238200 handmatig geannoteerde instanties van personen in meer dan 47300 afbeeldingen is EuroCity Persons bijna één order van grootte groter dan bestaande datasets voor de detectie van personen in verkeerssituaties. De dataset telt verder een groot aantal annotaties van de oriëntatie van personen (meer dan 211200). Vier *state of the art deep learning* methodes zijn zorgvuldig geoptimaliseerd om als baselines te dienen voor de nieuwe object detectie *benchmark*. In experimenten met vorige datasets worden het generalisatievermogen van deze detectors (getraind op de nieuwe dataset) geanalyseerd. Dit proefschrift onderzoekt ook het effect van de grootte van de training set, diversiteit van de dataset (dag vs. nacht, geografische regio), het detail van de dataset (i.e. beschikbaarheid van object oriëntatie informatie), en de annotatiekwaliteit op de prestatie van de detector.

De kwalitatieve en kwantitatieve analyses van foutbronnen voor het best presterende detector model tonen methodologische tekortkomingen aan voor drukke verkeerssituaties. Voor zulke situaties veroorzaakt de meest populaire greedy implementatie voor *non-maximum suppression* - die nodig is in de post-processing stap van de bestudeerde *deep learning* methodes - een tradeoff tussen *recall* en *precision*.

Omdat de robuustheid van de detectie en schatting van lichaamshouding ook belangrijk is voor situaties met veel personen, focust het derde deel van de proefschrift zich op het verbeteren van beide taken voor zulke situaties. *Non-maximum suppression* leren met een neuraal netwerk dat 'hoofd boxes' van voetgangers gebruikt als attributen om personen in groepen te onderscheiden verbetert de prestatie niet. Wel tonen de experimenten problemen met onenigheden tussen detectie en attribuutschatting (e.g. hoofd box estimation) voor voetgangers met een grote overlap. Een nieuwe methode voor het schatten van de lichaamshouding die gebruik maakt van *pairwise detections* wordt voorgesteld om deze onenigheid op te lossen voor paren van personen. Deze methode schat de lichaamshoudingen van zulke tweetallen tegelijk in een enkele *forward pass* van een 'diep' convolutioneel neuraal netwerk. Omdat de beschikbaarheid van datasets die lichaamshouding informatie geven en een redelijk aantal drukke scenes bevatten beperkt is, is de EuroCity Persons-dataset uitgebreid met extra afbeeldingen en lichaamshouding annotaties, publiekelijk beschikbaar gesteld als de EuroCity Persons Dense Pose-dataset. Deze dataset is de grootste lichaamshouding dataset die opgenomen is vanuit een bewegend voertuig. De experimenten met de nieuwe methode op deze dataset tonen een betere prestatie aan voor de voorspelling van lichaamshoudingen van een tweetal voetgangers dan de *state of the art* voor schatting van lichaamshouding in menigten.

Het laatste hoofdstuk van het proefschrift trekt conclusies uit de voorgaande hoofdstukken en bespreekt de vereiste modelprestatie voor zelfrijdende voertuigen. Verder redeneert het over de efficiëntie van de collectie, annotatie, en gebruik van data voor *deep learning* en presenteert het mogelijk toekomstig onderzoek met betrekking tot methodologische vorderingen en *end-to-end* training van de *pipeline* voor zelfrijdende voertuigen, inclusief de integratie van meerdere sensoren.

1

INTRODUCTION

Perception and the processing of sensory input by the brain enable humans to interact with a dynamic world. In particular, vision turned out to be very effective to understand our surroundings. From an early age we effortlessly not only perceive but also recognize and interpret. Things are categorized and thus receive a semantic meaning. The visual input alone provides rich information that allows the recognition of detailed attributes of objects. Based on experience and knowledge, we can even predict to some extent the future and the impact of our actions. Thus, we can plan our behavior and react to the outcome using our senses - thus closing the interaction loop with our surrounding world.

Due to the effortless and partially sub-conscious act of interpretation and recognition, one could easily underestimate the effort needed to recreate human visual capabilities. The research area of computer vision and pattern recognition seeks to recreate such skills that are needed to build automated systems in several domains that depend on visual input. Applications range from surveillance, over visual inspection systems in the industry to automatic image and video processing, as well as driver assistance systems, and fully automated driving. This thesis focuses on computer vision in the domain of intelligent vehicles.

Even for humans, driving a vehicle is a complex task. Apart from controlling the vehicle, all the information received from the dynamically changing surroundings has to be processed at the same time. A human driver builds a mental representation of the outside world to plan the future driving maneuvers. The representation not only comprises the road layout but also other traffic participants that can frequently appear and disappear - sometimes even unexpectedly. Other participants are categorized into different classes and localized relative to the ego vehicle. The task of classification and localization of other participants will be referred to as **detection** in the remainder of this thesis. A human driver not only detects other participants but also considers many of their attributes providing additional context information to understand their potential behavior. Perceived additional cues such as their line of sight, hand gestures (see Figure 1.1), or the gait cycle are automatically processed. E.g. the instantaneous **pose** of pedestrians is already a strong predictor for the potential moving directions or the motion

state like walking or standing. The accuracy of the human recognition system even allows the detection of other agents in very challenging scenarios. In crowds, single pedestrian instances are recognized even if most parts of a person are occluded.

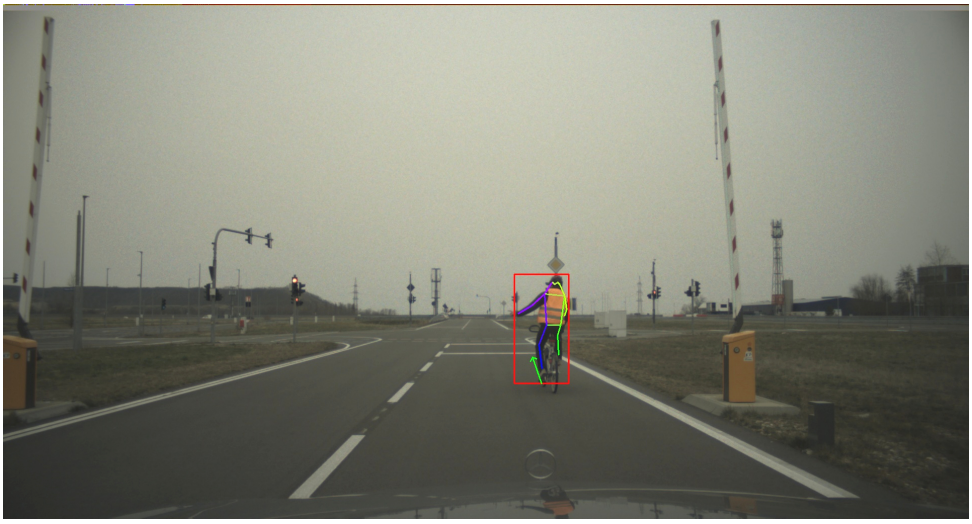


Figure 1.1: The rider's hand gesture indicates his wish to turn left. It may be used to predict his potential future moving direction. Visual perception results of a deep neural network running onboard the moving vehicle are depicted by the red bounding box regarding the detection, and by the arrow and the colored lines regarding the pose in terms of body orientation and the configuration of the joint points. The image was recorded by an onboard camera attached behind the windshield during performance tests at the Aldenhoven Testing Center.

This thesis addresses **vulnerable road users (VRUs)** [53] such as pedestrians and riders. The term **person** refers to pedestrians as well as riders in the following. Due to the high variability of the physical appearance of persons, it is difficult to handcraft a robust descriptive model for the appearance and further attributes of persons. This leads to the need for pattern recognition and machine learning techniques to replicate the recognition skills of a human driver. When comparing these computer vision methods with humans there is still a performance gap regarding the detection and pose estimation of VRUs. Recent deep learning approaches have contributed to narrow this gap [71, 178] and will be investigated and extended within this thesis for these tasks.

1.1. MOTIVATION, SCOPE, AND CHALLENGES

The current road safety situation for VRUs described in Section 1.1.1 motivates special attention to these object classes. Driver assistance systems already reduce accidents and fatalities caused by human failures [86]. A further transition to fully automated driving (more in Section 1.1.2) needs an even higher performance of perception methods and apart from object detection also relies on detailed object analysis, in particular pose estimation (see the scope of this thesis in Section 1.1.3). This is difficult to achieve especially for VRUs, as certain challenges are involved in detection and pose estimation

with images recorded from an onboard camera. These challenges are described in Section 1.1.4.

1.1.1. ROAD SAFETY AND HUMAN DRIVER WEAKNESSES

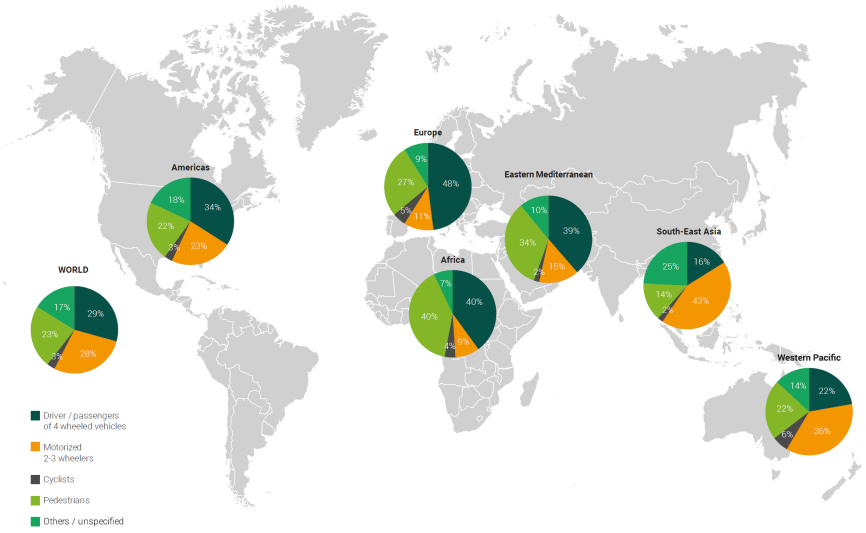


Figure 1.2: Distribution of deaths by road user type and WHO region in 2016 from [161]. The proportion varies drastically across different regions, e.g. in South-East Asia 14% of mortalities are among pedestrians in contrast to 40% in Africa.

The worldwide number of deaths in road traffic has increased to 1.35 million in 2016 according to the Global Status Report on Road Safety of the World Health Organization (WHO) [161]. The death rate, which is the number of deaths relative to the population size per year, remains constant despite the increasing motorization worldwide. Still, injuries caused by road traffic accidents are the leading cause of death for young people between the age of five and 29 [161]. The risk to die in road traffic varies drastically between countries and shows a correlation with the country income level. The death rate ranges from 8.3 per 100,000 citizens per year in high-income to 27.5 in low-income countries. There are also inequalities for whole regions. While Africa has the highest death rate of 26.6, the death rates in America and Europe have decreased between 2013 and 2016 resulting in 18 and 16.9 respectively.

There is also a variation in road users most affected that corresponds with the variations in death rates [161] (see Figure 1.2). VRUs are the most affected ones globally. Pedestrians and cyclists comprise 26%, while motorized two- and three-wheelers represent another 28%. In Africa, 44% of road traffic mortalities are among pedestrians and cyclists, which is the highest proportion of this group. In South-East Asia and the Western Pacific, the highest proportion of the mortalities is among riders of motorized two and three-wheelers, with 43% and 36% respectively. The high risk for VRUs motivates special attention for this group of traffic participants.

According to the WHO Status Report [161] the following legislative measures and traffic participant behaviors are important to improve road safety:

- Managing speed
- Reducing drunk driving
- Increase seat belt use
- Increase use of child restraint
- Build safer roads (e.g. regarding road layout)
- Use of safer vehicles (e.g. providing electronic stability or anti-lock braking system)

The progress in implementing these points varies between different countries worldwide. Such measures and changes in behavior are important, especially those affecting the human driver, who remains to be a major cause of accidents as shown by different studies [143, 152]. E.g. in 94% of accidents in the US between 2005 and 2007, the critical reason has been assigned to the driver [143].

Despite the capabilities in recognition, a human driver is less suitable for the following tasks according to [89]:

- Routine tasks
- Simple but time-critical tasks
- Vision at night and in adverse weather conditions
- Estimation of distance and speed differences
- Maintaining a safe and appropriate distance from other road users

Failing these tasks may result in critical mistakes during driving. Therefore, e.g. [89] proposes to use driver assistance systems for such tasks. These may contribute to road safety also for VRUs and may prevent traffic accidents.

1.1.2. FROM DRIVER ASSISTANCE SYSTEMS TO FULLY AUTOMATED DRIVING

Driver assistance systems are already available to improve road safety. The progress in pedestrian detection enabled the market introduction of active safety systems like the PRE-SAFE® brake [16], which is able to brake automatically in dangerous traffic situations. In autumn 2021, Mercedes-Benz has released the first level 3 S-Class model [113] according to the levels of automated driving as defined by the SAE [79] (see Figure 1.3). This conditional automation is a further step toward fully automated driving (level 5) and hits a major milestone, as the system takes over the responsibility of monitoring the driving environment. It is still restricted to traveling speeds below 60 km/h and highway driving in Germany. Thus, the recognition of VRUs plays a subordinate role, and the complexity of the road layout is lower compared to urban regions. For deployment of fully automated driving in urban regions the detection and analysis of VRUs takes a fundamental role. Hereby, the accuracy constraints are high. Pedestrian detection

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Copyright © 2014 SAE International. The summary table may be freely copied and distributed provided SAE International and J3016 are acknowledged as the source and must be reproduced AS-IS.

Figure 1.3: Levels of automated driving according to the SAE [79].

systems are bound to a tradeoff between generating false alarms and missing pedestrians. For a level 2 driver assistance system, a correct detection performance of e.g. 90% may be still acceptable, as long as the false alarm rate is essentially zero because there is a driver, who is responsible overall. With the advent of fully automated vehicles, performance needs to be significantly upped, as a driver is no longer necessarily available to intervene. No relevant VRU should be overseen while false alarms and resulting erroneous braking maneuvers still have to be avoided for comfort and safety reasons.

A fully automated vehicle has to be reliable also in challenging scenarios typical for urban areas, e.g. in crowded scenes with dense groups of pedestrians. Imagine a group of pedestrians waiting at a bus stop as shown in Figure 1.4. As each member in such a group can suddenly step out and enter the street, a reliable detection and analysis is equally important for all the pedestrians in the group, even for the ones further in the back, potentially occluded by other pedestrians.

Figure 1.4: A group of pedestrians waiting at a bus stop. As each of the pedestrians could start moving towards the street, reliable detection is equally important for all pedestrians.

Furthermore, a fully automated vehicle needs to predict the movement of surrounding VRUs and cars far in advance, in order to be able to brake and/or employ evasive maneuvers in time. Due to the high maneuverability of VRUs, any auxiliary context information that can reduce the uncertainty of movement prediction should be utilized. Using pose information of pedestrians for example can help increase the prediction horizon up to one second without increasing the false alarm rate [90, 91] (see Figure 1.5). As shown in Figure 1.1 pose information is also relevant for gesture recognition. In [135], estimating pose is even stated as a general requirement for the design of onboard pedestrian detection systems.

Benefits of Fully Automated Driving. Apart from increased road safety and comfort, fully automated driving will also have environmental impacts. There is an increasing

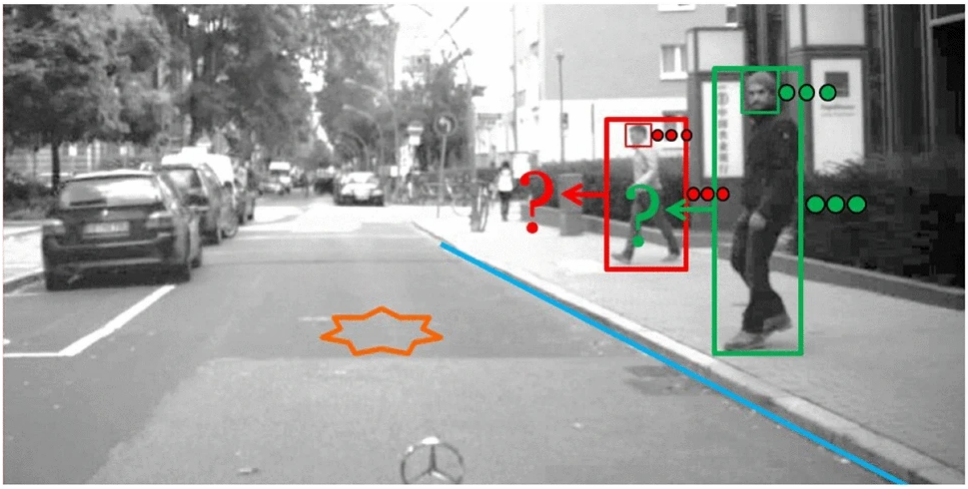


Figure 1.5: Path prediction of the pedestrian in [90] is based on context features, which can help increase the prediction horizon up to one second. The pedestrian's awareness is one of the context features and is based on the pedestrian's pose regarding the head orientation. It indicates if the pedestrian will stop at the curbside or cross the street. Figure from [90].

number of studies on that topic, see [140] for a survey. Cooperative driving of automated vehicles [159] could improve traffic flow, and thus reduce traffic congestions and emissions. Shared mobility services such as fully automated taxis could also reduce carbon emissions. Without the cost of a human driver, such taxis could be affordable enough to decrease the demand for private cars. A study by [49] estimates that one automated vehicle could replace around 11 conventional vehicles. Reducing the number of needed cars saves resources and spares our environment. Fully automated taxis could also solve the first-/last-mile problem [27] by providing accessibility to nearby transit stations (e.g. trains). Thus, usage of public transport could be encouraged. A case study by [54] estimates, that an electrified fleet of autonomous taxis could reduce greenhouse gas emissions by 60% in comparison with conventional vehicles. Hence, fully automated driving could play an important role in fighting man-made climate change. Reducing emissions by shared mobility could also improve air quality, which poses several health issues. According to the WHO [162], air pollution caused 7 million deaths globally in 2016 alone. The positive effects depend on usage of car sharing in conjunction with keeping the mobility level at the current level. In fact, automated vehicles might easily lead to an increased mobility demand, e.g. caused by people living further away from work encouraged by the availability of automated vehicles, which enable working during the commute. See [97] for a survey also listing negative effects of automated vehicles, e.g. investment which could be shifted away from public transport systems. Overall, the benefits towards fighting climate change depend on its utilization by society and the management of its introduction [140].

1.1.3. SCOPE OF THE THESIS

The scope of this thesis is the detection and pose estimation of VRUs with deep learning methods using monocular images. The targeted application domain is automated driving in urban traffic scenes. Hence, the camera sensors recording the images of the datasets used within this thesis are attached to moving platforms e.g. behind the windshield or on the vehicle roof. This causes specific challenges as described in the next section. Other sensors such as lidar and radar are out of the scope of this thesis.

Monocular detection and pose estimation usually is one part of a complex system for automated driving. In addition to the methods utilizing lidar or radar data (see for example [118]), it provides detection results as input for the functional chain. Subsequent modules take care of fusing the data of the different sensors and tracking of the detected objects. Following this, motion prediction as in [90, 91], which may also rely on additional context cues like the pose, takes care of predicting the future path of other traffic participants like VRUs. Situation analysis and motion planning depend on these predictions to plan a safe trajectory for the vehicle, which is then executed by the vehicle control module. Temporal information of course is essential for the functional chain, but out-of-scope of this thesis. A lot of valuable information e.g. for motion prediction can already be extracted from static, singular images by means of detection and pose estimation.

A specific focus of this thesis are person groups (see Chapter 5), which require special attention in urban areas as explained in the section before. The challenge for person detection and pose estimation is not a high overall number of persons in an image by itself, but rather a high person density. If the local density and hereby the visual overlap in the image is high, even two overlapping pedestrians may cause difficulties, due to mutual occlusions and methodical shortcomings explained in this thesis. To emphasize the fact that it is not only about the number of people but mainly the density, the term *dense traffic scenes* is used throughout the following work in addition to the more common terms *crowds* and *person groups*.

Regarding pose estimation this thesis considers two different representations for the pose of a person (see Figure 1.6). First, looking at generic objects in the computer vision domain, pose estimation refers to the task of estimating three orientation angles in Euclidean space (frequently in addition to the position). The focus of this thesis only lies on the estimation of the single yaw angle (see Chapter 3), as the other two angles can be assumed to be close to zero for upright objects like pedestrians [55]. The yaw angle is referred to as body orientation throughout this work. It may be used as a surrogate of the pose to determine the potential direction of movement and has the potential to support the initialization of tracks in tracking-by-detection approaches.

Second, human pose estimation often refers to the task of estimating the configuration of the human body regarding the position of certain keypoints, also called joint points. These joints comprise anatomical joints of a person such as ankles, knees, elbows, and shoulders, but often also the positions of the eyes and ears [19]. The localized joints can be used as an intermediate representation for gesture recognition and intention estimation [92]. Thus, cues such as the line of sight, hand gestures, or the gait cycle that are automatically perceived and processed by a human driver, can also be recognized.

In this thesis, new methods of both representation domains are presented. Figure 1.6 shows exemplary annotations for the body orientation as well as the joint point

annotations consisting of 17 joint points for the *EuroCity Persons Dense Pose* dataset presented in this work.

1.1.4. CHALLENGES OF PERSON DETECTION AND POSE ESTIMATION

Despite two decades of steady progress, person detection is still an open research problem. It often features as a canonical task to assess the performance of generic object detectors. Challenges particular to person detection and analysis from a moving vehicle with an onboard camera are described in this section. Figures 1.7 and 1.8 show samples for the challenges listed in the following.

- **High intra-class variance.** There is a wide variation in person appearance. Persons are non-rigid objects and their poses may be very articulated (see Figure 1.7a). Due to the non-rigid nature, the potentially cluttered background influences the overall appearance, in particular, if the person's location is represented by an enclosing bounding box. Sitting or even lying persons have a very different aspect ratio compared to standing persons. The clothing also varies a lot, not only due to different seasons, weather, and time of day but also due to personal style (see Figure 1.7b). Therefore, the intra-class variance is high and the difference in appearance between two persons (e.g. regarding pixel-wise intensities) may be even higher than between a person and other similar, out-of-class objects like shop-window mannequins. Such similar objects may sometimes only be discriminated from real persons based on the context information (see Figure 1.7c).

- **Occlusion, truncation.** Often, persons in urban areas are not fully visible. Obstacles like parking cars and vegetation cause occlusions. At the image border, persons are truncated. A special challenge is the occlusion in person groups with heavy mutual occlusions (see Figure 1.7d). These hinder the discrimination of single instances and may result in ambiguities, e.g. which body part belongs to which person if only parts of the limbs are visible. For a subsequent tracking approach, erroneous detections in groups result in wrong associations and erroneous velocity vectors. Such "ghost" velocities do not correspond to real movements, and may negatively influence the motion planning of the vehicle.

- **Low resolution.** The real-time requirements and the available computing power within a vehicle also limit the image resolution that may be processed. The pixel size of a person is inversely proportional to the distance to the camera. Doubling the distance means halving the pixel size for a fixed person size in meters. Thus, in particular, children who are far away appear very small within an image and are difficult to detect (see Figure 1.7e). Still, early detection of far-away objects is desirable to enable early planning and reaction.

- **Motion blur and other perturbations.** Due to the non-instantaneous exposure of the image sensor and a rolling shutter that is commonly used in automotive cameras, the recording of dynamic objects, and recording from a moving vehicle, objects may appear blurred. The higher the exposure time due to lower illumination (night, dawn, cloudy weather) the higher the motion blur (see Figure 1.8a). Dust,

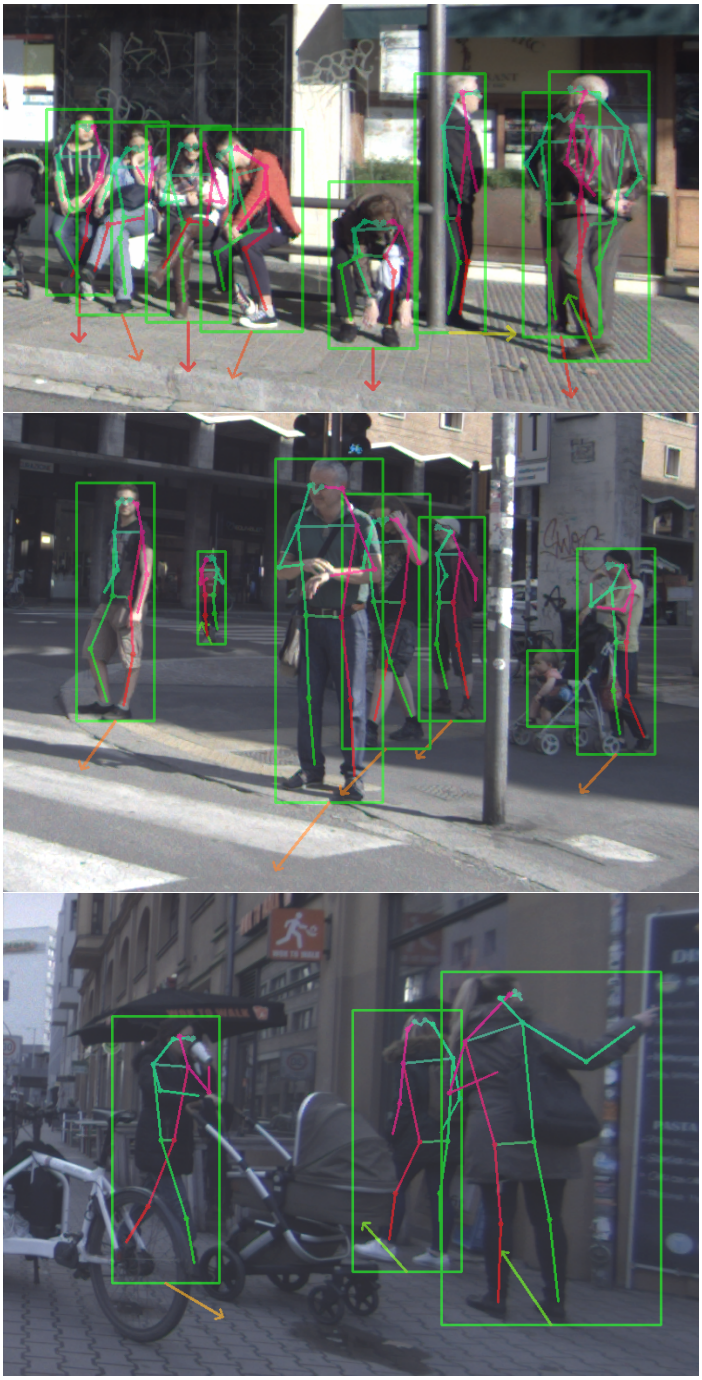


Figure 1.6: Examples for bounding box and pose annotations in terms of 17 joint points and the body orientation of the presented EuroCity Persons Dense Pose dataset.

dirt, or rain droplets on the windshield also reduce the image quality (see Figure 1.8b).

- **Illumination.** The illumination of the scene may differ a lot due to a different time of day, weather, season, or shadows caused by buildings and other obstacles. A low illumination causes low contrast, while a low standing sun blinds the camera if it is in or close to the line of sight (see Figure 1.8c). Rapid changes in the illumination, e.g. when entering or leaving a tunnel, also causes challenges as the regulation of the exposure time of the camera needs time to adjust. During this regulation, the recording is too bright or too dark. This also introduces a further variance in the appearance of persons.

Performance Gap. Deep learning had a large success in image classification (e.g. AlexNet [94]), which is the task of assigning an object category to a full image. It has been boosted by the availability of appropriate GPU hardware and big datasets such as ImageNet [35]. In contrast to classic machine learning approaches, the networks learn a feature representation from raw image pixels, instead of manual design and selection of appropriate features, e.g. edges. These powerful features extracted from raw pixels can be effectively used for other tasks besides image classification and lead to the successful incorporation of deep learning in the context of object detection [58, 59, 132]. This also turned out to improve the detection of VRUs, due to the notably high variation in appearance.

Despite the progress in detection with deep learning methods, a recent paper [178] argues that current pedestrian detection performance lags that of an attentive human by an order of magnitude, which also results from the challenges mentioned before. How can this performance gap be closed?

Datasets play a crucial role in today's computer vision research [35]. Corresponding benchmarks reveal strengths and weaknesses of existing approaches and are instrumental in guiding research forward. Still, [146] argues that even larger datasets are needed. Experiments on a 300 million images dataset show that the classification performance further increases logarithmically with the size of the training dataset. More data could prove useful for object detection as well [179].

1.2. OUTLINE AND CONTRIBUTIONS

The goal of this thesis is to enhance image-based person detection and pose estimation performance with deep learning methods. First, a new detection and orientation estimation method is presented and evaluated on the KITTI dataset, which is commonly used for benchmarking in automotive scenarios (Chapter 3). Second, as the size of the KITTI dataset is limited and deep learning methods could profit from bigger datasets, a new dataset for detection and orientation estimation is presented in conjunction with a thorough analysis of various deep learning methods (Chapter 4). Third, as crowded scenes turn out to be a major issue for detection performance, a further focus is put on optimizing the detection and pose estimation for groups of persons (Chapter 5). New datasets presented in this thesis are made publicly available for benchmarking and to stimulate further research.

(a) Variation in poses (sitting)

(b) Varying clothing

(c) Similarity with other objects

(d) Occlusion, truncation, groups

(e) Low resolution

Figure 1.7: Examples for different challenges in person detection and analysis.

(a) Motion blur

(b) Rain droplets on the windshield

(c) Low sun

Figure 1.8: Examples for different challenges in person detection and analysis.

In addition to the methodical Chapters 3, 4, and 5, the thesis presents previous work regarding person detection, pose estimation, and benchmarking in Chapter 2 and draws conclusions and discusses future work in Chapter 6.

1.2.1. JOINT DETECTION AND ORIENTATION ESTIMATION WITH 3D OBJECT PROPOSALS

Classical object detection approaches consist of a proposal and a classification stage [41]. Chapter 3 builds upon the R-CNN architecture [58], which profits from the success of deep convolutional neural networks for classification in the second stage. Still, detection performance is limited by the recall of the input proposals. Here, these proposals are optimized using stereo and lidar data. Regarding stereo, the chapter makes use of the "Stixel World" [121], which is a medium-level representation of stereo data that groups pixels of similar depth into vertical sticks of fixed width. Furthermore, the network architecture is extended by orientation estimation using a von Mises loss function, which is combined with a Biternion representation as in [9]. Results are presented on the frequently used KITTI benchmark. Chapter 3 is based on the work published in [15] (©2016 IEEE).

This chapter's main contributions are:

- First, the chapter proposes a novel deep CNN architecture called Pose-RCNN for joint object detection and orientation estimation based on the well-known R-CNN method [58]. It differs from other methods by modeling orientation regression with a carefully designed von Mises loss function based on a Biternion representation, while e.g. [24] applies a simple L1 regression. Whereas detection and orientation estimation was treated separately in most other works e.g. [52], the chapter presents a joint method for detection and orientation estimation by using one single CNN architecture.

- Second, two 3D proposal methods based on lidar and stixel information are presented. Compared to [24], a new proposal method based on stixel data is introduced in contrast to using raw pointcloud data. Combining lidar and stixel proposals improves the recall performance.

The chapter shows that the new Pose-RCNN architecture using the new proposal method achieves competitive results with state of the art approaches on the KITTI benchmark.

1.2.2. A NOVEL BENCHMARK FOR PERSON DETECTION IN TRAFFIC SCENES

The publicly available KITTI dataset used for benchmarking in Chapter 3 has certain limitations e.g. regarding dataset size and diversity of the recording locations. The small amount of person samples makes it prone to dataset bias and overfitting. Chapter 4 presents *EuroCity Persons* (ECP), a newly created highly diverse dataset for vision-based person detection collected onboard a moving vehicle in 31 cities of 12 European countries. A thorough performance evaluation and analysis is performed for several recent deep learning methods. In comparison to the R-CNN method built upon in Chapter 3 all of these methods integrated the proposal generation within the neural network itself.

The von Mises loss function used for the Pose-RCNN method in Chapter 3 is integrated into the best performing deep learning method Faster R-CNN [132] for evaluation of orientation estimation on the new dataset. Thus, despite the strong focus on detection performance, orientation estimation is also part of Chapter 4. It is based on the work published in [13] (©2019 IEEE).

The contributions are threefold:

- The EuroCity Persons dataset is introduced, which provides a large number of highly diverse, accurate, and detailed manual 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 4.2. It is made available for public benchmarking via website [a].
- Four deep learning approaches (Faster R-CNN [132], R-FCN [31], SSD [109] and YOLOv3 [128]) are optimized to serve as baselines for the new person detection benchmark. The chapter proves the generalization capabilities of detectors trained with the new dataset and thereby its usefulness. See Sections 4.3.1 and 4.3.2.
- Insights are gained and provided regarding the effect of several dataset characteristics on detector performance: the training set size, the dataset bias (day- vs. night-time, geographical region), the dataset detail (i.e. availability of object orientation information), and the annotation quality. Chapter 4 analyzes error sources and discusses the road ahead. See Sections 4.3.3 and 4.4.

1.2.3. DETECTION AND POSE ESTIMATION IN DENSE TRAFFIC SCENES

The approaches analyzed in Chapter 4 depend on a post-processing step to filter multiple detections per object. This step is called non-maximum suppression (NMS). Its goal is to reduce the initial detection set to end up with exactly one detection per object. Chapter 4 identifies this NMS as a major issue regarding detection performance as its configuration poses a trade-off between recall and precision. This hampers detection in dense traffic scenes and results in missed detections if the mutual overlap between persons is too high. Preceding work on detection in dense traffic scenes which lead to insights and methodical contributions of Chapter 5 is presented in Section 5.2. This section investigates the potential of further discriminative attributes, such as the head position, to aid the detection in groups of persons. A recent approach of [73] is adapted to learn the task of NMS from data using the head position as additional input. Section 5.2 shows issues with ambiguity in the estimation of discriminative attributes. Ambiguity is also an issue in object detection in general and pose estimation as well. When two persons are too close to each other, the detection result for a single proposal box in between these persons is influenced by features of both persons and during inference, it may be ambiguous which person and which pose or attribute should be the target for that proposal. Section 5.3 proposes a detection and pose estimation approach that tackles this ambiguity by jointly handling pairs of pedestrians. As the EuroCity Persons dataset presented in Chapter 4 is not annotated with joint points of persons, an extended pose

[a] https://eurocity-dataset.tudelft.nl

dataset based on the ECP dataset is presented. Chapter 5 is based on the work published in [14] (©2020 IEEE) and [12] (©2021 IEEE).

The contributions are twofold:

- First, the chapter proposes a new top-down pose estimation method to jointly estimate poses of pedestrian pairs in a single network. It relies on paired detections that improve the recall in groups. The new method is simple to integrate into existing network architectures for human pose estimation, yet effective and does not depend on a separate input hint or a post-processing stage for disambiguation of poses of pedestrian pairs. As the pairwise detection approach is similar to the concurrent work of [28], which was published just before [12], only the pairwise pose estimation and the combination of the two approaches is claimed as a contribution.
- Second, a new automotive dataset for pose estimation is created extending the original ECP dataset with additional images from the front-facing and two side-facing cameras. The detailed annotations including bounding boxes and poses of pedestrians and riders will be made available for public benchmarking.

2

Previous Work

This chapter presents the previous work on visual bounding box based person detection and pose estimation including relevant datasets and benchmarks with a focus on the intelligent vehicles domain. Work on other sensor modalities like radar and lidar is out of the scope of this thesis (see for example [118]).

2.1. Detection

Classic object detection based on hand-crafted features usually consists of two stages. Proposal boxes (also abbreviated as proposals in the following) that serve as candidate regions of interest (ROIs) are generated in the first stage. For example, the sliding window approach arranges boxes of different aspect ratios and sizes on a regular grid across the image [32]. More sophisticated methods (see [72] for an overview) reduce the needed amount of proposals and hereby the needed computation time by generating proposals that are more likely to contain objects. Thus, they enable the use of more complex methods in the second stage, which is the classification of proposal boxes. Classic object detection methods build upon hand-crafted features for classification. These have to be selected and optimized according to the task at hand. In the pedestrian detection domain, Deformable Part Models (DPM) using Histograms of Oriented Gradients (HOG) features [51, 116, 168], and Decision Forests using Integral Channel Features (ICF) [5, 37, 177, 180] were the established methods until a few years ago [6].

Successes of deep learning for image classification (e.g. AlexNet [94]) also lead to its incorporation in object detection. By training deep convolutional neural networks (CNN) like GoogleNet [148], VGG [141] and ResNet [68] on the ImageNet dataset [35] for classification, models learn to extract powerful features from raw pixels, which can be used effectively for other tasks like object detection [76].

A comparison of selected detection methods building upon feature maps of CNNs is shown in Table 2.1. They can be clustered into two-stage methods [31, 58, 132] that use a proposal stage and a downstream classification stage (like the classic detection methods), and one stage methods that go without the proposal stage [109, 127, 128]. The Regions

with CNN features (R-CNN) methods [58, 59, 132] are the basis for most current two-stage methods. Just like the classic detection methods, R-CNN [59] and its extension Fast R-CNN [58] 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 [72] for different color image based proposal methods like selective search [155], MCG [4], BING [26], Objectness [2], and Constrained Parameteric Min-Cut (CPMC) [21]. Recent works [24, 44, 63, 102, 104] showed performance improvements by taking advantage of 3D sensors including RGB-D camera and stereo camera for proposal generation. The works of [44] and [102] generate proposals based on the stixel world [121], which is a condensed stereo representation. The method of [24] improves the detection performance of Fast R-CNN with an energy formulation for proposal generation based on raw pointcloud data increasing the proposal recall.

Faster R-CNN [132] works 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 features are pooled only 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 [17, 100, 170, 184]. In SDP [170] features are pooled from different layers in dependence on the proposal size. MS-CNN [17] 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 minibatch during training consists of a sampled subset, which is usually several orders of magnitude smaller than the total amount of proposals. [58] and [51] 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 [139] is applicable.

One stage detection methods like You only look once (YOLO) [127], its extensions YOLOv2 [126], YOLOv3 [128], and others like the Single Shot MultiBox Detector (SSD) [109], or [130] 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 [126] 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 [128] 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 specialized for fast object detection, combining ideas of other CNNs [68, 141, 148] in particular the usage of residual blocks.

SSD [109] detects objects based on *default boxes*. These default boxes are similar to anchor boxes, but they are applied to 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 explicitly applied to select the boxes with the highest confidence loss similar to R-FCN. [130] introduces a recurrent neural network based on a VGG-16 architecture that improves the localization accuracy of one-stage methods. This is achieved by applying a recurrent rolling convolution on several feature layers.

Generative adversarial networks (GAN) [61] are also used for pedestrian detection. In [99] 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. Inspired by GANs, [75] trains a discriminator 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.

Table 2.1: Overview of recent deep learning detection methods. Methods evaluated in Chapter 4 are bold-faced. YOLOv3 and SSD use different feature maps for proposals of different scales.

two stage method	Fast R-CNN[58]	**Faster R-CNN**[132]	**R-FCN**[31]
region proposals	external	RPN	RPN
hard example mining	heuristic	heuristic	explicit
used feature maps	last	last	last

one stage method	YOLO [127]	**YOLOv3** [128]	**SSD**[109]
region proposals	gridbased	anchor boxes	default boxes
hard example mining	none	none	explicit
used feature maps	last	several	several

2.1.1. DETECTION IN DENSE TRAFFIC SCENES

The aforementioned detection methods like [58, 109, 128, 132] depend on non-maximum suppression (NMS) in a post-processing step to suppress multiple detections per object. Interestingly, many of the top-performing methods apply a simple greedy implementation based on a single intersection over union (IoU) threshold [10]. Selecting this threshold for suppression poses a tradeoff between recall and precision.

There are several approaches to improve the recall in particular in crowd situations, without losing precision. In Soft-NMS [10] detections are not discarded, but their class score is reduced if they overlap with another detection that has a higher confidence. The authors of [73] propose a network architecture to learn the NMS task using bounding box locations and class scores as input. Thus, the NMS could be trained in a fully end-to-end detection framework. In [108] a density value is estimated per prediction that is used instead of the single IoU threshold within the greedy NMS. A high density value leads to less suppression and a higher recall in groups. [173] builds upon this idea and additionally estimates a diversity value. This discriminative diversity value is estimated in an embedded feature space and is fed into the adapted NMS algorithm. Similarly, [169] estimates a discriminative feature in a geometric embedding. In [158], a special loss coined Repulsion Loss is used, to push detections of separate instances away from each other to lower the IoU between such detections. [28] tries to detect all objects in a group based on a single proposal. These set detections do not suppress each other within the NMS.

Other works estimate the head box in addition to the body box to support the NMS [175], or do not depend on anchor boxes/proposals as they directly estimate certain keypoints of the full bounding box like the corners [96] or additionally the center [39]. For keypoint localization, a convolutional heatmap approach is used, which is similar to scene segmentation [138] - the task of assigning a class to each pixel. Instance segmentation approaches like [33] that assign an instance to each pixel in addition to the class information, could also be used for detection in dense scenes. Still, instance segmentation is out of the scope of this thesis, as the main interest lies in the estimation of the full extent of the body bounding box, even if most parts are occluded. Instance segmentation focuses on the visible parts.

2.2. POSE ESTIMATION

First, Section 2.2.1 lists previous work for orientation estimation as surrogate of the pose of a person. The body orientation may be used to predict the probable direction of movement, while the head orientation gives information about a person's attention and awareness. Section 2.2.2 is about estimating joint points of persons with a focus on multi person pose estimation. It may be used as an intermediate representation for gesture recognition and intention estimation [92].

2.2.1. ORIENTATION ESTIMATION

The focus here is on work that estimates a single orientation angle. In particular for pedestrians it is reasonable to only estimate the body orientation/yaw angle, as the other two angles can be assumed to be close to zero [55] due to the mostly upright postures.

Early work uses hand-crafted features like HOG [7, 8, 23] for orientation estimation with SVMs [23] or Decision Trees/Ferns [7, 8]. These works model orientation estimation as a discrete classification problem with a varying number of orientation bins. The work of [42] also applies this discrete formulation but the trained classifiers then are used to infer a continuous orientation angle. This is also done in [52] to achieve the transition from the discrete to the continuous domain. Notably, [52] also jointly tracks the head and

body orientation over time with a Dynamic Bayesian Network.

Recently, deep neural networks are also used for estimating orientations of common objects in traffic scenarios [102, 145, 154]. In these works, orientation estimation is considered a multiclass classification problem. Beyer et al. [9] show that a correctly performed regression is a more natural way to address the problem of orientation estimation. They introduce Biternion Net which is capable of regressing fine-grained orientation angles using the von Mises loss. In [24] an L1 loss is used for estimating continuous orientation angles instead of estimating biternions in conjunction with a von Mises loss. They combine orientation estimation with detection in a jointly trained deep convolutional neural network.

2.2.2. MULTI PERSON POSE ESTIMATION

Multi person pose estimation for the estimation of joint points can be clustered into bottom-up and top-down approaches. Bottom-up approaches [19, 78, 82, 119, 122] first try to find all joints within an image, which are then clustered into instances. Early approaches solve the clustering by integer linear programming [78, 122]. In [19] part affinity heatmaps are estimated in addition to the joint heatmaps. The part affinities are used as edge weights in the graph-based clustering. In [119] pixelwise offset values are calculated pointing from one joint to another. These offsets are used for grouping. [82] proposes a graph convolutional network for clustering. Thus, clustering can be learned as part of an end-to-end framework. As stated in [101] and [124], invisible joints and the small local context used for joint estimation lead to inferior performance of bottom-up methods.

Top-down approaches first detect all persons within an image and then estimate the pose for every instance. Most works follow the heatmap-based approach of [150]. Mask R-CNN [67] learns both stages in a single end-to-end trainable network. Recent top-down methods profit from better person detectors or better network architectures [115, 147, 166]. Still, dense person group situations remain challenging for top-down methods. On the one hand, estimating the positions of occluded joints is difficult. On the other hand, image crops of detected persons contain parts of other persons as well. In some cases, the overlapping region between two persons is so high that the target pose is ambiguous. [60] proposes a solution for the handling of occluded joints training separate heatmap estimators for occluded and visible joints. Thus they train different experts for different occlusion states but not for disambiguation of multiple persons within a crop. [70] tries to solve the ambiguity for multiple persons by adding the position of a visible joint point for each person as an additional input. They depend on the results of a state of the art bottom-up pose estimation approach for these input hints. In AlphaPose+ [101] detections are handled independently within the single person pose estimation. A so-called joint candidate loss allows the estimation of all joints that are within an image crop. The disambiguation of poses of different persons is part of a post-processing stage. There, joint candidates from all heatmap estimations are extracted. In a global graph-based optimization procedure they can be reassigned to different detections based on the heatmap scores. As it is a fixed algorithm it can not be trained end-to-end within the framework. [124] depends on the initial pose results of AlphaPose+ [101], which are

refined by a graph convolutional network (GNN) depending on image features extracted from the base network of AlphaPose+. They also propose a variant of this GNN, where poses of pedestrian pairs are jointly refined.

2.3. DATASETS AND BENCHMARKING

Methodical progress in the computer vision domain is monitored on appropriate datasets that serve for benchmarking. Improved sensor technology, e.g. regarding the camera resolution, the need for more data in particular with deep learning methods, or other limitations of existing datasets may raise demands for the creation of new datasets. Previous work regarding datasets and performance analysis for person detection is presented in Section 2.3.1 and 2.3.2, while datasets specific for pose estimation are presented in Section 2.3.3.

2.3.1. PERSON DETECTION

A number of early datasets focus on pedestrian classification (e.g. Daimler-CB [114], CVC [57], and NICTA [117]) and detection (e.g. Daimler-DB [43], INRIA [32], ETH [46], and TUD-Brussels [160]). See [43] for an overview. Currently, KITTI [56] and Caltech [38] are the established pedestrian detection benchmarks. The latter has been extended by [178] with sanitised annotations. The Tsinghua-Daimler Cyclist (TDC) dataset [102] focuses on cyclists and other riders. In [77] a multi-spectral dataset for pedestrian detection is introduced, combining RGB and infrared modalities.

The Cityscapes dataset [29] was recorded in 50 cities during three seasons. Similar to earlier scene labeling challenges like Pascal VOC [48] and Microsoft COCO [107], it provides pixel-wise segmentations for a number of semantic object classes. The CityPersons dataset [179] 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 4.1 in the corresponding benchmarking Chapter 4 for an overview of the main person detection benchmarks in vehicle context. In terms of the annotation quantity and data diversity, CityPersons [179] and Tsinghua-Daimler Cyclist (TDC) [102] had, so far, the most to offer for the pedestrian and the riders class. Although Caltech [38] lists a large number of pedestrian annotations, only an unspecified subset of these annotations was done manually, the remainder was obtained by interpolation (the number of manual annotations is probably an order of magnitude smaller). In total there are about 2300 unique persons in this dataset. Training and evaluation on Caltech is typically performed on 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 from Strasbourg and Zurich, it covers only Germany, and recordings were not made throughout all seasons. Very recently, the Berkeley Deep Drive dataset (BDD) [171] was made available, which in total provides 100000 images recorded in a vehicle context. A white paper describing the dataset was announced.

Other person datasets relate to attribute recognition [11, 64, 137]. Notable for its sheer size is furthermore the recent Open Images V4 dataset [95], containing 15.4M bounding boxes on 1.9M images for 600 different categories.

2.3.2. PERFORMANCE ANALYSIS OF PERSON DETECTION

In [38], 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 [6] 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 [71] also deep models on the Caltech dataset are analyzed. False positives which are touching ground-truth samples are considered localization errors. 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 [178]. The authors focus on the boosted decision forests-based methods RotatedFilters [181] and Checkerboards [180]. 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 for qualitative failure reasons. In contrast, [125] applies an automatic failure analysis for ACF [36] 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 [17, 100, 176] building upon the work of Fast/Faster R-CNN are the top-performing methods on the Caltech dataset [178]. [176] 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 [17, 130, 165, 170, 184]. Apart from [130] all of these are two-stage methods building upon the work of Fast/Faster R-CNN.

[74] evaluates R-FCN, SSD, and Faster R-CNN on the generic object detection benchmark MSCOCO [107]. By varying the feature extractor, the image resolution, and other parameters various speed/accuracy trade-offs are examined.[a]

2.3.3. POSE ESTIMATION

The KITTI dataset [56] and the Tsinghua-Daimler Cyclist (TDC) dataset [102] provide a yaw angle describing the body orientation as pose surrogate. KITTI has been widely used for benchmarking orientation estimation.

Progress in deep learning based multi person pose estimation regarding joint point positions has been driven by datasets like MSCOCO [107], MPII [3] and AI Challenger

[a] This section refers to previous work in performance analysis and benchmark results before the publication of the EuroCity Persons detection and orientation benchmark [13] presented in Chapter 4. The ongoing release of even newer and better methods can be observed on the corresponding benchmark websites.

[163]. The CrowdPose [101] and OCPose [124] datasets focus on crowd situations with a high amount of pedestrians overlapping each other. These situations constitute a specific challenge for pose estimation. The images of the mentioned datasets have been collected using online search engines, Flickr and YouTube. In terms of automotive datasets, PedX[87] provides stereo images recorded from a moving vehicle including lidar annotated with 2D and 3D poses. Still, the diversity regarding context is rather low as only three urban intersections are covered. Recently, the TDUP dataset [157] has been announced, which will provide images recorded from a moving vehicle covering diverse urban traffic scenes in China. An overview of these datasets is shown in Table 5.1 within the corresponding Chapter 5.

3

Joint Detection and Orientation Estimation with 3D Object Proposals

3.1. Overview

As described in Chapter 1 detection and pose estimation of vulnerable road users are essential components for building a fully automated driving system. This chapter not only focuses on detection but also on the estimation of the body orientation of pedestrians and riders as surrogate of the pose.

Today, deep learning methods are able to capture complex context information by using powerful, multi-layer visual representations. The visual representations are extracted from a set of object proposals estimated by preceding proposal methods. The recall rate of the proposal methods is crucial because it specifies an upper bound for the overall detection performance. The detection performance of the Regions with CNN features (R-CNN) method on the KITTI benchmark [56] is limited due to the low recall performance of its proposal method, such as Selective Search [155]. Recent work [24] showed that the detection performance can be greatly improved by using more powerful 3D proposals from pointcloud data.

This chapter introduces Pose-RCNN, a combined approach for object detection and pose estimation based on a single R-CNN-like neural network. Pose estimation is carried out through an orientation regression network attached to an R-CNN architecture. The regression net is trained by using a carefully designed von Mises loss function [111] combined with a Biternion representation [9] of the orientation. Inspired by the good results achieved in [24], two different 3D proposal methods are presented: One originates from the stixel world [121], the other uses lidar point clouds. The proposed Pose-RCNN is evaluated on the KITTI dataset [56]. The results are competitive in both detection and orientation estimation. Both introduced proposal methods achieve similar recall performance as the state of the art and significantly outperform methods that only make

use of 2D image data. Figure 3.1 shows an example of detected objects with their bounding box and orientation regression.

This chapter is based on the work published in [15] (©2016 IEEE).

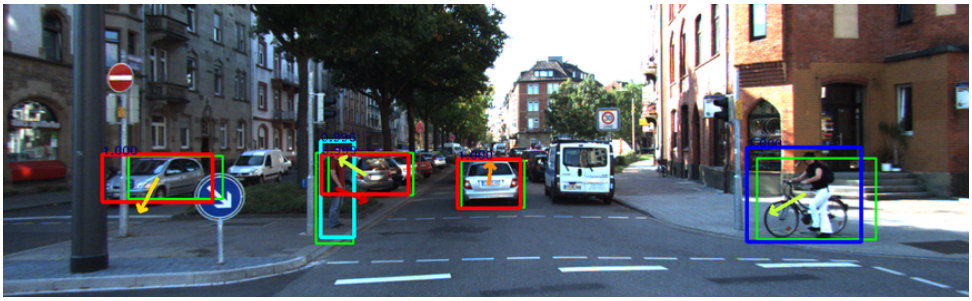


Figure 3.1: Example results on the KITTI dataset of the novel Pose-RCNN framework with bounding box proposals (green), and final detections (after bounding box regression) for cars (red), pedestrians (cyan), and cyclists (blue). Results from the orientation regression are shown by the arrow inside the detected bounding boxes.

3.2. PROPOSED APPROACH

3.2.1. LIDAR PROPOSAL GENERATION

3D object proposals are generated in a straightforward way by clustering an unorganized lidar scan of the 3D environment into smaller clusters. A particular approach to cluster a traffic scene is to remove ground points and group the rest using the nearest-neighbor clustering technique, as shown in Figure 3.2. Ground estimation is carried out through progressive morphological filter (PMF) [174], which distinguishes non-ground measurements such as buildings, vehicles, vegetations etc. from the ground plane. Subsequently, the non-ground lidar points are clustered by grouping nearest neighbors together using a kd-tree search structure [133]. In a last step, the 3D bounding box of each lidar cluster is projected onto the image plane in order to generate 2D object proposals. The 2D proposals are augmented again through spatial translation and scaling.

The recall rate of lidar proposals are highly affected by the parameter settings of the PMF and the nearest neighbor clustering. Here, this chapter evaluates two different parameter settings: The first one *Li1* attempts to rigorously keep the false negative rate as low as possible, whereas the second set *Li2* allows more smaller object clusters in order to increase the recall rate. Table 3.1 shows the detailed parameter settings of *Li1* and *Li2*.

Additionally, ground estimation by *Li2* is only performed on lidar points within a short range, since the laser scan hardly reaches the ground above a certain distance. In other words, ground points are clustered together with wide range objects by *Li2*. This helps us catch more wide range objects and thus increase the recall rate. The range threshold is set to be 20 meters.

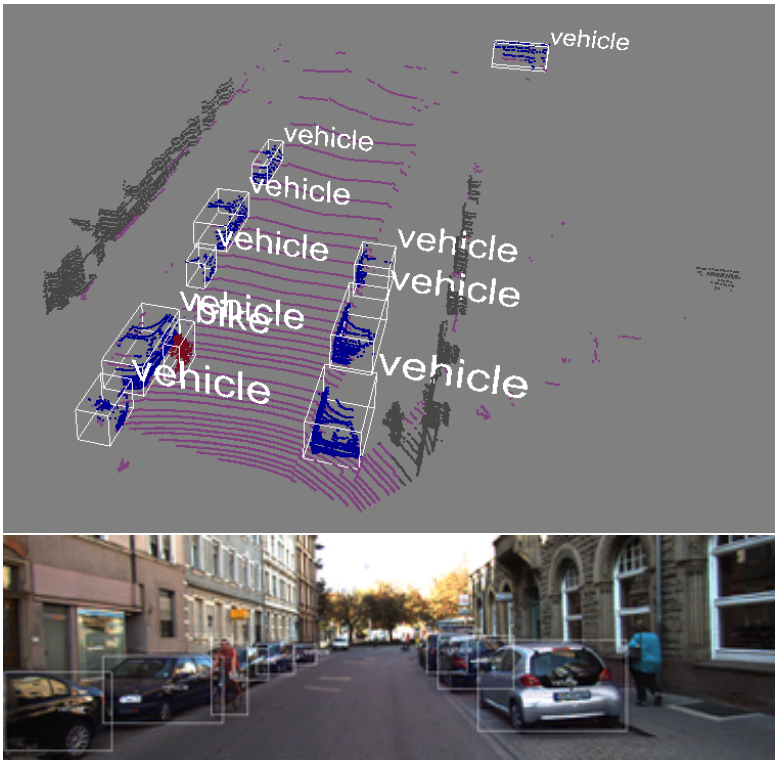


Figure 3.2: Example of lidar bounding box proposal through clustering. Top: Lidar data and 3D boxes resulting from the point clusters. Bottom: 2D proposal boxes resulting from the projection of the 3D boxes.

Table 3.1: Detailed parameter settings of lidar proposal.

Step	Parameter	*Li1*	*Li2*
Ground estimation	initial ground distance	0.2m	0.15m
	maximal ground distance	0.5m	0.15m
Euclidean clustering	cluster distance	0.3m	0.45m
	minimal number of points	50	10

3.2.2. STEREO PROPOSAL GENERATION

Similarly to [102], the stixel representation of the world [44] is used to generate proposals (see Figure 3.3). Stixels are calculated based on stereo data in a joint energy optimization that minimizes the variance of the depth within a stixel. Hence stixels are an efficient and sparse representation of objects having approximately vertical surfaces like vulnerable road users and cars. If the stixel calculation is supported by a ground plane estimation (i.e. in an automotive setup), the 3D position of the bottom of a stixel can be adjusted to match the ground plane. A priori knowledge of the size of possible objects is used to get proposals from the stixels. First, the stixels are filtered by their height in world coordinates that has to be within $[1.2m, 2.4m]$ and their distance that has to be less than $100m$. If an estimated ground plane is available, stixels that are more than $0.5m$ above the ground are also removed. In a second step, the width is adapted to match different aspect ratios. For each stixel, there will be one proposal per aspect ratio. This chapter evaluates two different parameter settings *SP* and *SPLJ*. The applied aspect ratios for both are 0.5, 1, and 2. The width of the stixels in the energy minimization is fixed to seven pixels for *SP* and three pixels for *SPLJ*. By *SPLJ*, each proposal is augmented through four additional proposals sampled in the surrounding. Therefore, the position of the proposals is adapted by 10% of the witdh to the left and right and 10% of the height to the top and bottom.

3.2.3. POSE-RCNN

Fast R-CNN [58] is extended by attaching a small orientation regression network on top of the ROI pooling layer. Based on the last convolution layer of a VGG16 [141] architecture, ROI-pooling is done in the same way as in the original Fast R-CNN version. A softmax probability for classification and per-class bounding-box offsets are estimated from the pooled feature vectors at the end. The orientation regression network estimates an additional per-class orientation angle from each pooled feature vector. Figure 3.4 shows the architecture of the proposed Pose-RCNN. It is essential for an orientation regression network to have a carefully-designed loss function and a “mathematically convenient" representation of orientation. This chapter uses the von Mises [111] distribution for designing a loss for the orientation regression as done in [9]. The von Mises distribution is an analog of the normal distribution for the circular domain which avoids the problem of angular discontinuity, and it is everywhere differentiable and thus optimal for gradient-based optimization. It has the form

$$g(\theta;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)}, \tag{3.1}$$

where θ is an angle, μ is the mean angle of the distribution, κ is the concentration parameter, and I_0 is the *modified Bessel function* of order 0. By inverting and scaling constants, one can derive the von Mises loss function which measures the probabilistic distance between a predicted angle θ and the target angle t as

$$L_{VM}(\theta|t;\kappa) = 1 - e^{\kappa(\cos(\theta-t)-1)}, \tag{3.2}$$

with κ as a hyper parameter controlling the shape of the used von Mises distribution. To predict a periodic value using a linear operation, [9] introduces the Biternion representation $\mathbf{y} = (\cos\theta, \sin\theta)$. By combining the von Mises loss function with the Biternion

Figure 3.3: Two qualitative results of the stereo proposal generation (three rows per example). First the complete set of stixels (top) is filtered by several constraints. The resulting proposals (bottom) have the height of the remaining stixels (middle). Their width is adapted to match given aspect ratios.

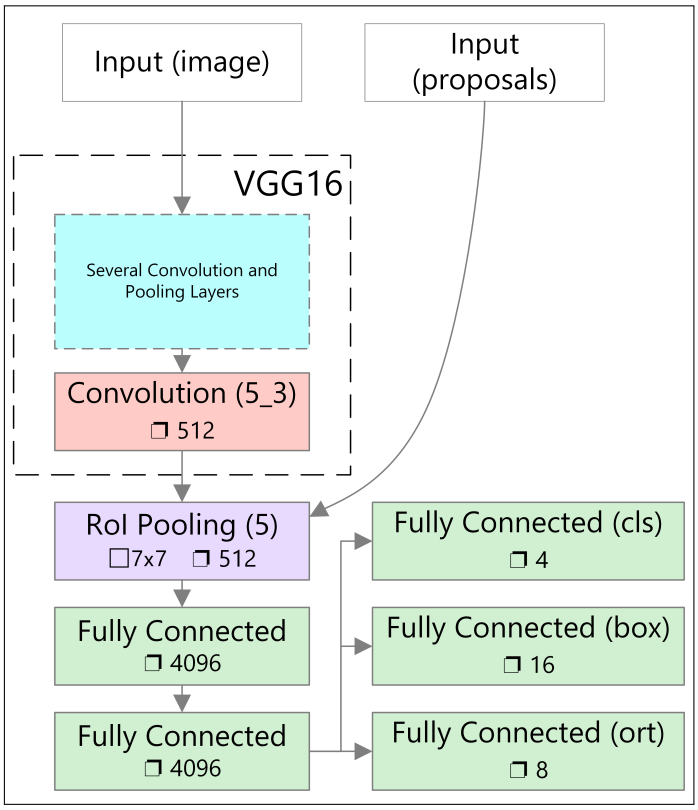


Figure 3.4: Net architecture of the proposed Pose-RCNN. ☐ denotes the size of the ROI pooling layer, and ❒ shows the layer depth. Inputs to the Pose-RCNN comprise an image and a number of bounding box proposals. Feature vectors per bounding box proposal are extracted through the ROI pooling layer, and they are mapped by three fully connected layers in order to generate outputs of the network. These include: softmax probability for classification, per-class bounding-box regression offsets, and per-class orientation regression.

representation and by using common trigonometric identities, the loss function of the orientation regression network and its gradient for back-propagation can be expressed by:

$$L_{VM}(\mathbf{y}|\mathbf{t};\kappa) = 1 - e^{\kappa(\mathbf{y}\cdot\mathbf{t}-1)}, \tag{3.3}$$

$$\frac{\partial L_{VM}}{\partial \mathbf{y}} = -e^{\kappa(\mathbf{y}\cdot\mathbf{t}-1)}\kappa\mathbf{y}. \tag{3.4}$$

The derivation shown here is based on the assumption that the biternion representation takes valid values, which means that $\|\mathbf{y}\| = \cos^2\theta + \sin^2\theta = 1$. Since the orientation regression network cannot guarantee the vector length, normalization is required before loss computation. Therefore, a normalization layer is added ensuring the estimations to be always on the unit circle. Given a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, the forward pass is simply a normalization, described by

$$\mathbf{y} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\mathbf{x}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}}. \tag{3.5}$$

For backward pass, the partial derivative of the loss with respect to each dimension of $\mathbf{x}$ is derived, which is

$$\frac{\partial L}{\partial x_j} = \sum_{i=1}^{d} \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial x_j},\ j \in \{1,2,\ldots,d\}, \tag{3.6}$$

while $\frac{\partial L}{\partial y_i}$ is the backpropagated gradient from the succeeding layer. Following basic derivation rules results in

$$\frac{\partial y_i}{\partial x_j} = \frac{\delta_{ij} - y_i y_j}{\sqrt{\mathbf{x}\cdot\mathbf{x}}}, \tag{3.7}$$

where $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$ represents the Kronecker delta.

3.3. EXPERIMENTS AND EVALUATION

3.3.1. EXPERIMENTAL SETUP

This section evaluates the proposed approach for proposal generation, object detection, and orientation estimation on the KITTI object benchmark, which consists of 7481 training images and 7518 test images that are captured by the left RGB camera mounted on the recording vehicle. For each left color image in the object dataset, a corresponding right color image is available, which enables computing disparities and stixels from stereo pairs. Lidar point clouds and calibration information between the lidar and the cameras are also provided. A total of 80256 labeled objects in common traffic scenes including cars, pedestrians, and cyclists are available in the public training dataset.

The different 3D proposal methods *SP*, *SPLJ*, *Li1*, *Li2* as well as certain combinations are evaluated. Hereby *SP-Li1*, *SPLJ-Li1*, and *SPLJ-Li2* are just the union of the corresponding proposals without any filtering of duplicates.

The framework and parameter settings provided by [24] and the different proposal methods are used for the training of the Pose-RCNN model. 50% of the images of the training dataset serve as validation set.

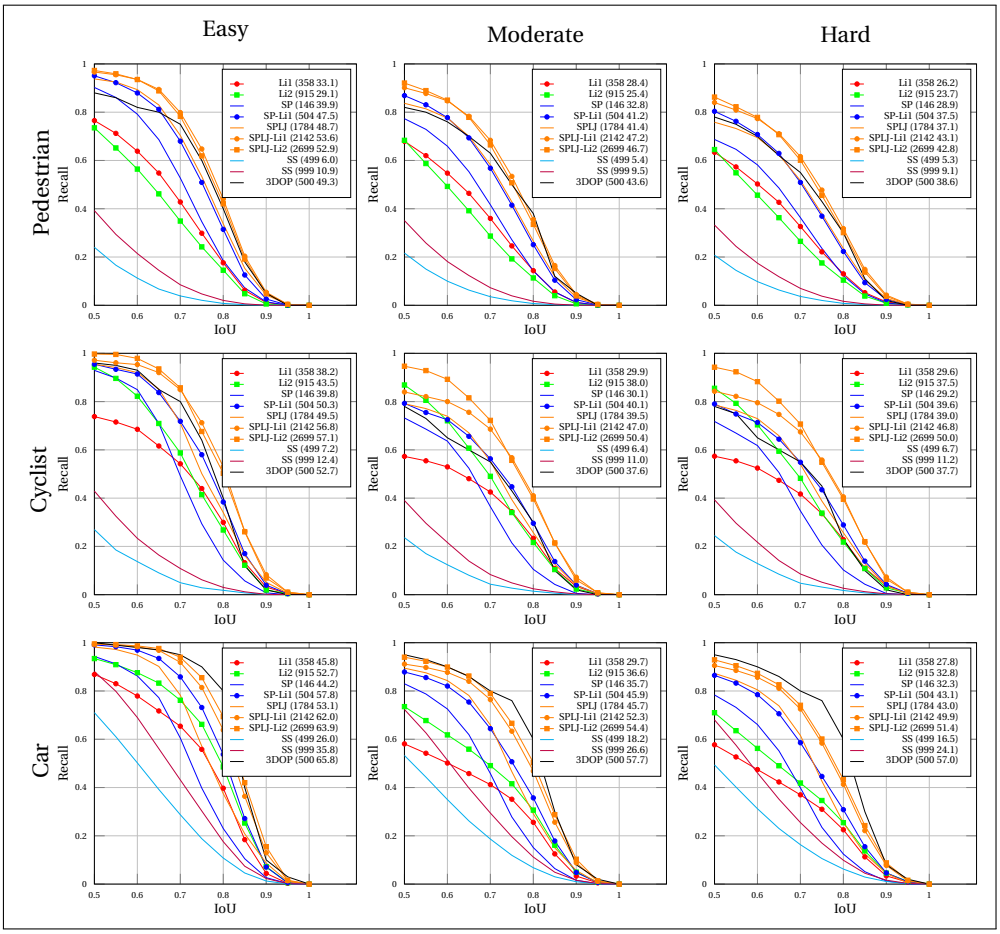


Figure 3.5: Recall as a function of the IoU threshold for several proposal methods. The average number of proposals per frame (first number) and the average recall (in %) are displayed inside the brackets. The *3DOP* curve is sampled from the original paper [24]. Rows show results for different classes, while columns show results for different test scenarios as defined in [56].

3.3.2. RESULTS

PROPOSAL METHODS

The new proposal methods are compared with several state of the art approaches – among others Selective Search (*SS*) [155] and *3DOP* [24]. Therefore, Figure 3.5 shows the Recall as a function of the IoU threshold between 0.5 and 1.0. Average recall (AR) defined in [72] is used as the metric for comparison. Note that the number of proposals is not the same for all methods. A fair comparison can be made between the results of *SP-Li1*, *3DOP* and *SS* for 500 proposals. *SP-Li1* achieves a higher AR than *3DOP* for the moderate and hard setting of the cyclist class and a lower AR for the other cases. *SS* and other state of the art methods like MCG [4] and BING [26] are outperformed (results for these are shown in [24]). The average recall of *SPLJ* and *Li2* is higher than for the corresponding settings *SP* and *Li1* while increasing the number of proposals (see Figure 3.5). Note how the combination of stereo and lidar proposals further boosts the average recall score for *SPLJ-Li1* and *SPLJ-Li2*. The combination of proposals of stereo and lidar data is further

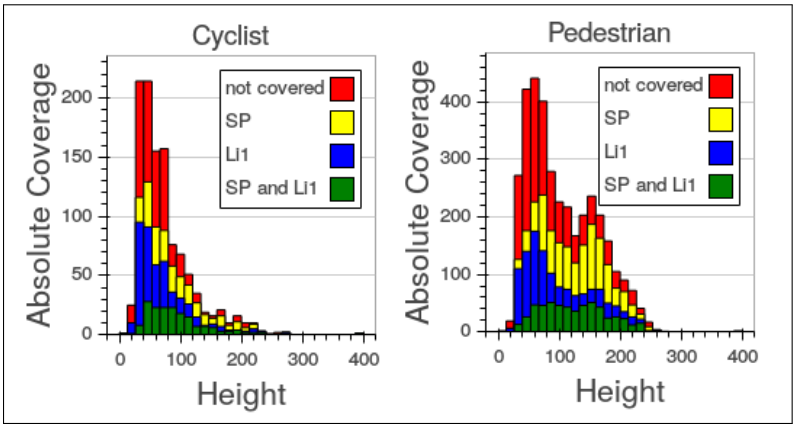


Figure 3.6: Absolute coverage of the ground truth samples of the pedestrian (left) and cyclist (right) classes (moderate difficulty). A sample counts as covered if there is a proposal with an IoU greater than 0.7. The samples are grouped by the method of the covering proposal and their height values. The absolute count of these groups is represented by the height of the bar segments.

analyzed in Figure 3.6. Although duplicates are not deleted when combining proposals, a lot of ground truth samples are only covered by one or more proposals of exactly one method. A great amount of small objects are only covered by lidar proposals. That is due to missing stereo data in the far range.

DETECTION AND ORIENTATION REGRESSION

The Pose-RCNN model trained with proposals of the *SPLJ-Li2* setting achieves the best results on the validation set. For evaluation on the test dataset, an additional Pose-RCNN model is trained with *SPLJ-Li2* proposals on the combined training and validation dataset. Table 3.2 and 3.3 show the respective test results. Especially for cyclists, Pose-RCNN performs very well. The model achieves the highest detection and orientation scores on the easy test scenario and competitive ones on the other scenarios.

Table 3.2: Average Precision (in %) on the test set of the KITTI benchmark for different classes and different test scenarios (Easy, Moderate, Hard) as defined in [56].

	Cars			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
ACF [36]	55.9	54.7	43.0	44.5	39.8	37.2	-	-	-
R-CNN [71]	-	-	-	61.6	50.1	44.8	-	-	-
DPM-VOC+VP [120]	75.0	64.7	48.8	59.5	44.9	40.4	42.4	31.1	28.2
3DOP [24]	**93.0**	88.6	79.1	81.8	67.5	64.7	78.4	68.9	61.4
SubCNN [165]	90.8	**89.0**	**79.3**	**83.3**	**71.3**	**66.4**	79.5	**71.1**	**62.7**
Pose-RCNN	88.4	75.8	66.6	77.5	63.4	57.5	**80.8**	68.8	60.4

Table 3.3: Average orientation Similarity (in %) on the test set of the KITTI benchmark for different classes and different test scenarios (Easy, Moderate, Hard) as defined in [56].

	Cars			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
DPM-VOC+VP [120]	72.3	61.8	46.5	53.6	39.8	35.7	30.5	23.2	21.6
3DOP [24]	**91.4**	86.1	76.5	72.9	59.8	57.0	70.1	58.7	52.4
SubCNN [165]	90.7	**88.6**	**78.7**	**78.5**	**66.3**	**61.4**	72.0	**63.7**	**56.3**
Pose-RCNN	88.3	75.4	66.1	74.0	59.9	54.3	**75.5**	62.9	55.5

3.4. DISCUSSION

Average orientation similarity as defined in [56] is strongly correlated with the average precision. The average precision score is even the upper bound for the average orientation similarity score. The ratio between the orientation and detection score can not be higher than one. For cars, 3DOP [24], SubCNN [165], and the new Pose-RCNN approach already achieve a ratio of nearly one. For cyclists and pedestrians, the ratio achieved by Pose-RCNN is higher than by 3DOP and SubCNN, which evinces the potential of the proposed approach. The improvement of the detection performance for example by improving the average recall of the proposals could automatically boost the average orientation similarity score.

4

A Novel Benchmark for Person Detection in Traffic Scenes

4.1. Overview

During the last two decades, an extensive amount of research has been spent on pedestrian detection [6, 38, 43, 71, 125, 178]. For several years, progress in this domain was monitored on benchmarks like Caltech [38] and KITTI [56], which was also used in the last Chapter 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 thousand samples) furthermore makes these benchmarks prone to dataset bias and to over-fitting [151]. Recently, CityPersons [179] was released with higher resolution images and a larger quantity of training data ($\approx$ 35000 person samples). Although these data additions are helpful, [179] 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 a few countries ($1-3$), and in daylight and dry weather conditions only; this hampers generalization to real world applications.

To address these limitations this chapter introduces a new dataset for vision-based person detection coined EuroCity Persons. The images for this dataset were collected onboard a moving vehicle in 31 cities of 12 European countries, see Figure 4.1. With over 238200 person instances manually labeled in over 47300 images, EuroCity Persons is nearly one order of magnitude larger than person datasets used previously for benchmarking, in terms of manual annotations (see Table 4.1). 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, this chapter addresses a number of questions in Section 4.3: 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?

This chapter is based on the work published in [13] (©2019 IEEE).

Figure 4.1: 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 the dataset.

Table 4.1: Comparison of person detection benchmarks in vehicle context

	Caltech [38]	KITTI [56]	CityPersons [179]	TDC [102]	**EuroCity Persons**
# countries	1	1	3	1	**12**
# cities	1	1	27	1	**31**
# seasons	1	1	3	1	**4**
# images day	**249884**	14999	5000	14674	40217
# pedestrians day	**289395**[a]	~9400[b]	31514	8919	183004
# riders day	-	~3300[b]	3502	**23442**	18216
# ignore regions day	57226[a]	~22600[b]	13172	-	**75673**
# orientations day	-	~12700[b]	-	-	**176879**
# images night	-	-	-	-	**7118**
# pedestrians night	-	-	-	-	**35309**
# riders night	-	-	-	-	**1564**
# ignore regions night	-	-	-	-	**20032**
# orientations night	-	-	-	-	**34393**
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 (the number of manual annotations probably are 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.

Figure 4.2: External view of the sensor vehicle buildup.

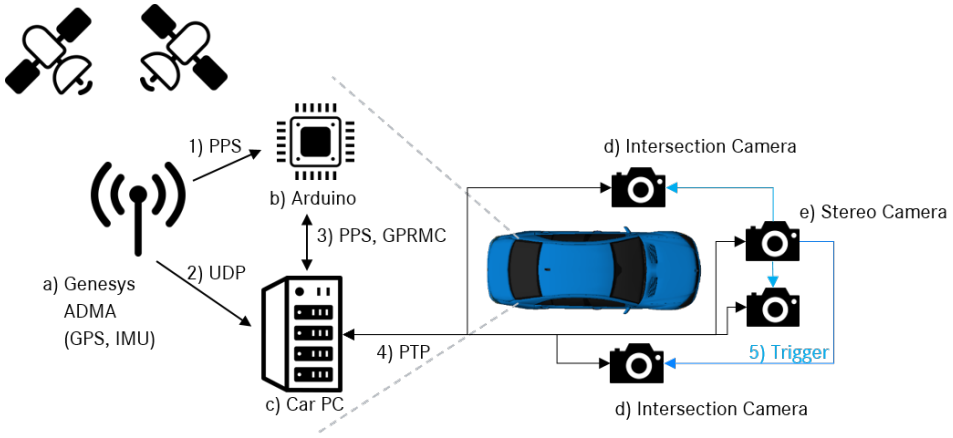


Figure 4.3: Schematic of the communication used to achieve time synchronization.

4.2. The EuroCity Persons Benchmark

4.2.1. Sensor Vehicle Buildup

For the recording of the EuroCity Persons dataset, a vehicle had to be built up, integrating the required sensors and other recording hardware, especially the car PC as major computing and storage unit. Figure 4.2 shows an external view of the car. While three lidars, five front-facing, and two side-facing cameras have been integrated in total, the focus of this thesis lies on two front-facing cameras mounted behind the windshield that form a stereo pair, and the two side facing intersection cameras mounted at the side mirrors. All sensors are connected with the car PC, which takes care of raw data processing, recording, and storage utilizing the robot operating system (ROS) [144]. For later use of the data, a precise localization in space, as well as time, is needed. The former is achieved by intrinsic and extrinsic calibration of the sensors, while the latter is fulfilled by the implemented time synchronization, which will be described in more detail below.

Calibration. The projection of the world on the image sensor is described by the pinhole camera model. The goal of the *intrinsic calibration* is the estimation of the free model parameters for each camera, e.g. the focal length and distortion parameters. These are calculated based on several recorded images of a Tsai grid [153] or a checkerboard pattern of known size at different positions relative to the camera. For a detailed explanation see [65].

The *extrinsic calibration* is needed to estimate the relative position and orientation of the cameras to each other and to the vehicle coordinate system. For this mapping between the coordinate systems of the sensors and the vehicle a tachymeter has been used. It forms the point of origin in a world coordinate system and measures the exact 3D position of a prism that is tracked by a laser beam. Measuring the 3D location in addition to the 2D image position of that prism at several positions enables the extrinsic 3D calibration of the cameras within the world coordinate system of the tachymeter

by solving the *Perspective-n-Point* problem. By measuring the 3D positions of the four wheels of the car with the tachymeter the world location and pose of the car is estimated. As the locations and poses of the cameras and the car are calculated in the same world coordinate system of the tachymeter the relative mapping between different sensors and the car is calculated by multiplying and inverting the respective calibration matrices.

Time Synchronisation. The implemented system (see Figure 4.3) ensures that the internal clocks of all sensors are precisely synchronized to the GPS time. The GPS signal is received by a Genesys ADMA, which provides an accurate car position in a global coordinate system by filtering its internal IMU data and the GPS data. The current location and time are sent via UDP to the car PC. An Arduino board serves as a relay to deliver a so-called pulse per second (PPS) signal of the ADMA to the serial car PC interface. The PPS signal on this interface is used to synchronize the car PC's clock to the GPS time via GPSd and the network time protocol (NTP). The clocks of the car PC and the cameras are then synchronized using the precision time protocol (PTP). Furthermore, the left camera of the stereo pair serves as the master camera and sends a trigger signal to all other cameras. Thus, all cameras start their image capturing at the same time.

Camera Specifications. The stereo and intersection cameras are state of the art two-megapixel cameras (1920 x 1024) with rolling shutter run at a frame rate of 20 Hz. They yield 16-bit rgb images; this high dynamic range is important for capturing scenes with strong illumination variation (e.g. night-time, low-standing sun shining directly into the camera). The EuroCity Persons dataset presented in this chapter consists only of images of the left stereo camera, while the EuroCity Persons Dense Pose dataset presented in Chapter 5 also utilizes images of the two intersection cameras. As the images are recorded without further compression, the raw data streams of the four megapixel cameras alone result in a data rate of ~320 MB/s - the total data rate of all sensors was between ~500 MB/s and ~700 MB/s. That produces sufficient computing load to require direct cooling by redirecting the car's air conditioning system into the trunk to avoid overheating the complete system. Recording at such a high bandwidth was furthermore enabled by an internal SSD RAID. The recorded data was ingested to slower, high-volume storage devices during and after the tours for transport and long-term storage.

4.2.2. DATASET COLLECTION

The images of the EuroCity Persons dataset were collected from the moving vehicle described in the section before in 31 cities of 12 European countries. To collect data from all seasons, the recordings were done during seven recording tours. This enabled an early beginning of data processing and data annotation right after the first tour. On average four to five cities have been visited per tour. The first tour began in Stuttgart, Germany on the 25th of October 2016 while the last tour to Italy ended on the 15th of June 2017. For every visited city, a route has been planned beforehand as close as possible to the city center and pedestrian areas, where most interesting traffic scenarios with vulnerable road users can be expected. In between the tours, the intrinsic calibration of the cameras had to be repeated, as varying temperatures within the car and the strong sun exposure cause a decrease in calibration accuracy over the course of time. Images of the left stereo camera were debayered and rectified after each tour. 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.

53 hours of image data were collected in total, for an average of 1.7 hours per city. To limit selection bias [151], every 80-th frame was extracted for the detection benchmark without further filtering. This means that a substantial fraction of the person annotations in the dataset are unique, although especially at traffic lights and in slow-moving traffic, the same persons might appear in different annotations. Even so, due to sparse sampling at every four seconds, image resolutions and body poses will differ.

4.2.3. DATASET ANNOTATION

Annotations comprise 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. The latter is discriminated between no occlusion, low occlusion (10%-40%), moderate occlusion (40%-80%), and strong occlusion (larger than 80%). 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, the riding person and its ride-vehicle are labeled with two separate bounding boxes. The ride-vehicle type is also annotated. 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 [178] and [179] 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 Figure 4.4) and to be comparable with the KITTI dataset, the classic bounding-box convention of labeling the outermost object parts is remained for the annotations. 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. This information is provided 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. Reasonably high demands are placed on accuracy. The amount of missed and hallucinated objects were each to lie within 1% of the annotated number. Annotators were

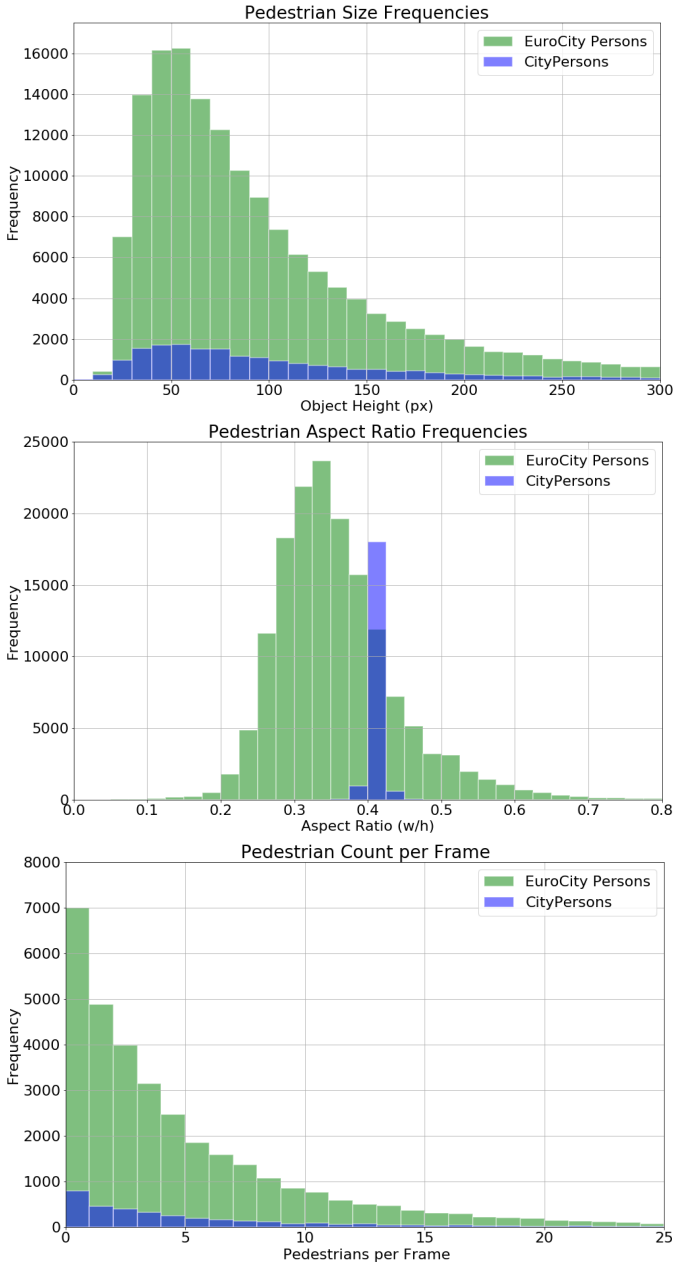


Figure 4.4: Statistics of EuroCity Persons and CityPersons for pedestrians of the training and validation datasets (top: height, middle: aspect ratio, and bottom: count per frame).

Figure 4.5: Screenshot of the Labrador annotation tool showing an annotated image recorded in Rome.

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.

4.2.4. ANNOTATION TOOLING

In practice, the efficiency of the manual annotations, as well as the quality control, also depends on appropriate tooling software. Here, *Labrador* has been used for both which is described in detail in [53]. Labrador provides a highly flexible, and configurable graphical user interface, and a plugin-based software architecture with a broad range of available plugins that can be positioned and customized according to the task at hand, e.g. bounding box based annotation or quality control. Figure 4.5 shows a screenshot of the software with bounding box and orientation annotations on an image recorded in Rome.

The *Document View* (top-left) lists the images within the sequence to be annotated and the already annotated objects per image. The sequence can also be navigated by the *Frame Slider* (top). A detailed list of the annotated objects is given by the *Entity Tree* (top-right), which also shows annotated tags, like the level of occlusion, which can be easily annotated using the buttons on the right of the *LabelEditor* (center). Apart from tag annotation, bounding boxes can be drawn in the *LabelEditor* plugin. A zoomed crop of the currently selected object is displayed by the *EntityViewer* (bottom-right), which provides functionality for the orientation annotation. A 3D model of a body is displayed right next to the crop. The task of the annotator is to align the orientation of this 3D model with the selected person. Hence, the annotator annotates the orientation angle of the person relative to the line of sight of the camera, as the annotation of the object based on the crop is independent of its position within the image. Regarding orientation estimation investigated in this thesis, which only depends on appearance information from single images, it is more feasible to estimate this relative angle. If the distance of the object and the camera calibration are known, the orientation within the coordinate system of the car may be calculated using simple trigonometry.

4.2.5. DATA SUBSETS

Various data subsets are defined on the overall EuroCity Persons dataset. First, it is split into 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 [56, 102, 179]:

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

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

Figure 4.6: 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% distribution.

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% respectively, as illustrated in Figure 4.6. During halts 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, sequences are only splitted at points in time where the recording vehicle had a speed larger than 7 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.

4.2.6. DATASET CHARACTERISTICS

See Table 4.1 and Figure 4.4 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 observed person count per frame, which, as shown in Figure 4.4 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 night-time 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, very diverse backgrounds and person appearances were recorded, see Figures 4.8, 4.9, 4.10 and 4.11.

4.2.7. EVALUATION METRICS

To evaluate detection performance, the miss-rate (mr) is plotted against the number of false positives per image (*fppi*) in log-log plots:

$$mr(c) = \frac{fn(c)}{tp(c) + fn(c)}, \tag{4.1}$$

$$fppi(c) = \frac{fp(c)}{\#img}, \tag{4.2}$$

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 [38, 48, 56, 179] 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. The log average miss-rate ($LAMR$) is defined as

$$LAMR = \exp\left(\frac{1}{9}\sum_{f}\log\left(mr(\operatorname*{argmax}_{fppi(c)\leq f} fppi(c))\right)\right), \tag{4.3}$$

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 $mr(\operatorname{argmax}_{fppi(c)\leq f} fppi(c))$. 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 this chapter follows 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 [56, 102] this chapter also utilizes the Average Precision (AP), which is defined as:

$$AP = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}} \max_{re(c)\geq r} pr(c), \tag{4.4}$$

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 the average orientation similarity (AOS) is used [56]:

$$AOS = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}} \max_{\tilde{r}:\tilde{r}\geq r} s(\tilde{r}), \tag{4.5}$$

where s is the orientation similarity given by:

$$s(r) = \frac{1}{|\mathscr{D}(r)|} \sum_{i \in \mathscr{D}(r)} \frac{1 + \cos \Delta_\theta^{(i)}}{2} \delta_i. \tag{4.6}$$

$\mathscr{D}(r)$ 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 [56], [179], 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 4.2.3, 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 of more than 0.5 with these regions are not considered as false positives.

4.2.8. 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 Subsection. Results are tallied online, either by name or anonymous. The frequency of submissions is limited.

4.3. EXPERIMENTS

All the baseline and generalization experiments (Sections 4.3.1 and Section 4.3.2) 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 4.3.3), unless stated otherwise.

4.3.1. BASELINES

As the top ranking methods on KITTI and Caltech use deep convolutional neural networks, the baselines are selected among these methods. Many recent pedestrian detection methods [17, 75, 99, 110, 165, 170, 184] 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 [179], 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 [127] is one of the first methods within this group. This section evaluates its latest extension **YOLOv3** [128], as in comparison with its predecessors, its design is promising regarding the detection of smaller objects. Within both groups methods with explicit hard example mining are also selected, namely **R-FCN** [31] and **SSD** [109].

Faster R-CNN, R-FCN and SSD are trained with the Caffe framework [80] using VGG-16 [141] as base architecture (as done for pedestrian detection in [75, 99, 110, 130, 179]; using ResNet as base architecture for Faster R-CNN did not improve experimental results, see supplemental material). YOLOv3 is trained with the Darknet framework and Darknet-53 [128] as base architecture. The base architectures are pre-trained on ImageNet [35].

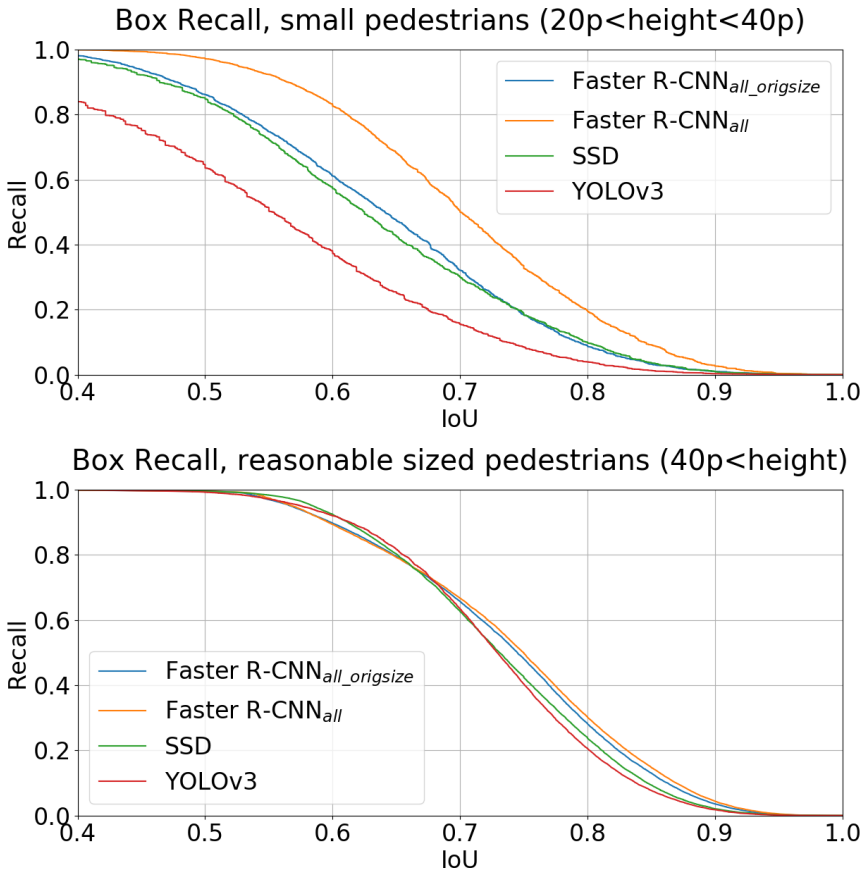


Figure 4.7: Recall vs. *IoU* for small pedestrians (top) and pedestrians of the "reasonable" test case (bottom) for the optimized anchor-boxes of Faster R-CNN and YOLOv3 and the SSD default boxes.

Adaptations and Training. The box recall for all methods is optimized as it is important for the overall detection performance. Improvements from [179] are applied for Faster

R-CNN and R-FCN namely 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, optimization of the default boxes of SSD and the anchor boxes of YOLOv3 results in similar recalls for all methods for the "reasonable" test case as shown in Figure 4.7. For Faster R-CNN and R-FCN an ignore region handling similar to [179] is implemented. Furthermore, training samples are filtered according to different test cases to train several Faster R-CNN models as summarized in Table 4.2. For all experiments with R-FCN, SSD and YOLOv3 samples are filtered that are more than 80% occluded or smaller than 20 *px* in height. SGD is used as backpropagation algorithm on the training dataset with a stepwise reduced learning rate. The model to be evaluated on the test dataset is selected on the validation dataset.

Table 4.2: 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 4.3: 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

Results. See Table 4.3 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% 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% faster during training and testing. As could be expected by the lower box recall shown in Figure 4.7, 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 Figures 4.8, 4.9, 4.10 and 4.11 for some illustrations of typical results with Faster R-CNN$_{all}$ (including night-time and rider results, not part of this section) and Figure 4.12 for miss-rate curves of the methods considered.

Failure Analysis. This section now analyzes the detection errors of the best-performer Faster R-CNN$_{all}$ qualitatively and quantitatively. Tables 4.4 and 4.5 illustrate 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 (the evaluation policy is strict and due to application considerations these are count as wrong; note, however, that depictions and reflections are annotated in the dataset, thus a more lenient policy to ignore false positives of these types is readily implemented).

Certain pedestrian poses and aspect ratios can lead to multiple detections for the same pedestrian as shown in the *Multidetections* category. Non-maximum suppression (NMS) is used by Faster R-CNN and other deep learning methods to suppress multiple detections. The used *IoU* threshold of 0.5 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% 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. The qualitative analysis shows 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.

The following quantitative analysis of false positives builds upon the ideas of oracle tests as in [178]. There, false positives touching ground-truth samples are regarded as localization error. Non-touching false positives are regarded as confusion of fore- and background. False positives types are analyzed for a finely discretized range of false

Figure 4.8: Qualitative detection results for true positives of Faster R-CNN$_{all}$ at *fppi* of 0.3 (green: pedestrians, blue: riders). Samples recorded during dry weather.

Figure 4.9: Qualitative detection results for true positives of Faster R-CNN$_{all}$ at *fppi* of 0.3 (green: pedestrians, blue: riders). Samples recorded during rainy weather and wintertime.

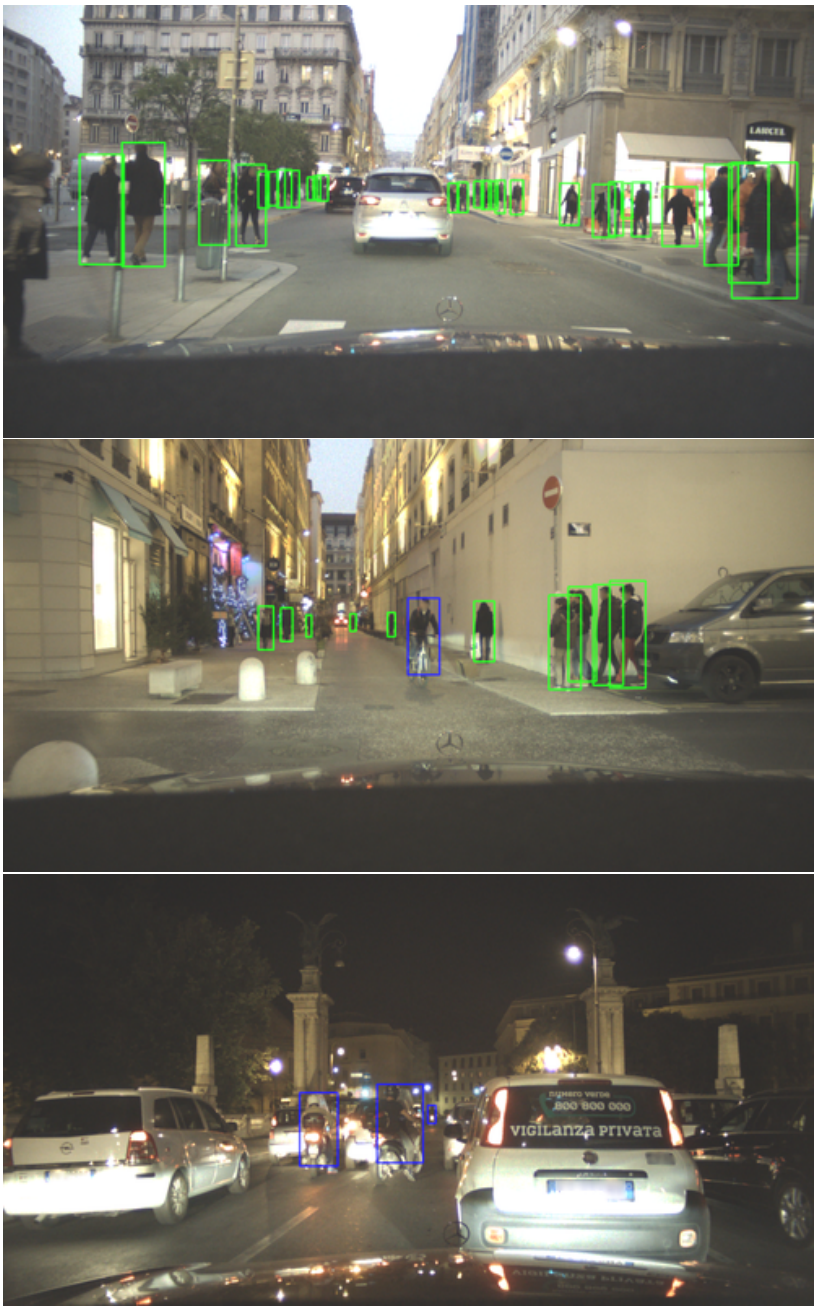


Figure 4.10: Qualitative detection results for true positives of Faster R-CNN$_{all}$ at *fppi* of 0.3 (green: pedestrians, blue: riders). Samples recorded during during dusk and night.

Figure 4.11: Qualitative detection results for true positives of Faster R-CNN$_{all}$ at *fppi* of 0.3 (green: pedestrians, blue: riders). Samples recorded during dusk and night.

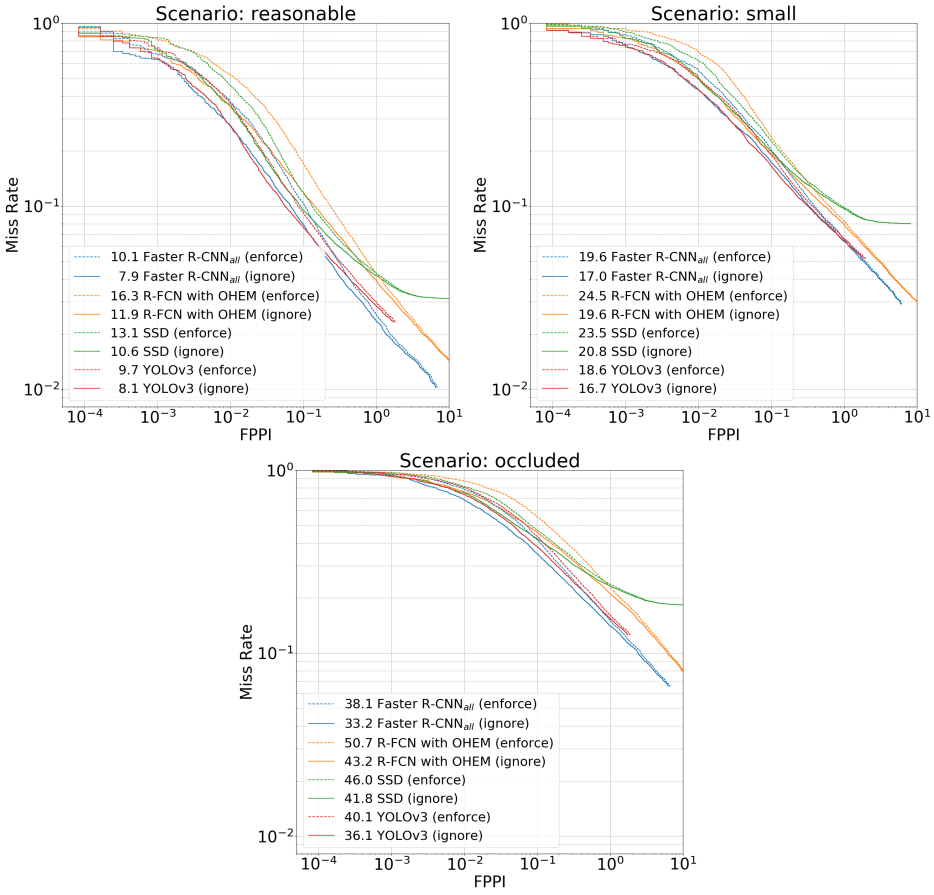


Figure 4.12: Miss-rate curves on the EuroCity Persons test set for the selected methods for the "reasonable" (top left), "small" (top right) and "occluded" (bottom) test case. The required *IoU* for a detection to be matched with a ground-truth sample is 0.5. For every method, the curves are shown for enforcing or ignoring precise class label with respect to neighboring classes.

Table 4.4: Error types in detection for Faster R-CNN$_{all}$ at 0.3 *fppi* (green: true positives, red: false positives, purple: false negatives, white: ground truth).

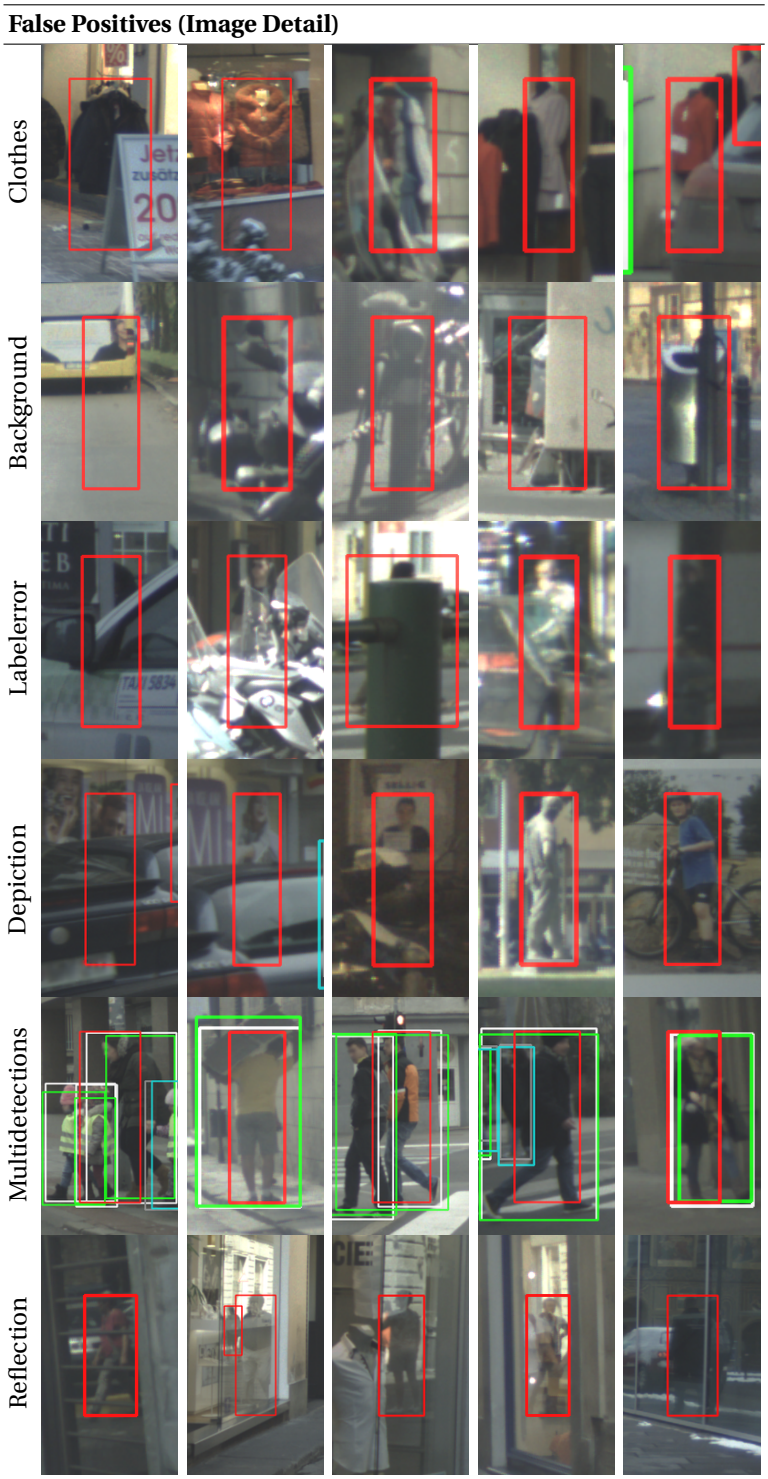


Table 4.5: Error types in detection for Faster R-CNN$_{all}$ at 0.3 *fppi* (green: true positives, red: false positives, purple: false negatives, white: ground truth).

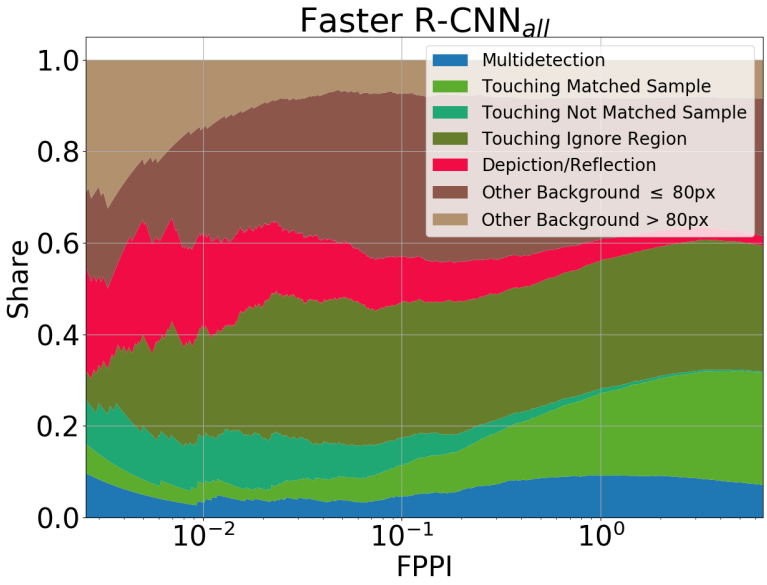


Figure 4.13: The contribution of various sources to the number of false positives of Faster R-CNN$_{all}$, depending on *fppi*

positive per image (*fppi*), see Figure 4.13. This section further subdivides the localization errors in four groups: multiple detections ($IoU > 0.5$ with ground-truth samples, as multiple assignments are penalized), 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. The fore- and background confusions are subdivided 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.

Figure 4.13 shows that localization errors account for about 60% of all errors at a high *fppi* of 6, decreasing to about 40% 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 [148] and upgrading to the latest GPU processor; this was outside the scope of this study.

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

4.3.2. 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 4.2.3). 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).

This section examines how the various datasets generalize with respect to the traffic-related ("target") datasets KITTI, CP, and ECP. For this, various training sets are considered (in isolation and with pre-training) and the performance of a reference model (i.e. the optimized Faster R-CNN baseline) is measured on a target evaluation set.

The OP dataset contains 3.2M individually labeled persons from 736433 images. Labeled groups of persons are used as ignore regions in the experiments. To compensate for the large person sizes the OP images are downscaled by a factor of 2 (OP512) or by a factor of 4 (OP256). The official KITTI training dataset is split into two equally sized, disjunct subsets to obtain separate training and validation datasets, as in [24].

All models derived from the individual KITTI, CP, OP, and ECP datasets are initialized with ImageNet [35]. 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 the generalization experiments are shown in Tables 4.6, 4.7 and 4.8. A first observation is that if no pre-training is used (rows 1, and 6-9 of Tables 4.6-4.8), 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 4.6 and 4.7). 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 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 4.7). 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 4.6 and 4.7), pre-training with ECP outperforms pre-training with OP significantly. For the KITTI validation set, Table 4.6 shows 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 4.7) this improvement is at least 0.9 and 1.5 in $LAMR$.

Table 4.6: Average Precision on the KITTI validation set for different training settings of Faster R-CNN. $A \rightarrow B$ denotes pre-training on A and finetuning on B.

	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

Table 4.7: Log average miss-rate ($LAMR$) on the CityPersons (CP) validation set for different training settings of Faster R-CNN. $A \rightarrow B$ denotes pre-training on A and finetuning on B.

	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

Table 4.8: Log average miss-rate (*LAMR*) on the EuroCity Persons (ECP) test set for different training settings of Faster R-CNN. $A \rightarrow B$ denotes pre-training on A and finetuning on B.

	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

Pre-training with ECP furthermore strongly outperforms pre-training with KITTI or CP across the board. For the KITTI validation set, there is 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 4.6). For the CP validation set, there is an improvement of 2.0, 9.3, and 5.9 in *LAMR* (rows 2 and 3 in Table 4.7). Note that the *LAMR* listed in [179] 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. Using the exact same settings as in [179] results in an even better *LAMR* of 12.2, which is improved by ECP pre-training to 10.2.

The benefit of pre-training with ECP on the official KITTI test set is analyzed by submitting to the evaluation server on the KITTI website. The pre-trained model on ECP achieved an average precision of 74.3 for the moderate setting. At the moment of the 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 4.8).

4.3.3. DATASET ASPECTS

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

Quantity. [146] shows a logarithmic relation between the amount of training data and the performance of deep learning methods. This relation is validated on the ECP benchmark. Therefore, the baseline methods are trained on different sized subsets which are randomly sampled from all cities. The detection results for the baseline methods with the use of different augmentation modes in dependence of the dataset proportion are shown in Figure 4.14. 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. The logarithmic relation between training set size and detection

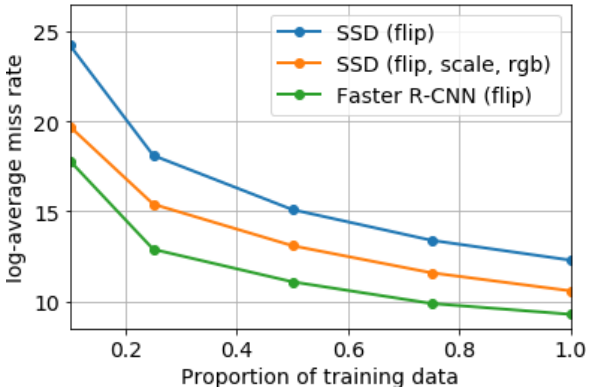


Figure 4.14: Detection performance (*LAMR*) of Faster R-CNN and SSD as a function of training set size

performance also holds on the ECP benchmark for Faster R-CNN and SSD.

Diversity. This section investigates whether overall geographical region introduces a dataset bias which influences person detection performance. For this, two datasets are constructed that are similar in terms of other influencing factors (i.e. season, weather, time of day, person count per frame):

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

These datasets are split into subsets for training, validation and testing as described in Section 4.6, such that the number of pedestrians in each training dataset is 15000. [34] shows that resampling of a dataset can be applied to evaluate the significance of benchmark results. The train-val-test blocks are permutated and the block length is varied (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 4.9. 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 for the Wilcoxon signed-rank test [34]. 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. The respective p-value is calculated, 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%, 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 4.10 shows detection results for the day-time, night-time and combined datasets. As the night-time dataset is only 20% of day data (Table 4.1) the number of training samples used for the day-time and combined

models is reduced accordingly for this experiment. Table 4.10 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.

Table 4.9: Effect of geographical bias on detection performance ($LAMR$) for the "reasonable" test case: central West Europe (WE) vs. central East Europe (EE). Datasets compiled to provide otherwise similar conditions. Results involve averages over different dataset splits.

	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

Table 4.10: Effect of day- vs. night-time condition on detection performance ($LAMR$) for the "reasonable" test case. Datasets compiled to provide otherwise similar conditions.

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

Table 4.11: 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

Detail. The importance of additional annotations for ignore regions, for riders, and for orientations is now examined. Table 4.11 shows results for a model trained without ignore region handling compared to the baseline method. In accordance with earlier findings [179], the detection performance deteriorates when not using ignore regions during training. For the "reasonable" and "small" test cases the $LAMR$ drops by about two points.

The baseline detection method is extended by an orientation estimation layer as in [15] (Two variants for the orientation loss are considered: L1 and Biternion loss). Hence, the network performs multi-tasking: classification, bounding box and orientation regression (see Figure 4.15 for qualitative results for the orientation estimation using the

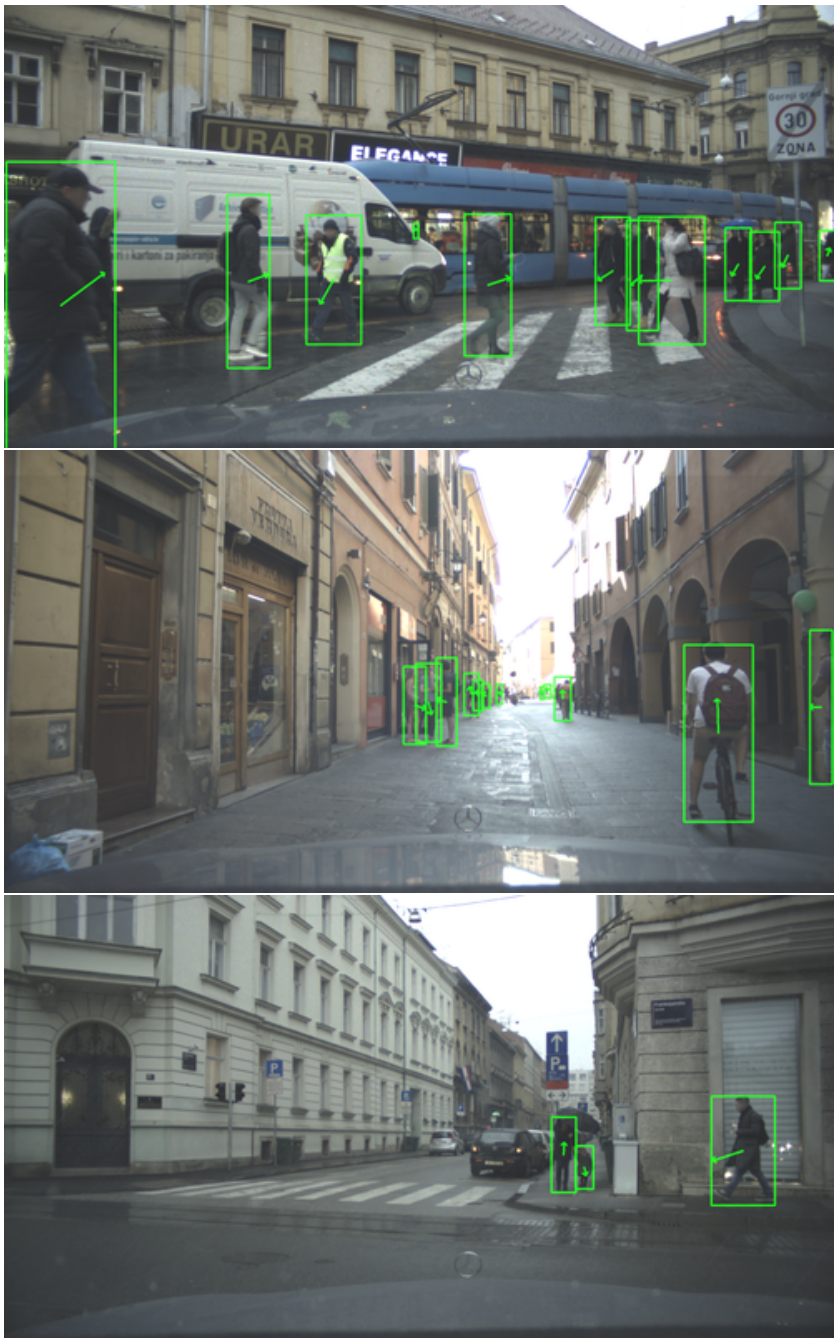


Figure 4.15: Qualitative results for orientation estimation. Top and center image show correct estimations. Bottom image contains a rare failure case (left person has orientation offset of about 180 degrees)

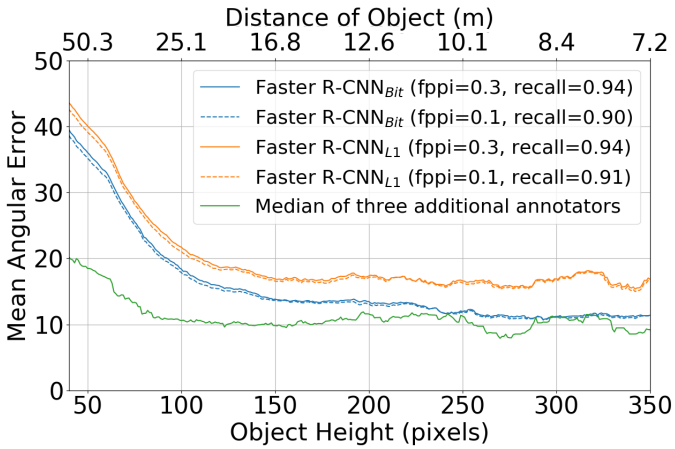


Figure 4.16: Person orientation estimation quality vs. object size (distance).

Biternion loss). As body orientation correlates with the aspect ratio 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 [58] which shows that training multiple tasks together can improve the overall result, the detection results decrease slightly for the multitask network with the Biternion loss as shown in Table 4.11. Figure 4.16 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 4.2.7 for the "reasonable" test case results in a score of 85.9 for the L1 loss and 86.7 for the Biternion loss.

Table 4.12: Effect of multi-class handling (pedestrian vs. riders) on detection performance ($LAMR$) for the "reasonable" test case. 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.

	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 evaluation protocol described in Section 4.2.7 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 the baseline method increases from 9.3 to 11.0, as shown in Table 4.12. By adding riders as an additional class, one observes that the pedestrian detection performance improves for the protocol which requires pedestrians

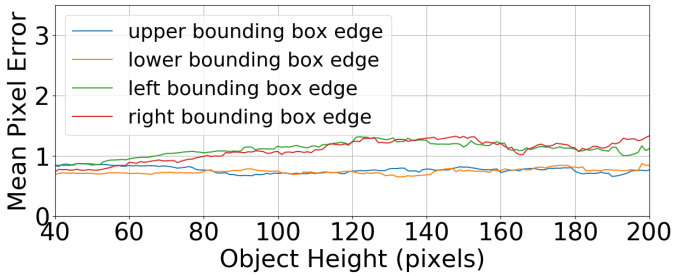


Figure 4.17: Mean pixel error between median of three additional annotators and the ECP dataset annotations, in dependence of object height p (averaged over the interval [p-20, p+20]).

to be classified as such (10.3 vs. 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 the training dataset.

Accuracy. Here, this section evaluates to what degree the annotation accuracy requirements from Section 4.2.3 were actually met in practice in the final EuroCity Persons annotations.

To estimate the amount of missed annotations, these are compared with the object detector output. At a *fppi* of 0.3 for Faster R-CNN$_{all}$ on the "reasonable" test case 230 missed annotations larger than 32 *px* are manually counted. However, the miss-rate for Faster R-CNN$_{all}$ at this *fppi* is about 10% for the small test scenario and about 30% for the occluded test scenario. Using the more conservative 30% figure, one can 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 48000 pedestrians in the test dataset, this corresponds to 0.7% missed annotations, which lies within the 1% quality requirement of Section 4.2.3.

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 1000 not occluded pedestrians was labeled again by three different persons. The average deviation between the median value of the three annotators and the corresponding Eurocity Persons annotation is analyzed in dependence of the object size. Figure 4.17 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). Figure 4.16 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*. This lies within the requirements of Section 4.2.3 as well.

In the following experiments the annotation quality of the training dataset is artificially disturbed, see Table 4.13. First, bounding boxes of instances and groups are randomly deleted to simulate the effect of missed objects during annotation ("delete"). Second, bounding boxes are moved by four pixels up or down and left or right ("jitter"). Third, (erroneous) ground-truth boxes are added 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, hallucinations are introduced that are more likely to resemble pedestrians, by running a SSD model of an early training stage on the training dataset (after 80000 iterations). The 11000 highest scoring false positives of these detections (corresponds to 10% of all pedestrians in the training dataset) are handled as regular ground truth boxes and added to the training dataset for the "false positives" experiment. Different levels of disturbances are examined by manipulating different amounts of bounding boxes. The effects for disturbances that are even worse than in the very first pilot study are also evaluated. The probability for a bounding box to be disturbed is given in the Table 4.13.

The detection performance of Faster R-CNN suffers from deleting and disturbing the bounding box locations. Deleting 25% of the bounding boxes results in a miss-rate of 11.3. Note that with 75% of the training samples a $LAMR$ of 10.0 is achieved (see Figure 4.14). 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% hallucinations that more resemble pedestrians ("false positives") result in a more significant drop in performance of 3.3 points.

Table 4.13: 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

4.4. DISCUSSION

A main outcome from the EuroCity Persons (ECP) 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 4.1), the considered state of the art deep learning methods (Faster R-CNN and SSD) do not saturate in detection performance.

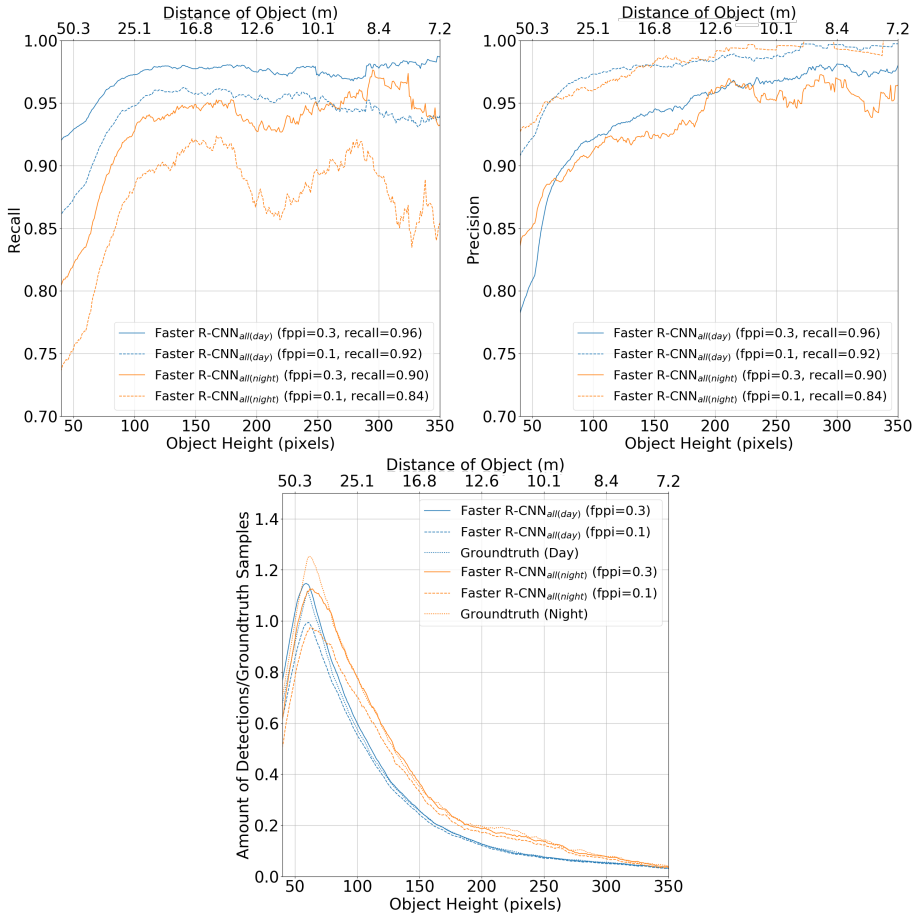


Figure 4.18: Recall (top left), precision (top right) and the associated per-image detection and ground-truth sample counts (bottom) vs. object height at two operating points for the Faster R-CNN variant at day- and night-time (each trained and tested separately on upscaled day- and night-time images 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*].

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 4.3.2, 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. Datasets were compiled for central West Europe vs. central East Europe, where other factors influencing performance were held similar. The analysis showed that the existence of a bias is statistically significant with a confidence interval of 99%.

Comparing day- and night-time detection performance, one observes from Table 4.10 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 Figure 4.18. 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. The ECP 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 [179] the experiments show that a correct ignore region handling has an impact on detection performance. It boosts performance by 1.5 points (see Table 4.11). This is a larger difference than that between the performances using 75% and 100% of the training data in Figure 4.14. This chapter goes beyond [179] and shows that it is beneficial to train specific detectors for classes that otherwise might be confused with the target class. In the 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 [179] 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 [43] was a method based on HOG features and SVM classifier. When comparing its performance with that of the best-performer in this chapter, 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 (Figure 8 in [43] vs. Figure 4.12 here).

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

production vehicles on the market. A number of factors improve performance in the vehicle application. First, other than assumed 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 Figure 4.18, the detection rate increases to 97% 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 [98]). 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 ([43] shows a reduction by up to 37%). 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% 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 Tables 4.4 and 4.5. A recent paper [178] 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. The ECP benchmark study shows that performance still improves with increasing training set size with a decent gradient (i.e. Figure 4.14). 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 4.13), 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. In particular the greedy non-maximum suppression that poses a tradeoff between recall and precision as shown in Tables 4.4 and 4.5 will be addressed in the following Chapter 5.

A number of methodical avenues could improve classification performance. Figure 4.18 and the baseline experiments show that small objects are still very challenging despite the great amount of small sized pedestrians present in the training dataset. Approximately 75% of the false positives at 0.3 *fppi* analyzed in Figure 4.13 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 analyzed in detail to find still remaining weaknesses and further possibilities for improvement. This chapter shows quantitatively in Figure 4.13 and qualitatively in Table 4.4 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. Training a detection network jointly for pedestrians and riders already shows 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 [90, 91] come into play which take into account additional pose information. In preparation for this, this benchmark includes the orientation estimation of the overall body, for which it was shown that it can be jointly trained with the detection task at minimal performance loss.

4.5. BENCHMARKING RESULTS SINCE RELEASE

This section analyzes the impact and use of the ECP dataset since its release for online, public benchmarking in March 2019 on the corresponding website[a]. As of April 2022, 1486 persons have registered online for use of the dataset and participation in the benchmark. The license terms of the benchmark[b] are strict to avoid any violation of privacy rights and to comply especially with the general data protection regulation (GDPR) of the European Union. Therefore, the dataset may only be used for scientific, non-commercial purposes and approximately half of the registrations had to be denied. Table 4.14 shows a distribution of the remaining 700 approved registrations in dependence of the registrants' countries. So far, persons from 60 different countries have been granted access, which proves a broad interest on a global scale. The highest numbers of registrations are from China, Germany, the USA, and India which already comprise nearly 50% of the total number of approved registrations.

Among the registrants, some have already participated in benchmarking and submitted detection results on the private test dataset, for which the annotations are kept secret. Table 4.15 shows the benchmark ranking as of April 2022. For online benchmarking, the "enforce" setting as in Table 4.12 is used. The additional "all" scenario is evaluated for all pedestrians greater than 20 pixels and less than 80% occluded.

Since the benchmark study of this chapter the performance on ECP has been further improved by [18, 66, 81, 167, 173]. The progress can be attributed to different aspects, mainly the data used for training [66], the backbone network to learn the feature representations [66, 81, 167, 173] and methodical extensions [18, 81, 167, 173]. Similar to the study in this chapter, the work of [66] investigates generalization capabilities of detectors by cross dataset evaluation on ECP, Caltech [38] and CityPersons [179]. By using further non-domain specific datasets for training, namely CrowdHuman [136] and

[a] https://eurocity-dataset.tudelft.nl
[b] https://eurocity-dataset.tudelft.nl/eval/license/ecplicense

Table 4.14: Number of approved registrations per country for the EuroCity Persons benchmark.

Country	Number of registrations	Country	Number of registrations
China	159	Finland	4
Germany	93	Denmark	4
USA	57	Portugal	3
India	31	Norway	3
Suisse	23	Morocco	3
Netherlands	23	Indonesia	3
South Korea	20	Czech Republic	3
France	20	Peru	2
Italy	18	New Zealand	2
Spain	17	Mexico	2
Japan	17	Malaysia	2
Australia	15	Israel	2
Canada	14	Ecuador	2
United Kingdom	13	Cyprus	2
Taiwan	13	Croatia	2
Romania	13	Algeria	2
Turkey	11	Ukraine	1
Singapore	10	Tunisia	1
Russia	10	Thailand	1
Vietnam	8	Syria	1
Brazil	8	Sri Lanka	1
Austria	8	Slovakia	1
Poland	7	Pakistan	1
Egypt	7	Nigeria	1
Ireland	6	Lithuania	1
Belgium	6	Kazakhstan	1
Sweden	5	Ghana	1
Greece	5	Georgia	1
Iran	4	Ethiopia	1
Hungary	4	Estonia	1

Widerperson [182], they achieve better results with a HRNet backbone [156] and a Cascaded R-CNN head [18]. This Cascaded R-CNN method [18] trains a cascade of R-CNN detectors of increasing quality by varying the IoU threshold for the matching of positive samples during training. Thus, every detector in this cascade generates detections of higher accuracy while also depending on higher accuracy input from the previous stage. Throughout the cascade, more and more nearby false positives are rejected. APD [173] estimates an additional attribute per object in an embedded feature space that serves to discriminate instances in dense scenes. ResNet-50 [68] is one of the backbones used. The work of [167] also builds upon the ResNet-50 architecture. They design special modules coined Deformable Convolution with Attention Module (DCAM) that are incorporated into the ResNet-50. The deformable convolution may adapt the receptive field being used, while the attention module reweights the feature maps, which guides the attention towards pedestrian regions. In combination, this is targeted towards better performance for occluded pedestrians.

Deep neural networks which are commonly used for object detection are usually not designed for this detection task. The domain of neural architecture search [40] tries to automatically optimize the architecture itself for the task at hand. In the case of detection, this is difficult, as repeating pretraining on ImangeNet for image classification after every adaptation is computationally infeasible. The authors of SP-NAS [81] propose a method to iteratively morph a network in a way that frequent pretraining on ImageNet can be avoided. They initialize this network search with a ResNet-50 architecture for the ECP dataset and combine the resulting network SPNet with an FPN head [105] and the Cascaded R-CNN head [18] resulting in the first and fourth rank on the ECP benchmark.

Table 4.15: Log average miss-rate ($LAMR$) on the test set of the EuroCity Persons benchmark for different test data splits of the top performing methods submitted to the benchmark website.

	Test Case			
	reasonable	small	occluded	all
SPNet with Cascade [81]	4.2	9.5	21.6	13.9
Pedestron [66]	5.1	11.2	25.4	16.2
Attribute-aware pedestrian detection (APD) [173]	5.3	12.4	26.8	17.3
SPNet with FPN [81]	5.5	12.1	24.6	16.5
DAGN [167]	5.9	14.2	26.3	17.5
Method based on Cascade R-CNN [18]	6.6	13.6	31.3	19.3
Method based on Cascade R-CNN [18]	8.6	16.8	37.9	23.0
YOLOv3 [13]	9.7	18.6	40.1	24.2
Faster R-CNN [13]	10.1	19.6	38.1	25.1
SSD [13]	13.1	23.5	46.0	29.6
PyTorch Faster-RCNN [132]	14.1	29.6	43.9	30.9
R-FCN (with OHEM) [13]	16.3	24.5	50.7	33.0

Apart from the number of registrations and benchmark submissions, the citation count also shows the interest and relevance of the ECP benchmark. As of April 2022, it has been referenced 175 times according to Google Scholar. In particular, its size and diversity are acknowledged in several recent surveys (e.g. [185] and [1]).

5

Detection and Pose Estimation in Dense Traffic Scenes

5.1. Overview

The focus of this chapter is detection and pose estimation of vulnerable road users in dense traffic scenes, which cause challenges due to significant, mutual occlusions in the presence of multiple persons as analyzed in the last chapter. The detection methods evaluated there like R-FCN or Faster R-CNN have profited from incorporating the proposal generation in an end-to-end learning strategy. Still, every proposal is associated with at least one object during training, and therefore there is only one detection per proposal during inference. To ensure that every object can be detected, there are usually (a lot) more proposals than objects within an image. E.g. the YOLOv3 method evaluated in Chapter 4 has been configured with 120,000 prior boxes per image, which serve as proposal boxes. As the average number of persons per image on the ECP dataset is 5.0, this means the number of prior boxes per image is 24,000 times higher than the number of persons on average. 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 (see Figure 5.1). 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 a successive, decoupled step of non-maximum suppression (NMS). Interestingly, most top-performing methods of the common generic object detection benchmarks depend on a simple greedy non-maximum suppression (greedy NMS) [10]. This greedy NMS poses a problem for overlapping objects e.g. in pedestrian groups (see Figure 5.1). When selecting the *IoU* threshold there is a tradeoff between recall and precision as shown in the last chapter (e.g. Tables 4.4 and 4.5).

This threshold is often set to 0.5 for pedestrians, meaning if two pedestrians have a higher mutual IoU only one will be detected assuming perfectly localized detections. Such

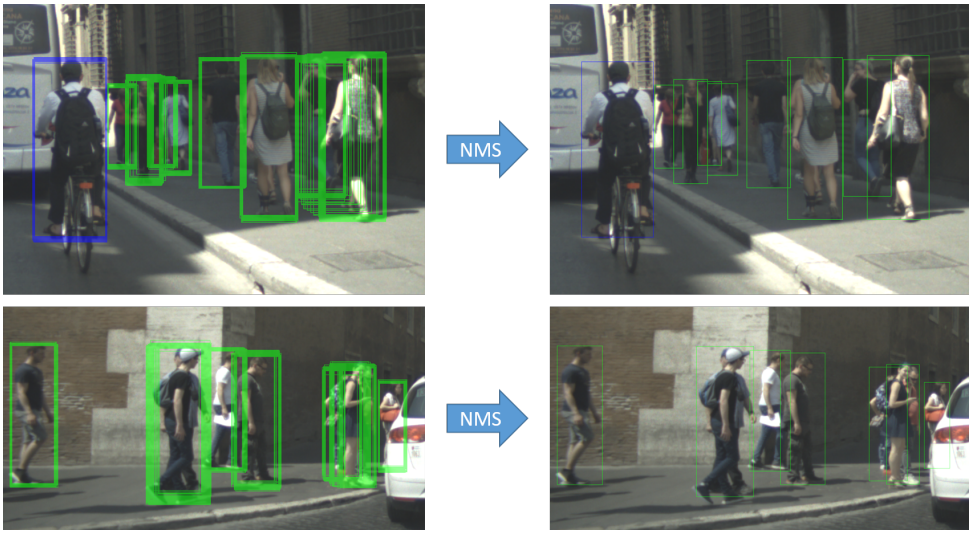


Figure 5.1: Detection results before and after applying the NMS (left and right column). Proposal boxes are classified independently of each other resulting in multiple detections for the same object (left). The greedy NMS suppresses such multiple detections. In this example here, it is configured with an IoU threshold of 0.5, which results in missing detection boxes for pedestrians with an IoU higher than 0.5 (bottom right). There are only seven detections but nine pedestrians present.

pedestrians with a higher mutual IoU than 0.5 are defined as *pedestrian pair*s throughout this chapter. If there are multiple pedestrians with an IoU higher than 0.5, only the two pedestrians with the highest mutual IoU are regarded as pairs.

Estimation of Discriminative Attributes and NMS Adaptations. Section 5.2 presents work on using discriminative attributes for learning the task of the NMS with the GossipNet architecture [73] to replace the greedy NMS. This work preceded the methodical contributions of this chapter published in [12].

GossipNet [73] is a neural network trained to rescore detections. Filtering detections based on their final score renders a further NMS stage unnecessary. Still, GossipNet solely relies on bounding box locations and confidence scores as input. Looking at the instance segmentation domain, [33] proposes a discriminative loss function to estimate a vector in an embedded feature space per pixel in addition to the semantic class confidence. This vector is used as a discriminative feature to distinguish instances. [169] shows how this idea may also be applied for proposal-based object detection. For every prior box, a feature in a geometric embedding is estimated. A high distance in this embedding indicates different objects and thus supports the NMS. Section 5.2 follows this idea but uses explicit attributes that may be used as discriminative features. It considers the body orientation, the level of occlusion, and the position of the head box as discriminative attributes. Figure 5.4 shows the intuition of why these attributes are discriminative. The high IoU threshold used for suppression on the right results in too many detections and a low precision. Still, it shows that a difference in these attributes for different detections

indicates the presence of several instances.

In Section 5.2.1, the YOLOv3 detector [128] is extended by task uncertainty weighting similar to [93]. The work on this extension has been published in [14] (©2020 IEEE). Predictors are added to estimate the attributes per prior box in addition to the bounding box regression and class confidence, and the performance is evaluated. Section 5.2.3 applies a ceiling analysis to evaluate the discriminative potential of the different attributes based on ground truth annotations. The head box attribute shows high potential and is therefore incorporated into the GossipNet architecture [73] in Section 5.2.4. Similarly, [175] estimates the head box in addition to the body box, and thus supports the NMS but does not use the head box to learn the NMS task as in this chapter.[a]

Pairwise Detection. Estimating discriminative attributes and the GossipNet experiments show issues with the inherent ambiguity in proposal-based detection approaches in dense traffic scenes. In such scenarios, pedestrians cover similar regions within an image, which may result in a similar IoU with a proposal. The prediction is done based on features that result from both pedestrians. During training, minor differences in the IoU may cause the association with different pedestrians for two nearly identical proposals. These two proposals depend on nearly identical features within the network. If a proposal box is in between two pedestrians, it is rather ambiguous which person should be the detection target, and also which head location should be predicted. During inference, this may result in head box locations right in between two other heads or in body boxes between two pedestrians. [158] takes care of this issue with a repulsion loss that pushes detections of different objects away from each other. [108] proposes to estimate the density as an additional attribute, which is defined as the highest IoU with any other pedestrian. This renders the problem less ill-posed when the value to be estimated is identical e.g. as two overlapping pedestrians share the same density. The density value is used as threshold instead of the single IoU threshold within the greedy NMS. Estimating the density is some kind of inverse approach to estimating the number of pedestrians at an image location. Solving this counting problem would also ease the NMS task. Section 5.3.2 takes a very explicit approach to this counting problem and also to solve the ambiguity issue, by directly estimating both pedestrians of a pedestrian pair based on a single proposal. This is similar to the set detection approach of [28]. They propose to predict a full set of objects based on a single proposal. As [28] has been a concurrent work to this thesis which has been published before, Section 1.2.3 does not claim the pairwise detection approach as a major contribution.

Pairwise Pose Estimation. This section estimates the position of 17 joint points as surrogate of the pose instead of the body orientation as in Chapter 3. There are basically two different approaches to human pose estimation. Top-down approaches first detect all persons in an image and estimate the pose of each person in a second stage, whereas bottom-up approaches first try to find all joints within an image, which are then clustered into instances. Top-down approaches are still leading on the MSCOCO dataset [107] frequently used for benchmarking, but of course depend on the performance of the underlying detector. A missed detection also results in a missing pose.

[a] GossipNet experiments are part of the Master's thesis "Pedestrian detection in autonomous driving by techniques optimized for crowds with deep neural networks" by Phillip Czech, supervision by Markus Braun, Ruhr-Universität Bochum, Germany, 2020 [30].

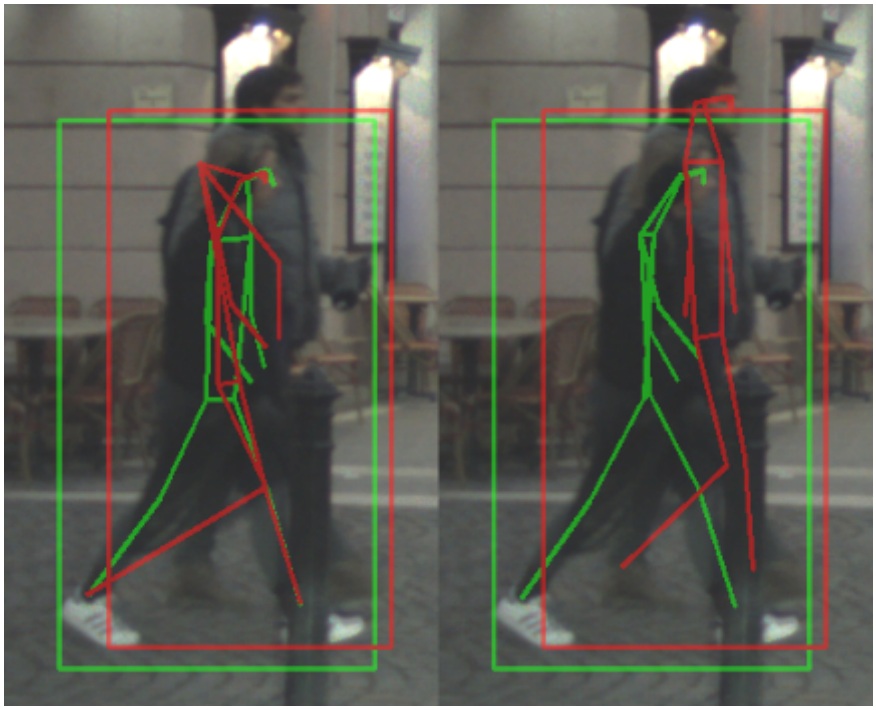


Figure 5.2: Qualitative pose estimation result of AlphaPose+ [101] (left) and the presented Simple Pair Pose method (right) for a pedestrian pair. While AlphaPose+ confuses joints of the two pedestrians and fails to estimate the pose of the pedestrian in the back (red), Simple Pair Pose provides a quite accurate pose estimation for both pedestrians. The focus of this chapter lies on such pair situations in dense urban traffic scenes.

Regarding pose estimation in groups, the cropped detections used as input often contain parts of other persons. Similar to the ambiguity in proposal based detection, the target pose sometimes becomes ambiguous [70], in particular, if the overlap of persons is very high as in the pair situations. [70] solves the disambiguation by adding an additional input hint for the target pose, while other methods optimize poses of multiple persons in a post-processing step [101, 124]. The pose estimation approach presented here makes use of the paired detections and jointly estimates the two poses within a single network. The new approach makes use of a heuristic to determine the z-ordering of the ground truth annotations and solves the disambiguation of poses by training separate experts for front pedestrians and back pedestrians in pairs according to the z-ordering. It does not depend on a complex post-processing step as [101, 124]. For an exemplary result see Figure 5.2.

As the ECP detection dataset does not provide joint point annotations, a pose annotated dataset coined EuroCity Persons Dense Pose (ECPDP) is created (see Table 5.1 for a comparison with other human pose datasets). The ECPDP extends the ECP dataset by additional images from two side-facing cameras that have been synchronously recorded. These additional images increase the availability of crowded scenes, which is the focus for

the selection of the 47k images that form the new ECPDP dataset. The dataset provides pose annotations for pedestrians and riders (see Figure 1.6). Still, the focus is on pedestrians throughout this chapter as heavy mutual occlusions are more frequent among these.

The work on combining pairwise detection with pairwise pose estimation and the creation of the EuroCity Persons Dense Pose (ECPDP) has been published in [12] (©2021 IEEE).

Table 5.1: Overview of human pose datasets including the new **ECPDP** dataset.

Dataset	**ECPDP**	**TDUP** [157]	**PedX** [87]	**MSCOCO** [107]
Domain	Automotive	Automotive	Automotive	General
# Images	47k	21k	5k (stereo)	200k
# Person Poses	279k	93k	14k	250k
Avg. Persons/Img	**5.9**	4.4	2.8	1.3

Dataset	**MPII** [3]	**AI Chall.** [163]	**CP** [101]	**OP** [124]
Domain	General	General	General	General
# Images	25k	**300k**	20k	9k
# Person Poses	40k	**700k**	80k	18k
Avg. Persons/Img	1.6	2.3	4.0	2.0

5.2. ESTIMATION OF DISCRIMINATIVE ATTRIBUTES AND NMS ADAPTATIONS

This section addresses the NMS which is often done by the greedy implementation shown in Algorithm 1. This algorithm suppresses detections based on confidence values and the body bounding box positions. It loops over all detections in descending order of their scores, while every detection removes all remaining detections with an IoU greater than a given threshold N_t (red part).

The first part of this section extends YOLOv3 by additional prediction heads for several discriminative attributes. Similar to [83] the overall detection performance could profit from these attributes as they relate to the detection task and could provide additional explicit information about the appearance of pedestrians. In the second part, the NMS task is learned with the GossipNet architecture [73] incorporating the head box information to replace the greedy NMS algorithm. As a further baseline comparison, the greedy NMS is adapted to make use of head boxes. Instead of using the IoU of the body boxes (red part in Algorithm 1), the IoU of the head boxes is utilized.

5.2.1. IMPROVING AND EXTENDING YOLOV3

This section first revisits YOLOv3 [128] which is used as underlying 2D object detector throughout this chapter. In comparison with Faster R-CNN the one-stage architecture of YOLOv3 facilitates its extension in this chapter and the detection performance of the

Algorithm 1: Greedy NMS algorithm (adapted from [10])

Input: $B = \{b_1, \ldots, b_n\}, S = \{s_1, \ldots, s_n\}, N_t$
B is the list of initial detection boxes
S contains corresponding detection scores
N_t is the NMS threshold

begin
  $D \leftarrow \{\}$
  while $B \neq empty$ **do**
    $m \leftarrow \text{argmax}\, S$
    $M \leftarrow b_m$
    $D \leftarrow D \bigcup M; B \leftarrow B - M$
    for b_i *in* B **do**
      if $IoU(M, b_i) \geq N_t$ **then**
        $B \leftarrow B - b_i; S \leftarrow S - s_i$
      end
    end
  end
end
return D, S

two methods is similar (see Table 4.3). YOLOv3 even outperforms Faster R-CNN when the same input resolution is used. Similar to [93] it is taken care that all losses match a probabilistic log-likelihood formulation to make use of task uncertainty weighting as proposed in [83]. Weighting tasks by their uncertainty may improve the performance of every single task as shown in [83] for scene segmentation and pixel-wise distance estimation. Then, further tasks and losses are added to the object detection network to estimate discriminative attributes with a single jointly trained model.

Task Uncertainty Weighting. YOLOv3 extends the Darknet53 architecture and predicts bounding boxes based on three feature layers that are downscaled by a factor of 8, 16, and 32 respectively. Each cell within these feature layers encodes prior boxes of different aspect ratios that are centered within the cell. Given an input image $\boldsymbol{x}$ the convolutional neural network f parameterized with w predicts four coordinate offsets $f^w_{loc}(\boldsymbol{x})$ for the full body box and c class scores $f^w_{cls}(\boldsymbol{x})$ per prior box p. In contrast to [128] the objectness classification is skipped and the four bounding box edges are directly regressed as in [69]. The class likelihood is calculated by

$$p(y|f^w_{cls}(\boldsymbol{x})) = \text{softmax}(f^w_{cls}(\boldsymbol{x})). \tag{5.1}$$

Similar to [93] the regressed bounding box values are modeled to follow a multivariate normal distribution. A diagonal covariance matrix with identical entries σ_{loc} is used and the minimization of the negative log-likelihood results in an L2 loss $\mathcal{L}_{loc}(w)$. Regarding classification a standard cross-entropy loss $\mathcal{L}_{cls}(w)$ is applied, which is the log-likelihood of the probability function in Eq. (5.1). The detection losses match the regression and classification losses described in [83], so task uncertainty weighting can be applied for

the total loss

$$\mathcal{L}(w) = \frac{1}{\sigma_{cls}^2}\mathcal{L}_{cls}(w) + \log\sigma_{cls} + \frac{1}{2\sigma_{loc}^2}\mathcal{L}_{loc}(w) + \log\sigma_{loc} \tag{5.2}$$

with the aleatoric, homoscedastic uncertainty weights σ_{loc} and σ_{cls} optimized during training. During training, all person samples those bounding boxes that have an IoU > 0.5 with a prior box are associated as positive training targets. Prior boxes with no associated sample only contribute to the classification loss.

Estimating Further Discriminative Features. The baseline detector is extended estimating three discriminative attributes for every prior box, namely the orientation of the body (yaw only), the level of occlusion of the complete bounding box, and the bounding box for the head. Similar to Chapter 3 this is done by adding further prediction layers in parallel to the two detection layers as shown in Figure 5.3.

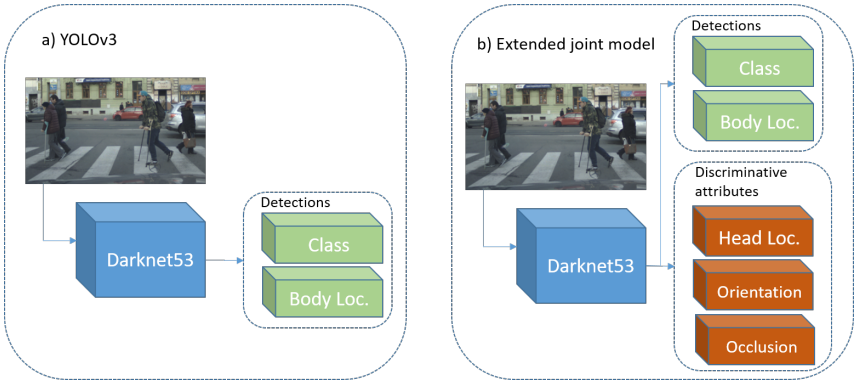


Figure 5.3: Original a) and extended YOLOv3 network architecture b) with additional prediction heads for discriminative attributes.

This section describes the occlusion estimation as a discrete classification problem. In general, the amount of occlusion by other objects or infrastructure is a continuous value between zero and 100%, which suggests a formulation as a regression problem. Still, the ECP dataset provides four discrete levels of occlusions, which facilitated the annotation process as explained in Section 4.2.3. Therefore, the same softmax formulation as in Eq. (5.1) is used for the occlusion prediction of the four occlusion classes and similar to Eq. (5.2) the homoscedastic uncertainty weight σ_{occ} is used for the occlusion loss $\mathcal{L}_{occ}(w)$.

The formulation of the head box estimation is identical to the estimation of the bounding box of the full body. As before, four coordinate offsets per prior box for the head box are estimated with a further localization loss $\mathcal{L}_{head}(w)$ in addition to the localization loss $\mathcal{L}_{loc}(w)$ and a further uncertainty weight σ_{head}.

For the body orientation, the same von Mises based formulation as in Chapter 3 is used. Therefore, a biternion is estimated for the orientation angle resulting in an additional loss $\mathcal{L}_{orient}(w)$. This orientation loss is weighted using a fixed manual weight λ_{orient} as in Chapter 3.

The total loss is the sum of the weighted singular losses and the logarithms of the

uncertainty weights. If an attribute is not labeled for a training sample associated with a prior box, no loss is added for the corresponding task.

Metrics. The detection performance is evaluated using the log average miss rate (LAMR) as described before in Chapter 4. For benchmarking, the three different data subsets *reasonable*, *small*, and *occluded* have been defined in Section 4.2.5. In many pair situations, one of the two pedestrians has a rather low while the other has a high level of occlusion. Hence, pedestrians of a pair would be divided into the *reasonable* and the *occluded* subsets as defined in ECP. To have a common subset for pairs and hereby a targeted evaluation for dense traffic scenes another subset named *relevant* is added. It consists of all pedestrians of at least 40 pixels in height and less than 80% occlusion.

5.2.2. EXPERIMENTS

This section builds upon the YOLOv3 TensorFlow implementation provided by [93] for the experiments. As in Chapter 4 nine prior box sizes are calculated with the dimension clustering proposed in [128] on the training split of the ECP dataset and distributed on the three output layers. Hence, the same prior box recall as before is achieved, which is about 100% for an IoU of 0.5. The networks are trained to discriminate between pedestrians and riders. Still, the focus is on the evaluation of the former as pedestrians are a lot more frequent. Flipping and crop and scale augmentation have been used in all trainings. Predictions are filtered with a greedy non-maximum suppression depending only on the IoU between bounding boxes parametrized with an IoU threshold of 0.5. Further discriminative attributes are not used for filtering in this section. An edge-aware debayering is used for the images of ECP, which reduces artifacts along edges. It improved the visual appearance but did not show any influence on detection performance with YOLOv3. Experiments are run and evaluated on day-time data only.

First, a detection-only model (named *Extended*) is trained on the ECP training dataset. The Darknet53 part of the extended YOLOv3 network - using the detection losses and task-uncertainty weighting described in Section 5.2.1 - is initialized with weights optimized for classification on ImageNet [35]. The network is trained for 800,000 iterations with an initial learning rate of 1e-5, which is decreased by a factor of 0.1 after 300,000 and 600,000 iterations. A focal loss [106] weighting with $\gamma = 2.0$ instead of the standard cross-entropy loss is used, as it improves the detection performance. The best-performing model with the lowest LAMR on the reasonable scenario on the validation subset is selected and evaluated on the ECP test dataset.

The joint models that additionally estimate discriminative attributes are trained with the same settings and strategy. The only difference is the addition of the losses for the different tasks. Three different variants of joint models are trained with an increasing number of tasks. The first one additionally estimates the head box. The next model additionally estimates the body orientation, while the third joint model is trained with all three additional tasks.

Detection Results. Detection results for the *Extended* model and different variants of joint models are shown in Table 5.2. The *Extended* model achieves a LAMR of 6.9 in contrast to 8.1 of the original YOLOv3 benchmark model evaluated in Section 4.3.1, where the Darknet implementation of [128] was used. The joint models including head box and body orientation estimation achieve a similar detection performance. Adding

Table 5.2: LAMR detection results for pedestrians on the different scenarios of the day test subset. All values are given in percentage points (lower values are better). Discriminative attributes are added to the Extended model. The model of the last row estimates all three discriminative attributes.

Model	reasonable	occluded	relevant
YOLOv3 (from ECP benchmark Section 4.3.1)	8.1	36.1	17.3
Extended	**6.9**	31.9	**15.1**
+ Head	7.0	**31.4**	15.2
+ BodyOrientation	7.1	32.2	15.5
+ Occlusion	8.0	34.9	16.7

the last additional task of occlusion estimation slightly degrades the overall detection performance. As before in Chapter 4 and in contrast to [58] and also [83], adding further tasks does not improve the overall detection performance despite the correlation of the tasks.

The hypothesis that further discriminative attributes might be used to improve the NMS in dense traffic scenes is qualitatively verified. Figure 5.4 exemplarily shows two results for the last joint model for two different NMS thresholds - the default threshold of 0.5 on the left and 0.9 on the right. Despite the presence of additional attributes, the greedy NMS still is based on body bounding boxes and class confidences only. The NMS run with a threshold of 0.9 suppresses fewer predictions resulting in a higher recall but also a lower precision caused by multiple detections for the same objects. The difference in the estimated discriminative attributes (occlusion, body orientation) on the right are an indicator of the presence of several persons and could prove beneficial for the detection performance if the NMS used this information to (not) suppress instances on the left. Thus, the recall for objects with differing discriminative attributes in such dense scenarios could be improved.

5.2.3. DISCUSSION AND ANALYSIS: ARE ATTRIBUTES DISCRIMINATIVE?

The extended YOLOv3 model achieves improved performance results in comparison with the YOLOv3 baseline model evaluated on the ECP benchmark in the last Chapter 4. Estimating additional attributes does not further improve the detection performance in the experiments with the greedy NMS, which still only uses the bounding box and classification scores as input. Still, it has been qualitatively verified that further attributes could be discriminative for different instances of pedestrians and thus helpful for NMS. Therefore, these differences in the three attributes body orientation, level of occlusion, and head boxes are analyzed quantitatively for pairs of pedestrians in the training set using the ground truth annotations. This so-called ceiling analysis reveals the potential beneficial effect these attributes could provide for the NMS, if they would be used there to discriminate instances and if their estimation would be perfect. As it uses ground truth annotation it shows the upper bound of potential benefits. In the remainder of this section, we refer to this analysis shown in Figure 5.5.

The absolute difference in body orientation between two pedestrians of a pair is below 25 degrees in most cases. This results from the fact that paired pedestrians usually share the same path moving together, thus sharing the same direction and orientation of the

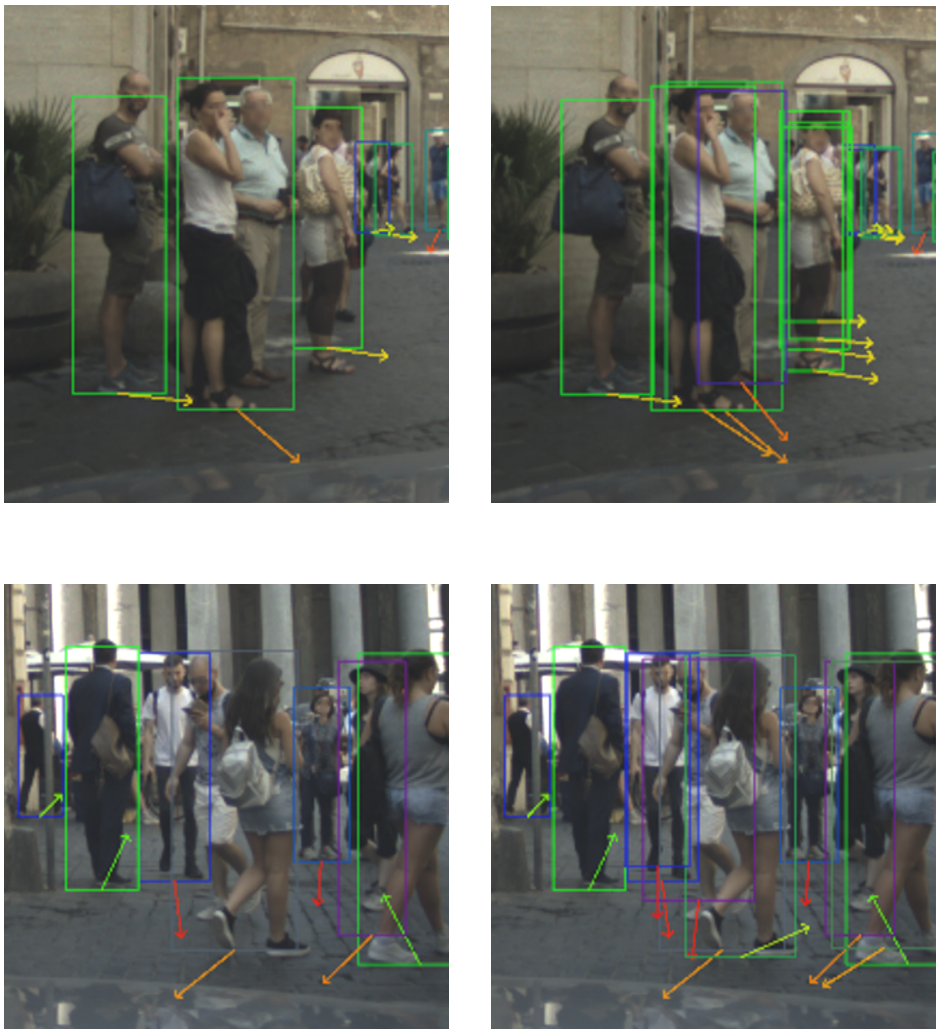


Figure 5.4: Exemplary, qualitative detection results on two images (top, bottom) including discriminative attributes after running the greedy NMS with different IoU thresholds - the default threshold of 0.5 on the left and a threshold of 0.9 on the right. Estimated body orientation is depicted by the arrows, while the color is also changed in dependence on the direction. The color of the bounding box depicts the level of occlusion. Green means no occlusion, while purple means heavily occluded. The high threshold of 0.9 is used to show that a difference in the estimated attributes may be discriminative for different persons and thus indicate the presence of several persons. For the default NMS threshold, there is only one detection for the woman and the man in the first row. The denser predictions on the right show differing estimations of the occlusion levels that are discriminative for the two persons. In the second example the same can be observed for the estimated body orientation. Once again there is only one detection for the man and woman on the left, still, the differing orientations in the right column for the dense predictions indicate the existence of several persons.

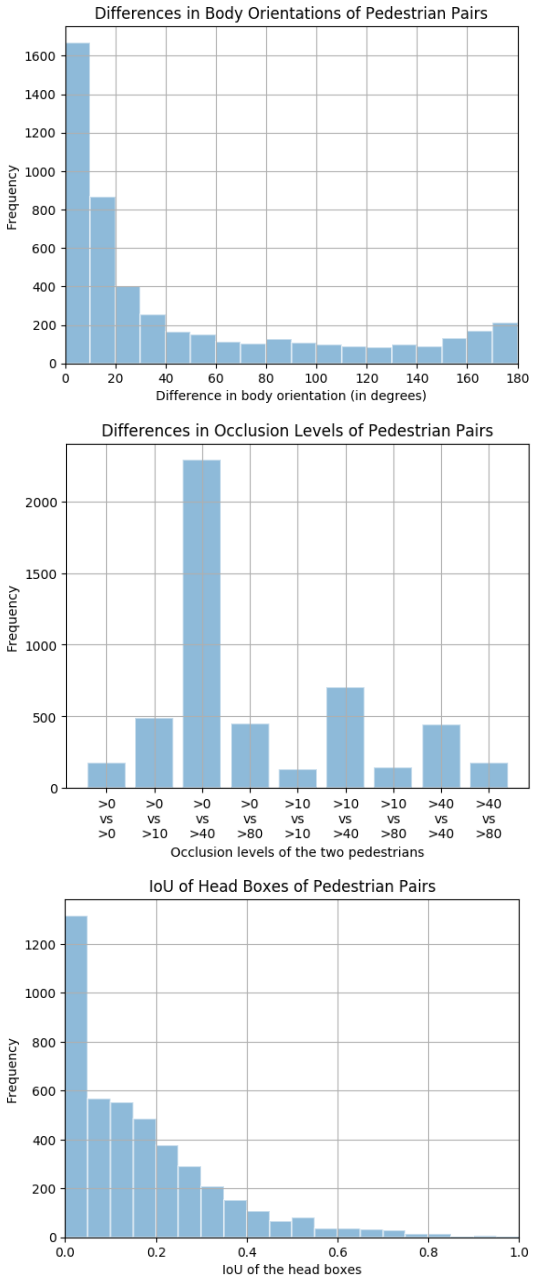


Figure 5.5: Histograms for the difference of the three discriminative attributes body orientation (top), level of occlusion (middle), and IoU of the head boxes (bottom) between the two pedestrians of pedestrian pairs on the training subset of the ECP dataset. There are four different classes representing the level of occlusion resulting in 10 possible combinations for two pedestrians (middle).

body. As the accuracy of the body orientation estimation often is in a similar range (see Figure 4.16), that attribute is not sufficiently discriminative for pedestrians within pairs in many cases.

The level of occlusion has a higher potential for discrimination. At most 10% of occlusion for one pedestrian and at least 40% of occlusion for the second pedestrian is the most frequent case. For the cases, where both pedestrians of a pair have the same level of occlusion, this attribute is also not sufficiently discriminative.

Regarding the IoU of the head boxes of pedestrians within pairs, the IoU is below 0.5 in most cases despite the body boxes that have an IoU greater than 0.5. Even if pedestrians are very close to each other, the heads are still mostly visible in many cases. Therefore, work like [129] even focuses on head detection alone to detect pedestrians as heads are more likely to be visible. This can be also observed in the example of a crowd scenario showing ground truth annotations for head and body boxes in Figure 5.6. Therefore, the focus is on using the head box as discriminative attribute in the following section.

5.2.4. LEARNING NMS WITH DISCRIMINATIVE FEATURES

This section replaces the default greedy NMS and learns the NMS task using the GossipNet [73] architecture to improve the detection performance in dense traffic scenes. The GossipNet is trained to reduce the predictions to exactly one detection per object by re-scoring the input predictions. Herefore, all neighboring predictions are processed together by the neural network. The loss function is designed to enforce a single detection per object. It depends on an association of predictions and ground truth samples to assign class labels for the predictions. This step is called matching in the remainder of this section. The raw predictions of the extended YOLOv3 model from the last section (including the head box attribute) without applying any NMS serve as input to the GossipNet.

First, this section recapitulates the network architecture of GossipNet and the loss function. Second, the matching of predictions and ground truth samples needed to calculate the loss is adapted to improve the overall performance. In a third step, GossipNet is extended to make use of the head box information as additional input, which shall serve as a discriminative attribute and additional indicator for the correct amount of predictions. This extended approach is compared with a greedy NMS that depends on the head boxes instead of the body boxes.

GossipNet Architecture and Loss Function [73]. The predictions are jointly processed by a recurrent network architecture consisting of several identical blocks. Each prediction is represented by a d-dimensional *information vector* that is initialized with zeros for the first block. Predictions with an IoU above a threshold (0.2 proposed in [73]) are paired to enable information exchange between neighboring predictions. The input of each block consists of the current information vector per prediction and pairwise features of paired predictions. Several convolutional layers and max-pooling layers within each block process the information vector. Due to the possible information exchange between information vectors of different predictions the architecture is called GossipNet. The output of each block is the processed information vectors that are passed as input to the next block. After the last block, several fully connected layers rescore every prediction and assign a new confidence g_i based on the information vectors. The pairwise features only depend on the body box and the initial confidence. No further information like image

Figure 5.6: Exemplary visualization of the ground truth body (top) and head box annotations (bottom). In most cases, the level of occlusion of head boxes is lower than for the full body boxes. Reproduced from [30] with author's permission.

features is used to train the network.

The final predictions are sorted based on their new confidences g_i and matched with ground truth objects based on the IoU. Every ground truth object can only be matched once in contrast to the training of the YOLOv3 detector, where every ground truth object can be matched by several prior boxes that surpass an IoU threshold. The result of the matching defines the class labels $y_i \in \{-1, 1\}$. The following logistic loss function is used for training

$$L(g, y) = \frac{1}{N} \sum_{i=1}^{N} w_{y_i} \cdot \log(1 + \exp(-g_i \cdot y_i)) \tag{5.3}$$

with N as number of predictions and w_{y_i} as weight to handle the imbalance of samples as there are a lot more negative than positive samples.

These weights are calculated as following

$$w_1 = \frac{\gamma}{\mathbb{E}} \text{ and } w_{-1} = \frac{1-\gamma}{1-\mathbb{E}} \tag{5.4}$$

with $\mathbb{E}$ as the expected amount of positive predictions and $\gamma \in [0, 1]$ as additional hyperparameter to influence the balancing.

Pre-matching. The described matching and assignment of class labels within the loss calculation results in the training goal to select exactly one prediction per object. Still, it is undefined which prediction should be selected from the initial set of predictions generated by the detector. The original confidences of this detector s_i are used within the pairwise features, but not within the matching. In general, the detector not only generates predictions with high confidence for an object but also low confident ones. This is not an issue with the default greedy NMS, which sorts the predictions in descending order of their confidence resulting in a suppression of the low confident predictions. Yet, the vanilla loss function of [73] allows the selection (meaning up-scoring) of low confident predictions. Since this is not necessarily preferable, this section replaces the vanilla matching with a so-called pre-matching. Predictions are sorted based on their original confidence s_i instead of g_i before matching with ground truth samples. Hereby, the GossipNet is trained to select the matching prediction with the highest confidence instead of any prediction within the set. This is more similar to the greedy NMS algorithm that would also select the predictions with high confidences. The proposed GossipNet with pre-matching will be called GossipNet$_{pre}$ in the following.

Incorporating Head Box Information. Providing the head locations as an additional information source, should help to understand the constellation of the pedestrians within the scene. The head locations are incorporated as pairwise features in addition to the pairwise body box and confidence features. Analogously to the body box features, these are:

- the IoU of the head boxes of the paired detections
- pixelwise distances in x- and y-direction and the Euclidean distance of the head box centers normalized by the head box size of the detection that is rescored by the network
- ratios of the widths and heights of the head boxes

- ratio of the aspect ratios of the head boxes

The proposed GossipNet with pre-matching and incorporated head box information will be called GossipNet$_{pre+head}$ in the following.

GreedyHeadNMS Baseline. The head boxes may not only be used to learn the NMS, but they may be used in the greedy NMS itself as in [175]. A simple version is implemented as additional baseline for comparison that uses the head box IoU instead of the body box IoUs. The IoU threshold is adapted by grid search to reach optimal performance.

5.2.5. EXPERIMENTS

The experiments use the daytime data of the ECP dataset and focus on pedestrians only. Predictions from the extended YOLOv3 model including head box estimation only trained for pedestrians are used during training, validation, and testing of the GossipNet based approaches. This section investigates three different variants, namely the vanilla version GossipNet$_{vanilla}$ [73] and the extensions GossipNet$_{pre}$ and GossipNet$_{pre+head}$ that make use of the proposed pre-matching and in the case of the latter also incorporates head box information. The runtime of GossipNet, in general, depends on the number of paired predictions. As no NMS is applied for filtering the YOLOv3 detections, the predictions are filtered based on their confidence. Only predictions with a confidence higher than 0.2 are used. That is sufficient to reach the last reference point of 1.0 false-positives-per-image when applying a greedy NMS.

Training Settings. As for the training of the *Extended* model, Adam optimizer [88] is used as backpropagation algorithm. The batch size is two and all three different GossipNet variants are trained for 300,000 iterations with a learning rate of $7 \cdot 10^{-5}$. As in [73] the recurrent architecture consists of 16 blocks. The information vector has a dimension of 128. The hyperparameter γ for balancing positive and negative samples in the logistic loss is set to 0.48. The best model is selected on the validation dataset and evaluated on the test dataset.

GossipNet Results. The results of the vanilla GossipNet, the GossipNet with the pre-matching approach, and with the usage of additional head box features are shown in Table 5.3.

Table 5.3: LAMR detection results for pedestrians on the different scenarios of the day test subset for different models. All values are given in percentage points. GossipNet$_{vanilla}$ represents the vanilla version of [73], while GossipNet$_{pre}$ makes use of the proposed pre-matching and GossipNet$_{pre+head}$ makes use of the proposed pre-matching also incorporating head box information for the pairwise features.

Model	reasonable	occluded	relevant
YOLOv3 + Head boxes (greedy NMS)	7.4	31.1	15.4
GossipNet$_{vanilla}$ [73]	23.5	49.2	32.0
GossipNet$_{pre}$	8.9	32.5	17.3
GossipNet$_{pre+head}$	12.4	34.7	21.0
YOLOv3 + Head Boxes (greedy NMS) (2nd seed)	**7.2**	**31.3**	**15.3**
GreedyHeadNMS	7.7	32.2	16.5

Compared to the greedy NMS results shown in the last section, the performance of

GossipNet$_{vanilla}$ is a lot worse for all three test scenarios, e.g. 16.6 percentage points for the relevant scenario. Figure 5.7 shows the analysis of the confidences before and after running the GossipNet$_{vanilla}$. The first plot shows a heatmap for a single, exemplary image. There are five predictions with a high confidence g_i that can be regarded as the predictions selected by the GossipNet$_{vanilla}$. All of these had a low confidence s_i below 0.35, despite the fact that there would be a lot of high confident predictions. This trend is also shown by the lower plot in Figure 5.7 depicting all confidences for the whole test subset. All predictions with a high confidence s_i get a low confidence by the GossipNet$_{vanilla}$. This verifies the hypothesis that the matching dependent on the final GossipNet$_{vanilla}$ confidences results in the selection of low confident predictions. This also results in up-scoring of low-confident predictions that are false positives. As the LAMR is quite sensitive to false positives, this also leads to the bad overall detection performance.

[73] presented a slightly improved performance of GossipNet$_{vanilla}$ in comparison with the greedy NMS on two different datasets. In both cases, they still depend on the greedy NMS. For experiments on the first dataset, they pre-filter predictions with a greedy NMS configured with an IoU threshold of 0.8, to make them fit on the GPU. For the second dataset, they use predictions of a Faster R-CNN method, filtering predictions based on confidence before the NMS similar to the experiments here. Still, the RPN within the Faster R-CNN already runs a NMS to filter proposals, while YOLOv3 runs without any NMS to generate the predictions used in this section. Thus, in both cases, low scoring predictions might be already removed in [73] by running a NMS. The experiments here go completely without any greedy NMS and the low performance is solved by the pre-matching.

The variant with pre-matching is a lot better and nearly achieves the greedy NMS performance, but still lacks behind by one or two percentage points. The analysis of the pre- and post-confidences in Figure 5.8 shows the desired behavior. There are a lot of predictions that keep the high confidences throughout the GossipNet$_{pre}$. Interestingly, GossipNet$_{vanilla}$ as well as the GossipNet$_{pre}$ take strong binary decisions, which is shown by the concentration on confidences of 0.0 and 1.0. The full range between 0.2 and 1.0, which can be observed for the YOLOv3 confidences, is condensed around these two values.

The last variant uses additional pair features for the head boxes with the same pre-matching. It does not improve performance but worsens the results. This is interesting, as GossipNet$_{pre+head}$ could have learned to simply ignore the additional head features, as all other features are completely identical. Still, the GossipNet$_{pre+head}$ is trained to use the head features by the loss function, but in inference, it does not help the overall detection performance. The next paragraph further analyzes this issue and a potential cause, why the head boxes are not beneficial.

GreedyHeadNMS. Predictions from an extended YOLOv3 model trained with a different seed (second-last row in Table 5.3) have been used for the GreedyHeadNMS experiments. First, the threshold to be used for filtering predictions based on the IoU of the head boxes is optimized using grid search. Results for varying IoU thresholds are shown in Figure 5.9. Best performance for all scenarios is achieved with a threshold of zero. This means only if heads do not overlap at all, predictions do not suppress each other. The

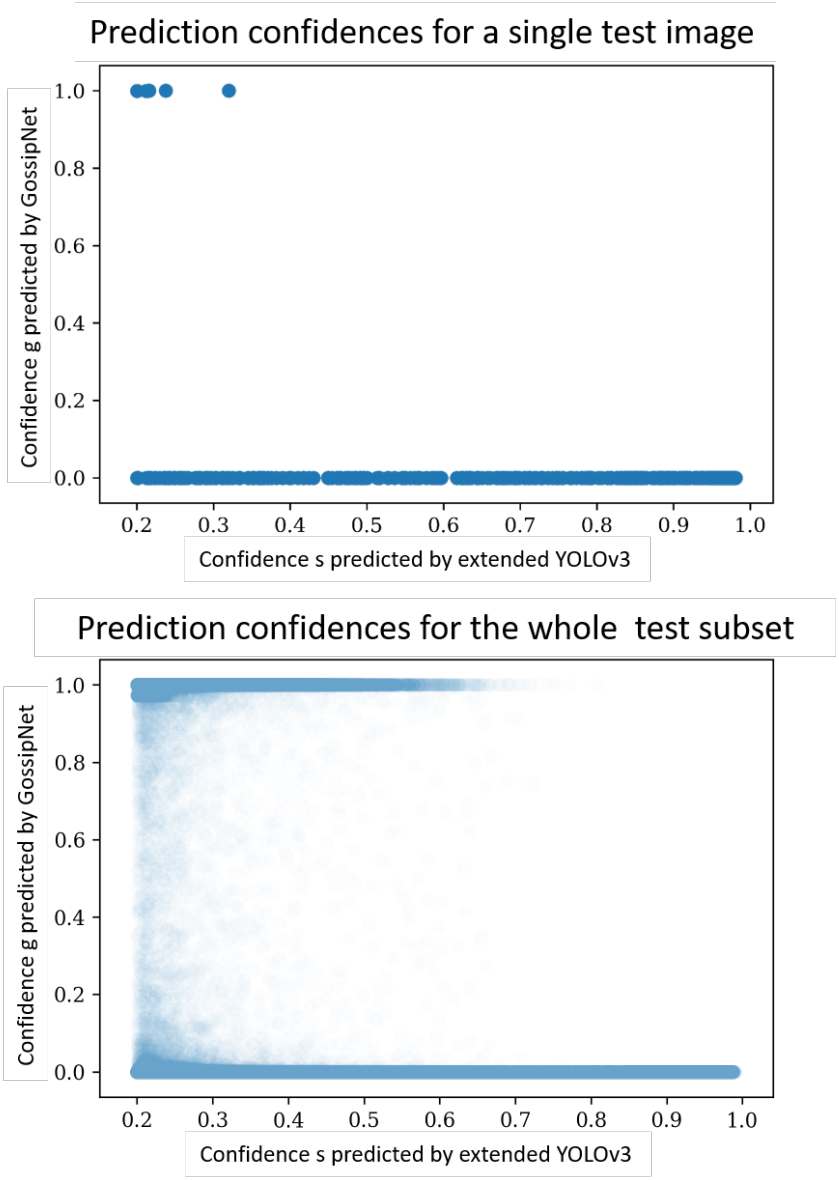


Figure 5.7: Confidences before (x-axis) and after (y-axis) applying the GossipNet$_{vanilla}$ as heatmap for a single exemplary image (top) and the full test dataset (bottom). Reproduced from [30] with author's permission.

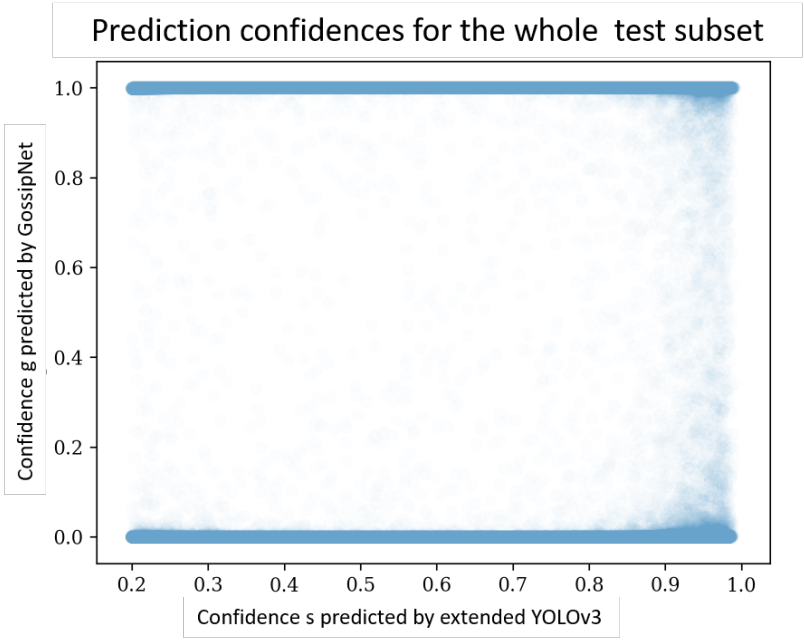


Figure 5.8: Confidences before (x-axis) and after (y-axis) applying the GossipNet$_{pre}$ with pre-matching as heatmap for the full test dataset. Reproduced from [30] with author's permission.

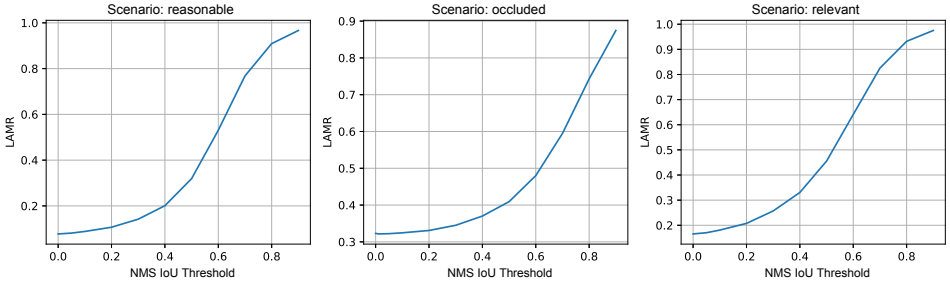


Figure 5.9: Detection results for the extended YOLOv3 with head box estimation model applying the Greedy-HeadNMS with varying IoU thresholds for different scenarios on the ECP day test dataset (left: reasonable, middle: occluded, right: relevant). While the default greedy NMS is usually run with an IoU threshold of 0.5, the GreedyHeadNMS achieves best performance for all scenearios with an IoU threshold of 0.0. This means that only if head boxes do not overlap at all predictions do not suppress each other. Reproduced from [30] with author's permission.

Figure 5.10: Qualitative detection results for the extended YOLOv3 with head box estimation model applying the GreedyHeadNMS for the *relevant* scenario and a *fppi*-rate of 0.3. The box colors depict false negatives (yellow), false positives (red), true positives (blue), matched ground truth annotation (green), and ignored ground truth annotation (gray). Reproduced from [30] with author's permission.

ground truth analysis in Figure 5.5 has shown that for pedestrian pairs the mutual head IoU is often greater than zero. To increase the recall for these cases a higher threshold had to be used. Still, the localization of the head varies even for a single pedestrian between several predictions. Thus, a higher threshold would result in multiple detections for single objects and false positives. With the threshold of zero, the GreedyHeadNMS is nearly on par with the default greedy NMS (see Table 5.3).

The qualitative results shown in Figure 5.10 reveal another issue that is related to the ambiguity in object detection. For an IoU threshold of zero, there are false positives right in between a pair of pedestrians. These predictions have a high confidence. As both pedestrians around these false positives cover a similar region within the image, features from both pedestrians influence the final prediction. In some cases, it might be ambiguous which head had to be estimated based on a prior box, which might result in ambiguous predictions right in between two pedestrians.

5.2.6. DISCUSSION: AMBIGUITY IN ATTRIBUTE ESTIMATION IN DENSE TRAFFIC SCENES

In this section, GossipNet was extended to learn the task of NMS. It did not outperform the greedy NMS. Still, the presented pre-matching extension to the GossipNet achieves similar results. Head box information has been incorporated into the architecture, but its usage results in worse performance. The qualitative results of the GreedyHeadNMS experiments show badly localized head boxes right in between the heads of two other pedestrians. As the prior boxes cover regions that contain features from both pedestrians, it might be ambiguous for the regression head, which pedestrian's head location should be estimated, resulting in erroneous predictions. The GossipNet variants can not depend on explicit image features, but only on the features provided in the information vector. Thus, such erroneous attribute estimations may also result in erroneous predictions of the GossipNet$_{pre+head}$, when it is trained to use the head information. Overall the experiments showed that ambiguity is not only an issue for body box detection itself as described in the introduction of this chapter but also for the estimation of discriminative features. This ambiguity and the ceiling analysis of the ground truth annotations in Figure 5.5 limit the potential of discriminative attributes in improving the detection performance in dense traffic scenes.

To further improve the performance of the GossipNet$_{vanilla}$ [73] proposes to incorporate image features in future works, which could also be beneficial for the proposed variants within this section. By doing so the network could directly estimate the correct number of objects based on appearance information. 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 truly end-to-end approach.

5.3. PAIRWISE DETECTION AND POSE ESTIMATION

This section presents the new *Simple Pair Pose* (SPP) for top-down pose estimation to solve inherent ambiguity issues in proposal based object detection. It consists of two parts (see Figure 5.11). For the *pairwise detection*, the detection head of the YOLOv3

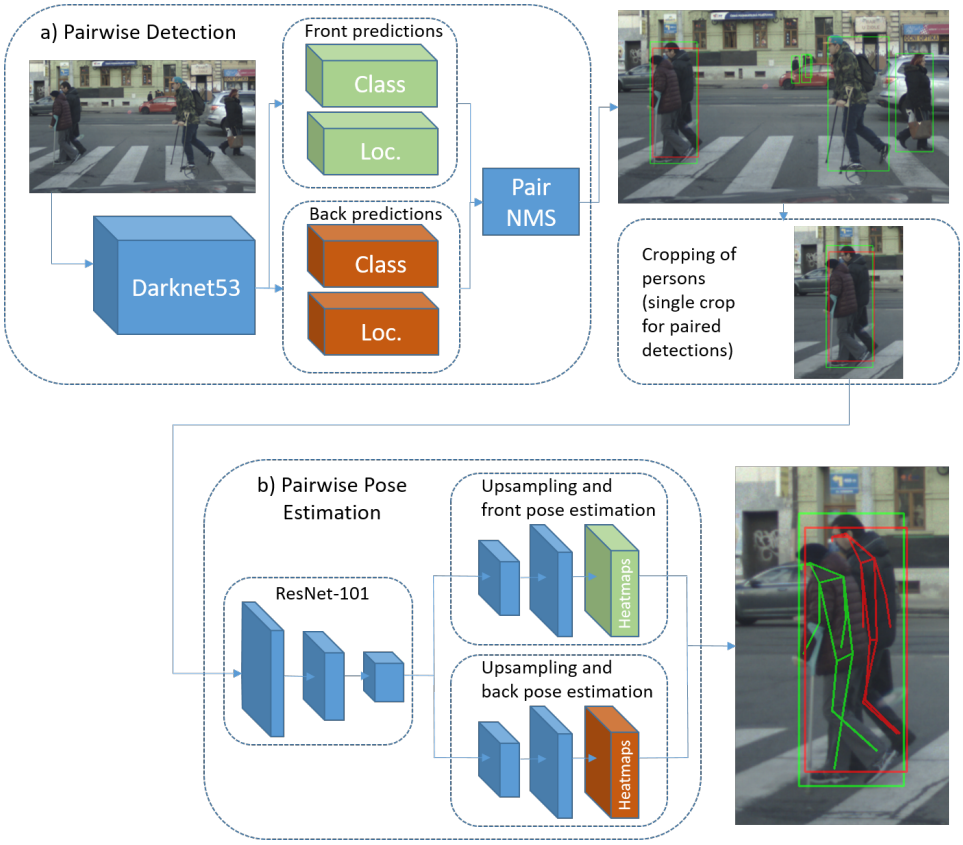


Figure 5.11: Overview of the Simple Pair Pose (SPP) method consisting of a pairwise detection a) and pose estimation method b).

[128] detector is duplicated to predict pedestrian pairs based on single proposals and to improve the recall in groups. In the *pairwise pose estimation* part, the single person pose estimation network described in [101] is extended to jointly estimate the poses of paired detections.

5.3.1. RECAPITULATION OF NMS ISSUES

The focus of Section 5.3 is on the following two issues regarding detection in crowds for proposal based detection approaches like [58, 109, 128, 132]. First, for a single pedestrian there are usually several overlapping proposals. The pedestrian is used as training target for all proposals that are associated e.g. based on an IoU threshold. During inference, this many to one mapping results in multiple detections per pedestrian that have to be suppressed by the NMS. Depending on the IoU threshold of the NMS, not all pedestrians within a crowd may be detected.

Second, within a group scenario, a single proposal often overlaps with several pedes-

trians. Still, many approaches only select a single person with the highest overlap as target for every proposal. In inference this may result in some kind of ambiguity. When a proposal is placed between two pedestrians the final detection may be influenced by both pedestrians and has a low localization accuracy [158].

To solve this issue a pair of pedestrians is predicted based on a single proposal. As such paired detections do not suppress each other within the adapted NMS, the recall in dense traffic scenes is improved. The concurrent work of [28] predicts the full set of all associated objects based on a single proposal. They duplicate the predictor head of a feature pyramid network [105] consisting of a classification and localization part. During training, all predictions from a single proposal are matched with the associated ground truth annotations minimizing an earth mover distance loss [28]. A *Set NMS* is applied where predictions from the same proposal do not suppress each other. Hence, the *Set NMS* depends on the information which predictions result from which proposal.

5.3.2. PAIRWISE DETECTION

This section builds on the extended YOLOv3 detection method described in Section 5.2.1. There, for every prior box, the detection head estimates four coordinate offsets in the localization part and the confidences for the different classes. The NMS uses an IoU threshold of 0.5 as in the last Chapter 4 resulting in a low recall for pedestrian pairs.

For such pairs, the pedestrian with the lower bounding box edge is defined to be the *front pedestrian*, whereas the other one is the *back pedestrian*. Following a flat world assumption this corresponds to the z-ordering in the traffic scene. The ordering for pedestrians with an equal lower bounding box edge or contradicting occlusion levels is manually annotated.

The idea of pairwise detection is implemented by duplicating the detection head of YOLOv3 as shown in part a) of Figure 5.11. Thus, for every prior box two predictions including two bounding box regressions and classifications are estimated. In the work of [28] a similar structure results from setting the set cardinality to two. Though, in the approach here the loss of the set prediction is defined more explicitly as described in the following. For the first prediction head the target is always the front pedestrian, while the second prediction head is responsible for estimating the back pedestrian. The matching is not permutated as in [28]. Thus, separate experts are trained for both cases disambiguating the detection task for pedestrian pairs, as it is defined beforehand which pedestrian has to be detected by which head. In [28] this has to be learned implicitly.

As before, the bounding box regression loss ($\mathscr{L}_{loc}$) is modeled to follow a normal distribution as in [93] and the classification to follow a softmax loss ($\mathscr{L}_{cls}$), which enables uncertainty weighting [83]. The total loss for a prior box associated with a pedestrian pair is

$$\begin{aligned}\mathscr{L}(w) = &\mathscr{L}^f_{cls}(gt^f, w) + \mathscr{L}^f_{loc}(gt^f, w) \\ &+ \mathscr{L}^s_{cls}(gt^b, w) + \mathscr{L}^s_{loc}(gt^b, w)\end{aligned} \tag{5.5}$$

with gt^f, gt^b as the ground truth annotations of the front and back pedestrian, w as the weights of the network, and $\mathscr{L}^f, \mathscr{L}^s$ as the losses of the first and second prediction head.

If a pedestrian is not part of a pair, it is defined to be a front pedestrian by default. In this case the regression loss for the second prediction head is zero, and its target class

is background. For inference, similar to [28] the NMS is adapted in a way that front pedestrians do not suppress back pedestrians estimated based on the same proposal (*Pair NMS* in Figure 5.11). If both class confidences of a front and back prediction from the same proposal are above a certain threshold, this is defined as a *paired detection*.

5.3.3. PAIRWISE POSE ESTIMATION

This section follows the top-down multi person pose estimation approach: In general, detections are cropped from the input image and a single person pose estimation (SPPE) network estimates the n heatmaps of the n joints. If a crop contains several pedestrians it may be ambiguous which pose has to be estimated [70]. This is also caused by imperfectly localized detection boxes.

To avoid confusions of front and back joints of the front and back pedestrian, heatmaps for both pedestrians are jointly estimated in a single forward pass. In [101] a pose head consisting of two upsampling modules and a final convolutional layer is attached to a ResNet-101 backbone to estimate the pose heatmaps. This pose head is duplicated to jointly estimate front and back heatmaps as shown in part b) of Figure 5.11. It is possible to split the paths later (or even earlier) within the network, e.g. by only duplicating the final layer. The point to branch may be empirically selected, while an earlier split increases runtime.

During training, ground truth boxes gt^f and gt^b of pedestrian pairs are combined to a single pair box gt^p enclosing the two boxes. This combined box is used to crop the image to ensure that the context of both pedestrians is fully available. The overall heatmap loss is the sum of the separate heatmap losses for the joints of the front and the back pedestrian. Training separate experts for estimating the heatmaps of front and back pedestrians disambiguates the target pose. As before in the detection method, single pedestrians are defined to be front pedestrians by default, for which the single box is used for cropping the image. The heatmap loss for the back joints is zero in this case. During inference, paired detections of the pairwise detector are combined as in the training before cropping, while single detections are kept as they are. The best front and back poses from the heatmaps are extracted using spatial argmax. Hence, the new method does not depend on a post-processing step to handle poses of pedestrian pairs. As all computations are shared apart from the duplicated pose head, the runtime for pedestrian pairs is lower in comparison with estimating the two poses based on separate image crops.

5.4. THE EUROCITY PERSONS DENSE POSE DATASET

5.4.1. DATA SELECTION

For the ECP detection benchmark in Chapter 4 images only of the front facing camera attached behind the windshield have been utilized. A fixed sample rate has been used to extract and annotate images to avoid any selection bias.

For the EuroCity Persons Dense Pose (ECPDP) dataset the main focus is shifted to crowded scenes. In addition to the front facing camera, two side facing cameras with a higher horizontal field of view of 85° had been attached at the left and right door mirrors (see Figure 4.2). As described in Chapter 4, they feature the same resolution and have

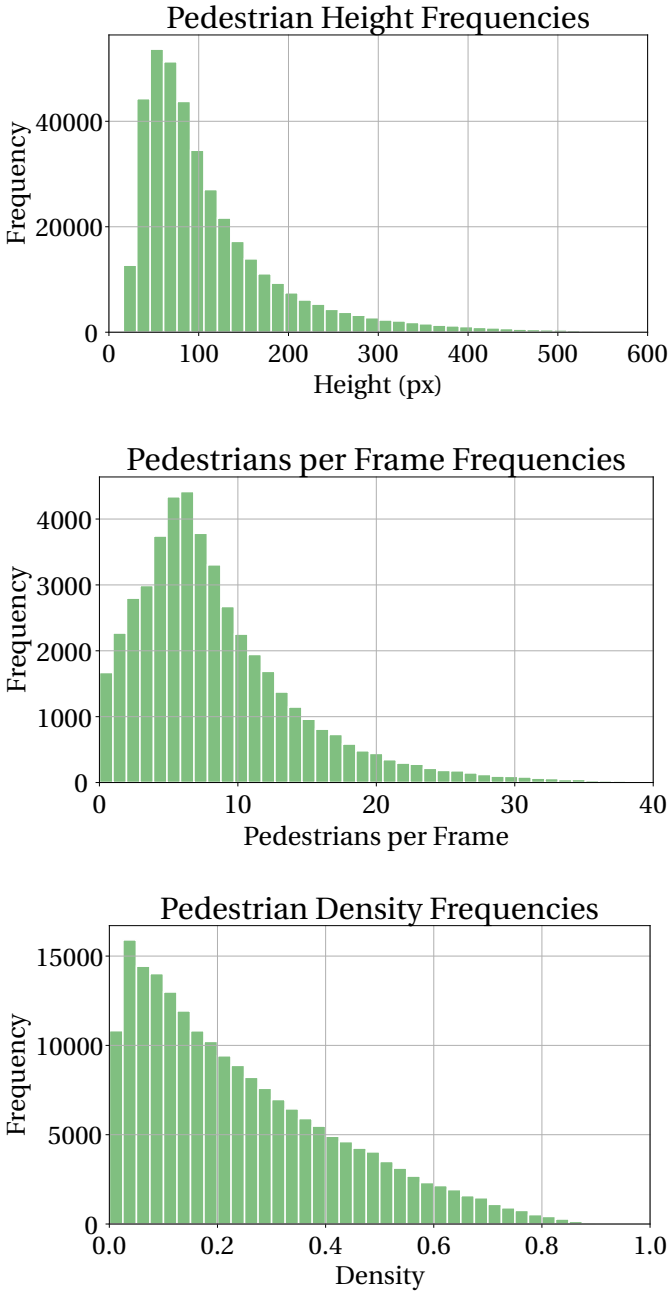


Figure 5.12: Frequencies of various pedestrian configurations within the EuroCity Persons Dense Pose (ECPDP) dataset (top: height, middle: pedestrians per frame, bottom: density). The density of pedestrians (bottom) is only shown for overlapping boxes, meaning densities greater zero. The density of a pedestrian is defined as the highest IoU with any other pedestrian within the same image.

been synchronously triggered with the front facing camera. Adding images of the side facing cameras increases the amount of crowded scenes for the ECPDP dataset.

Hence, images with a high number of persons are selected from the front as well as side facing cameras. For images already contained in ECP this is done based on the number of box annotations. For the remaining images a Faster R-CNN [132] model trained on ECP is run to detect crowded scenes.

Overall, 30,704 images are selected from the front facing camera, of which 14,438 are already part of the ECP dataset. Further 8,263 images from the left and 8,008 images from the right camera are added to the final image set consisting of 46,975 images in total. It is ensured that the train-val-test split of the new ECPDP dataset is aligned with the train-val-test split of ECP described in Section 4.6.

5.4.2. DATASET ANNOTATION

Apart from poses the annotation protocol mimics that of ECP [13]. For every image all pedestrians and riders of at least 20 pixels in height are annotated with tight bounding boxes of the complete extent. If a person is not fully visible, the extent is estimated. In that case, the level of occlusion and truncation is annotated. Groups of persons that are not distinguishable are annotated with boxes enclosing the groups, serving as ignore regions during evaluation. Ignore regions are also class-specific for pedestrians and riders. If the class can not be discriminated by the annotator, it is labeled as generic person ignore region. In addition, annotations comprise the complete poses consisting of 17 joint points as in MSCOCO [107] for persons that are greater than 60 pixels in height. For every joint, it is indicated if it is fully visible, self-occluded or occluded.

5.4.3. DATASET STATISTICS

Data distributions for pedestrians of the new dataset are shown in Figure 5.12. Due to the data selection targeted on crowded scenes, there is a peak around six pedestrians per frame. The overlap between pedestrians is analyzed, as mutual occlusions of pedestrians cause major challenges even for recent deep learning approaches. As in [108], the density of a pedestrian is defined as the highest IoU with any other pedestrian in this scene. As defined before, if the density is greater than 0.5, the two pedestrians form a pair. The amount of pairs in the ECPDP dataset is about one percentage point higher than in the ECP dataset (5.9% in contrast to 5.0%). Regarding riders, only 1.2% of these have a mutual IoU greater 0.5. Therefore, the focus is on pedestrians only in the pairwise experiments.

Compared to other automotive datasets, the new ECPDP dataset provides the largest number of pose annotated persons (cf. Table 5.1). Furthermore, the ECPDP contains the largest average number of persons per image overall. Thus, it enables the targeted evaluation of pose estimation in dense urban traffic scenes. Detailed statistics of the dataset subsets are shown in Table 5.4.

5.4.4. METRICS

The detection performance is evaluated applying the log average miss rate (LAMR) as in [13] on the relevant subset defined in this chapter. The object keypoint similarity (OKS) from [107] is used to evaluate pose estimation accuracy. The pairwise pose evaluation matches objects based on their IoU and measures the average OKS for true positives.

Table 5.4: Statistics of the subsets of the new ECPDP dataset regarding the number of images and the amount of boxes, poses and ignore regions of pedestrians and riders and the number of generic person ignore regions that may contain pedestrians as well as riders.

	train	val	test	**total**
# images	29,570	5,150	12,255	**46,975**
# pedestrian boxes	251,654	47,530	99,529	**398,713**
# pedestrian poses	167,066	30,960	65,698	**263,724**
# pedestrian ignore	17,140	3,394	7,255	**27,789**
# rider boxes	21,617	3,624	8,458	**33,699**
# rider poses	10,164	1,704	3,737	**15,605**
# rider ignore	943	150	347	**1,440**
# person ignore	9,158	1,783	3,605	**14,546**

In [107], objects are matched based on their OKS instead of the IoU, as not all of the bottom-up methods provide bounding boxes. They calculate the average precision (AP) for different OKS matching thresholds. The same evaluation procedure is applied for the overall pose estimation performance that serves as baseline for benchmarking on the new pose dataset. Instead of calculating the AP the LAMR is used. The LAMR implementation of [13] is adapted matching objects based on their OKS instead of the IoU. Samples without pose annotations or that are not part of an evaluation subset serve as ignore instances and are still matched based on the IoU if there is no other non-ignore instance that exceeds the OKS threshold for matching.

5.5. EXPERIMENTS

The experiments first focus on training and evaluation of the pairwise detection method for pedestrians. Then, the training setup for the pairwise pose estimation is described, and results of the pose estimation for pedestrian pairs are shown. Finally this section shows the overall pose estimation performance on the complete ECPDP test dataset that serves as baseline on the new pose benchmark. Riders are also included in the training of all models. Still, as rider pairs are rare, only the front prediction head and the front pose heatmaps are trained with riders and the evaluation focuses on pedestrians.

5.5.1. PAIRWISE DETECTION TRAINING

This section builds upon the experiments for the extended YOLOv3 model in Section 5.2.2. The same nine prior sizes optimized on the ECP training dataset are used. Flipping and crop and scale augmentation are applied in all trainings. For the *Base* model the training of the *Extended* model evaluated in Section 5.2.2 with a single prediction head is continued on the training subset of the new ECPDP dataset. It is trained for 50 epochs, reducing the initial learning rate of 1e-5 after 30 and 44 epochs by a factor of 0.1.

Finally, the pairwise detection network is also trained on ECPDP. The *Extended* model is used for initialization. The weights of the additional convolutional filters of the second prediction head for estimating the classification and bounding box regression of the back pedestrian are randomly initialized. Best results are achieved with a fixed weighting for

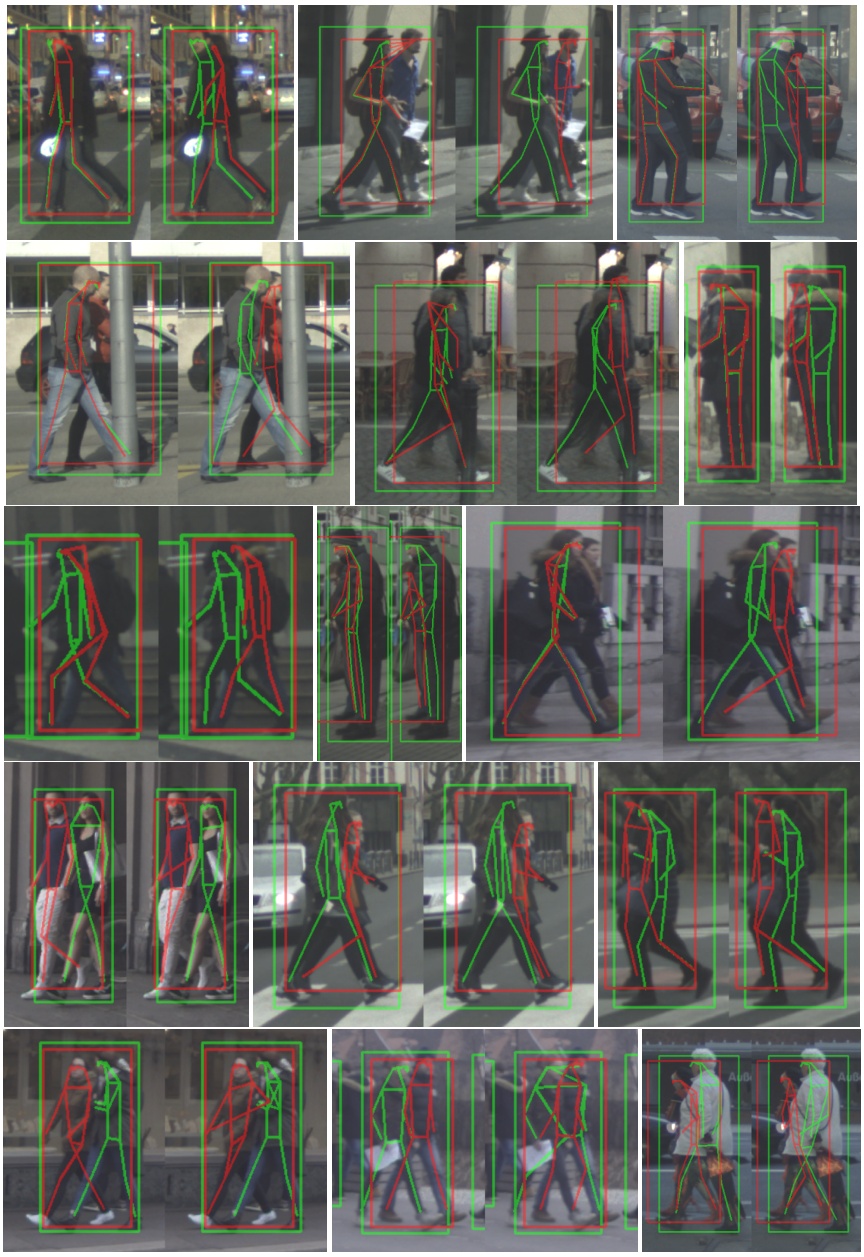


Figure 5.13: Qualitative results of AlphaPose+ (left within each pair of image crops) and the new pairwise pose estimation (right within each pair of image crops) for back pedestrians (red) and front pedestrians (green) of valid paired detections (green and red bounding boxes). The first three rows show samples where the presented method surpasses AlphaPose+, while the last two rows show error cases.

the losses of the second prediction head instead of uncertainty weighting [83] that is used for the losses of the first prediction head. The *Pair* model is also trained for 50 epochs with the same learning rate strategy as the *Base* model.

5.5.2. PAIRWISE DETECTION RESULTS

The detection performance for pedestrians is evaluated on the *relevant* subset of the ECPDP test dataset.

A version of the *Pair* model discarding all back predictions (coined *Pair w/o back*) is also evaluated. This removes the influence of back predictions that on the one hand increase the recall in groups but on the other hand decrease precision due to false positives. A greedy NMS with an IoU threshold of 0.5 is applied for this version of the *Pair* model and the *Base* model. The full *Pair* model including back predictions makes use of the adapted NMS as described in Section 5.3.2.

Quantitative results are shown in Table 5.5. The LAMR of the two *Pair* model variants is 0.8 points higher than of the *Base* model. The additional prediction head slightly reduces the overall detection performance. Still, the recall for pedestrian pairs can be increased by the back predictions. Despite the NMS threshold of 0.5, the recall of the *Base* model for pairs at a false positive per image (fppi) rate of 1.0 is also greater 50%. This is caused by imperfectly localized predictions that are not suppressed by the greedy NMS.

Results of valid paired detections are shown in Figure 5.13. In Figure 5.14 the recall is shown for different density ranges of the test samples for a fppi rate of 1.0. The *Pair* model achieves the highest recall for density ranges above 0.5.

Table 5.5: Detection results for pedestrians of the *relevant* subset on the ECPDP test subset. All values are given in percentage points. $\text{Rec}_{>0.5f:x}$ is the recall for pedestrians of pairs with a mutual IoU greater 0.5 for a given false positive per image (fppi) rate of x.

Model	LAMR	$\text{Rec}_{>0.5f:0.1}$	$\text{Rec}_{>0.5f:1}$
Base	**28.2**	51.7	62.1
Pair w/o back	29.0	49.1	62.0
Pair	29.0	**55.5**	**70.0**

5.5.3. PAIRWISE POSE TRAINING

For better comparability with AlphaPose+ [101] the new pairwise pose estimation method is integrated into the provided source code[c]. This comprises the duplication of the pose head consisting of two upsampling modules and a final convolutional layer attached to a ResNet-101 [68] backbone, combining pedestrian pairs before cropping, and skipping the graph based optimization of AlphaPose+. (Only duplicating the final convolutional layer has lead to inferior results in previous experiments.) By using the framework of AlphaPose+, the straightforward integrability into other methods is verified. The joint candidate loss proposed in [101] is not provided in their framework. For training of the AlphaPose+ baseline and the new pairwise pose estimation the mean squared error is used as heatmap loss. The training settings are identical for both methods. The person

[c] https://github.com/MVIG-SJTU/AlphaPose/tree/pytorch

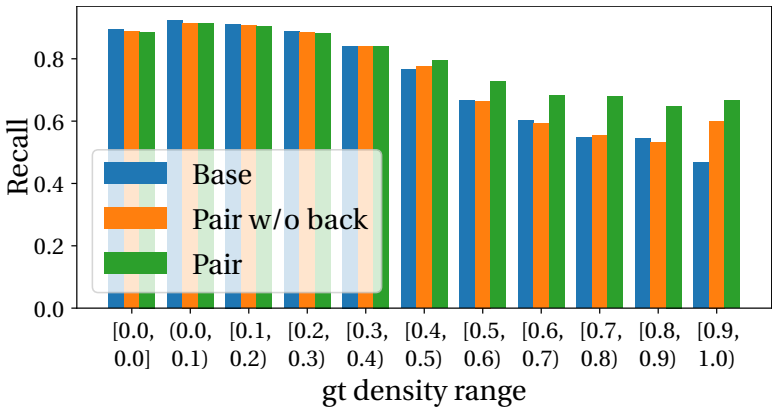


Figure 5.14: Recall of pedestrians normalized per bin for the three detection models in dependence of the density of the test samples. The density is defined as the highest IoU with any other test sample.

crops from the input image are rescaled to 320 x 256. The output heatmap resolution is 80 x 64. The training duration is 110 epochs, reducing the initial learning rate of 1e-4 to 1e-5 after 80 epochs.

Table 5.6: Pose results in terms of mean OKS (median in brackets) for detected pedestrian pairs of the ECPDP test subset. All values are given in percentage points.

Model	OKS^f_{vis}	OKS^f_{all}	OKS^b_{vis}	OKS^b_{all}
AlphaPose+	85.6 (93.3)	83.8 (88.9)	68.7 (75.8)	65.5 (68.7)
SPP	**86.9** (**94.3**)	**84.9** (**89.7**)	**75.9** (**81.9**)	**68.3** (**71.7**)

5.5.4. PAIRWISE POSE RESULTS

For evaluation of the new *SPP* method on the test dataset, the pairwise pose model is run on the detections of the *Pair* detection model including back predictions. Paired pedestrian predictions are combined first to jointly estimate the front and back pose. The predictions of the *Pair* model are also used as input for AlphaPose+ for better comparability and as YOLOv3 is also the underlying detection method in [101].

The evaluation focuses on the pair scenarios. Table 5.6 shows mean and median OKS values on the 346 correctly detected pedestrian pairs. The two estimated poses are associated with the front and back ground truth poses optimizing the overall OKS value. OKS^f and OKS^b are the OKS values for front and back pedestrians. All joints or only visible joints are taken into account for OKS_{all} and OKS_{vis}. The *SPP* model performs best for front as well as back pedestrians of pairs. Most significant improvement can be observed for back pedestrians, which is 7.2 percentage points for the mean OKS evaluated on visible joint points (OKS^b_{vis}) and 2.8 points on all joints (OKS^b_{all}). AlphaPose+ performs similarly for poses of front pedestrians. In the qualitative results in Figure 5.13 there are several cases where AlphaPose+ confuses poses of the front with the back pedes-

trians. This is caused by missing joint candidates for the pedestrians in the back, whereas *SPP* profits from the expert knowledge for the back pedestrians. Figure 5.15 shows the OKS improvement of the new method in comparison with AlphaPose+ for these back pedestrians in dependence of the IoU between the paired detections. Apart from the last bin that only contains seven samples, a higher IoU between the detections results in a higher average improvement by *SPP*. This can be expected as a higher overlap between detections may also induce more difficulties in discriminating the two pedestrians within the pose estimation. This higher overlap can be also caused by a low localization accuracy of the pair detector, e.g. when the pair detector itself confuses extents of front and back pedestrians. The *SPP* method suffers less from these localization errors of the underlying detector as the two boxes are combined and the disambiguation is solved by the different experts.

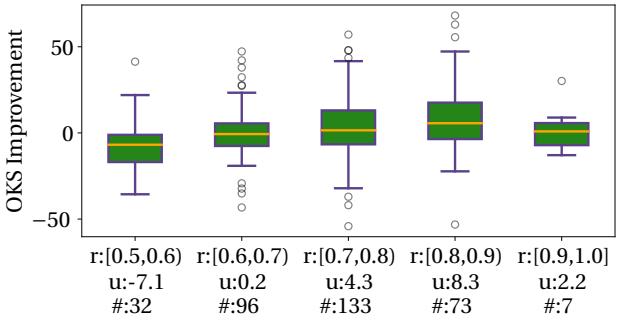


Figure 5.15: Mean OKS^{b}_{all} improvement u for back pedestrians of the *SPP* method in comparison with AlphaPose+ in dependence of the density of the paired predictions binned over different density ranges r with # as the number of samples per bin.

Table 5.7: Overall pose performance on the ECPDP test subset. All values are given in percentage points. For the LAMR L^{t}_{o} test samples occluded up to o% are matched based on an OKS threshold t. Pairwise training is only applied for pedestrians in the SPP method.

Model	Class	Scores	$L^{0.5}_{40}$	$L^{0.75}_{40}$	$L^{0.5}_{80}$	$L^{0.75}_{80}$
AlphaPose+	Ped.	Box	33.9	56.7	41.1	64.0
AlphaPose+	Ped.	Pose	29.8	49.3	36.1	56.2
SPP	Ped.	Box	32.0	56.3	39.8	63.8
SPP	Ped.	Pose	**28.9**	**48.8**	**35.9**	**56.0**
AlphaPose+	Rider	Pose	**11.2**	**19.0**	**13.7**	**23.1**
SPP	Rider	Pose	11.5	**19.0**	14.1	**23.1**

5.5.5. OVERALL POSE RESULTS

For benchmarking purposes on the new dataset, Table 5.7 shows the OKS based LAMR (abbreviated as L in the following) on all test samples annotated with poses for pedestrians and riders. Two different OKS thresholds are used for matching: 0.5 for $L^{0.5}$ and 0.75

for $L^{0.75}$ respectively. L_{40} is calculated for persons less than 40% occluded and L_{80} for less than 80% occlusion. As before, the detections from the *Pair* model are used for inference of AlphaPose+ and the new pairwise pose model. For pedestrians, results are compared using the confidences from the box detector and the confidences from the pose estimation, where heatmap scores are added to the initial class scores. Confidences from the pose estimation result in better performance for both models. The methodical focus is on pedestrian pair situations. As the amount of pedestrian pair situations is low in comparison with all test samples the overall performance is only slightly better than for AlphaPose+, e.g. by 0.9 points for $L_{40}^{0.5}$. The performance for pedestrians up to 80% occlusion is similar. Note that the experiments do not make use of pairwise training for riders in the *SPP* method due to the low relative amount of pairwise rider situations. Therefore, the results for riders in Table 5.7 for *SPP* does not show any improvement over AlphaPose+.

6

Conclusion and Future Work

This thesis addressed the topic of visual person detection and pose estimation for automated driving in urban traffic scenes with deep learning. On one hand, new methods have been presented leveraging the capability of deep neural networks to learn powerful features from raw pixel data. Thus, detection and pose estimation performance has been improved overall despite the challenges listed in Section 1.1.4, in particular the high intra-class variance of persons and dense traffic scenes. On the other hand, new datasets have been created accompanying the work of this thesis, which were made publicly available. Especially the *EuroCity Persons* (ECP) detection dataset enabled a detailed evaluation of deep learning methods. Pre-training on ECP also led to an improved detection performance on other datasets due to its high diversity and quantity of person annotations.

6.1. Conclusion

This section draws conclusions along the chapters of the thesis, while also paying attention to the question “Which performance is needed for automated driving?".

Joint Detection and Orientation Estimation. Chapter 3 presented a novel approach called *Pose-RCNN* for joint object detection and orientation estimation. The proposed approach exploits deep learning to jointly perform object bounding box regression, classification, and orientation estimation. It is supported by a combination of 3D object proposals from stereo and lidar measurements. On the KITTI benchmark, the proposal generation in combination with the joint detection and orientation approach outperforms other state of the art approaches for the *Easy* test scenario of the bicycle class. It achieves an average precision of 80.8, while the second-best performing method SubCNN [165] achieves 79.5. Regarding orientation estimation, the average orientation similarity is 75.5 which is 3.5 percentage points more than achieved by SubCNN. As explained in Section 3.4, average precision is an upper bound for the average orientation similarity. For pedestrians and riders, *Pose-RCNN* gets closer to this upper bound than the other state of the art methods, which shows the high potential of the joint detection and orientation estimation using the von Mises loss.

ECP Benchmark Results. As the KITTI 2D detection benchmark has aged since its recording a decade ago, the new ECP dataset has been created and presented in Chapter 4. It takes annotations of persons in urban traffic scenes to a new level in terms of quantity, diversity, and detail compared to datasets used for person detection in traffic scenes before the publication of the ECP dataset [13]. Four state of the art deep learning approaches (Faster R-CNN, R-FCN, SSD, and YOLOv3) were thoroughly optimized to serve as baselines for the new person detection benchmark. The recall of the proposal boxes has been optimized by adapting the scales and aspect ratios for all methods. For Faster R-CNN and R-FCN, the network architecture has been adapted reducing the feature stride by removing a max-pooling layer. Furthermore, upscaled images have been used as input as it improved performance. A variant of Faster R-CNN performed best overall, with a log-average-miss-rate of 7.9, 17.0, and 33.2 on the "reasonable", "small" and "occluded" test cases, respectively. The better performance of this two-stage method in comparison with the one-stage methods YOLOv3 and SSD comes at the cost of a slower runtime (1.7 fps vs. 3.8 fps of YOLOv3). In contrast to the *Pose-RCNN* method, Faster R-CNN integrated the proposal generation into the network itself. Nowadays, external proposal sources are no longer used in most approaches [39, 96, 109, 127, 132], which is a necessary step towards end-to-end learning. For one of the experiments on the ECP benchmark the orientation estimation of *Pose-RCNN* has been integrated into Faster R-CNN.

The experiments showed 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. Furthermore, the experiments on transfer learning showed that detectors pre-trained with the ECP dataset and fine-tuned on another 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 KITTI and CityPersons helped only marginally, if at all, when evaluating on the ECP test dataset. These results can also be attributed to the diversity of the ECP dataset.

The experiments also showed that night-performance is only a few percentage points lower than day-time performance. Experimental results on the ECP dataset 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.

The error analysis has shown that similar objects like depictions and reflections result in confusion with real VRUs. Utilizing the already annotated reflections and depictions as additional classes during training could improve the discrimination thereof, similar to the discrimination of pedestrians and riders. Samples with small resolution are still challenging despite the great amount of small-sized pedestrians present in the ECP dataset. Approximately 75% of the false positives at 0.3 fppi analyzed in Figure 4.13 are smaller than 80 pixels. Despite the fact that there are methods tuned for smaller objects like [17], the performance for the *small* scenario still falls behind the *reasonable* test scenario on the ECP benchmark as shown in Section 4.5. This also holds for the *occluded* scenario.

Further improvement will in part still come from additional data. The analysis of the effect of annotation accuracy on detection performance could be useful to plan future annotation efforts.

Group Handling (NMS Adaptations and Pairwise Estimation). Still, it is not data alone that is needed for improvement. The benchmark experiments showed methodical weaknesses in particular in dense, crowded scenes. The greedy NMS used in a post-processing step by the analyzed methods implies a trade-off between recall and precision. Chapter 5 addressed detection and pose estimation in such crowded scenes. The first part aimed for improving the detection performance by using discriminative attributes within the NMS. Therefore, YOLOv3 has been extended by uncertainty estimation and several prediction heads for the different attributes, i.e. the level of occlusion, the body orientation, and the head box position. Extending YOLOv3 did slightly lower the detection performance from 6.9 to 8.0 points LAMR, when estimating all three additional attributes. The multitask networks using additional prediction heads for the orientation and head box position only achieve a similar performance. A ceiling analysis based on ground truth annotations showed that the level of occlusion and the body orientation are less discriminative for the number of instances than the head box position. Using the head box instead of the body box within the greedy NMS did not improve detection performance. In further experiments, the GossipNet architecture [73] was trained for the task of NMS to replace the greedy NMS also incorporating the head box position as an additional input feature. Despite several adaptations e.g. regarding the loss function, the performance of the greedy NMS could not be achieved. This is caused by ambiguities involved in detection and estimation of further attributes. If a proposal box is in between overlapping persons, features from both pedestrians influence the inference result and it may be ambiguous whose position or attribute should be estimated based on that single proposal. As overlapping persons also pose challenges for pose estimation, the ambiguities in detection and pose estimation were tackled in the second part of Chapter 5 by jointly handling pairs of pedestrians.

To this end, the new *Simple Pair Pose* method for top-down human pose estimation has been created. The extended, underlying YOLOv3 detector improves the recall in groups by jointly detecting pairs of pedestrians. The issue of ambiguities is solved by training separate prediction heads for the pedestrian in the front and back. Experimental results for the new pose estimation method that jointly predicts poses for both pedestrians of these pairs have been shown. As all computations are shared apart from the final duplicated layers, it reduces the runtime for paired detections in comparison with separate pose estimation. Yet, implicitly training different experts for poses of front and back pedestrians is very effective and surpasses the AlphaPose+ method used for comparison. Regarding pedestrian pairs, *Simple Pair Pose* achieves a mean object keypoint similarity (OKS) of 75.9 for visible joints of pedestrians in the back, while AlphaPose+ achieves 68.7. Despite the focus on pairwise constellations, the pose estimation performance on the full test dataset is similar to AlphaPose+. The approach could be easily integrated into other heatmap-based single-person pose estimation approaches than AlphaPose+. It could also be used as input for the recent graph-based method [124] that relies on input poses from AlphaPose+. The graph convolutional network proposed by [124] refines pose results and thus could further improve the performance of *Simple Pair Pose*. The idea

to jointly predict poses of pedestrian pairs to solve ambiguities may also be applied to jointly predict other attributes (e.g. head box position, body orientation).

The new *EuroCity Persons Dense Pose* (ECPDP) dataset has been created, which provides the largest number of pose annotated persons in comparison with other automotive datasets and the largest average number of persons per image. Thus, it will serve for benchmarking of pose estimation methods on dense urban scenarios.

Performance Considerations. The introduction chapter of this thesis discussed the performance gap between machine learning approaches and an attentive human as shown in [178] (see Section 1.1.4). Regarding orientation estimation, this performance gap is nearly closed for pedestrians greater 200 pixels as shown in Figure 4.16, which compares human annotation accuracy with estimation results. In terms of detection performance, deep learning approaches are now also closing in on the human performance. A recent paper [85] achieves 2.2 points LAMR on the Caltech-USA dataset, while the human baseline is at 0.8. This is also the case regarding detection performance on the ECP dataset. The qualitative analysis in Chapter 4 already showed that the remaining errors are indeed "hard" samples even for a human, and there has been further progress since then as shown in Section 4.5. For a quantitative evaluation if there is a remaining performance gap the experiments of [178] could also be repeated on the ECP dataset. Still, such experiments would not answer the question of which visual detection performance is needed for the different SAE levels of automated driving (shown in Figure 1.3). The performance requirements differ in dependence on the desired level. For example, level three and four in contrast to level five are still limited to specified driving modes such as certain regions or road types, which may put lower requirements on detection performance. For all levels, detections may not need to be perfect as discussed in Section 4.4. First, errors are not equally important. They matter more at close distance to the vehicle, which coincides with a better detection performance for nearby objects. Second, the functional chain, in particular the follow-up tracking module, may suppress false positives and recover false negatives over time. Third, automated vehicles usually not only rely on a single camera but multiple sensors including radar and lidar. Therefore, it is difficult to answer the initial question only considering the visual detection performance alone, as the answer depends on the overall sensor setup within the car and the performance of the perception methods running on other sensors and the processing by the following functional chain. Still, it seems to be sufficient to a certain degree looking at what is already available in market. E.g. Waymo provides a fully automated taxi service in San Francisco even without a safety driver. The service is geofenced, meaning it is limited to a certain area as their vehicles also depend on a high-definition map. Long-term statistics will hopefully show fewer accidents in comparison with human drivers and a hereby increased road safety for VRUs, which would provide some kind of empirical proof that detection performance is already sufficient to replace a human driver. Still, even if those automated vehicles already drive more safely than humans, further improvement will be needed to further increase the safety.

6.2. FUTURE WORK

This section first discusses potential methodical improvements of *Pose-RCNN* and the need for uncertainty estimation. Then, data efficiency aspects regarding collection,

annotation, and usage of data for deep learning including benefits for *Pose-RCNN* and *Simple Pair Pose* are addressed. Finally, this section concludes with a discussion of end-to-end training of the full functional chain including the integration of multiple sensors.

The proposal generation of *Pose-RCNN* may be further improved by reducing the number of proposals while keeping a high recall. This would further improve the detection performance while reducing the needed runtime. Still, nowadays external proposal sources are no longer needed. Faster R-CNN [132] integrated the proposal generation into the network, which is a lot faster due to the realization with convolutional layers. Other works skip the separate proposal stage, like SSD [109] or YOLOv3 [127], or come with no proposals at all [39, 96].

The von Mises loss formulation for orientation regression shows promising results with *Pose-RCNN* and with Faster R-CNN on the ECP dataset in Chapter 4. Still, it is not only important to get an accurate estimate of person attributes like the orientation or pose but also information on how reliable the estimation is. This motivates stronger attention to this uncertainty estimation in the future, also triggered by the work of Kendall [84]. The uncertainty estimation will be integrated into the networks themselves. This research branch can already be observed e.g. for the von Mises loss in the work of [123] building upon the *Pose-RCNN* publication [15] (Chapter 3). [123] models orientation estimation with a mixture of von Mises distributions. For every distribution, a mean, concentration, and confidence value is estimated by the network. This solves some ambiguity issues in orientation estimation, as confusions of opposite directions are quite common, in particular for low resolutions. A single von Mises distribution can only represent a single direction by its single mean value, while a mixture can distribute the confidence on mixture components with different, even opposing mean values.

The large number of person annotations of the ECP dataset leads to an improved detection performance that does not saturate yet. Further diverse training datasets will add further performance boosts in the future. Still, as the gain reduces logarithmically as the amount of training data increases it will be less efficient and more difficult to collect appropriate data. There has to be more focus on the collection of difficult cases, which are more helpful in the training of detectors. That is especially the case for rare classes. For the ECP dataset rare classes are e.g. children (as analyzed in [14]), duties (e.g. police officers, firemen, ...), and persons showing non-standard poses, like traffic control gestures or turning gestures of bicyclists. For the application of fully automated driving, good performance for these rare cases is just as crucial as for other frequent classes. There are mainly two potential ways to collect targeted data for difficult or rare cases. First, more sophisticated collection strategies are needed. Instead of sub-sampling recordings for annotation at a fixed rate, the recordings may be analyzed automatically with machine learning support selecting frames with difficult or rare cases. See [131] for a recent survey on this domain of active learning. For the selection of difficult cases, uncertainty estimation also plays an important role, as a high uncertainty indicates a difficult sample. Note that this machine learning support has been explicitly avoided for the creation of the ECP dataset to avoid any dataset bias. This trade-off between dataset bias and efficient data collection has to be considered in the future. Second, more data could be created with generative models as done in [62] to improve classification performance for special persons like police officers, construction workers and school

guards.

Even with sophisticated methods, it might be difficult to get sufficient data. Graphics processing units and huge amounts of data have enabled the rise of deep learning, but it is debatable if a large number of training samples can be achieved for rare cases or when domain adaptation is needed as investigated in [66]. Humans are capable of interpreting their environment even if there are unknown objects that have never been seen before based on their skill to transfer experience and knowledge to unseen object categories. The same skill might be needed for rare classes. This is targeted by the domain of single-shot learning [50] or even zero shot learning [164]. Needing less training data also means less manual annotation effort. While the ECP dataset has been annotated completely manually, this is no longer feasible for even bigger datasets like Argoverse [22] and the Waymo Open Motion dataset [47], as the cost of hand labeling is too high [47]. E.g. [47] uses offboard lidar based 3D object detection for label generation.

Another possibility to increase data efficiency is self-supervised representation learning, which is a subcategory of unsupervised learning. Methods of that category automatically generate pseudo-labels [45] from unlabeled data with no or little manual effort for pre-training of deep neural networks. See [45] for an overview of different mechanisms to generate such labels. Recently, [149] has been very successful with self-supervised training and outperformed another baseline for classification on the ImageNet dataset without using ImageNet labels for pre-training (finetuning is done with labels on a smaller portion of the dataset). Finally, graphic rendering has already been used for the training of deep neural networks [112]. A further increase in photo-realism e.g. within simulation frameworks for intelligent vehicles will be beneficial for visual deep learning approaches.

A higher data efficiency will also benefit the presented approaches *Pose-RCNN* and *Simple Pair Pose*. As these approaches are mostly modular, further improvements can also be achieved by replacing modules by newer methods. The orientation head of *Pose-RCNN* and the pose head of *Simple Pair Pose* may be combined with other feature extractors like HRNet [156], for example. The usage of HRNet [156] has already led to performance improvements on the ECP detection benchmark as analyzed in Section 4.5. The pairwise pose estimation could also be used with future methods optimized for detection in dense traffic scenes. Such future detection methods will probably skip the NMS as a separate post-processing step. Recently, e.g. transformer networks have been used for object detection [20]. The attention mechanism used within the transformer processes global information across the full image, which is difficult to achieve with small local receptive fields of convolutional layers. [20] achieves competitive results with an optimized Faster R-CNN method on the MSCOCO dataset. Furthermore, it does not depend on a NMS as a post-processing step, as the transformer network learns to handle relations between different objects and directly estimates unique detections. Replacing the NMS or integrating it into the network itself like in GossipNet [73] will move us towards the goal of fully end-to-end training.

When looking at the intelligent vehicles domain, the trend towards fully end-to-end training does not only affect detection. In classic approaches, the modules detection, tracking, prediction, and planning are decoupled. The tracking-by-detection paradigm, where framewise independent detections are associated and filtered over time, raises the following issue. Detection performance is evaluated with metrics like LAMR and AP.

These are metrics to evaluate single frame performance and they treat every pedestrian equally. Still, for the follow-up tracking approach, a burst of false positives or false negatives might be worse than missing pedestrians on single frames now and then, which can be recovered by filtering. Separate training of the detection module usually targets an optimized detection metric, which does not always result in improved tracking performance. End-to-end training provides a training signal for the perception module, which persons have to be detected for optimal performance of the full functional chain. Apart from that, only using predicted boxes as input for the tracker may once again cause ambiguities e.g. in the extraction of features for re-identification of objects in dense scenes [183]. Therefore, in the future, the space and time components will be integrated more frequently within single approaches as in [25], where LSTMs with attention modules are used to integrate the time component in detection. The work of [103] presents an end-to-end model combining detection and prediction. Regarding the full functional chain, the works of [172] and [134] even go one step further. They use voxelized lidar and an HD/raster map to integrate detection, prediction, and planning in a full end-to-end trainable framework.

To increase the reliability of fully automated vehicles, usage of redundant information from different sensors with different strengths and weaknesses will be required, while there is no use of multimodalities yet in the works of [172] and [134]. E.g. the View-of-Delft dataset [118] provides a multi-sensor benchmark comparing detection performance for the usage of radar and lidar pointclouds. In classic approaches, lidar, radar, and visual detection are often done separately and fused in a later stage, e.g. within the tracking module. Early fusion of features of lidar and RGB as in [142] may also improve the overall performance.

The experiments of the ECP benchmark have shown that 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) are promising.

Summarizing most of this section, future work will probably show a full integration of multiple tasks (detection, pose estimation), uncertainty estimation, context information (geolocation, time of day), multiple sensor modalities, temporal information, and the complete functional chain (including detection, tracking, prediction, and planning) within end-to-end trainable frameworks. The ECP benchmark and the ECPDP dataset will stimulate research towards finding the best approach for person detection and pose estimation in traffic scenes within such a framework - an approach that will hopefully run onboard future intelligent vehicles to save lives.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support and inspiration of many people. First and foremost, I would like to thank my advisor and promotor Prof. Dr. Dariu M. Gavrila. Your invaluable advice, insightful suggestions, and critical questions enabled a high quality of work. With your guidance, I learned how to do research.

Further, I would like to express my appreciation to my co-promotor Julian F. P. Kooij and the committee members for their feedback and their effort spent in reviewing this thesis.

A lot of colleagues at Mercedes Benz also played a decisive role in the creation of this thesis. I am very grateful to Dr. Ulrich Kreßel, for his profound belief in my work, the offered opportunities, and for supporting my ideas and goals. Many thanks to Prof. Dr. Fabian Flohr, my former local supervisor at Mercedes Benz. You encouraged me to hang on even in the most difficult and turbulent times. Your contributions to the projects of this thesis, in particular the ECP dataset, are invaluable and I will never forget the exciting experience of our common vehicle presentation in Berlin Tempelhof for Angela Merkel and Li Keqiang. I would also like to extend my deepest gratitude to Sebastian Krebs, for his unwavering support, and also invaluable contributions to the success of this thesis. I always enjoyed working together with you, no matter if it was about integrating soft- and hardware into our vehicle, test drives in the context of the EU project PROSPECT, or recording the ECP data throughout Europe including trips to the Pantheon or the cathedral of Florence. The interesting and humorous coffee sessions with you helped me to keep up, and I am grateful for our friendship and frequent activities outside of work. I am very grateful to Markus Roth, for his practical suggestions and his knowledge and expertise in particular with the Robot Operating System. The common hacking sessions in the car reverse engineering badly documented hardware has been an amazing experience. Our adventurous trips to the mountains and on (and under) the water of the Soča (together with Sebastian) are quite memorable. I also had the great pleasure of working with several master students, in particular Yikang Wang and Phillip Czech. Thanks for your contributions to this thesis. Many thanks to Florian Kraus, for interesting and joyful conversations and your deep learning expertise e.g. regarding the mathematical foundations of loss functions. Special thanks to Dr. Qing Rao for the collaboration on one of my first publications and Julian Wiederer for nice conversations, great hospitality in Stuttgart, and intensive squash sessions. Thanks also go to Dr. Andreas Fregin for helpful tooling in the application of ROS in our sensor vehicle and Hidde Boekema for the Dutch translation of the summary.

I would like to thank my friends, who never let me down. In particular, I am very grateful to Roland Wörz and Eray Özmü for their unwavering friendship since school. I also want to express my deepest gratitude to my dear parents, for their outstanding role in my life. I know, my deceased father, who also had a strong passion for technical innovations, would be proud. The immeasurable support and encouragement of my

mother, and also my brothers, cannot be overestimated.

Finally to my own family - the center of my universe. The completion of this thesis would not have been possible without the infinite support, encouragement, tolerance, and optimism of my wife, Anna. Thank you for always being at my side during that chapter in my life. Two years ago, we have been joined on our journey by our daughter Linda.I could not have imagined the boundless joy that comes from your smile and laughter, your excitement in discovering the world around you, or from holding you in my arms. Last month, my infinite love and joy grew even bigger when we welcomed our daughter Elise. I have now concluded the chapters of this thesis. With the deepest love for my family, I look forward to future chapters in our life.

CURRICULUM VITÆ

Markus BRAUN

24-09-1989	Born in Heilbronn, Germany.

EDUCATION

2015-Present	Ph.D. in the Cognitive Robotics department Delft University of Technology *Thesis:* Visual Detection and Pose Estimation of Vulnerable Road Users for Automated Driving
2012–2015	MSc Computer Science Karlsruhe Institute of Technology, Germany
2009–2012	BSc Computer Science Karlsruhe Institute of Technology, Germany

PROFESSIONAL EXPERIENCE

2015–Present	Mercedes Benz AG, Germany Machine Learning Engineer in the Pattern Recognition team.
2014	Microsoft Development Center Norway, Norway Three months internship as Software Developer in the Microsoft SharePoint team in Oslo.

List of Publications

7. **M. Braun**, F. Flohr, S. Krebs, U. Kreßel, D. M. Gavrila, *Simple Pair Pose - Pairwise Human Pose Estimation in Dense Urban Traffic Scenes*, Proc. of the IEEE Intelligent Vehicles Symposium, 2021, pp.1545–1552.

   Author contributions: M. Braun implemented and evaluated the proposed method and wrote the paper, M. Braun created the dataset together with F. Flohr, S. Krebs and U. Kreßel supported the recordings and the preparation of the dataset, D. M. Gavrila provided guidance and supervision.

6. **M. Braun**, S. Krebs, D. M. Gavrila, *ECP2.5D - Person Localization in Traffic Scenes*, Proc. of the IEEE Intelligent Vehicles Symposium, 2020, pp.1694-1701.

   Author contributions: M. Braun implemented and evaluated the proposed baseline method for monocular 2.5D localization, M. Braun and S. Krebs created and implemented the up-lifting method used to create the ECP2.5D dataset, S. Krebs implemented the ego-motion correction of lidar pointclouds, M. Braun and S. Krebs wrote the paper, D. M. Gavrila provided guidance and supervision.

5. S. Krebs, **M. Braun**, D. M. Gavrila, *Generating 3D Person Trajectories from Sparse Image Annotations in an Intelligent Vehicles Setting*, Proc. of the IEEE Intelligent Transportation Systems Conf., 2019, pp.783-788.

   Author contributions: S. Krebs designed, implemented, and evaluated the proposed system, and wrote the paper, M. Braun provided technical support and feedback, D. M. Gavrila provided guidance and supervision.

4. **M. Braun**, S. Krebs, F. Flohr, D. M. Gavrila, *EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes*, IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), 2019, vol. 41, no. 8, pp.1844-1861.

   Author contributions: M. Braun implemented, modified, and evaluated the benchmark methods and experiments. M. Braun, S. Krebs, and F. Flohr built up the sensor vehicle and recorded the data during several journeys. M. Braun created the detection benchmark with help from F. Flohr and S. Krebs. F. Flohr implemented the server framework to be used for online benchmarking, M. Braun wrote the journal article with help from S. Krebs and D. M. Gavrila. D. M. Gavrila furthermore provided guidance and supervision.

3. A. Fregin, M. Roth, **M. Braun**, S. Krebs, F. Flohr, *Building a computer vision research vehicle with ROS*, Proc. of the ROSCon, 2017.

   Author contributions: All authors have contributed to building several computer vision research vehicles with ROS. A. Fregin has created the publication slides.

2. **M. Braun**, Q. Rao, Y. Wang, and F. Flohr, *Pose-RCNN: Joint object detection and pose estimation using 3D object proposals*, Proc. of the IEEE Intelligent Transportation Systems Conf., 2016, pp.1546-1551.

Author contributions: M. Braun implemented and evaluated Fast-RCNN methods with stixel and lidar proposals. Q. Rao implemented Lidar proposal generation. Master student Y. Wang implemented the method Pose-RCNN together with M. Braun and F. Flohr. All authors contributed in writing the paper. F. Flohr provided guidance and supervision.

1. X. Li, F. Flohr, Y. Yang, H. Xiong, **M Braun**, S. Pan, K. Li, and D. M. Gavrila, *A new benchmark for vision-based cyclist detection*, Proc. of the IEEE Intelligent Vehicles Symposium, 2016, pp.1028-1033.

   Author contributions: X. Li implemented, modified, and evaluated state of the art methods, F. Flohr implemented and evaluated the method SP-FRCN and created the dataset together with Y. Yang. Further F. Flohr has built up the test vehicle for recording the dataset. H. Xiong and M. Braun helped with the evaluation of the experiments. S. Pan, K. Li, and D. M. Gavrila provided guidance and supervision.

BIBLIOGRAPHY

[1] S. Agarwal, J. O. D. Terrail, and F. Jurie. “Recent advances in object detection in the age of deep convolutional neural networks”. In: *arXiv preprint arXiv:1809.03193* (2018).

[2] B. Alexe, T. Deselaers, and V. Ferrari. “Measuring the Objectness of image windows”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 34.11 (2012), pp. 2189–2202.

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. “2D human pose estimation: New benchmark and state of the art analysis”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3686–3693.

[4] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. “Multiscale combinatorial grouping”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 328–335.

[5] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool. “Seeking the strongest rigid detector”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 3666–3673.

[6] R. Benenson, M. Omran, J. Hosang, and B. Schiele. “Ten years of pedestrian detection, what have we learned?” In: *European Conference on Computer Vision (ECCV)Workshop*. 2014, pp. 613–627.

[7] B. Benfold and I. Reid. “Guiding visual surveillance by tracking human attention”. In: *Proc. of the British Machine Vision Conf. (BMVC)*. 2009, pp. 1–11.

[8] B. Benfold and I. Reid. “Unsupervised learning of a scene-specific coarse gaze estimator”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2011, pp. 2344–2351.

[9] L. Beyer, A. Hermans, and B. Leibe. “Biternion Nets: Continous head pose regression from discrete training labels”. In: *German Conference on Pattern Recognition (GCPR)*. 2015, pp. 157–168.

[10] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. “Soft-NMS – Improving object detection with one line of code”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2017, pp. 5561–5569.

[11] L. Bourdev, S. Maji, and J. Malik. “Describing people: A poselet-based approach to attribute classification”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2011, pp. 1543–1550.

[12] M. Braun, F. B. Flohr, S. Krebs, U. Kreßel, and D. M. Gavrila. “Simple Pair Pose - Pairwise human pose estimation in dense urban traffic scenes”. In: *Proc. of the IEEE Intelligent Vehicles Symposium*. 2021, pp. 1545–1552.

[13] M. Braun, S. Krebs, F. B. Flohr, and D. M. Gavrila. “EuroCity Persons: A novel benchmark for person detection in traffic ccenes”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 41.8 (2019), pp. 1844–1861.

[14] M. Braun, S. Krebs, and D. M. Gavrila. “ECP2.5D - Person localization in traffic scenes”. In: *Proc. of the IEEE Intelligent Vehicles Symposium*. 2020, pp. 1694–1701.

[15] M. Braun, Q. Rao, Y. Wang, and F. Flohr. “Pose-RCNN: Joint object detection and pose estimation using 3D object proposals”. In: *Proc. of the IEEE Intelligent Transportation Systems Conf.* 2016, pp. 1546–1551.

[16] J. J. Breuer, A. Faulhaber, P. Frank, and S. Gleissner. “Real world safety benefits of brake assistance systems”. In: *International Technical Conference on the Enhanced Safety of Vehicles (ESV)*. 2007.

[17] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos. “A unified multi-scale deep convolutional neural network for fast object detection”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 354–370.

[18] Z. Cai and N. Vasconcelos. “Cascade R-CNN: Delving into high quality object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6154–6162.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. “Realtime multi-person 2d pose estimation using part affinity fields”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7291–7299.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. “End-to-end object detection with transformers”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 213–229.

[21] J. Carreira and C. Sminchisescu. “CPMC: Automatic object segmentation using constrained parametric Min-Cuts”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 34.7 (2012), pp. 1312–1328.

[22] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. “Argoverse: 3d tracking and forecasting with rich maps”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 8748–8757.

[23] C. Chen and J.-M. Odobez. “We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 1544–1551.

[24] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. “3D object proposals for accurate object class detection”. In: *Adv. in Neural Information Processing Systems (NIPS)*. 2015, pp. 424–432.

[25] X. Chen, Z. Wu, and J. Yu. “TSSD: Temporal Single-Shot Detector Based on Attention and LSTM”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018, pp. 1–9.

[26] 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. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3286–3293.

[27] Z. Chong, B. Qin, T. Bandyopadhyay, T. Wongpiromsarn, E. Rankin, M. Ang, E. Frazzoli, D. Rus, D. Hsu, and K. Low. “Autonomous personal vehicle for the first-and last-mile transportation services”. In: *IEEE International Conference on Cybernetics and Intelligent Systems (CIS)*. 2011, pp. 253–260.

[28] X. Chu, A. Zheng, X. Zhang, and J. Sun. “Detection in crowded scenes: One proposal, multiple predictions”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 12214–12223.

[29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. “The Cityscapes dataset for semantic urban scene understanding”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 3213–3223.

[30] P. Czech. “Pedestrian detection in autonomous driving by techniques optimized for crowds with deep neural networks”. Supervision by Markus Braun. MA thesis. Ruhr-Universität Bochum, Germany, 2020.

[31] J. Dai, Y. Li, K. He, and J. Sun. “R-FCN: Object detection via region-based fully convolutional networks”. In: *Adv. in Neural Information Processing Systems (NIPS)*. 2016, pp. 379–387.

[32] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2005, pp. 886–893.

[33] B. De Brabandere, D. Neven, and L. Van Gool. “Semantic instance segmentation for autonomous driving”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7–9.

[34] J. Demšar. “Statistical comparisons of classifiers over multiple data sets”. In: *Journal of Machine Learning Research* 7 (2006), pp. 1–30.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “ImageNet: A large-scale hierarchical image database”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255.

[36] P. Dollár, R. Appel, S. Belongie, and P. Perona. “Fast feature pyramids for object detection”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 36.8 (2014), pp. 1532–1545.

[37] P. Dollár, Z. Tu, P. Perona, and S. Belongie. “Integral channel features”. In: *Proc. of the British Machine Vision Conf. (BMVC)*. 2009, pp. 91.1–91.11.

[38] P. Dollár, C. Wojek, B. Schiele, and P. Perona. “Pedestrian detection: An evaluation of the state of the art”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 34.4 (2012), pp. 743–761.

[39] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. “CenterNet: Keypoint triplets for object detection”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2019, pp. 6569–6578.

[40] T. Elsken, J. H. Metzen, and F. Hutter. “Neural architecture search: A survey”. In: *Journal of Machine Learning Research* 20.1 (2019), pp. 1997–2017.

[41] M. Enzweiler. “Compound Models for Vision-Based Pedestrian Recognition”. PhD thesis. Institut für Technische Informatik, Ruprecht-Karls-Universität Heidelberg, Germany, 2011.

[42] M. Enzweiler and D. M. Gavrila. “Integrated pedestrian classification and orientation estimation”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2010, pp. 982–989.

[43] M. Enzweiler and D. M. Gavrila. “Monocular pedestrian detection: Survey and experiments”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 31.12 (2009), pp. 2179–2195.

[44] M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke. “Efficient Stixel-based object recognition”. In: *Proc. of the IEEE Intelligent Vehicles Symposium*. 2012, pp. 1066–1071.

[45] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales. “Self-supervised representation learning: Introduction, advances and challenges”. In: *IEEE Signal Processing Magazine* 39.3 (2022), pp. 42–62.

[46] A. Ess, B. Leibe, and L. Van Gool. “Depth and appearance for mobile scene analysis”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2007, pp. 1–8.

[47] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov. “Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2021, pp. 9710–9719.

[48] 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”. In: *International Journal of Computer Vision* 111.1 (2015), pp. 98–136.

[49] D. J. Fagnant and K. M. Kockelman. “The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios”. In: *Transportation Research Part C: Emerging Technologies* 40 (2014), pp. 1–13.

[50] L. Fei-Fei, R. Fergus, and P. Perona. “One-shot learning of object categories”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 28.4 (2006), pp. 594–611.

[51] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. “Object detection with discriminatively trained part-based models”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 32.9 (2010), pp. 1627–1645.

[52] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila. “A probabilistic framework for joint pedestrian head and body orientation estimation”. In: *IEEE Trans. on Intelligent Transportation Systems (TITS)* 16.4 (2015), pp. 1872–1882.

[53] F. B. Flohr. “Vulnerable road user detection and orientation estimation for context-aware automated driving”. PhD thesis. Informatics Institute (IVI), University of Amsterdam, Netherlands, 2018.

[54] J. H. Gawron, G. A. Keoleian, R. D. De Kleine, T. J. Wallington, and H. C. Kim. “Deep decarbonization from electrified autonomous taxi fleets: Life cycle assessment and case study in Austin, TX”. In: *Transportation Research Part D: Transport and Environment* 73 (2019), pp. 130–141.

[55] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. “Vision meets robotics: The KITTI dataset”. In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1231–1237.

[56] A. Geiger, P. Lenz, and R. Urtasun. “Are we ready for autonomous driving? The KITTI vision benchmark suite”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2012, pp. 3354–3361.

[57] D. Gerónimo, A. Sappa, A. López, and D. Ponsa. “Adaptive image sampling and windows classification for on-board pedestrian detection”. In: *Proc. of the International Conf. on Computer Vision System*. 2007.

[58] R. B. Girshick. “Fast R-CNN”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2015, pp. 1440–1448.

[59] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 580–587.

[60] T. Golda, T. Kalb, A. Schumann, and J. Beyerer. “Human pose estimation for real-world crowded scenarios”. In: *Proc. of the International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*. 2019, pp. 1–8.

[61] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets”. In: *Adv. in Neural Information Processing Systems (NIPS)*. 2014, pp. 2672–2680.

[62] Z. Guo, R. Zhi, W. Zhang, B. Wang, Z. Fang, V. Kaiser, J. Wiederer, and F. Flohr. “Generative model based data augmentation for special person classification”. In: *Proc. of the IEEE Intelligent Vehicles Symposium*. 2020, pp. 1675–1681.

[63] S. Gupta, R. B. Girshick, P. Arbeláez, and J. Malik. “Learning rich features from RGB-D Images for object detection and segmentation”. In: *European Conference on Computer Vision (ECCV)*. 2014, pp. 345–360.

[64] D. Hall and P. Perona. “Fine-grained classification of pedestrians in video: Benchmark and state of the art”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 5482–5491.

[65] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge University Press, 2003.

[66] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao. “Generalizable pedestrian detection: The elephant in the room”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11328–11337.

[67] K. He, G. Gkioxari, P. Dollár, and R. Girshick. “Mask R-CNN”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2017, pp. 2961–2969.

[68] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[69] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang. “Bounding box regression with uncertainty for accurate object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2888–2897.

[70] S. Hong, H. Park, J. Park, S. Cho, and H. Park. “HintPose”. In: *arXiv preprint arXiv:2003.02170* (2020).

[71] J. Hosang, M. Omran, R. Benenson, and B. Schiele. “Taking a deeper look at pedestrians”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 4073–4082.

[72] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. “What makes for effective detection proposals?” In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 38.4 (2016), pp. 814–830.

[73] J. Hosang, R. Benenson, and B. Schiele. “Learning non-maximum suppression”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4507–4515.

[74] 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. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3296–3297.

[75] S. Huang and D. Ramanan. “Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4664–4673.

[76] M.-Y. Huh, P. Agrawal, and A. A. Efros. “What makes ImageNet good for transfer learning?” In: *arXiv preprint arXiv:1608.08614* (2016).

[77] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon. “Multispectral pedestrian detection: Benchmark dataset and baseline”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1037–1045.

[78] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. “DeeperCut: A deeper, stronger, and faster multi-person pose estimation model”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 34–50.

[79] S. International. *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. j3016*. Tech. rep. 2016.

[80] 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. of the ACM international conference on Multimedia*. 2014, pp. 675–678.

[81] C. Jiang, H. Xu, W. Zhang, X. Liang, and Z. Li. “SP-NAS: Serial-to-parallel backbone search for object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11863–11872.

[82] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo. “Differentiable hierarchical graph grouping for multi-person pose estimation”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 718–734.

[83] A. Kendall, Y. Gal, and R. Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7482–7491.

[84] A. Kendall, Y. Gal, and R. Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7482–7491.

[85] A. H. Khan, M. Munir, L. van Elst, and A. Dengel. “F2DNet: Fast focal detection network for pedestrian detection”. In: *arXiv preprint arXiv:2203.02331* (2022).

[86] A. Khan, C. D. Harper, C. T. Hendrickson, and C. Samaras. “Net-societal and net-private benefits of some existing vehicle crash avoidance technologies”. In: *Accident Analysis & Prevention* 125 (2019), pp. 207–216.

[87] W. Kim, M. S. Ramanagopal, C. Barto, M.-Y. Yu, K. Rosaen, N. Goumas, R. Vasudevan, and M. Johnson-Roberson. “PedX: Benchmark dataset for metric 3D pose estimation of pedestrians in complex urban intersections”. In: *IEEE Robotics and Automation Letters (RA-L)* 4.2 (2019), pp. 1940–1947.

[88] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *International Conference on Learning Representations (ICLR)*. 2015.

[89] W. König. “Guidelines for user-centered development of DAS.” In: *Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort*. 2016, pp. 781–796.

[90] J. F. Kooij, F. Flohr, E. A. Pool, and D. M. Gavrila. “Context-based path prediction for targets with switching dynamics”. In: *International Journal of Computer Vision* 127.3 (2019), pp. 239–262.

[91] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. “Context-based pedestrian path prediction”. In: *European Conference on Computer Vision (ECCV)*. 2014, pp. 618–633.

[92] I. Kotseruba, A. Rasouli, and J. K. Tsotsos. “Benchmark for evaluating pedestrian action prediction”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 1258–1268.

[93] F. Kraus and K. Dietmayer. “Uncertainty estimation in one-stage object detection”. In: *Proc. of the IEEE Intelligent Transportation Systems Conf.* 2019, pp. 53–60.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Adv. in Neural Information Processing Systems (NIPS)*. 2012, pp. 1097–1105.

[95] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale”. In: *International Journal of Computer Vision* 128.7 (2020), pp. 1956–1981.

[96] H. Law and J. Deng. “CornerNet: Detecting objects as paired keypoints”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 734–750.

[97] C. Legacy, D. Ashmore, J. Scheurer, J. Stone, and C. Curtis. “Planning the driverless city”. In: *Transport Reviews* 39.1 (2019), pp. 84–102.

[98] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. “Dynamic 3d scene analysis from a moving vehicle”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2007, pp. 1–8.

[99] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. “Perceptual generative adversarial networks for small object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1951–1959.

[100] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan. “Scale-aware Fast R-CNN for pedestrian detection”. In: *IEEE Transactions on Multimedia* 20.4 (2018), pp. 985–996.

[101] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu. “CrowdPose: Efficient crowded scenes pose estimation and a new benchmark”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10863–10872.

[102] 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. of the IEEE Intelligent Vehicles Symposium*. 2016, pp. 1028–1033.

[103] M. Liang, B. Yang, W. Zeng, Y. Chen, R. Hu, S. Casas, and R. Urtasun. “PnPNet: End-to-end perception and prediction with tracking in the loop”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11553–11562.

[104] D. Lin, S. Fidler, and R. Urtasun. “Holistic scene understanding for 3D object detection with RGBD cameras”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2013, pp. 1417–1424.

[105] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. “Feature pyramid networks for object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2117–2125.

[106] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. “Focal loss for dense object detection”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2017, pp. 2980–2988.

[107] 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: *European Conference on Computer Vision (ECCV)*. 2014, pp. 740–755.

[108] S. Liu, D. Huang, and Y. Wang. “Adaptive NMS: Refining pedestrian detection in a crowd”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 6459–6468.

[109] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. “SSD: Single shot multibox detector”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 21–37.

[110] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. “What can help pedestrian detection?” In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6034–6043.

[111] K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley & Sons, Inc., 2008.

[112] M. Martinez, C. Sitawarin, K. Finch, L. Meincke, A. Yablonski, and A. Kornhauser. "Beyond Grand Theft Auto V for training, testing and enhancing deep learning in self driving cars". In: *arXiv preprint arXiv:1712.01397* (2017).

[113] Mercedes Benz. *Introducing DRIVE PILOT: An Automated Driving System for the Highway*. Tech. rep. 2019.

[114] S. Munder and D. M. Gavrila. "An experimental study on pedestrian classification". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 28.11 (2006), pp. 1863–1868.

[115] A. Newell, K. Yang, and J. Deng. "Stacked hourglass networks for human pose estimation". In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 483–499.

[116] W. Ouyang and X. Wang. "Single-pedestrian detection aided by multi-pedestrian detection". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 3198–3205.

[117] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. "A new pedestrian dataset for supervised learning". In: *Proc. of the IEEE Intelligent Vehicles Symposium*. 2008, pp. 373–378.

[118] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila. "Multi-class road user detection with 3+1D radar in the View-of-Delft dataset". In: *IEEE Robotics and Automation Letters (RA-L)* 7.2 (2022), pp. 4961–4968.

[119] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. "PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 269–286.

[120] B. Pepik, M. Stark, P. Gehler, and B. Schiele. "Multi-view and 3D deformable part models". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 37.11 (2015), pp. 2232–2245.

[121] D. Pfeiffer and U. Franke. "Towards a global optimal multi-layer Stixel representation of dense 3D data". In: *Proc. of the British Machine Vision Conf. (BMVC)*. 2011, pp. 51.1–51.12.

[122] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. "DeepCut: Joint subset partition and labeling for multi person pose estimation". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4929–4937.

[123] S. Prokudin, P. Gehler, and S. Nowozin. "Deep directional statistics: Pose estimation with uncertainty quantification". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 534–551.

[124] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui. "Peeking into occluded joints: A novel framework for crowd pose estimation". In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 488–504.

[125] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi. “An exploration of why and when pedestrian detection fails”. In: *Proc. of the IEEE Intelligent Transportation Systems Conf.* 2015, pp. 2335–2340.

[126] J. Redmon and A. Farhadi. “YOLO9000: Better, faster, stronger”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6517–6525.

[127] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. “You only look once: Unified, real-time object detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788.

[128] J. Redmon and A. Farhadi. “YOLOv3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

[129] E. Rehder, H. Kloeden, and C. Stiller. “Head detection and orientation estimation for pedestrian safety”. In: *Proc. of the IEEE Intelligent Transportation Systems Conf.* 2014, pp. 2292–2297.

[130] 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. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5420–5428.

[131] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang. “A survey of deep active learning”. In: *ACM Comput. Surv.* 54.9 (2021), pp. 1–40.

[132] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *Adv. in Neural Information Processing Systems (NIPS)*. 2015, pp. 91–99.

[133] R. Rusu. “Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments”. PhD thesis. Computer Science department, Technische Universität München, Germany, 2009.

[134] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun. “Perceive, predict, and plan: Safe motion planning through interpretable semantic representations”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 414–430.

[135] B. Schiele and C. Wojek. “Camera based pedestrian detection.” In: *Handbook of Driver Assistance Systems: Basic Information, Components and Systems for Active Safety and Comfort*. 2016, pp. 525–545.

[136] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun. “CrowdHuman: A benchmark for detecting human in a crowd”. In: *arXiv preprint arXiv:1805.00123* (2018).

[137] G. Sharma and F. Jurie. “Learning discriminative spatial representation for image classification”. In: *Proc. of the British Machine Vision Conf. (BMVC)*. 2011, pp. 1–11.

[138] E. Shelhamer, J. Long, and T. Darrell. “Fully convolutional networks for semantic segmentation”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 39.4 (2017), pp. 640–651.

[139] A. Shrivastava, A. Gupta, and R. Girshick. “Training region-based object detectors with online hard example mining”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 761–769.

[140] Ó. Silva, R. Cordera, E. González-González, and S. Nogués. “Environmental impacts of autonomous vehicles: A review of the scientific literature”. In: *Science of The Total Environment* (2022), p. 154615.

[141] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *International Conference on Learning Representations (ICLR)*. 2015.

[142] V. A. Sindagi, Y. Zhou, and O. Tuzel. “MVX-Net: Multimodal VoxelNet for 3D object detection”. In: *Proc. of the International Conf. on Robotics and Automation (ICRA)*. 2019, pp. 7276–7282.

[143] S. Singh. *Critical reasons for crashes investigated in the national motor vehicle crash causation survey*. Tech. rep. 2015.

[144] Stanford Artificial Intelligence Laboratory et al. *Robotic Operating System*. Version ROS Indigo Igloo. July 22, 2014.

[145] 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. of the International Conf. on Computer Vision (ICCV)*. 2015, pp. 2686–2694.

[146] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. “Revisiting unreasonable effectiveness of data in deep learning era”. In: *Proc. of the International Conf. on Computer Vision (ICCV)*. 2017, pp. 843–852.

[147] K. Sun, B. Xiao, D. Liu, and J. Wang. “Deep high-resolution representation learning for human pose estimation”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5693–5703.

[148] 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. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[149] N. Tomasev, I. Bica, B. McWilliams, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic. “Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?” In: *arXiv preprint arXiv:2201.05119* (2022).

[150] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. “Efficient object localization using convolutional networks”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 648–656.

[151] A. Torralba and A. A. Efros. “Unbiased look at dataset bias”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2011, pp. 1521–1528.

[152] J. R. Treat, N. Tumbas, S. McDonald, D. Shinar, R. D. Hume, R. Mayer, R. Stansifer, and N. Castellan. *Tri-level study of the causes of traffic accidents: final report. Executive summary*. Tech. rep. 1979.

[153] R. Tsai. “A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses”. In: *IEEE Journal on Robotics and Automation* 3.4 (1987), pp. 323–344.

[154] S. Tulsiani and J. Malik. “Viewpoints and keypoints”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1510–1519.

[155] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. “Selective search for object recognition”. In: *International Journal of Computer Vision* 104.2 (2013), pp. 154–171.

[156] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. “Deep high-resolution representation learning for visual recognition”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 43.10 (2021), pp. 3349–3364.

[157] S. Wang, D. Yang, B. Wang, Z. Guo, R. K. Verma, J. Ramesh, C. Weinrich, U. Kreßel, and F. B. Flohr. “UrbanPose: A new benchmark for VRU pose estimation in urban traffic scenes”. In: *Proc. of the IEEE Intelligent Vehicles Symposium*. 2021, pp. 1537–1544.

[158] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. “Repulsion loss: Detecting pedestrians in a crowd”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7774–7783.

[159] Z. Wang, Y. Bian, and S. E. Shladover. “A survey on cooperative longitudinal motion control of multiple connected and automated vehicles”. In: *IEEE Intelligent Transportation Systems Magazine* 12.1 (2020), pp. 4–24.

[160] C. Wojek, S. Walk, and B. Schiele. “Multi-cue onboard pedestrian detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 794–801.

[161] World Health Organization. *Global status report on road safety 2018*. Tech. rep. 2018.

[162] World Health Organization. *World health statistics 2021*. Tech. rep. 2021.

[163] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, et al. “AI Challenger: A large-scale dataset for going deeper in image understanding”. In: *arXiv preprint arXiv:1711.06475* (2017).

[164] Y. Xian, B. Schiele, and Z. Akata. “Zero-shot learning - the good, the bad and the ugly”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4582–4591.

[165] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. “Subcategory-aware convolutional neural networks for object proposals and detection”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 924–933.

[166] B. Xiao, H. Wu, and Y. Wei. “Simple baselines for human pose estimation and tracking”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 466–481.

[167] H. Xie, W. Zheng, and H. Shin. “Occluded pedestrian detection techniques by Deformable Attention-Guided Network (DAGN)”. In: *Applied Sciences* 11.13 (2021), p. 6025.

[168] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li. “Robust multi-resolution pedestrian detection in traffic scenes”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 3033–3040.

[169] C. Yang, V. Ablavsky, K. Wang, Q. Feng, and M. Betke. “Learning to Separate: Detecting heavily-occluded objects in urban scenes”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 530–546.

[170] 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. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2129–2137.

[171] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. “BDD100K: A diverse driving dataset for heterogeneous multitask learning”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2636–2645.

[172] W. Zeng, S. Wang, R. Liao, Y. Chen, B. Yang, and R. Urtasun. “DSDNet: Deep structured self-driving network”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 156–172.

[173] J. Zhang, L. Lin, Y. Li, Y.-c. Chen, J. Zhu, Y. Hu, and S. C. Hoi. “Attribute-aware pedestrian detection in a crowd”. In: *IEEE Transactions on Multimedia* (2019).

[174] K. Zhang, S.-C. Chen, D. Whitman, M.-L. Shyu, J. Yan, and C. Zhang. “A progressive morphological filter for removing nonground measurements from airborne LiDAR data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 41.4 (2003), pp. 872–882.

[175] K. Zhang, F. Xiong, P. Sun, L. Hu, B. Li, and G. Yu. “Double anchor R-CNN for human detection in a crowd”. In: *arXiv preprint arXiv:1909.09998* (2019).

[176] L. Zhang, L. Lin, X. Liang, and K. He. “Is Faster R-CNN doing well for pedestrian detection?” In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 443–457.

[177] S. Zhang, C. Bauckhage, and A. B. Cremers. “Informed Haar-like features improve pedestrian detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 947–954.

[178] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. “Towards reaching human performance in pedestrian detection”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 34.4 (2018), pp. 973–985.

[179] S. Zhang, R. Benenson, and B. Schiele. “CityPersons: A diverse dataset for pedestrian detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3213–3221.

[180] S. Zhang, R. Benenson, and B. Schiele. “Filtered channel features for pedestrian detection”. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1751–1760.

[181] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. “How far are we from solving pedestrian detection?” In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1259–1267.

[182] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo. “Widerperson: A diverse dataset for dense pedestrian detection in the wild”. In: *IEEE Transactions on Multimedia* 22.2 (2019), pp. 380–393.

[183] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu. “FairMOT: On the fairness of detection and re-identification in multiple object tracking”. In: *International Journal of Computer Vision* 129.11 (2021), pp. 3069–3087.

[184] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu. “Scale-adaptive deconvolutional regression network for pedestrian detection”. In: *Asian Conf. on Computer Vision (ACCV)*. 2016, pp. 416–430.

[185] Z. Zou, Z. Shi, Y. Guo, and J. Ye. “Object detection in 20 years: A survey”. In: *arXiv preprint arXiv:1905.05055* (2019).